


UNIVERSITY OF SHEFFIELD

# How do transcribed enhancers confer genetic risk in Amyotrophic Lateral Sclerosis?

by

Laura J. Harrison

October 2024

Thesis submitted to the School of Biosciences, Faculty of Science, University of Sheffield for the degree of Doctor of Philosophy

# *Acknowledgements*

Firstly, I would like to thank my supervisors Dr. Dan Bose and Dr. Johnathan Cooper-Knock for their continued support. I feel incredibly fortunate to have had Dan as a supervisor, I couldn't have asked for a better, or more supportive, supervisor. Thank you, Dan, for always making time for me and for fostering my scientific curiosity with your infectious passion for research. Thanks to Dan, I feel like I have grown so much as an independent researcher, thank you for instilling a self-confidence in me that I never knew was possible.

I would like to thank all Bose lab members, past and present: Dr. Petra Celadova, Dr. Nicola Carruthers, Dr. Katie Gelder, Katy Boswell, Dr. Archna Shah, Grace Gilbert, Rabiatul Adawiyah and Owen Godwin. It has been a joy to develop as a scientist surrounded by so many kind souls. Starting my PhD in 2020 was intimidating but I never had to worry about finding a support network - each of you have been there for me when I needed it most. I couldn't ask for a better group of people to call my friends. Thank you, Petra, for always putting kindness above anything else. Thank you, Nicola, for being a long-time friend, you have forever been my number one cheerleader; I'm so glad I followed you to the Bose lab. Thank you, Katie, for the late-night chats, you have helped build my confidence, and I appreciate your kindness beyond words. Thank you, Katy, for all the laughs over the years. Thank you, Archna, for your wisdom and mentorship, and letting me know when I'm being too hard on myself. Thank you, Grace, for always encouraging me, forever the pessimist, to show myself kindness. Thank you, Bia, for always reigniting my passion for science - your excitement is contagious.

I would like to thank John and the Cooper-Knock group for the guidance given throughout my PhD. Thank you, John, for consistently sharing your enthusiasm for science, and a special thank you Dr. Tobi Moll for your ongoing mentorship and readiness to lend a helping hand.

A big thank you to the E20 lab and the RNA group, with special thanks to my advisors Dr. Emma Thomson and Dr. Ian Sudbery for their academic and pastoral support and timely advice. I would also like to thank Dr. Ang Li for the advice over the years, and Dr. Pete Daniels for all the laughs and nonsense.

I would like to thank my DiMeN DTP cohort — it has been a joy to share this PhD journey with all of you, from our Google Meet introduction in 2020 to now. Special thanks to Dr. Callum Portman Ross for always being my partner in crime.

Thank you to the Sheffield Bioinformatics core for their technical assistance throughout my PhD.

I would like to thank Dr David Turton, without your guidance I would've never taken a chance on research. Thank you for always believing in me.

I would like to thank Jack Poole for putting up with the late nights and weekends stolen by tissue culture. Thanks for all the coffees and pep talks. You have always encouraged me to chase my ambitions and strive for the best. Above all, I'm sorry that I never grew an ear on a mouse like you wanted.

To my family — thank you for your constant support. Dad, thank you for your encouragement to pursue science and always sharing your scientific insight from the newest podcast you're listening to. Mum, thank you for your encouragement and always lending an ear. Emma, thank you for the late-night phone calls and always pushing me to dream big. Matthew, thank you for being quintessentially you and providing a much-needed distraction.

I would also like to thank my friends, near and far, for their ongoing support and for providing an often much-needed escape. I have made some incredible friends during my PhD experience. I would especially like to thank Mark Gainford, Tayla Shakespeare and Sam Lewin. Thank you, Mark, for your unwavering support - you've been there for me since day one, you're my rock and I couldn't have done this without you. Thank you, Tayla, for always showing me the best qualities in myself, for the laughter (and the tears), and always making time for me. Thank you, Sam, for always helping me see the best in a bad situation, you unbreakable optimism is admirable and I treasure our weekend walks.

Lastly, my sincere thanks to Dr. Emily Brookes and Professor Janine Kirby for taking the time to examine my thesis.

# *Abstract*

Spatial interactions between transcriptional regulatory elements establish gene regulatory networks. Enhancers are fundamental to such regulatory networks, conferring precise spatiotemporal gene expression patterns. Thus, it is not surprising that variation within enhancers may underpin the genetic architecture of disease. In fact, the vast majority of single nucleotide polymorphisms (SNPs) associated with human complex diseases map to the non-coding genome, with causal disease variants particularly enriched within enhancers. The transcription of enhancer RNAs (eRNAs), the non-coding RNA species encoded from active enhancers, is a key factor contributing to enhancer activity. Not only may enhancer transcription data provide insight into how enhancer activity changes upon mutation, but enhancers containing disease-associated variants may yield aberrant eRNAs capable of disrupting transcriptional regulation to culminate in disease phenotype.

By mapping putative Amyotrophic Lateral Sclerosis (ALS)-associated enhancers, we evaluate their activity by assessing chromatin accessibility, enhancer-associated chromatin modifications, and eRNA transcription, in the context of disease. By employing epigenomic and transcriptomic analyses, coupled with reporter assays to assess the impact of ALS-associated SNPs, we gain insights into the transcriptional and regulatory consequences of disease-associated SNPs on enhancer function. Using computational RNA structure prediction, enhancers containing ALS-associated variants are predicted to yield eRNAs with altered structure, and therefore potentially function. In disease contexts, we further explore how both distal and intragenic ALS-associated enhancers affect not only gene expression but differential messenger RNA (mRNA) isoform expression. Therefore, investigating the effects of ALS-associated SNPs on enhancer activity, eRNA expression and eRNA structure not only provide great insights into the pathogenic mechanisms underlying sporadic ALS, but may also contribute to a broader understanding of how eRNAs function independently from their associated enhancers.

# List of Figures


1.1 Waddington's epigenetic landscape . . . 3
1.2 Multi-level control of the the central dogma . . . 4
1.3 TADs as units of higher-order genome organisation . . . 5
1.4 Enhancers are fundamental to gene regulatory networks . . . 7
1.5 Enhancer architecture is fundamental to enhancer potential . . . 8
1.6 Histone modifications can differentiate regulatory elements . . . 10
1.7 Active enhancers are transcribed into heterogeneous non-coding RNAs . . . 12
1.8 eRNA interactions with chromatin modifying enzymes at enhancers . . . 15
1.9 R-loops and DNA G-quadruplexes at enhancers . . . 19
1.10 Model of eRNA-mediated control of transcriptional condensate formation 21
1.11 3' UTR functionality is controlled by *cis*-regulatory elements . . . 24
1.12 The role of APA in controlling gene expression . . . 26
1.13 Complex disease variants map to *cis*-regulatory elements . . . 27
1.14 Integrating GWAS associations with eQTLs may provide power to understanding regulatory genotypes underpinning disease . . . 29
1.15 Enhancer variants can alter chromatin architecture . . . 32
1.16 Enhancer variants can disrupt TF binding sites . . . 33
1.17 Enhancer variants introduce *de novo* TF binding sites . . . 35
1.18 Rare variant burden analysis pipeline . . . 41
1.19 Chromosomal location of ALS-associated enhancer variants relative to the *CAV1/CAV2* locus . . . 42
1.20 ALS-associated rare variants are enriched in *CAV1/CAV2* enhancers . . . 43

3.1 Chromosome topology around the *CAV1/CAV2* gene locus in *in vitro* differentiated motor neurons . . . 98
3.2 CTCF binds at TAD boundaries . . . 100
3.3 Chromosome topology of the *CAV1/CAV2* regulatory network in *in vitro* differentiated motor neurons . . . 101
3.4 Candidate *CAV1/CAV2* enhancers largely localise to the *CAV1/CAV2* TAD . . . 102
3.5 Candidate *CAV1/CAV2* enhancers localise proximal to TAD boundaries . 103
3.6 ALS-associated SNPs in *CAV1/CAV2* candidate enhancers cluster proximal TAD boundaries . . . 105
3.7 ALS-associated SNPs within *CAV1/CAV2* enhancers largely localise to the *CAV1/CAV2* TAD . . . 106
3.8 The relationship between inter- and intra-TAD enhancer location and number of ALS-associated SNPs in enhancers . . . 108

3.9 Localisation of *CAV1/CAV2* enhancers relative to the *CAV1* transcription start site . . . 110
3.10 Frequency of *CAV1/CAV2* enhancers and ALS-associated SNPs relative to genomic features . . . 112

4.1 ATAC-seq workflow . . . 121
4.2 CUT&RUN workflow . . . 123
4.3 Relative distance between peaks in SH-SY5Y ATAC-seq publicly available ATAC-seq . . . 125
4.4 Chromatin accessibility at candidate *CAV1/CAV2* enhancers in SH-SY5Y cells . . . 126
4.5 Chromatin accessibility at candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons . . . 128
4.6 *CAV1* and *CAV2* gene expression in iPSC-derived motor neurons . . . 129
4.7 Comparison of chromatin accessibility at candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons and SH-SY5Y cells . . . 130
4.8 Number of ALS-associated SNPs in candidate *CAV1/CAV2* enhancers in open and closed chromatin states . . . 132
4.9 Relative distance between peaks in H3K27ac CUT&RUN and publicly available ChIP-seq . . . 134
4.10 Enrichment of H3K27ac at candidate *CAV1/CAV2* enhancers in SH-SY5Y cells . . . 135
4.11 Enrichment of H3K4me1 at candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons . . . 137
4.12 Enrichment of H3K27ac at candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons . . . 138
4.13 Characterisation of candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons . . . 139
4.14 Characterisation of active *CAV1/CAV2* enhancers in Amyotrophic Lateral Sclerosis . . . 141
4.15 Characteristic enhancer-associated histone modifications at the *CAV1/CAV2* proximal enhancer in iPSC-derived motor neurons . . . 142
4.16 Enhancer-associated histone modifications are enriched at ALS-associated *CAV1/CAV2* enhancers . . . 143
4.17 Number of ALS-associated SNPs in candidate *CAV1/CAV2* enhancers enriched for histone modifications characteristic of enhancers . . . 144
4.18 Comparison of enhancer-associated characteristics at candidate *CAV1/CAV2* enhancers in SH-SY5Y cells and patient-derived motor neurons . . . 147

5.1 TT-seq workflow . . . 153
5.2 Dynamics of RNA structure in the free energy landscape . . . 155
5.3 Gene count differences for TT-seq data before and after spike-in normalisation . . . 157
5.4 Correlation between WT SH-SY5Y TT-seq replicates . . . 158
5.5 TT-seq captures more unstable intronic transcripts than total RNA-seq . 159
5.6 TT-seq reads map to the non-coding genome . . . 159
5.7 TT-seq coverage over ABC enhancer predictions . . . 161
5.8 TT-seq coverage over shuffled control regions . . . 162

5.9 Comparison of TT-seq coverage over ABC enhancer predictions and shuffled control regions . . . 163
5.10 TT-seq identifies transcription from enhancers . . . 164
5.11 eRNA transcription at *CAV1/CAV2* candidate enhancers in SH-SY5Y cells 165
5.12 Transcribed *CAV1/CAV2* candidate enhancers contain ALS-associated SNPs . . . 166
5.13 The relationship between chromatin accessibility at candidate *CAV1/CAV2* enhancers and eRNA transcription . . . 168
5.14 The relationship between H3K27ac deposition at candidate *CAV1/CAV2* enhancers and eRNA transcription . . . 170
5.15 The relationship between H3K27ac deposition at candidate *CAV1/CAV2* enhancers and eRNA transcription . . . 171
5.16 eRNA transcript assembly in WT SH-SY5Y cells . . . 173
5.17 MFEden of shuffled sequences with given GC-content . . . 176
5.18 eRNA structure and stability predictions for the $^{DRR}$eRNA . . . 179
5.19 eRNA structure and stability predictions for the eRNAs produced by the *TAL1* super-enhancer . . . 180
5.20 Native eRNA MFE predictions compared to shuffled randomised control MFE values for example eRNAs . . . 181
5.21 eRNA structure and stability predictions for the novel ALS-associated antisense eRNA in SH-SY5Y cells . . . 183
5.22 eRNA MFE predictions compared to shuffled randomised control for native SH-SY5Y antisense eRNA and eRNA with induced ALS-associated SNP . . . 184
5.23 eRNA transcription at *CAV1/CAV2* enhancers in human-derived lymphoblastoid cells . . . 186
5.24 Levels of nascent transcription over candidate *CAV1/CAV2* enhancers in lymphoblastoid cells . . . 187
5.25 eRNA transcription at ALS-associated *CAV1/CAV2* enhancers in lymphoblastoid cells . . . 188
5.26 eRNA transcript assembly in lymphoblastoid cell lines . . . 189
5.27 eRNA transcription at healthy control-associated enhancers . . . 190
5.28 eRNA structure and stability predictions for eRNAs identified only in healthy control-derived lymphoblastoid cells . . . 191
5.29 MFE predictions of native eRNAs assembled from SC3504 TT-seq data compared to shuffled randomised control MFE values . . . 192
5.30 TT-seq coverage at healthy control-associated enhancers . . . 193
5.31 eRNA transcription at ALS patient-associated enhancers . . . 195
5.32 eRNA structure and stability predictions for eRNAs identified only in ALS patient-derived lymphoblastoid cells . . . 196
5.33 eRNA structure and stability predictions for eRNAs identified only in ALS patient-derived lymphoblastoid cells . . . 197
5.34 MFE predictions of native eRNAs assembled from ALS patient-derived lymphoblastoid TT-seq data compared to shuffled randomised control MFE values . . . 198
5.35 TT-seq coverage at ALS patient-associated enhancers . . . 199

5.36 eRNA structure and stability predictions for the sense eRNA identified only in ALS patient-derived lymphoblastoid cells containing the chr7:115,994,269:C>T mutation . . . 200
5.37 MFE predictions of native ALS patient-specific eRNA assembled from SP3388 TT-seq data compared to shuffled randomised control MFE values 201

6.1 CAV1 in APP proteolysis . . . 216
6.2 CRISPRi enables repression of gene expression via targeting of *cis*-regulatory elements . . . 219
6.3 Enhancer characteristics at the *CAV1/CAV2* proximal enhancer in iPSC-derived motor neurons . . . 221
6.4 Enhancer characteristics at the *CAV1/CAV2* proximal enhancer in SH-SY5Y cells . . . 222
6.5 eRNA transcription at the *CAV1/CAV2* proximal enhancer in lymphoblastoid cells . . . 223
6.6 Epigenomic landscape surrounding proximal and distal model *CAV1/CAV2* enhancers . . . 225
6.7 dCas9-Zim3 SH-SY5Y cells express Cas9 . . . 226
6.8 CRISPRi of candidate *CAV1* enhancers resulted in reduced *CAV1* expression . . . 227
6.9 The impact of *CAV1* enhancer knockout on *CAV1* expression . . . 229
6.10 *CAV1* transcription characteristics in lymphoblastoid cells . . . 231
6.11 *De novo* transcription in SH-SY5Y cells at the *CAV1* gene locus . . . 233
6.12 SH-SY5Y cells show preference for the proximal polyA site . . . 234
6.13 The impact of *CAV1* enhancer knockout on *CAV1* eRNA expression . . . 235
6.14 *De novo* transcription in lymphoblastoid cells at the *CAV1* gene locus . . . 237
6.15 *CAV1* eRNA expression in lymphoblastoid cells . . . 238
6.16 Enrichment of H3K27ac CUT&RUN signal across the model proximal *CAV1* enhancer . . . 240
6.17 The impact of CBP inhibition on *CAV1* mRNA and eRNA expression . . . 241
6.18 ALS-associated SNPs alter the regulatory function of the model proximal *CAV1/CAV2* enhancer . . . 244
6.19 MEF2 binding motifs within the model proximal *CAV1* enhancer . . . 246
6.20 CTCF binding at the model distal *CAV1* enhancer . . . 247
6.21 FOXO1::ETS binding motifs within the model distal *CAV1* enhancer . . . 248

7.1 *CAV1* transcript length increases with age . . . 264
7.2 *CAV1* expression decreases with *CAV1* transcript length in ALS patients 265
7.3 AU-rich element in the *CAV1* 3' UTR is disrupted by ALS-associated SNP 266
7.4 Isolated *CAV1* 3' UTR expression in SH-SY5Y and lymphoblastoid cell lines . . . 267
7.5 Isolated *CAV1* 3' UTR expression in iPSC-derived motor neurons . . . 268
7.6 Isolated *CAV1* 3' UTR expression in tissue . . . 269
7.7 Genomic location of ABC candidate enhancers . . . 271
7.8 T7E1 assay for gRNA validation in HEK293FT cells . . . 273
7.9 T7E1 assay for gRNA validation in SH-SY5Y cells . . . 274
7.10 Plasmid maps of HDR templates following whole plasmid sequencing . . . 276

7.11 Successful incorporation of the SV40 PAS into *CAV1* sequence in HEK293FT cells . . . . . 277

8.1 Models of enhancer-promoter interaction . . . . . 287

# List of Tables


2.1 Plasmids utilised for doctoral work . . . 50
2.2 Composition of 50$\mu$l PCR reaction . . . 51
2.3 Cycling conditions for PCR reaction . . . 51
2.4 Sanger sequencing primers . . . 53
2.5 Site-directed mutagenesis primers and PCR conditions Primers designed using NEBaseChanger . . . 54
2.6 Oligonucleotides for cloning sgRNAs into lentiGuide-Puro vectors for generation of CRISPRi SH-SY5Y cell lines . . . 56
2.7 Oligonucleotides for cloning sgRNAs into pX459 for generation of genetically modified SH-SY5Y cell lines . . . 58
2.8 Site-directed mutagenesis primers and PCR conditions . . . 59
2.9 Mammalian cell lines utilised for doctoral work . . . 60
2.10 Clinical information of healthy control subjects and ALS patients from whom lymphoblastoid cell lines were derived . . . 63
2.11 Antibodies used for Western Blotting . . . 68
2.12 Primers used for gRNA validation . . . 71
2.13 Primers used for RT-qPCR . . . 74
2.14 Cycling conditions for RT-qPCR reaction . . . 75
2.15 Antibodies used for CUT&RUN . . . 78
2.16 Clinical information on both healthy control subjects and ALS patients from whom iPSCs were derived . . . 87
2.17 Sources of published datasets . . . 89

5.1 Predicted stability of $^{DRR}$eRNA . . . 179
5.2 Predicted stability of *TAL1* eRNAs . . . 180
5.3 Predicted stability of ALS-associated eRNAs in SH-SY5Y cells . . . 183
5.4 Predicted stability of eRNAs in healthy control-derived lymphoblastoid cells . . . 192
5.5 Predicted stability of upstream eRNAs in ALS patient-derived lymphoblastoid cells . . . 196
5.6 Predicted stability of downstream eRNAs in ALS patient-derived lymphoblastoid cells . . . 197
5.7 Predicted stability of ALS-associated eRNAs in ALS patient-derived lymphoblastoid cells . . . 200

B.1 Enhancer regions associated with *CAV1* and *CAV2* expression, and tissue associations . . . 309
B.2 ALS-associated SNPs within candidate *CAV1/CAV2* enhancer regions . . . 313

# Nomenclature

$3'UTR$ 3' untranslated region

$3C$ Chromosome conformation capture

$3D$ Three-dimensional

$4sU$ 4-thiouridine

$A\beta$ Amyloid-$\beta$

$ABC$ Activity-By-Contact

$AD$ Alzheimer's disease

$ALS$ Amytrophic lateral sclerosis

$APA$ Alternative polyadenylation

$APP$ Amyloid-beta precursor protein

$AR-eRNA$ Androgen receptor-regulated eRNA

$ARE$ AU-rich element

$ATAC-seq$ Assay for Transposase-Accessible Chromatin using sequencing

$BBB$ Blood-brain barrier

$BCA$ Bicinchoninic Acid

$BDNF$ Brain-derived neurotrophic factor

$Bvht$ Braveheart

$CAV1$ Caveolin-1

$CAV2$ Caveolin-2

$CBP$ CREB-binding protein

$CDK9$ Cyclin-dependent kinase 9

$CDS$ Coding sequence

$ChIP$ Chromatin immunoprecipitation

$ChIP-seq$ Chromatin immunoprecipitation followed by sequencing

$CNS$ Central nervous system

$CRISPRi$ CRISPR interference

$CTCF$ CCCTC-binding factor

$CTD$ C-terminal domain

$CUT\&RUN$ Cleavage under targets and release using nuclease

$DMEM$ Dulbecco's Modified Eagle's Medium

$DNase-seq$ DNase I hypersensitive sites sequencing

$DSIF$ DRB sensitvity inducing factor

$DTT$ Dithiothreitol

$EMSA$ Electrophoretic mobility shift assay

$eQTL$ Expression quantitative trait loci

$eRNAs$ Enhancer RNAs

$ESCs$ Embryonic stem cells

$ETS$ E26 transformation specific

$ExAC$ Exome aggregation consortium

$FBS$ Foetal bovine serum

$FISH$ Fluorescence in situ hybridization

$FTD$ Frontotemporal dementia

$G-quadruplex$ Guanine quadruplex

$Gbp$ Giga base pair

$GWAS$ Genome-wide association studies

$H3K122ac$ Acetylation of lysine 122 on histone H3

$H3K27ac$ Acetylation of lysine 27 on histone H3

$H3K27me3$ Trimethylation of lysine 27 on histone H3

$H3K4me1$ Monomethylation of lysine 4 on histone H3

$H3K4me3$ Trimethylation of lysine 4 on histone H3

$H3K64ac$ Acetylation of lysine 64 on histone H3

$H4K16ac$ Acetylation of lysine 64 on histone H3

$HD$ Huntington's disease

$HDR$ Homology-directed repair

$HSD$ Honestly Significant Difference

$HSF1$ Heat shock factor-1

$HuR$ Human antigen R

$IEG$ Immediate early gene

$Indel$ Insertion-deletion

$LB$ Luria broth

$m6A$ N6-methyladenosine

$m^7G$ 7-methylguanylate

$Mbp$ Mega base pair

$MF$ Maximum fluorescence

$MFEden$ Minimum free energy density

$miRNA$ microRNA

$MNDA$ Motor Neurone Disease Association

$mRNA$ Messenger RNA

$MTC$ METTL3/METTL14/WTAP m6A methyltransferase complex

$ncRNAs$ Non-coding RNAs

$NELF$ Negative elongation factor

$NeuN$ Neuronal nuclear antigen

$NHEJ$ Non-homologous end joining

$P-TEFb$ Positive transcription elongation factor b

$pAG-MNase$ Protein A-protein G-micrococcal nuclease

$PAS$ Polyadenylation site

$PCR$ Polymerase chain reaction

$PD$ Parkinson's disease

$PIC$ Preinitiation complex

$PNK$ Polynucleotide kinase

$poly(A)$ Polyadenosine

$PRC2$ Polycomb repressive complex 2

$PTM$ Post-translational modification

$PWM$ Positional weight matrix

$RA$ rheumatoid arthritis

$RBP$ RNA-binding protein

$RNAPII$ RNA polymerase II

$RNP$ Ribonucleoprotein

$ROS$ Reactive oxygen species

$RPMI$ Roswell Park Memorial Institute

$RRM$ RNA recognition motif

$rSAP$ Shrimp Alkaline Phosphatase

$RT-qPCR$ Reverse transcription quantitative real-time PCR

$SDS$ Sodium dodecyl sulfate

$sgRNA$ Single guide RNA

$siRNA$ Small interfering RNA

$SLAM-seq$ Thiol(SH)-linked alkyation for the metabolic sequencing of RNA

$SNP$ Single nucleotide polymorphism

$SOD1$ $Cu^2+/Zn^2+$ superoxide dismutase

$ss$ Single-stranded

$SV40$ Simian virus 40

$T2DM$ Type-2 diabetes mellitus

$T7E1$ T7 endonuclease I

$TAD$ Topologically associating domain

$TAE$ Tris-acetate-EDTA

$TAR-L$ TAR RNA-like

$TF$ Transcription factor

$TLR4$ Toll-like receptor 4

$TSS$ Transcription start site

$TT-seq$ Transient transcriptome sequencing

$TT-seq$ Transient transcriptome sequencing

$TTP$ Tristetraprolin

$TTS$ Transcription termination site

$WT$ Wildtype

$YY1$ Yin yang 1

$Zim3-dCas9$ pHR-UCOE-SFFV-Zim3-dCas9-P2A-Hygro

# Contents

**Acknowledgements** i

**Abstract** iii

**List of Figures** iv

**List of Tables** ix

**Nomenclature** x

**1 Introduction** 1

1.1 Genomes: From coding to non-coding . . . . . 1

1.1.1 Increase in non-coding DNA as a function of organismal complexity 1

1.1.2 The non-coding genome: Home to regulatory elements . . . . . 3

1.2 The three-dimensional genome . . . . . 4

1.3 Transcriptional control: As determined by regulatory elements . . . . . 6

1.4 Enhancers: key regulators of transcription . . . . . 6

1.5 Histone modifications as identifiers of regulatory elements . . . . . 9

1.6 eRNAs: Functions, sequence and structure . . . . . 11

1.6.1 Characterisation of eRNAs . . . . . 12

1.6.2 The effects of eRNAs on the chromatin environment . . . . . 13

1.6.2.1 eRNAs activate histone acetylation . . . . . 13

1.6.2.2 The effects of eRNAs on repressive histone methylation . 13

1.6.3 eRNAs as direct transcriptional regulators . . . . . 15

1.6.3.1 Effects on RNA polymerase II pause release . . . . . 15

1.6.3.2 eRNAs and transcription factor binding . . . . . 16

1.6.4 The role of eRNA structure in regulating eRNA function . . . . . 17

1.7 Transcriptional condensates and eRNAs . . . . . 18

1.7.0.1 Post-transcriptional modification of eRNAs . . . . . 21

1.8 3' UTRs: The fine-tuners of post-transcriptional control . . . . . 22

1.8.1 Cleavage and polyadenylation . . . . . 23

1.8.2 3' UTR function: mRNA stability, sublocalisation and translational control . . . . . 23

1.8.3 Alternative polyadenylation: The relevance of 3' UTR isoforms . . 25

1.9 The non-coding genome and disease: Regulatory elements matter . . . . . 25

1.9.1 Genome Wide Association Studies . . . . . 25

1.9.2 Enhancers in disease . . . . . 28

1.9.2.1 Alterations in chromatin architecture . . . . . 30
1.9.2.2 Alterations in transcription factor binding . . . . . 31
1.9.2.2.1 Disruption of TF binding sites . . . . . 31
1.9.2.2.2 Introduction of *de novo* TF binding sites . . . . 34
1.9.3 eRNAs in disease . . . . . 35
1.9.3.1 Transcription factor binding impacts eRNA-mediated pathogenesis . . . . . 35
1.9.3.2 Neurotoxic protein aggregate formation as mediated by RNA structure . . . . . 36
1.9.4 3' UTRs in disease . . . . . 36
1.10 Amyotrophic Lateral Sclerosis: A complex disease underpinned by non-coding genetic variants . . . . . 37
1.10.1 ALS as a complex disease with little known heritability . . . . . 37
1.10.1.1 Functional annotation of GWAS datasets . . . . . 38
1.10.1.2 Rare variant burden analysis . . . . . 40
1.10.2 Caveolin-1: Function and disease relevance . . . . . 44
1.10.2.1 Caveolin-1 in the DNA damage response . . . . . 45
1.11 Hypotheses . . . . . 46
1.12 Aims and motivations of this research . . . . . 47

**2 Materials and Methods 49**
2.1 Molecular cloning . . . . . 49
2.1.1 Buffers and media . . . . . 49
2.1.2 Molecular biology kits . . . . . 49
2.1.3 Plasmids . . . . . 50
2.1.4 Polymerase chain reaction (PCR) . . . . . 50
2.1.5 Agarose gel electrophoresis . . . . . 51
2.1.6 Transformation of bacterial cells . . . . . 51
2.1.7 Plasmid purification and Sanger sequencing . . . . . 52
2.1.8 Preparation of plasmids for luciferase reporter assay . . . . . 54
2.1.8.1 Site-directed mutagenesis . . . . . 54
2.1.8.2 Cloning enhancer DNA into luciferase reporter plasmid . 55
2.1.9 Preparation of plasmids for CRISPR interference . . . . . 55
2.1.10 Preparation of plasmids for CRISPR gRNA validation and generation of genetically modified cell line . . . . . 58
2.1.11 Preparation of donor template for generation of genetically modified cell line . . . . . 58
2.1.11.1 Generation of donor template for induction of ALS-associated SNPs . . . . . 59
2.2 Mammalian Cell culture . . . . . 60
2.2.1 Cell lines . . . . . 60
2.2.2 Sources of healthy control- and patient-derived cell lines . . . . . 62
2.2.2.1 Ethical permissions . . . . . 63
2.2.3 Maintenance of mammalian cells . . . . . 63
2.2.3.1 SH-SY5Y cell culture . . . . . 63
2.2.3.2 HEK293FT cell culture . . . . . 63
2.2.3.3 Lymphoblastoid cell culture . . . . . 64

2.2.3.4 CRISPRi SH-SY5Y cell culture . . . 64
2.2.4 Routine culturing conditions . . . 64
2.2.4.1 Subculturing SH-SY5Ys and their derivative cell lines . . 64
2.2.4.2 Subculturing HEK293FT cells . . . 65
2.2.4.3 Subculturing lymphoblastoid cell lines . . . 65
2.2.5 Small molecule treatment . . . 65
2.2.5.1 CBP/p300 inhibition . . . 65
2.3 Transfection and transduction of plasmid DNA . . . 65
2.3.1 Electroporation-based transfection . . . 65
2.3.1.1 Transfection of SH-SY5Y cells for luciferase reporter assays 66
2.3.1.2 Transfection of SH-SY5Y cells for generation of a modified cell line . . . 66
2.3.2 Lentiviral packaging . . . 66
2.3.3 Lentiviral concentration . . . 67
2.3.4 Transduction of SH-SY5Y cells . . . 67
2.3.5 Antibiotic selection . . . 67
2.4 Validation of stable expression of Zim3-dCas9 in SH-SY5Y cells . . . 68
2.4.1 Western blot . . . 68
2.4.1.1 Buffers . . . 68
2.4.1.2 Antibodies . . . 68
2.4.1.3 Cell harvest and bicinchoninic acid (BCA) protein assay . 69
2.4.1.4 SDS-PAGE electrophoresis . . . 69
2.4.1.5 Protein transfer . . . 69
2.4.1.6 Membrane probing and visualisation . . . 69
2.5 Validation of gRNAs for generation of genetically modified cell lines . . . 70
2.5.1 Liposome-based transfection . . . 70
2.5.2 gRNA validation . . . 70
2.5.2.1 T7 endonuclease I assay . . . 70
2.6 Luciferase reporter assays . . . 71
2.6.1 Visualisation of data . . . 72
2.7 Quantitative reverse transcription polymerase chain reaction (RT-qPCR) 72
2.7.1 RNA extraction . . . 72
2.7.2 Reverse transcription of RNA . . . 73
2.7.3 RT-qPCR . . . 73
2.7.4 Visualisation of data . . . 75
2.8 Hi-C . . . 75
2.8.1 Bioinformatic analysis . . . 75
2.8.2 Liftover of genomic annotations . . . 76
2.8.3 Visualisation of data . . . 76
2.9 Assay for transposase-accessible chromatin with sequencing (ATAC-seq) . 76
2.9.1 Generation of ATAC-seq libraries . . . 76
2.9.2 Bioinformatic analysis . . . 76
2.9.2.1 Quality control and adapter trimming . . . 76
2.9.2.2 Alignment . . . 77
2.9.2.3 Post-alignment quality control . . . 77
2.9.2.4 Peak calling . . . 77
2.9.3 Validation of data . . . 77

2.9.4 Comparison to published datasets . . . 77
2.10 Cleavage under targets and release using nuclease (CUT&RUN) sequencing 78
2.10.1 Generation of CUT&RUN libraries . . . 78
2.10.2 Bioinformatic analysis . . . 78
2.10.2.1 Quality control and adapter trimming . . . 78
2.10.2.2 Alignment . . . 78
2.10.2.3 Post-alignment quality control . . . 79
2.10.2.4 Peak calling . . . 79
2.10.2.5 Validation of data . . . 79
2.11 Chromatin immunoprecipitation sequencing (ChIP-seq) . . . 79
2.11.1 Use of published datasets . . . 79
2.12 Transient transcriptome sequencing (TT-seq) . . . 80
2.12.1 Generation of TT-seq libraries . . . 80
2.12.1.1 4sU labelling and extraction of 4sU-labelled RNA . . . 80
2.12.1.2 RNA fragmentation . . . 81
2.12.1.3 Biotinylation of 4sU-labelled RNA . . . 81
2.12.1.4 Diagnostic dot blot . . . 81
2.12.1.5 Streptavidin pull-down of biotinylated 4sU-labelled RNA 82
2.12.1.6 Library preparation and next generation sequencing (NGS) 83
2.12.2 Bioinformatic analysis . . . 83
2.12.2.1 Quality control . . . 83
2.12.2.2 Unique molecular identifier (UMI) extraction . . . 83
2.12.2.3 Adapter trimming . . . 83
2.12.2.4 Alignment . . . 83
2.12.2.5 Post-alignment quality control and normalisation . . . 84
2.12.2.6 Transcript assembly . . . 84
2.12.2.7 RNA structure prediction . . . 85
2.13 RNA-sequencing (RNA-seq) . . . 85
2.14 Cap analysis of gene expression sequencing (CAGE-seq) . . . 86
2.15 Visualisation of next-generation sequencing data . . . 86
2.16 Computational genomics . . . 87
2.16.1 Published datasets . . . 87
2.16.1.1 Clinical information . . . 87
2.16.1.2 Published dataset sources . . . 88
2.16.2 Calculating read coverage over genomic features . . . 90
2.16.2.1 Visualisation of data . . . 90
2.16.3 Per gene exonic versus intronic read coverage . . . 90
2.17 Identification of polyadenylation sites . . . 91
2.17.1 Transcription factor binding motifs . . . 91
2.17.2 Identification of AU-rich motifs . . . 91
2.17.3 GNF atlas . . . 92

**3 Genome architecture of the *CAV1/CAV2* enhancer landscape: Structural insights into disease 93**
3.1 Introduction . . . 93
3.1.1 Mapping enhancer-promoter interactions . . . 93
3.1.2 The Activity-By-Contact model . . . 94

3.1.3 Enhancer cooperativity . . . 95
3.1.4 Chapter aims . . . 96
3.2 Chromosome topology of the *CAV1/CAV2* regulatory network . . . 97
3.2.1 *CAV1* and *CAV2* are likely co-regulated . . . 97
3.2.2 CTCF binds at TAD boundaries . . . 98
3.2.3 *CAV1/CAV2* candidate enhancers localise within the *CAV1/CAV2* TAD . . . 101
3.2.4 Intra-TAD enhancer localisation . . . 102
3.2.5 Enhancer location and disease relevance . . . 104
3.2.5.1 Enhancer density and SNP localisation . . . 106
3.3 Genomic location of candidate *CAV1/ CAV2* enhancers and ALS-associated SNPs . . . 109
3.3.1 Candidate *CAV1/CAV2* enhancers localise proximal to the *CAV1* transcription start site . . . 109
3.3.2 ALS-associated SNPs localise proximal to the *CAV1* transcription start site . . . 109
3.3.3 Tissue-specificity of enhancers and genomic location . . . 111
3.3.4 Genomic location of ALS-associated SNPs . . . 111
3.4 Discussion . . . 113
3.4.1 Chromatin architecture of the *CAV1/CAV2* enhancer landscape . 113
3.4.1.1 Co-regulation of *CAV1* and *CAV2* . . . 113
3.4.1.2 TADs and the implications for long-range enhancer-promoter interactions . . . 113
3.4.1.3 Disease relevance of chromatin architecture . . . 115
3.4.2 Conclusion . . . 117

**4 Chromatin signatures of *CAV1/CAV2* enhancer landscape: Epigenetic regulation and disease implications 119**
4.1 Introduction . . . 119
4.1.1 Methodologies for enhancer identification . . . 119
4.1.1.1 Mapping of accessible chromatin . . . 120
4.1.1.2 Mapping of enhancer characteristics . . . 121
4.1.2 Chapter aims . . . 123
4.2 Chromatin accessibility at enhancers . . . 124
4.2.1 Chromatin accessibility of enhancers in SH-SY5Y cells . . . 124
4.2.1.1 Validation of ATAC-seq in SH-SY5Y cells . . . 124
4.2.1.2 Chromatin accessibility of *CAV1/CAV2* enhancers in SH-SY5Y cells . . . 125
4.2.2 Chromatin accessibility of enhancers in iPSC-derived motor neurons 127
4.2.3 Comparatative analysis of chromatin accessibility at *CAV1/CAV2* enhancers in SH-SY5Y cells and iPSC-derived motor neurons . . . 129
4.2.4 Chromatin accessibility at *CAV1/CAV2* enhancers containing ALS-associated SNPs . . . 131
4.3 Mapping enhancer-associated characteristics . . . 133
4.3.1 Enhancer-associated characteristics in SH-SY5Y cells . . . 133
4.3.1.1 Validation of CUT&RUN . . . 133
4.3.1.2 Enhancer-associated characteristics of *CAV1/CAV2* enhancers in SH-SY5Y cells . . . 134

4.3.2 Enhancer-associated characteristics in iPSC-derived motor neurons . . . 136
4.3.2.1 Characterisation of *CAV1/CAV2* enhancers in iPSC-derived motor neurons . . . 139
4.3.2.2 Enhancer characteristics in Amyotrophic Lateral Sclerosis 140
4.3.2.3 Enhancer characteristics at *CAV1/CAV2* enhancers containing ALS-associated SNPs . . . 143
4.3.2.4 Comparison of enhancer characteristics between SH-SY5Y cells and iPSC-derived motor neurons . . . 144
4.4 Discussion . . . 148
4.4.1 Chromatin accessibility at *CAV1/CAV2* enhancers . . . 148
4.4.2 Enhancer-associated histone modifications at *CAV1/CAV2* enhancers . . . 148
4.4.3 Comparative analysis of *CAV1/CAV2* enhancer activity in neuronal-like model systems . . . 150
4.4.4 Conclusion . . . 151

**5 Characterisation of *CAV1* enhancer transcriptional activity: eRNAs and disease 152**
5.1 Introduction . . . 152
5.1.1 Mapping enhancer transcription . . . 152
5.1.2 RNA structure . . . 154
5.1.2.1 RNA structure prediction . . . 155
5.1.3 Chapter aims . . . 156
5.2 Using nascent transcription data to identify active enhancers . . . 156
5.2.1 TT-seq normalisation and replicate correlation . . . 156
5.2.2 TT-seq is capable of capturing transient non-coding transcripts . . . 158
5.2.3 TT-seq captures enhancer transcription . . . 159
5.2.4 eRNA transcription at candidate *CAV1/CAV2* enhancers . . . 164
5.2.4.1 eRNA transcription at candidate *CAV1/CAV2* enhancers containing ALS-associated SNPs . . . 165
5.3 Integration of eRNA transcription data with other enhancer characteristics 166
5.3.1 eRNA transcription and chromatin accessibility . . . 167
5.3.2 eRNA transcription and histone modification . . . 169
5.4 Identification of novel eRNAs in SH-SY5Y cells . . . 171
5.5 eRNA structure and stability prediction . . . 174
5.5.1 Minimum free energy of RNA secondary structures . . . 174
5.5.1.1 Normalisation of MFE . . . 174
5.5.2 Thermodynamic stability prediction . . . 176
5.5.3 Comparison of predicted MFE values between native and shuffled sequences . . . 177
5.5.4 RNA structure and stability prediction of example eRNAs . . . 178
5.5.5 RNA structure and stability prediction of eRNAs in SH-SY5Y cells to model ALS-associated SNPs . . . 181
5.6 eRNA transcription in human-derived lymphoblastoid cells . . . 184
5.7 Identification of novel enhancer RNAs in human-derived lymphoblastoid cells . . . 188
5.7.1 eRNA transcription in healthy control-derived lymphoblastoid cells 189

5.7.2 eRNA transcription in ALS patient-derived lymphoblastoid cells . 194
5.7.2.1 ALS patient-specific assembled eRNA transcripts . . . . . 199
5.8 Discussion . . . . . 202
5.8.1 TT-seq as means to capture eRNA transcription . . . . . 202
5.8.2 Integration of eRNA transcription data with additional enhancer epigenomic features . . . . . 202
5.8.3 Assembly of *de novo* eRNA transcripts . . . . . 205
5.8.4 eRNA structure and stability prediction . . . . . 206
5.8.5 Predicted structure and stability of benchmark eRNAs . . . . . 207
5.8.6 The impact of ALS-associated SNPs on predicted eRNA structure and stability . . . . . 208
5.8.7 eRNAs in human-derived lymphoblastoid cells . . . . . 209
5.8.7.1 eRNAs in healthy control-derived lymphoblastoid cells . . 209
5.8.7.2 eRNAs in ALS patient-derived lymphoblastoid cells . . . 211
5.8.7.3 Genotype-specific eRNA structure and stability in ALS . 212
5.8.8 Conclusion . . . . . 212

**6 How do Amyotrophic Lateral Sclerosis SNPs impact regulatory activity? 214**
6.1 Introduction . . . . . 214
6.1.1 CAV1 neuroprotective function . . . . . 214
6.1.2 Caveolin-1 isoforms . . . . . 216
6.1.3 The pathological impacts of *CAV1* enhancer mutations . . . . . 217
6.1.4 CRISPR interference . . . . . 218
6.1.5 Chapter aims . . . . . 219
6.2 Selection of candidate *CAV1* enhancers . . . . . 220
6.2.1 Proximal enhancer . . . . . 220
6.2.2 Distal model enhancer . . . . . 223
6.2.3 Genomic location of the model enhancers . . . . . 224
6.3 ALS-associated enhancers and their impact on *CAV1* transcription . . . . 225
6.3.1 How does repression of ALS-associated enhancers affect the expression of *CAV1*? . . . . . 225
6.3.1.1 Validation of dCas9 expression in dCas9-Zim3 SH-SY5Y cells . . . . . 225
6.3.1.2 CRISPRi of model *CAV1/CAV2* significantly reduces *CAV1* expression . . . . . 226
6.3.2 How does knockout of the distal *CAV1* enhancer impact *CAV1* transcriptional characteristics? . . . . . 228
6.3.3 *CAV1* transcriptional characteristics in patient-derived lymphoblastoid cells . . . . . 230
6.4 eRNA transcription from the proximal enhancer . . . . . 231
6.4.1 Identification of *de novo* transcription start sites . . . . . 231
6.4.2 Quantification of eRNA expression from the model proximal *CAV1/CAV2* enhancer . . . . . 234
6.4.2.1 Quantification of eRNA expression in SH-SY5Y cells . . . 234
6.4.2.2 Quantification of eRNA expression in lymphoblastoid cells 236
6.5 The impact of CBP inhibition on *CAV1* mRNA and eRNA transcription 239
6.6 The impact of ALS-associated SNPs of *CAV1* enhancer regulatory potential 242

6.7 Prediction of the impact ALS-associated SNPs have on protein binding motifs at enhancers . . . 245
6.7.1 Transcription factor motifs in the model distal *CAV1/CAV2* enhancer . . . 246
6.8 Discussion . . . 249
6.8.1 Establishing a functional link between the model *CAV1* enhancers and *CAV1* gene expression . . . 249
6.8.2 Do *CAV1* enhancers play a role in *CAV1* mRNA isoform choice? . 249
6.8.2.1 Differential *CAV1* mRNA isoform expression . . . 249
6.8.2.2 CAV1$\alpha$ phosphorylation . . . 251
6.8.2.3 Increased *CAV1* expression in ALS patient-derived lymphoblastoid cells . . . 251
6.8.3 eRNA transcription from the proximal enhancer provides insight into enhancer cooperativity . . . 252
6.8.4 CBP is a key regulator of enhancer activity at the proximal *CAV1* enhancer . . . 253
6.8.5 ALS-associated SNPs have differential impacts on proximal enhancer activity . . . 254
6.8.6 Predicted impacts of ALS-associated SNPs on TF binding at model enhancers . . . 256
6.8.6.1 Predicted impacts of ALS-associated SNPs on the proximal enhancer . . . 256
6.8.6.2 Predicted impacts of ALS-associated SNPs on the distal enhancer . . . 257
6.8.7 Conclusion . . . 259

**7 Disentangling *CAV1* enhancer and 3' UTR function 261**
7.1 Introduction . . . 261
7.1.1 Alternative polyadenylation yields differential 3' UTR isoforms . . 261
7.1.2 The impact of genetic variants in 3' UTRs . . . 262
7.1.3 Chapter aims . . . 263
7.2 Characteristics of the CAV1 3' UTR . . . 263
7.2.1 Characterising *CAV1* 3' UTR isoform expression . . . 263
7.2.2 The impact of ALS-associated SNPs on post-transcriptional regulatory elements in the *CAV1* 3' UTR . . . 265
7.3 Isolated 3' UTR expression . . . 266
7.3.1 Isolated *CAV1* 3' UTR expression in model cell lines . . . 266
7.3.2 Isolated *CAV1* 3' UTR expression in patient derived motor neurons 267
7.3.3 Tissue-specific isolated *CAV1* 3' UTR expression . . . 268
7.4 UTR-associated enhancers . . . 270
7.5 Disentangling the regulatory functions of the candidate *CAV1* enhancer from the *CAV1* 3' UTR . . . 271
7.5.1 sgRNA validation . . . 272
7.5.1.1 sgRNA validation in SH-SY5Y cells . . . 273
7.5.2 Generation of homology-directed repair template for insertion of SV40 PAS . . . 274
7.5.3 Test transfection of SV40 polyadenylation signal into *CAV1* in HEK293FT cells . . . 276

7.6 Discussion . . . 278
7.6.1 Differential *CAV1* 3' UTR expression . . . 278
7.6.2 The impact of ALS-associated SNPs on AU-rich elements . . . 279
7.6.3 3' UTR expression isolated from cognate CDS expression . . . 279
7.6.4 Isolated *CAV1* 3' UTR expression . . . 280
7.6.5 Enhancers in 3' UTRs and disease relevance . . . 282
7.6.6 Generation of cell lines to study the *CAV1* proximal enhancer independent of *CAV1* mRNA . . . 283
7.6.7 Conclusion . . . 284

**8 Discussion** **286**
8.1 Enhancer selectivity and specificity . . . 286
8.2 Model cell lines for studying enhancer activity . . . 289
8.2.1 SH-SY5Y cells as a neuronal model . . . 289
8.2.1.1 Differentiation of SH-SY5Y cells . . . 289
8.2.1.2 Transcriptional similarities between SH-SY5Y cells and neurons . . . 290
8.2.2 Model systems to capture the cell type-specific impact of *CAV1* enhancer mutations . . . 291
8.2.3 Enhancer cooperativity in model cell lines . . . 292
8.3 ALS-associated SNPs produce distinct effects on *CAV1* expression . . . 293
8.3.1 Mutations in the model proximal enhancer . . . 293
8.3.1.1 Impacts of ALS-associated SNPs on eRNA structure and function . . . 293
8.3.2 DNA sequence-dependent impacts of ALS-associated SNPs on enhancer activity . . . 294
8.3.2.1 Assessing DNA-dependent enhancer functionality by luciferase reporter assay . . . 295
8.3.3 Mutations in the model distal enhancer . . . 296
8.3.3.1 ALS-associated mutations may result in a compensatory upregulation in *CAV1* expression . . . 296
8.3.3.2 SNPs can impact *CAV1* isoform usage . . . 298
8.3.4 The impact of intra-enhancer location of SNP impact . . . 299
8.4 Future perspectives . . . 299
8.4.1 Established techniques for further enhancer interrogation . . . 299
8.4.2 Complexity of *CAV1/CAV2* enhancer regulatory mechanisms . . . 300
8.5 Concluding remarks . . . 301

**A Molecular Cloning** **302**
A.1 Proximal *CAV1/CAV2* enhancer G-block . . . 302
A.2 SV40 polyadenylation signal G-block . . . 305
A.3 Donor template for generation of genetically modified cell lines . . . 305

**B Bioinformatic analysis** **309**
B.1 Candidate *CAV1* enhancers . . . 309
B.2 eRNA seqeunces . . . 317
B.2.1 $^{DRR}$eRNA . . . 317
B.2.2 *TAL1* sense eRNA . . . 317

B.2.3 *TAL1* antisense eRNA . . . 318
B.2.4 *eCAV1_WT* . . . 318
B.2.5 *eCAV1_ALS* . . . 318
B.2.6 *eCAPZA2_antisense* . . . 318
B.2.7 *eCAPZA2_sense* . . . 318
B.2.8 *eCAV1_3363_us* . . . 319
B.2.9 *eCAV1_3388_us* . . . 319
B.2.10 *eCAV1_3363_ds* . . . 319
B.2.11 *eCAV1_3388_ds* . . . 320
B.2.12 *eTES_3388* . . . 320
B.3 Generation of shuffled RNA sequences . . . 320
B.3.1 Generation of shuffled RNA sequences of a given length and GC% 320
B.3.2 Generation of shuffled RNA sequences to calculate MFE density . 322
B.4 Generation of quantile bins within TAD domains . . . 324

**Bibliography** **325**

# Chapter 1

# Introduction

## 1.1 Genomes: From coding to non-coding

### 1.1.1 Increase in non-coding DNA as a function of organismal complexity

Prior to the advent of the genomic era, it was largely assumed that with increasing organismal complexity came an increased genome size, and in turn, an increased number of protein-coding genes. However, neither gene content, nor genome size, appear directly related to biological complexity (Mirsky and Ris, 1951, Moore, 1984, Thomas, 1971). Following assembly of the complete human reference genome in 2022, it became apparent that despite the 3 Giga base pair (Gbp) size of the human genome, it is predicted to contain only 19,969 protein-coding genes (Nurk et al., 2022). Comparing the human genome to that of the simple unicellular eukaryote Saccharomyces cerevisiae, which is only 12 Mega base pairs (Mbp) in size and contains a predicted 5,885 protein-coding genes, it is evident that despite the ∼250-fold increase in genome size from yeast to humans, there is only a ∼3-fold increase in the number of coding genes (Goffeau et al., 1996).

This apparent incongruence between perceived organismal complexity, genome size and number of genes may be largely resolved by considering the non-coding genome, which constitutes a significant proportion of the genomes of more complex organisms (Hahn and Wray, 2002, King and Wilson, 1975, Lander et al., 2001). The non-coding genome, namely genomic DNA that does not encode proteins, was formerly considered to be made up of "junk" DNA with little to no function (Doolittle and Sapienza, 1980, Ohno, 1972, Orgel and Crick, 1980). However, since the relative abundance of non-protein-coding DNA within the genome increases consistently with organismal complexity (Taft et al.,

2007), this raises the question, how would genomes laden with large amounts of inert "junk" DNA dictate biological complexity?

Some hypothesised that more complex eukaryotes simply possess a higher propensity to tolerate non-functional genomic DNA in contrast to prokaryotes, which select against the accumulation of non-coding DNA to maintain rapid DNA replication and cellular division rates (Gilbert et al., 1986). Yet, this does little to explain the organismal differences between distinct species with substantial genomic conservation. Alternatively, it was suggested that evolutionary changes among more complex organisms were a product of changes in gene regulation as opposed to changes in protein-coding genes, establishing intricacy of gene regulation as the principal factor contributing to organismal complexity. Further, this perspective posited that the accumulation of non-coding DNA in the genomes of higher organisms therefore likely serves a regulatory function (Britten and Davidson, 1969, King and Wilson, 1975).

Broadly, organismal complexity can be defined by the diversity of distinct cell states constituting the organism, and the spatiotemporal complexity and specialisation of these individual cell states (Bonner, 1988, McShea, 1996). Waddington's epigenetic landscape serves as a metaphor for development and cell fate choice, illustrating the transition of a cell from an undifferentiated state to various discrete, distinct, terminally differentiated cell fates. The epigenetic landscape diagrammatically depicts development, with cells, as represented by balls, rolling down an undulating surface of bifurcating valleys, where each valley represents a potential cell fate, with this choice modified by both internal and external stimuli (Figure 1.1) (Waddington, 1957). Hence, specification of differentiated cell state is largely mediated by external signals, with consequential alterations in gene expression fundamental to cell differentiation. Any given state of differentiation requires the integrated activation of sets of noncontiguous genes, these patterns of activity, and the response to environmental and developmental stimuli, must require a method of control (Britten and Davidson, 1969).

The necessity for developmental regulatory control in higher eukaryotes elucidates the significance of the extensive non-coding genome within such organisms. Such that the non-coding genome houses much of the regulatory elements required for development, differentiation and organismal complexity (Hahn and Wray, 2002, Lander et al., 2001, Sandelin et al., 2004, Taft et al., 2007). If we revisit Waddington's epigenetic landscape, regulatory elements within the non-coding genome may be imagined abstractly as a force biasing the cell to follow a particular path across the landscape, with cell fate decisions pre-determined by the activities of these regulatory regions (Figure 1.1).

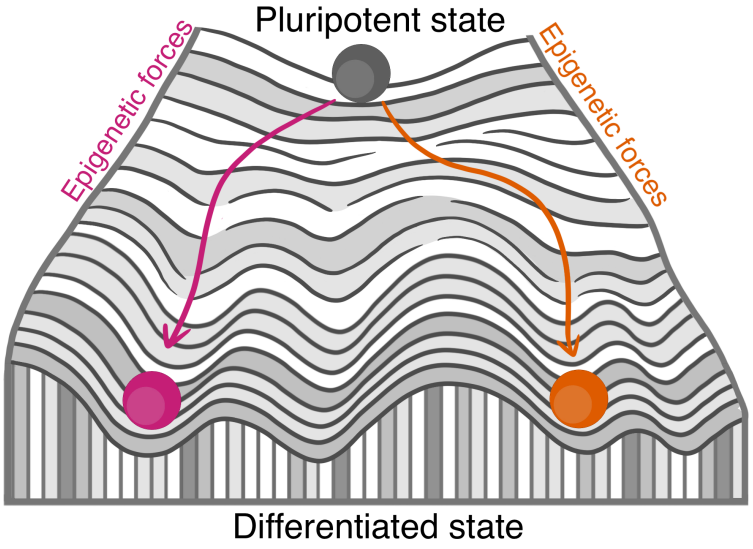



FIGURE 1.1: Waddington's epigenetic landscape
Waddington diagrammatically depicted cell development as a ball rolling down a surface, with the path followed by the ball analogous to the developmental history of the cell, largely determined by epigenetic influences. (Waddington, 1957).

### 1.1.2 The non-coding genome: Home to regulatory elements

Decoding genetic information into phenotype unfolds over a series of distinct steps, as outlined in the Central Dogma (Crick, 1970). This entails the conversion of linear DNA sequence information into RNA, followed by the translation of RNA into proteins. Gene expression changes underlying inter- and intraspecies divergence are to a great extent genetically encoded, as dictated by *cis*-regulatory DNA sequences which determine the transcriptional activity of genes, and the splicing and stability of the resulting messenger RNAs (mRNAs) (Figure 1.2) (reviewed in Cramer, 2019).

The evolutionary expansion and rewiring of gene regulatory networks that enable intricate transcriptional and post-transcriptional control of gene expression, and ultimately protein synthesis, underpins biological complexity. The rapid evolution of transcriptional regulatory elements, such as enhancers (Klein et al., 2018, Villar et al., 2015), and alternative cleavage and polyadenylation of the 3' untranslated regions (3' UTRs) of mRNAs (Chen et al., 2012a, Derti et al., 2012, Hoque et al., 2013, Tian et al., 2005), which function as master regulators of post-transcriptional control (reviewed in Mayr, 2019), provide great complexity to gene regulatory circuits.

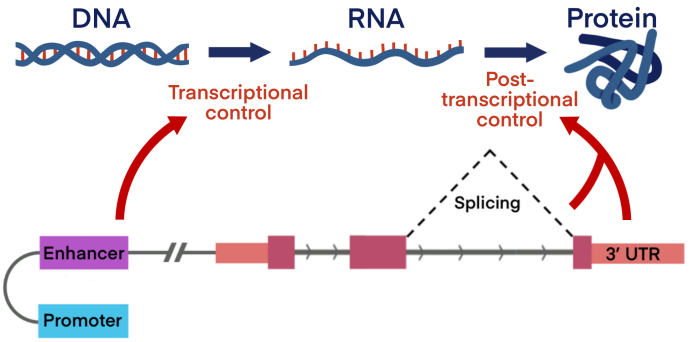



FIGURE 1.2: Multi-level control of the the central dogma
Transcriptional and post-transcriptional regulatory mechanisms, as determined by the activity of *cis*-regulatory elements, bestows stringent control over the Central Dogma at multiple levels.

## 1.2 The three-dimensional genome

The intricacy of gene regulation requires precise control over the three-dimensional (3D) organisation of the genome, from the compaction of DNA into chromatin fibres up to large complex chromosomes. The dynamics of DNA wrapping around the nucleosome core, formed of a octamer of histone proteins, provides a structural framework for transcriptional control in eukaryotes (reviewed in Wolffe and Guschin, 2000). Whereby, modulation of the chromatin architecture can promote or impede gene transcription dependent on structural context, by facilitating or restricting accessibility of DNA, respectively (Chipev and Wolffe, 1992, Schild et al., 1993, Sera and Wolffe, 1998, Wolffe, 1989). Chromatin fibres are largely organised into euchromatin - regions of open, accessible, transcriptionally active chromatin, and heterochromatin - regions of condensed chromatin that is consequently less transcriptionally active (reviewed in Morrison and Thakur, 2021). The open structure of euchromatin facilitates the binding of the transcriptional machinery to DNA, fostering active transcription. Hence, euchromatin is predominantly found at actively transcribing genes and active regulatory elements such as promoters and enhancers, which, in turn, regulate the access of transcription factors (TFs) and chromatin remodelers to DNA, contributing to the further opening of chromatin and initiation of transcription (reviewed in Calo and Wysocka, 2013, Venkatesh and Workman, 2015).

Beyond chromatin accessibility, transcriptional control requires physical interactions between genes and regulatory element DNA, often over long distances of linear genomic sequence, highlighting the necessity for chromosomes to fold into complex 3D structures, allowing regulatory elements to interact with each other in 3D space. Chromosomes are compartmentalised into neighbourhoods of preferential local chromatin interactions,

termed topologically associating domains (TADs) (Figure 1.3) (Dixon et al., 2012, Nora et al., 2012, Sexton et al., 2012).

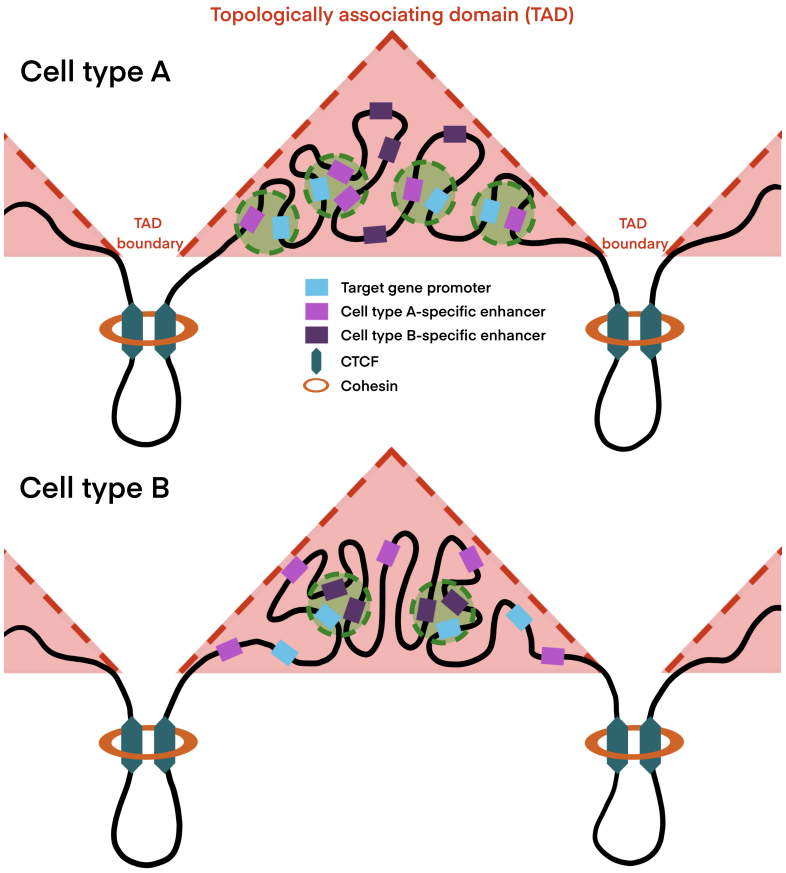



FIGURE 1.3: TADs as units of higher-order genome organisation
TADs function as self-interacting domains that structurally insulate adjacent domains, limiting chromatin interactions among loci within the same TAD. While there is little variance in TAD location between distinct cell types, variable cell type-specific regulatory interactions occur within the boundaries of TAD, as demonstrated by differential enhancer-promoter interactions between cell type A and cell type B, denoted by dashed green circles.

TADs represent functionally and physically isolated, self-associating domains that house co-regulated genes, and contribute to regulation of gene expression by bringing distant *cis*-regulatory elements into physical proximity (Dixon et al., 2012, Nora et al., 2012, Sexton et al., 2012, Shen et al., 2012). TAD boundaries, as maintained by interactions between the insulator binding protein CCCTC-binding factor (CTCF) and the cohesin complex (Dixon et al., 2012, Kagey et al., 2010, Phillips-Cremins and Corces, 2013, Zuin et al., 2014), also contribute to gene regulation by ensuring *cis*-regulatory sequences interact almost exclusively with target genes located within the confines of the TAD (Dixon et al., 2012, Nora et al., 2012, Sexton et al., 2012). TAD organisation is largely cell type invariant (Dixon et al., 2015, 2012, Nora et al., 2012). However, while TAD positioning remains stable between distinct cell types, local alterations in chromatin

structure and contact frequency occur within domains. Chromatin organisation within TADs is dynamic, with cell type-specific regulatory events, namely enhancer-promoter interactions, shaping local high-order chromatin structure (Dixon et al., 2015, 2012, Nora et al., 2012).

## 1.3 Transcriptional control: As determined by regulatory elements

To a great extent, the intricate regulation of gene expression in metazoans occurs during the transcription of genes into mRNA by the molecular machine, RNA polymerase II (RNAPII). The DNA-dependent synthesis of RNA by RNAPII is under strict control, dictated by the combinatorial sequence-specific binding of TFs at promoter elements that nucleate the assembly of the preinitiation complex (PIC) and integrate stimulatory and repressive signals from TFs bound at distal enhancer and silencer elements, respectively. The fidelity of transcription is meticulously regulated by the concerted activity of transcriptional activators and repressors bound at regulatory regions, in coordination with transcriptional co-regulators. Such factors govern the assembly of the transcriptional machinery at the promoter, either through direct protein-protein interactions or by influencing chromatin via nucleosome remodelling and deposition of histone modifications (reviewed in Cramer, 2019, Smale and Kadonaga, 2003).

## 1.4 Enhancers: key regulators of transcription

Core promoters alone, while sufficient for the initiation of transcription, generally only exhibit low basal transcriptional activity (reviewed in Shlyueva et al., 2014). Distally located *cis*-regulatory enhancer elements are required to fine-tune gene expression patterns, influencing the activity of RNAPII at target loci, at distances of up to megabases away from the transcription start site (TSS) (Banerji et al., 1983, 1981, Fulco et al., 2016, Gillies et al., 1983, Lettice et al., 2002, Long et al., 2020). Enhancers constitute a pivotal component of the eukaryotic gene regulatory landscape, conferring precise spatiotemporal control over gene expression in response to developmental and environmental stimuli (Figure 1.4). Thus, their functionality is highly specific to distinct cell lineages and stimuli (reviewed in Heinz et al., 2015, Shlyueva et al., 2014).

The functional potential of an enhancer is influenced by the interplay between features intrinsically encoded in enhancer DNA, such as distance from target loci (Zuin et al., 2022), and *trans*-acting factors, such as interactions with TFs and coregulators, in the

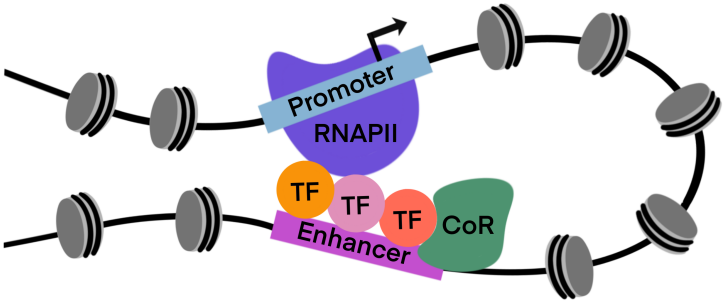



FIGURE 1.4: Enhancers are fundamental to gene regulatory networks
Enhancers underpin gene regulation, orchestrating precise gene expression patterns in response to internal and external stimuli. Enhancers control the expression of their target genes by contacting the target gene promoter and recruiting TFs, transcriptional coregulators, and RNAPII

context of the local chromatin environment. Simplistically, enhancer activity is a function of TF binding and the packaging of enhancer DNA into chromatin (Miller and Widom, 2003). Broadly, chromatin accessibility can be considered a master regulator of enhancer potential, with enhancer potential also reliant on a myriad of factors that control chromatin structure, for example DNA methylation and the post-translational modification (PTM) of histones (reviewed in Calo and Wysocka, 2013). Together, histone PTMs and DNA methylation confer epigenetic control over the chromatin environment, and therefore gene expression, by altering local chromatin density and thereby accessibility of DNA to the basal transcriptional machinery, modulating the transcriptional potential of enhancers (Figure 1.5) (Cedar and Bergman, 2009).

Analogous to higher-order chromatin organisation into TADs providing the structural framework for defining regions of enhancer activity, finer-scale chromatin interactions connect enhancers and promoters within, and in some cases, beyond TADs, enabling precise enhancer-mediated gene regulation. Hence, the hierarchical organisation of chromatin regulates transcription via modulation of 3D regulatory interactions, notably through CTCF-anchored enhancer-promoter chromatin looping and TAD structure (Amano et al., 2009, Dixon et al., 2012, Fudenberg et al., 2016, Haarhuis et al., 2017, Kagey et al., 2010, Nora et al., 2012, Sanborn et al., 2015, Schwarzer et al., 2017, Wutz et al., 2017). The linear proximity of enhancers to their target loci impacts enhancer activity, with enhancer strength generally increasing with decreasing genomic distance from the target gene promoter (Zuin et al., 2022). The compartmentalisation of chromatin into self-interacting TADs and the submegabase range of CTCF-mediated enhancer-promoter interactions may contribute to this effect (Bonev et al., 2017, Zuin et al., 2022).

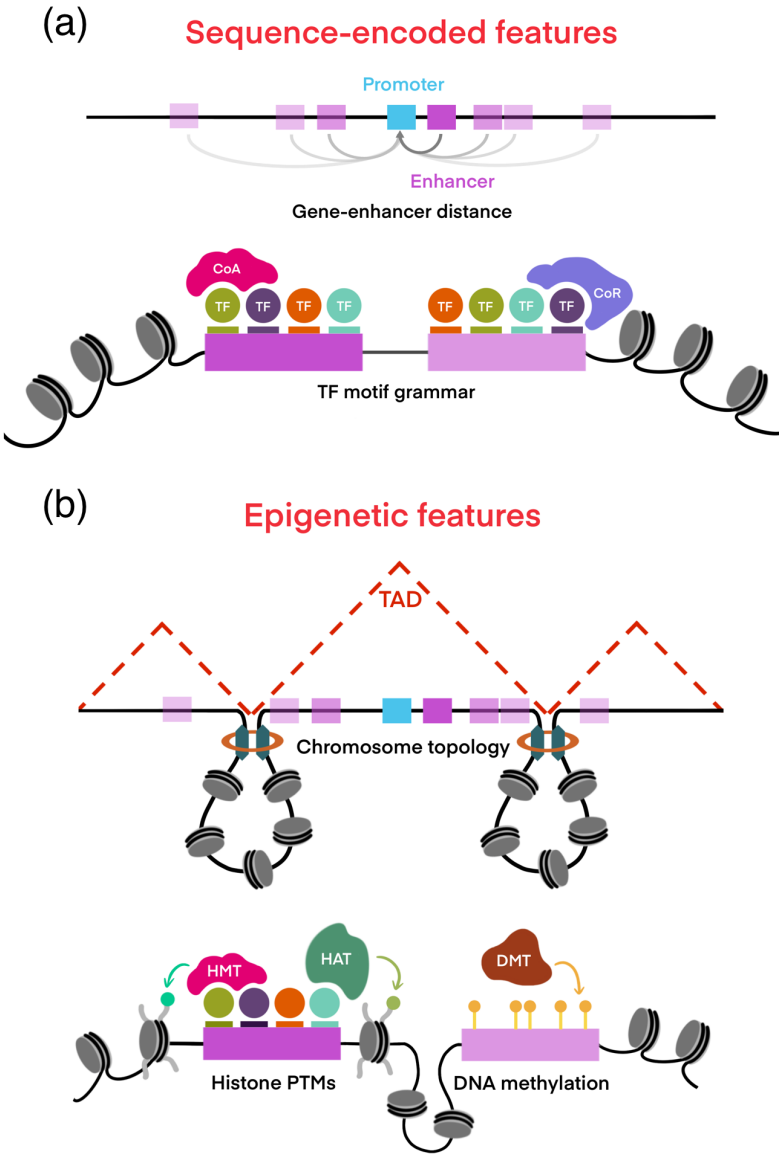



FIGURE 1.5: Enhancer architecture is fundamental to enhancer potential
(a) The functional potential of an enhancer is influenced by features encoded intrinsically within the enhancer, such as distance from the target gene and TF motif grammar, and by *trans*-acting factors, such as TFs and coregulators. (b) Central to enhancer architecture is chromatin accessibility, with the above factors largely only relevant in the context of open chromatin. Hence, factors controlling chromatin accessibility, such as DNA methylation and histone post-translational modifications (PTMs), also contribute to enhancer potential.

Around 1,600 human TFs have been identified (Lambert et al., 2018). The nucleosomal barrier to TF binding ensures concentrated TF occupancy at enhancers, where the organisation of TF binding sites allows integration of multiple signals to accomplish spatiotemporal specificity (Zinzen et al., 2009). Pioneer factors are TFs with the capacity to access nucleosome-bound DNA, establishing open chromatin, thereby facilitating

enhancer competence for binding of additional TFs (Cirillo et al., 2002, Zaret and Carroll, 2011). Pioneer factors can often invoke direct functions to physically unwrap DNA from histones (Michael et al., 2020), or indirect functions, further recruiting nucleosome remodelers (King and Klose, 2017). An alternative, non-exclusive model suggests that cooperativity of multiple TFs could outcompete and ultimately evict nucleosomes to bind to DNA (Mirny, 2010).

TFs themselves do not directly alter transcription but instead serve as adapters that guide cofactors to their binding sites. It is no surprise that enhancers, as major sites of combinatorial TF assembly, are also occupied by transcriptional cofactors, specifically coactivators. Cofactors known to function at enhancers include the Mediator complex (Aranda-Orgilles et al., 2016, Moyo et al., 2020); nucleosome remodelers, such as the SWI/SNF complex (Hu et al., 2011); histone modifiers including the histone methyltransferases MLL3/MLL4 (Hu et al., 2013) and the histone acetyltransferases CREB-binding protein (CBP) and p300 (Creyghton et al., 2010, Heintzman et al., 2007, Rada-Iglesias et al., 2011).

## 1.5 Histone modifications as identifiers of regulatory elements

The PTM of histone tails is indicative of distinct chromatin states, typically correlated either with activation or repression of transcription, dependent on the nature of chemical modification and its location on the histone octamer (Strahl and Allis, 2000). While active promoters and enhancers are associated with regions of open chromatin, and hence typically devoid of nucleosomes, these regulatory elements are flanked by nucleosomes decorated with histone PTMs which are informative for their identification and characterisation genome-wide (reviewed in Calo and Wysocka, 2013).

Generally, enhancers and promoters are predominantly characterised by the mono- and trimethylation of lysine 4 on histone H3 (H3K4me1 and H3K4me3), respectively, with both being additionally marked by acetylation of lysine 27 on histone H3 (H3K27ac) upon activation (Creyghton et al., 2010, Heintzman et al., 2007, Rada-Iglesias et al., 2011). Further, reciprocal modification of histone H3 lysine 27 is informative of enhancer activation state, with acetylation and trimethylation indicative of transcriptionally active and repressive chromatin states, respectively (Figure 1.6)(Allfrey et al., 1964, Creyghton et al., 2010, Francis et al., 2004, Rada-Iglesias et al., 2011, Ringrose et al., 2004, Zentner et al., 2011). Poised enhancers, which are typically transcriptionally inactive but poised for activation in response to external stimuli or developmental cues, are distinguished by

enrichment of trimethylation of lysine 27 on histone H3 (H3K27me3), co-occurring with H3K4me1, and characterised by their capacity to acquire chromatin signature of active enhancers, namely H3K27ac, upon activation (Rada-Iglesias et al., 2011, Zentner et al., 2011). Further histone PTMs are linked to active enhancers, particularly those implicated in nucleosome dynamics and chromatin opening, such as acetylation of lysines 64 and 122 on histone H3 and lysine 16 on histone H4 (H3K64ac, H3K122ac and H4K16ac, respectively) (Di Cerbo et al., 2014, Pradeepa et al., 2016, Simon et al., 2011, Taylor et al., 2013, Tropberger et al., 2013), broadly implicating chromatin decompaction and transcription as key features of active enhancers.

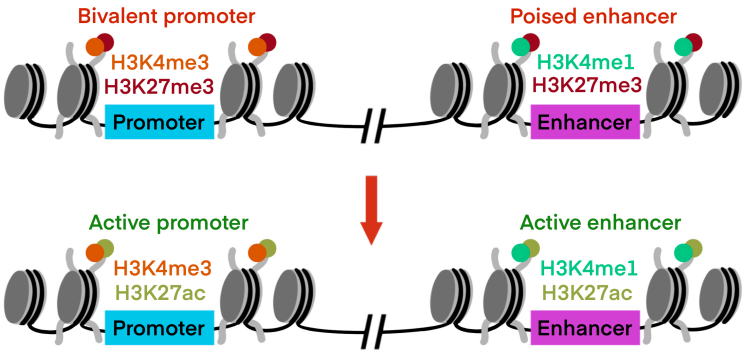



FIGURE 1.6: Histone modifications can differentiate regulatory elements
Simplistically, the post-translational modification on histone H3 at lysine 27 (H3K27) can discern active from inactive regulatory elements, or those poised for later activation. Both inactive promoters and poised enhancers are flanked by histones carrying H3K27me3. Upon activation both promoters and enhancers lose the transcriptionally repressive H3K27 modification to gain the activating modification, H3K27ac. Historically, promoters and enhancers have been differentiated by the levels of methylation on histone H3 lysine 4, with H3K4me3 and H3K4me1 considered characteristic of promoters and enhancer, respectively.

Such histone modifications can be mapped genome-wide by exploiting chromatin immunoprecipitation (ChIP)-based sequencing technologies such as ChIP sequencing (ChIP-seq) and Cleavage Under Targets and Release Using Nuclease (CUT&RUN), using specific antibodies, providing predictive power in the identification of promoters and enhancers (Heintzman et al., 2009, 2007, Mikkelsen et al., 2007, Rada-Iglesias et al., 2011, Skene and Henikoff, 2017).

With the deposition of histone acetylation a fundamental characteristic of active enhancers, the binding of CBP/p300 - the histone acetyltransferases that catalyse H3K27ac deposition - is widely used as a hallmark for the identification of active enhancers (Creyghton et al., 2010, Heintzman et al., 2007, Rada-Iglesias et al., 2011). In fact, the

acetyltransferase activity of CBP/p300 bestows dynamic control over enhancer activation and the consequential stimulation of target gene expression, required for recruiting the PIC and RNAPII at enhancers (Hilton et al., 2015, Miao et al., 2022, Narita et al., 2021).

## 1.6 eRNAs: Functions, sequence and structure

*The following subsection is adapted from Harrison and Bose (2022)*

Over the last decade, the transcription of enhancer RNAs (eRNAs) – nascent RNA species pervasively transcribed from active enhancers – has emerged as a key factor that contributes to enhancer activity (De Santa et al., 2010, Hah et al., 2011, Kaikkonen et al., 2013, Kim et al., 2010, Li et al., 2013). The transcriptional landscape at enhancers is complex, generating eRNAs with diverse structures, directionality and stability (Figure 1.7). Due to this heterogeneity, the categorisation of enhancer-derived transcripts poses a challenge, and the definitions of eRNAs and characteristically similar long non-coding RNAs (lncRNAs) are blurred. Furthermore, some regulatory lncRNAs function in a transcript-independent manner (Engreitz et al., 2016), whereas others function via the act of transcription itself (reviewed in Kaikkonen and Adelman, 2018). eRNAs tend to be relatively short-lived, transcribed at very high rates but also quickly degraded, such that their half-lives are often of the order of minutes (Rabani et al., 2014, Schwalb et al., 2016). Nevertheless, eRNA production is deeply intertwined within enhancer regulatory networks and is thus implicated in the transcriptional control of gene expression.

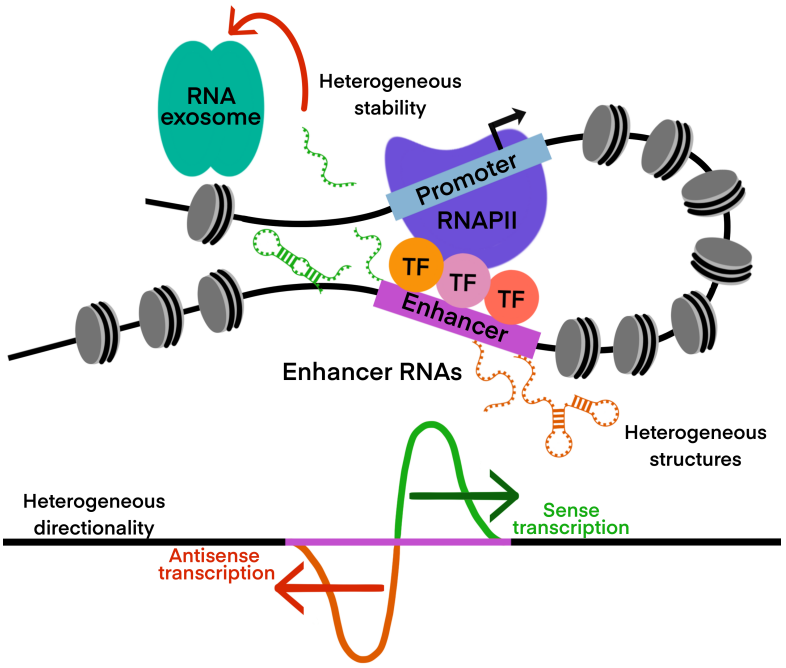



FIGURE 1.7: Active enhancers are transcribed into heterogeneous non-coding RNAs eRNAs represent a diverse family of non-coding RNAs. Enhancer transcription yields bidirectional RNAs with heterogeneous structures, stability (as largely determined by degradation by the RNA exosome) and functions.

### 1.6.1 Characterisation of eRNAs

The majority of identified 'classical' eRNAs are short (~200nt-2kb), bidirectionally transcribed transcripts that are 5' 7-methylguanylate ($m^7G$) capped, non-polyadenylated and non-spliced (Andersson et al., 2014, De Santa et al., 2010, Kim et al., 2010, Kristjánsdóttir et al., 2020, Schwalb et al., 2016). Longer (up to 4-5kb), polyadenylated and spliced, unidirectionally transcribed eRNAs – often referred to as enhancer-associated lncRNAs – have also been identified (De Santa et al., 2010, Djebali et al., 2012, Gil and Ulitsky, 2018, Hah et al., 2013, Koch et al., 2011, Vučićević et al., 2015). For shorter transcripts, eRNA length is dictated by the Integrator complex, which processes and controls the length of a number of short RNA species (Lai et al., 2015, Lykke-Andersen et al., 2021). It is unclear whether the differences between classical eRNAs and enhancer-associated lncRNAs confer any distinct functions, but distinct transcripts are likely to display differences in transcript stability (Gil and Ulitsky, 2018). The sensitivity of RNA transcripts to degradation by the RNA exosome decreases as the distance between the TSS and the polyadenylation site increases (Ntini et al., 2013). Thus, the production of short eRNAs may prevent the complete assembly of the polyadenylation machinery on RNAPII, resulting in suboptimal polyadenylation and subsequently RNA exosome intervention

(Ntini et al., 2013). This may explain the propensity for longer eRNAs to be polyadenylated while the majority of short eRNAs are non-polyadenylated and unstable, subject to rapid degradation. The transience of short, non-polyadenylated eRNA sequences likely limits their spatial range of action, whereas the absence of such strict temporal restriction may allow longer, polyadenylated eRNAs to survey the nuclear space in search of target sequences, potentially in *trans* (De Santa et al., 2010, Kim and Shiekhattar, 2016, Tsai et al., 2018).

### 1.6.2 The effects of eRNAs on the chromatin environment

#### 1.6.2.1 eRNAs activate histone acetylation

Like enhancer elements themselves, eRNAs can contribute to gene expression by regulating access of transcriptional machinery at defined genomic loci. Upon eRNA depletion, a reduction in chromatin accessibility at promoters is observed, suggesting that eRNAs may facilitate nucleosome rearrangement and/or recruitment of chromatin remodellers (Mousavi et al., 2013). Notably, CBP/p300 inhibition abrogates enhancer activation, coupled with the rapid attenuation of eRNA expression levels, highlighting CBP/p300-dependent acetylation as a key driver in enhancer activation (Narita et al., 2021). CBP directly interacts with eRNAs at active enhancers, an interaction that stimulates the acetyltransferase activity of CBP and results in RNA-dependent modulations of H3K27ac (Bose et al., 2017, Carullo et al., 2020). These interactions underlie changes in CBP activity, modulating histone acetylation and thus fine-tuning target gene expression. Notably, limited sequence conservation is observed in the eRNAs that bind to CBP; specificity instead arises from the binding of specific, local populations of eRNAs to CBP (Bose et al., 2017). These results suggest that the precise sequence of an eRNA may be less important than the presence of the eRNA transcript itself and hint at a general mechanism for eRNA function.

#### 1.6.2.2 The effects of eRNAs on repressive histone methylation

Although H3K27ac is generally considered an activating histone modification, tri-methylation of the same residue (H3K27me3) is a crucial repressive modification during development, ensuring genes that require silencing are packaged into heterochromatin. H3K27me3 is generated by polycomb repressive complex 2 (PRC2) (Ferrari et al., 2014). In embryonic stem cells (ESCs), PRC2 binds at repressed developmental regulator genes (Tanay et al., 2007) and promiscuously interacts with nascent RNAs at essentially all active genes (Beltran et al., 2016, Wang et al., 2017b). RNA binding effectively antagonises

the histone methyltransferase activity of PRC2, alleviating chromatin of repressive histone modifications and activating gene expression (Cifuentes-Rojas et al., 2014, Kaneko et al., 2013). eRNAs may share this function: knockdown of eRNAs transcribed from the distal enhancer of the gonadotropin hormone $\alpha$-subunit (*Cga*) gene converts H3K27 from an active acetylated state to a repressive methylated state at both the enhancer and the associated proximal promoter (Pnueli et al., 2015). Moreover, the eRNA *CARMN*, which plays a crucial role in cardiovascular cell fate determination and differentiation, physically interacts with PRC2 to mediate epigenetic regulation of the cardiac commitment programme (Ounzain et al., 2015), akin to other cardiogenic lncRNAs, such as braveheart (*Bvht*) and *Fendrr* (Grote et al., 2013, Klattenhoff et al., 2013).

PRC2 preferentially binds RNA motifs composed of short repeats of consecutive guanine residues; the natural abundance of this RNA motif explains the promiscuous binding activity of PRC2 (Wang et al., 2017a). Further, PRC2-DNA binding sites are significantly enriched for sequences that code for PRC2-binding RNA motifs, potentially mediating site-specific RNA interactions. In line with its preference for guanine tracts, PRC2 binds to guanine quadruplex (G-quadruplex) structures in RNA with high affinity (Wang et al., 2017a). This posits a model in which secondary structures in eRNAs promote active transcription from enhancers, firstly by inhibiting DNA methylation and secondly by sequestering PRC2 from nucleosome substrates, leading to H3K27me3 depletion and subsequently gene activation (Beltran et al., 2019, Mao et al., 2018). Thus, DNA demethylation and PRC2 removal may be encoded within the DNA sequence to promote gene activation at associated loci (Wang et al., 2017a,b).

Together, these studies of H3K27ac and H3K27me3 highlight the potential of eRNAs to interact with enzyme complexes responsible for generating histone modifications – both activating and repressive – to generate a chromatin environment that is conducive to active transcription (Figure 1.8). At active enhancers, although eRNAs may bind to PRC2 to inhibit its histone methyltransferase activity (Beltran et al., 2016, Cifuentes-Rojas et al., 2014), they may also stimulate the acetyltransferase activity of CBP (Bose et al., 2017). This suggests that eRNAs promote a feed-forward loop, whereby enhancer activation yields eRNA production that, in turn, maintains open chromatin at the enhancer to reinforce enhancer activation. Moreover, without nascent eRNA transcripts antagonising PRC2 activity, H3K27me3 and thus repressive chromatin is successfully maintained at inactive enhancers (Beltran et al., 2016, Cifuentes-Rojas et al., 2014).

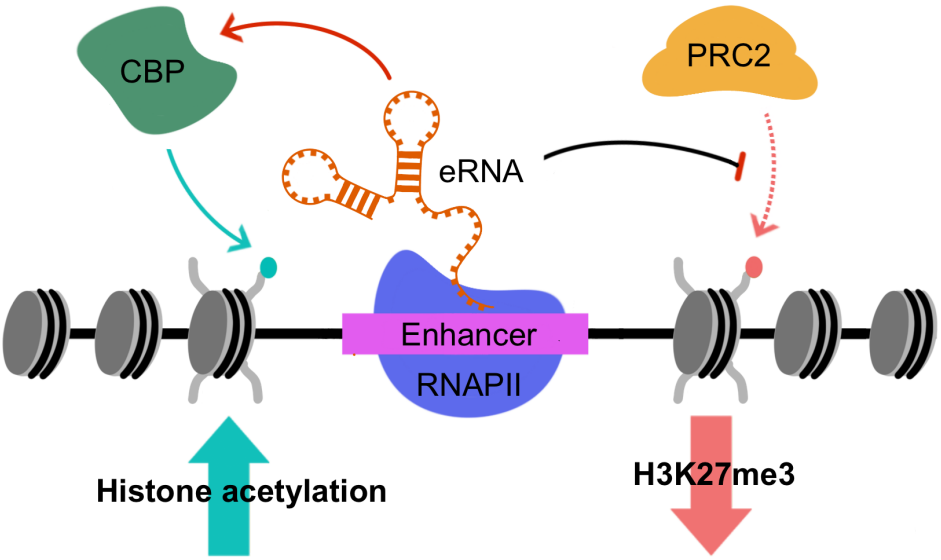



FIGURE 1.8: eRNA interactions with chromatin modifying enzymes at enhancers
Enhancer RNAs (eRNAs) are RNA species that are pervasively transcribed from active enhancers. They interact with histone modifying enzyme complexes at enhancers resulting in altered patterns of histone modifications at these sites. For example, eRNA interactions with PRC2 and p300/CBP inhibit the deposition of repressive H3K27me3 PTMs and promote the deposition of activating H3K27ac PTMs, respectively, helping to maintain an open chromatin state at enhancers and promoting further eRNA synthesis

### 1.6.3 eRNAs as direct transcriptional regulators

#### 1.6.3.1 Effects on RNA polymerase II pause release

RNAPII pausing is a genome-wide regulatory mechanism by which the pausing and subsequent release of RNAPII into productive elongation acts as a key rate-limiting step for transcription. eRNAs can facilitate the release of paused RNAPII, thus promoting the transition into productive elongation, via interactions with complexes of negative elongation factor (NELF) and the positive transcription elongation factor b (P-TEFb) (Schaukowitch et al., 2014, Shii et al., 2017). P-TEFb is a multiprotein complex composed of cyclin-dependent kinase 9 (CDK9) and cyclin T1. eRNAs interact with P-TEFb by directly binding to cyclin T1; the resultant activation of P-TEFb promotes the phosphorylation of NELF, DRB sensitivity inducing factor (DSIF) and Serine 2 of the RNAPII C-terminal domain (CTD), via the pause-releasing kinase activity of CDK9 (Fujinaga et al., 2004, Shii et al., 2017, Wada et al., 1998, Yamaguchi et al., 1999, Zhao et al., 2016). P-TEFb-catalysed phosphorylation sites on the NELF-A subunit of the NELF complex overlap with eRNA binding sites identified on this subunit (Gorbovytska et al., 2022). Phosphorylation of such sites is essential for RNAPII pause release (Lu et al., 2016) and is hypothesised to induce a conformational change in NELF conducive

to NELF release from the paused elongation complex. In line with this, eRNA binding within the vicinity of P-TEFb target sites may trigger a similar conformational change that bypasses P-TEFb activity altogether. Consequently, NELF phosphorylation by P-TEFb and eRNA binding may similarly bestow a negative charge to NELF that facilitates a conformational change and, ultimately, results in dissociation of NELF from the paused elongation complex (Gorbovytska et al., 2022).

Such interactions between eRNAs and NELF have been shown to be crucial for rapid induction of neuronal immediate early genes (IEGs) in neurons. In this context, eRNA binding to NELF facilitates the release of paused RNAPII in response to neuronal stimulation (Gorbovytska et al., 2022, Schaukowitch et al., 2014). Replacing wildtype (WT) NELF-E with an RNA recognition motif (RRM) deletion mutant in neurons causes a reduction in NELF binding at IEG promoters and mRNA induction (Schaukowitch et al., 2014). In this case, eRNAs bind directly to a canonical RNA binding motif in the NELF-E subunit. However, NELF-E binds to a variety of RNA sequences with little or no sequence or structural constraint (Yamaguchi et al., 2002), suggesting that mutations in the eRNAs would have minimal effect on the efficiency of eRNA binding to NELF. This is another example of the eRNA transcript itself rather than a particular RNA sequence being the driver of eRNA function.

#### 1.6.3.2 eRNAs and transcription factor binding

As well as regulating transcription initiation directly, eRNAs can interact with a variety of transcriptional activators and repressors (Oksuz et al., 2023). For example, eRNAs can bind to the multi-functional TF yin yang 1 (YY1), which functions as both a transcriptional activator and a repressor (reviewed in Shi et al., 1997). eRNAs bind to YY1, increasing its occupancy at enhancers, via the zinc fingers of YY1 and with no strong RNA sequence preference (Sigova et al., 2015, Wai et al., 2016). Nascent RNA transcribed from both enhancers and promoters may help to capture dissociating YY1, increasing its occupancy (Sigova et al., 2015). This 'TF trapping' activity may be a common mechanism of action for eRNAs, as eRNAs also increase the binding of other TFs such as c-Jun and NF-$\kappa$B at their target loci (Huang et al., 2018, Shii et al., 2017, Spurlock et al., 2017). Transcriptional control may therefore involve an additional positive-feedback loop in which YY1 and other TFs stimulate local enhancer transcription, and newly transcribed eRNAs reinforce local TF occupancy, contributing stability to gene expression in vivo (Sigova et al., 2015). YY1 can also recruit PRC2 to specific loci, promoting the deposition of repressive H3K27me3 (Atchison et al., 2003, Caretti et al., 2004, Pan et al., 2013, Wilkinson et al., 2006). However, RNA interactions with PRC2 via the EZH2 subunit effectively inhibit its methyltransferase activity

(Cifuentes-Rojas et al., 2014). Thus, eRNAs may both promote YY1, and consequently PRC2, occupancy but may also downregulate the activity of recruited PRC2, providing another example of eRNAs fine-tuning transcriptional outputs.

### 1.6.4 The role of eRNA structure in regulating eRNA function

Although multiple studies have demonstrated functional roles for eRNAs in transcriptional regulation, the role played by specific eRNA sequences and conserved eRNA structures is less well defined. Following the classification of ∼10,000 newly detected nascent RNAs, including 3,257 classified as eRNAs, ∼10% were predicted to be structured (Schwalb et al., 2016). However, the roles played by these RNA structures is relatively poorly understood. Below we highlight examples in which the function of an eRNA has been linked to its structure.

The $^{DRR}$eRNA (also known as MyoD upstream non-coding RNA or MUNC) is transcribed from the enhancer regions of the myogenic master regulator MyoD (Cichewicz et al., 2018, Tsai et al., 2018). $^{DRR}$eRNA directs cohesin loading in *trans* to promote chromatin opening during myogenesis, implicating $^{DRR}$eRNA as a regulator of muscle cell differentiation. $^{DRR}$eRNA is recruited to the *Myog* locus in *trans*, where it colocalizes with nascent transcripts, positing the formation of RNA:RNA interactions between the $^{DRR}$eRNA and *Myog* nascent transcripts as a potential in *trans* target recognition mechanism (Tsai et al., 2018). Chemical probing studies enabled modelling of $^{DRR}$eRNA secondary structure, revealing the presence of multiple well-defined functional domains, with different domains mediating distinct features of $^{DRR}$eRNA promyogenic activity (Przanowska et al., 2022). Notably, $^{DRR}$eRNA does not regulate the expression of all of its target genes by a common mechanism; instead, distinct combinations of effector domains are required for the induction of different target genes, thus demonstrating that the functional complexity of $^{DRR}$eRNA is integrated within its complex secondary structure (Przanowska et al., 2022).

eRNAs can regulate RNAPII pause release by binding to NELF. They exert their dissociative function on NELF in a manner that is more dependent on eRNA length than overt structure. To trigger the dissociation of NELF from the paused elongation complex, a single eRNA containing unpaired guanosines (but not G-quads, in contrast to PRC2) forms multiple simultaneous allosteric contacts with several RNA binding sites on NELF. In the absence of sequence and structure dependency, the length of the eRNA, and therefore the degree of eRNA flexibility, is key, facilitating folding of the eRNA substrate into a structure that favours simultaneous binding to both the NELF-E and NELF-A subunits (Gorbovytska et al., 2022).

However, although eRNA binding to NELF-E (Rao et al., 2006, Yamaguchi et al., 2002) does not require structure specificity, RNA structure is important for eRNA binding to P-TEFb, where eRNA-mediated activation of P-TEFb promotes target gene transcription via phosphorylation of Serine 2 of the RNAPII CTD. (Richter et al., 2002, Zhao et al., 2016). Here, a TAR RNA-like (TAR-L) motif in the androgen receptor-regulated eRNA (*AR*-eRNA) shares a similar secondary structure with the 3′ end of the small nuclear RNA *7SK* (also known as *RN7SK*) (Zhao et al., 2016), an established inhibitor of P-TEFb (Nguyen et al., 2001, Yang et al., 2001). Competitive binding of the *AR*-eRNA displaces *7SK* and activates P-TEFb (Zhao et al., 2016). As *AR*-eRNA and *7SK* exhibit a common secondary structure, the interactions between them and P-TEFb are probably due to structural recognition rather than sequence specificity.

eRNAs can also co-transcriptionally assemble into DNA/RNA hybrid structures, termed R-loops (Figure 1.9), in which the nascent RNA anneals to the template strand of the DNA duplex (Pefanis et al., 2015, Wulfridge and Sarma, 2021, Yan et al., 2019). R-loop formation is associated with enhancer-like epigenetic signatures, including an open chromatin state, the deposition of characteristic histone modifications and unmethylated DNA (Ginno et al., 2013, 2012, Kuznetsov et al., 2018, Nadel et al., 2015, Sanz et al., 2016). R-loop formation is determined by nucleic acid sequence, with G-rich RNA hybridising with C-rich DNA in the template strand (Ratmeyer et al., 1994), and is favoured by the presence of G-quadruplexes on the non-template strand of the DNA duplex (Ginno et al., 2013, 2012, Kuznetsov et al., 2018, Li and Manley, 2005, Lim and Hohng, 2020, Ratmeyer et al., 1994, Wulfridge and Sarma, 2021). G-quadruplexes are also enriched within enhancer DNA (Hou et al., 2021) and it is thought that G-quadruplexes help to stabilise the R-loops formed by eRNAs and contribute to maintaining the nucleosome-depleted regions necessary for sustained enhancer activity (Hou et al., 2021, Maizels and Gray, 2013, Wulfridge and Sarma, 2021).

## 1.7 Transcriptional condensates and eRNAs

Phase-separated transcriptional condensates have been shown to form at enhancers, generating high local concentrations of transcriptional regulators and promoting their synergistic function (Hnisz et al., 2017, Shrinivas et al., 2019). The disordered domains that promote condensate formation are often sites of sequence-independent interactions with RNA and form the basis of a prevalent category of non-canonical RNA binding motifs (Castello et al., 2016, He et al., 2016, Hentze et al., 2018). Unsurprisingly, therefore, transcriptional condensates are enriched in and dependent on RNA, which scaffolds and modulates their morphology and composition (Banani et al., 2016). Many properties of

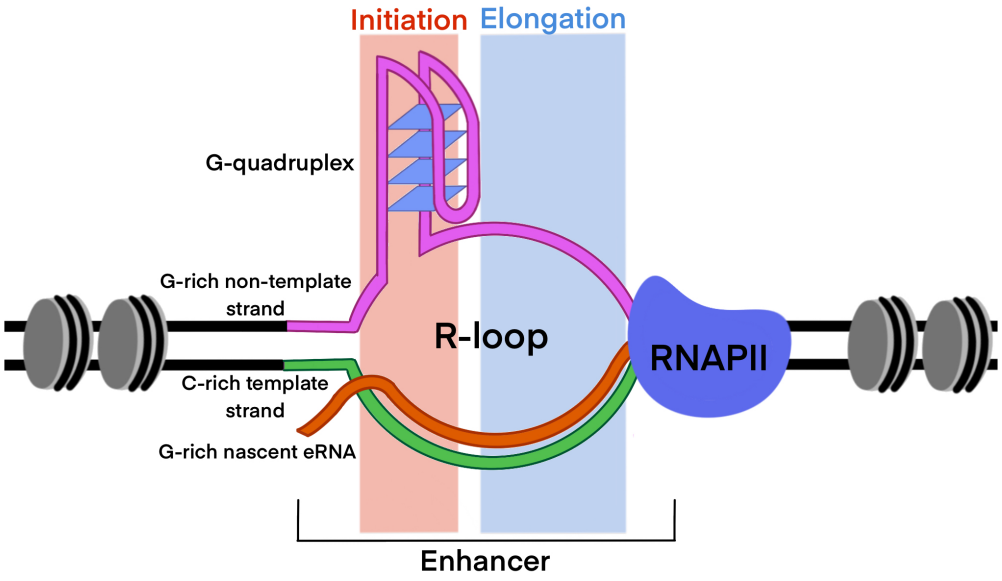



FIGURE 1.9: R-loops and DNA G-quadruplexes at enhancers
DNA/RNA hybrid structures, termed R-loops, are regions in which a nascent RNA anneals to the template strand of the DNA duplex. R-loop formation occurs preferentially at consecutive clusters of guanine residues in the non-template DNA strand within the R-loop initiation zone. Downstream of this, high-density G-rich sequences in the R-loop elongation zone facilitate extension of the RNA:DNA hybrid. R-loop formation is also dependent on the stability of the resulting RNA:DNA hybrid and the unhybridised single-stranded (ss)DNA stretch. G-richness in the non-template strand imparts greater stability to R-loop structures; in particular, clusters of G tracts can fold into stable G-quadruplex structures that can stabilise the displaced DNA strand.

RNA might promote their involvement in the formation of phase-separated biomolecular condensates. This compartmentalisation of RNAs and proteins into sub-organellar condensates may compensate for the lack of sequence specificity in RNA-protein interactions. Notably, high intrinsic disorder in proteins is a prerequisite for condensate formation and many factors that bind to eRNAs do so through disordered regions (Boija et al., 2018, Cho et al., 2018, Roden and Gladfelter, 2021).

Assembly of the transcriptional machinery into biomolecular condensates at enhancers is a tightly-regulated process that is dependent on both RNA sequence features, such as RNA identity, length and modifications, and RNA properties intrinsically encoded in their sequence, such as RNA structure (reviewed in Roden and Gladfelter, 2021). For example, the negative charge of RNA promotes complex coacervation – the demixing of oppositely-charged macromolecular species such as polymers and proteins (Drobot et al., 2018, Frankel et al., 2016, Pak et al., 2016). Such dependence on electrostatic interactions may explain the largely sequence-independent actions of diverse RNAs in phase-separated condensates (Drobot et al., 2018, Frankel et al., 2016). However, phase separation characteristics may not be entirely independent of sequence. Poly(A), poly(C) and poly(U) RNAs are unlikely to form persistent tertiary structures and tend to form

dynamic, liquid-like condensates when interacting with disordered protein models. In contrast, poly(G)-rich RNA sequences form highly stable G-quadruplexes (Petrovic and Polavarapu, 2008) and have a higher propensity to form a fractal network of structures characteristic of kinetically arrested gel-like condensates (Boeynaems et al., 2019, Sciortino et al., 1993). Condensate formation may also be influenced by additional, partially sequence-encoded RNA features, such as RNA modifications and RNA:RNA interactions (Langdon et al., 2018, Ries et al., 2019).

A model is emerging in which the transition from transcriptional initiation into elongation is driven by the shuttling of RNAPII between distinct transient condensates at gene promoters and within gene bodies, whereby condensates function to concentrate factors required for initiation and elongation, respectively (Cramer, 2019, Guo et al., 2019). Such self-compartmentalisation of transcription occurs in a phosphoform-specific manner dependent on regulatory CDKs (Guo et al., 2019). Phosphorylation of the RNAPII CTD promotes the exchange of RNAPII from condensates involved in transcription initiation, containing the Mediator co-activator complex (Kornberg, 2005), to those involved in RNA processing and splicing (Guo et al., 2019). Considering the established role for RNAs in phase separation, eRNAs could potentially play a functional role in phosphorylation-mediated transcriptional condensate partitioning. eRNAs interact with the Mediator complex to regulate its chromatin localisation, connecting transcriptional activators bound at enhancers with promoter-bound RNAPII (Lai et al., 2013, Taatjes, 2010, Wang et al., 2011a). Although no direct interactions of eRNAs with Mediator subunits have been observed, small interfering RNA (siRNA)-mediated knockdown of eRNAs results in decreased enhancer-promoter interactions and decreased occupancy of Mediator and RNAPII at target promoters (Hsieh et al., 2014, Lai et al., 2013). Conversely, direct activating interactions have been identified between eRNAs and the cyclin T1 component of P-TEFb (Zhao et al., 2016). Hyperphosphorylated RNAPII CTD can be incorporated into phase-separated condensates formed by the intrinsic disorder of P-TEFb. Cyclin T1 promotes phase separation, compartmentalising the kinase with the substrate CTD for robust hyperphosphorylation and subsequent transcriptional elongation and RNA processing (Lu et al., 2018). Overall, overlapping functions of eRNAs and condensates thus suggest roles for eRNAs in the partitioning behaviour of both initiation and elongation condensates.

Although RNAs play a broad role in determining the formation of membraneless compartments, it has been suggested that they may be less significant for transcriptional condensate formation (Decker et al., 2022). Super-enhancers foster the crowding of TFs and transcriptional coactivators at densities conducive to transcriptional condensate formation (Shrinivas et al., 2019). However, nuclear RNA degradation has little effect on super-enhancer condensates, suggesting that eRNAs are not required for the integrity or

maintenance of transcriptional condensates (Decker et al., 2022). Although eRNAs may not play a structural role in the assembly of transcriptional condensates, they may have a regulatory function in their formation and dissolution (Decker et al., 2022, Henninger et al., 2021). Early in transcription, condensate formation is driven by favourable electrostatic interactions arising from the low levels of non-coding RNAs, including eRNAs, synthesised from enhancers and promoter-proximal regions by RNAPII (Figure 1.10) (Adelman and Lis, 2012, Core and Adelman, 2019, Henninger et al., 2021, Kim et al., 2010, Seila et al., 2008). RNAs may scaffold multivalent interactions between condensate components implicated in different stages of transcription, coupling distinct steps in transcription for synergistic activation (Ma et al., 2021, Narita et al., 2021), alongside directly influencing condensate dynamics to control transcriptional regulation (Henninger et al., 2021). Upon release of RNAPII into productive elongation, the synthesis of long genic RNAs leads to repulsive-like electrostatic effects that result in condensate dissolution (Banerjee et al., 2017, Henninger et al., 2021, Milin and Deniz, 2018), suggesting that short-lived eRNAs are more likely to contribute to the early stages of condensate formation than their later dissolution.

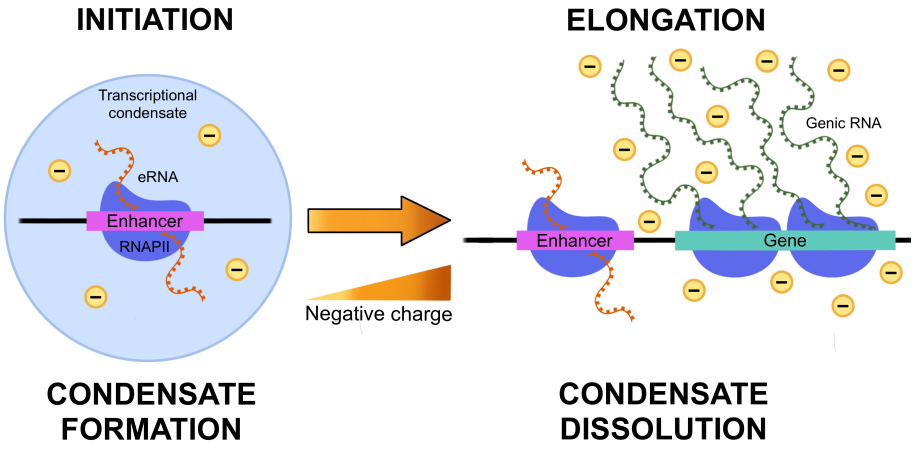



FIGURE 1.10: Model of eRNA-mediated control of transcriptional condensate formation
Early in transcription initiation, nascent eRNAs transcribed from enhancers stimulate the formation of transcriptional condensates via phase separation. Upon transition into transcriptional elongation, the production of long genic RNAs results in electrostatic repulsion that stimulates dissolution of the transcriptional condensates.

#### 1.7.0.1 Post-transcriptional modification of eRNAs

The deposition of chemical modifications on eRNAs, such as N6-methyladenosine (m6A), may affect transcriptional condensate assembly. Both the METTL3/METTL14/WTAP m6A methyltransferase complex (MTC) and the demethylase ALKBH5 associate with enhancers (Lee et al., 2021, Xu et al., 2022). Consequently, m6A is pervasively deposited

on nascent eRNAs (Lee et al., 2021), and eRNA abundance is dynamically regulated by m6A methylation (Lee et al., 2021, Xu et al., 2022). Notably, m6A signals on eRNAs are highly correlated with features of active enhancers (Lee et al., 2021), suggesting that they act to promote gene expression.

Mechanistically, m6A-modified eRNAs recruit nuclear m6A reader proteins such as hnRNPG (RBMX) and YTHDC1 to active enhancers, promoting productive enhancer transcription. The protective functions of m6A readers shield eRNAs from processing by the endonuclease subunit of the Integrator complex, INTS11, which would otherwise cleave nascent RNAs, resulting in premature transcriptional termination (Xu et al., 2022). In addition, eRNA transcripts, together with YTHDC1, assemble into nuclear phase-separated condensates that can augment BRD4 coactivator condensates to modulate downstream gene expression (Lee et al., 2021).

m6A modification levels are largely hard-wired in nascent RNA sequences via the presence of 'GGACT' motifs, consistent with the canonical RRACH methylation motifs seen in mRNAs (Dominissini et al., 2012, Lee et al., 2021, Meyer et al., 2012). With the deposition of m6A, and potentially other chemical modifications, being dependent on short degenerate sequence motifs, such motifs may act as fundamental elements in nascent RNAs for the recruitment of reader proteins to chromatin, bestowing some level of sequence-encoded control over the transcription process.

## 1.8 3' UTRs: The fine-tuners of post-transcriptional control

Post-transcriptionally, mRNAs progress through multiple regulatory steps in the nucleus and the cytoplasm that control their processing, localisation, stability and translation (reviewed in Moore, 2005). This post-transcriptional fine-tuning of gene expression is, in part, genetically encoded, through alternative polyadenylation, with differential 3' UTR isoform expression exerting influence over the binding of *trans*-acting factors, such as microRNAs (miRNAs) and RNA-binding proteins (RBPs) to the 3' UTR, directly impacting the stability and translation of target mRNAs (Chen et al., 2001, Humphreys et al., 2005, Ji et al., 2011a, Lykke-Andersen and Wagner, 2005a, Nelson et al., 2004, Tian et al., 2005).

### 1.8.1 Cleavage and polyadenylation

mRNA polyadenylation, that is, the formation of mature 3' ends, is a critical cellular step in the synthesis of most eukaryotic mRNAs. Polyadenylation is a two-step process consisting firstly of endonucleolytic cleavage of pre-mRNA downstream from the coding sequence (CDS), followed by addition of the polyadenosine (poly(A)) tail at the 3' end (reviewed in Wahle and Kühn, 1997). Over 70% of human mRNAs contain more that one polyadenylation site (PAS) (Derti et al., 2012, Hoque et al., 2013), producing alternatively polyadenylated mRNAs with alternative 3' UTRs. Alternative polyadenylation (APA) is controlled by the combinatorial effects of *cis*-acting elements, including not only the strength of the PAS but also epigenetic factors such as DNA methylation and histone modifications (Barski et al., 2007, Cowley et al., 2012, Ji et al., 2011b, Kaczmarek Michaels et al., 2020, Legendre and Gautheret, 2003, Monk et al., 2011, Spies et al., 2009, Wood et al., 2008), and *trans*-acting factors such as 3' processing machinery and RBPs (Bava et al., 2013, Katz et al., 2010, Kühn et al., 2009, Shi et al., 2009).

### 1.8.2 3' UTR function: mRNA stability, sublocalisation and translational control

Translational regulation is tightly coupled to mRNA turnover and stability. 3' UTRs house multiple *cis*-acting regulatory sequences that bestow control over mRNA stability. These regulatory regions promote binding of *trans*-acting factors, including RBPs and miRNAs, impacting mRNA stability, and ultimately translation efficiency (Figure 1.11) (reviewed in Mayr, 2017).

AU-rich elements (AREs) in 3' UTRs are primarily associated with mRNA destabilisation and translational repression (Chen et al., 2001, Kruys et al., 1989, Meijlink et al., 1985, Shaw and Kamen, 1986). AREs are directly recognised by a diverse variety of RBPs, with specific binding of RBPs destabilising, or in some cases, stabilising, their target mRNAs. AREs serve as binding sites for members of the tristetraprolin (TTP) RBP family and human antigen R (HuR), exhibiting negative and positive regulatory effects on mRNA stability, respectively (Brennan and Steitz, 2001, Chen et al., 2001, Lai et al., 2000, Lykke-Andersen and Wagner, 2005b, Peng et al., 1998, Sandler et al., 2011). Hence, ARE recognition and competitive binding of RBPs to mRNA transcripts are deterministic of mRNA fate (reviewed in García-Mauriño et al., 2017).

miRNAs comprise a family of small single-stranded non-coding RNAs that function as negative regulators of gene expression through their imperfect base-pairing with target sites located within the 3' UTR. miRNA-mediated recruitment of effector proteins

including the Argonaute complex to the target mRNA, drive mRNA degradation or translational repression (reviewed in Bartel, 2004, O'Brien et al., 2018), with more than 60% of mammalian protein-coding genes predicted to undergo miRNA-driven regulation (Friedman et al., 2009).

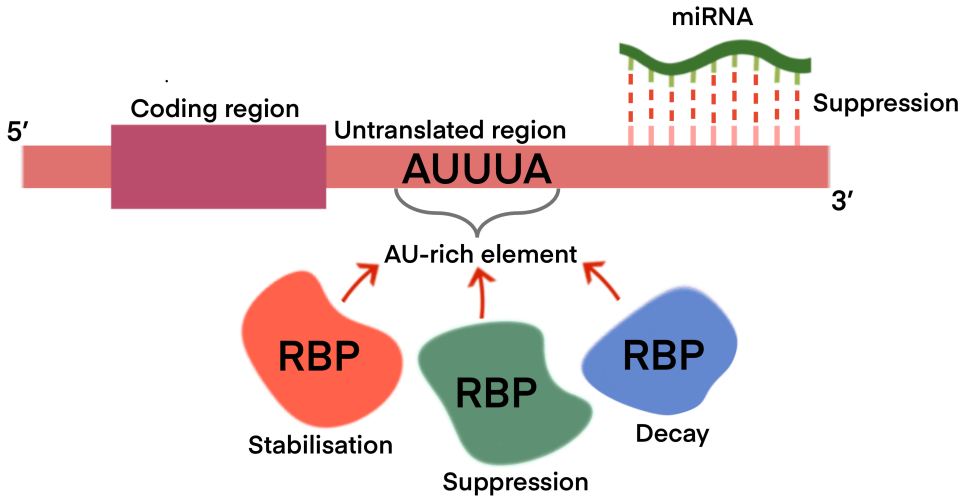



FIGURE 1.11: 3' UTR functionality is controlled by *cis*-regulatory elements
3' UTRs contain regulatory sequences, including AREs and miRNA-binding sites, that control mRNA turnover, stability, localisation and translation

The complex folding of the 3' UTR RNA can positively or negatively impact proximal regulatory sequences. RNA structures can constitute determinants of RBP binding site recognition, or occlude binding of miRNAs to their target sites. Whereby, steric hindrance can cause direct competition between RBPs and miRNAs for target sites in 3' UTRs. Such antagonism has been suggested in the case of HuR binding, which prevents the access of miRNAs to nearby and overlapping binding sites, thus alleviating miRNA-driven repression (Kundu et al., 2012, Mukherjee et al., 2011). This highlights the intricacies of post-transcriptional regulatory mechanisms, as tuned by the co-evolution of 3' UTR structures, regulatory sequences and trans-acting RBPs.

Proteins frequently exhibit an asymmetrical distribution within specific subcellular compartments, aligning with their designated functions. This spatial arrangement is often influenced by the prior localisation of the cognate mRNA, with their local translation facilitating rapid protein synthesis to ensure precise and robust temporal regulation in response to external stimuli (Lécuyer et al., 2007). Overall mRNA stability can impact localisation, whereby mRNAs with a long half-life may diffuse passively to a subcellular compartment (Chang et al., 2004, Forrest and Gavis, 2003, Semotok et al., 2005, Tushev et al., 2018). Additionally, *cis*-regulatory localisation elements within the 3' UTR can permit the active transport of mRNAs. Site-specific binding of RBPs within the

3' UTR promotes recruitment of effector proteins that repress translation and mediate interactions with motor proteins. The resultant ribonucleoprotein (RNP) complex is then transported by the cytoskeleton to its destination, where the mRNA undergoes spatiotemporal-specific derepression to enable translation (reviewed in Besse and Ephrussi, 2008).

### 1.8.3 Alternative polyadenylation: The relevance of 3' UTR isoforms

APA can promote quantitative alterations in gene expression, with a generalised relationship observed between 3' UTR length and mRNA stability. Dependent on choice of proximal versus distal PASs, APA produces mRNA isoforms that can either contain or lack the full complement of regulatory elements within their 3' UTRs. Shorter 3' UTR isoforms can contain less AREs and miRNA binding sites, potentially serving to increase mRNA stability and translation compared to their longer 3' UTR counterparts (Halees et al., 2008, Legendre and Gautheret, 2003, Sandberg et al., 2008). Indeed, mRNA transcripts containing shorter 3' UTRs have been observed to produce higher levels of protein (Figure 1.12) (Mayr and Bartel, 2009, Sandberg et al., 2008).

Dynamic alterations in APA profiles are observed in cell proliferation (Sandberg et al., 2008), differentiation and development (Ji et al., 2009, Lianoglou et al., 2013). Ubiquitously expressed genes utilise APA to achieve cell and tissue type-specific gene expression (Lianoglou et al., 2013), yielding distinct APA profiles in different cell and tissue types (Zhang et al., 2005). Further, some tissue types demonstrate a bias for the global production of short or long 3' UTR isoforms (Li et al., 2012b, Ramsköld et al., 2009, Shepard et al., 2011, Smibert et al., 2012, Ulitsky et al., 2012, Zhang et al., 2005). While neuronal cells are biased towards expression of long 3' UTR isoforms, with preferential usage of distal PASs (Guvenek and Tian, 2018, Miura et al., 2013, Zhang et al., 2005), immune cells display global preference for proximal PASs, producing relatively shorter 3' UTR isoforms (Cheng et al., 2020, Singh et al., 2018). This highlights the prevalence of APA as a mechanism underpinning cell type-specific post-transcriptional regulation.

## 1.9 The non-coding genome and disease: Regulatory elements matter

### 1.9.1 Genome Wide Association Studies

Unlike Mendelian disorders, complex diseases arise from the intricate interplay of genetic and environmental factors, with each contributing to disease susceptibility. Lacking a

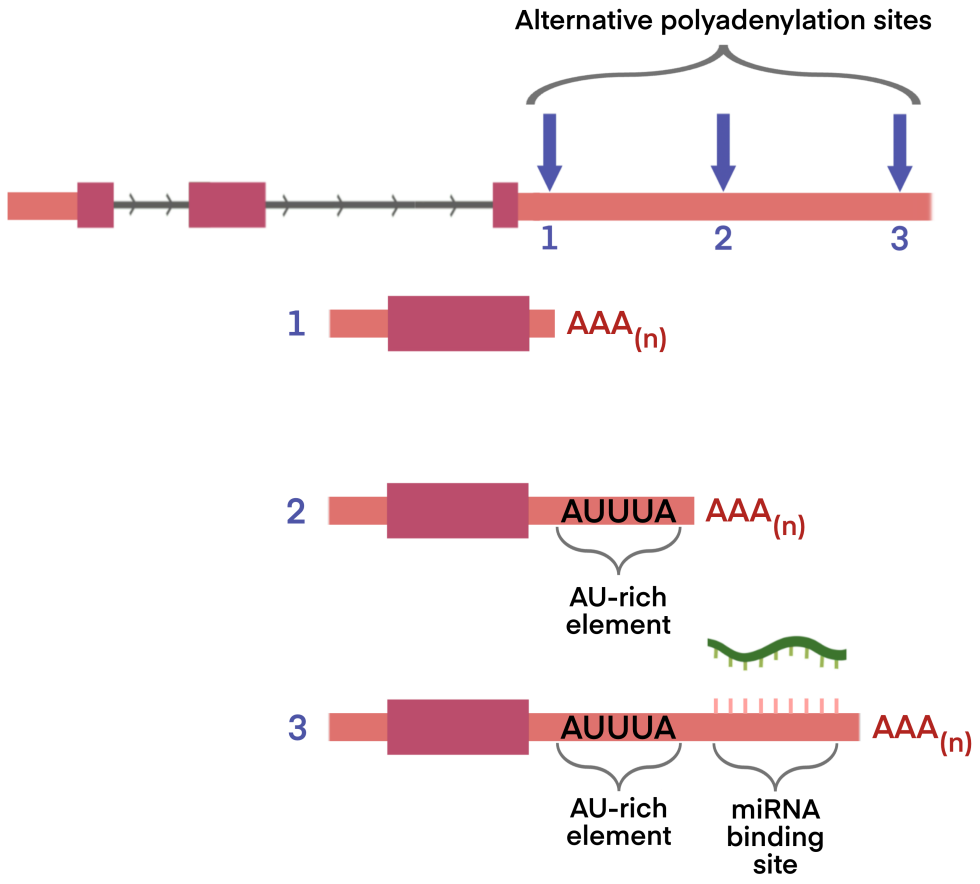



FIGURE 1.12: The role of APA in controlling gene expression
Choice of PAS, deterministic of differential 3' UTR isoform expression, bestows control over 3' UTR-mediated regulation. Inclusion or exclusion of regulatory elements in long and short 3' UTR isoforms, respectively, may contribute to mRNA stability, sublocalisation and translation

single causative variant, genome-wide association studies (GWAS) aim to identify single nucleotide polymorphisms (SNPs) that contribute to disease aetiology via cumulative contributions. GWAS in human populations have reproducibly identified thousands of SNPs significantly associated with hundreds of complex traits and diseases (Hindorff et al., 2009, Welter et al., 2014). Yet, SNPs discovered by GWAS account for only a fraction of the genetic variation of complex human traits, with many causal variants remaining undetected due to their small individual effect sizes (Hindorff et al., 2009, Yang et al., 2010).

In contrast to single gene disorders, which are largely driven by protein-coding variants that often result in substantial phenotypic effects, complex traits are caused primarily by non-coding variants which presumably affect gene regulation (Pickrell, 2014, Welter et al., 2014). The vast majority of SNPs associated with human complex diseases by GWAS, and many of those associated with cancer, map to the non-coding genome,

with causal disease variants enriched within *cis*-regulatory elements, particularly promoters and enhancers (Bal et al., 2022, ENCODE Project Consortium, 2012, Farh et al., 2015, Hindorff et al., 2009, van Arensbergen et al., 2019). Many complex trait alleles may therefore act by altering regulatory elements to influence gene regulation in a cell type-specific manner, specifically affecting cell types most relevant to disease phenotype (Figure 1.13) (Farh et al., 2015, Finucane et al., 2015, Roadmap Epigenomics Consortium et al., 2015, Trynka et al., 2013). The widespread localisation of genetic variants associated with disease in non-coding regulatory DNA has implications for defining the associations between genotype and disease phenotype. Whereby, connecting enhancers harbouring GWAS SNPs to promoters of distant genes highlights the abundance of plausible causal genes that may explain genotype-phenotype associations. Such association data may act to unify seemingly unrelated genetic variants which drive pathogenicity through common cellular networks thereby resulting in related phenotypes (Maurano et al., 2012).

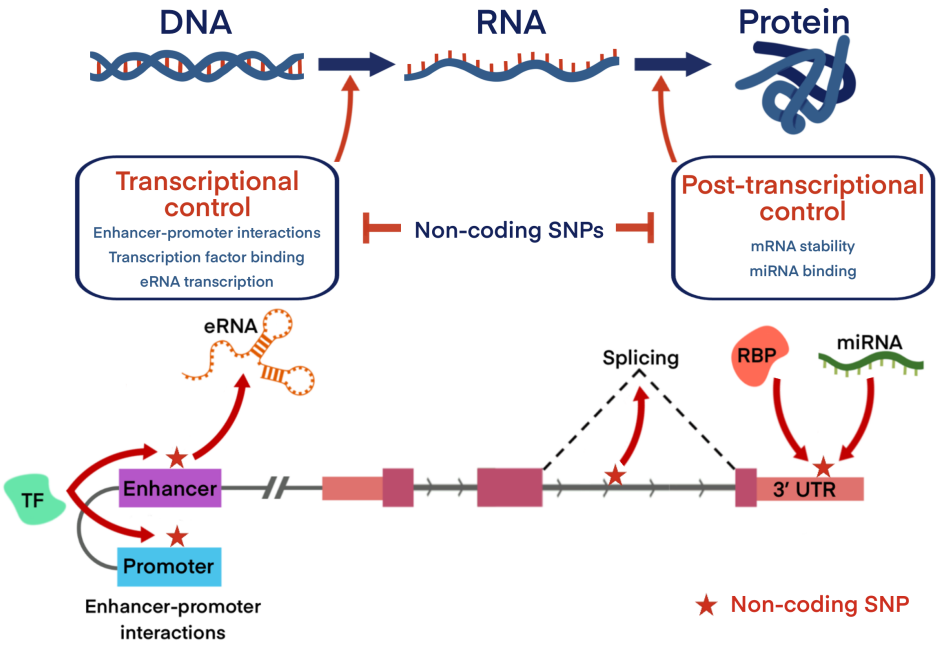



FIGURE 1.13: Complex disease variants map to *cis*-regulatory elements
Non-coding SNPs can alter transcriptional control, by impacting enhancer-promoter interactions and transcription of functional eRNAs. Non-coding SNPs can also alter post-transcriptional control, impacting messenger RNA stability by dysregulated RNA-binding protein or microRNA binding within the 3' untranslated region.

Due to the prevalence of GWAS signals within regulatory elements, notably with ~90% of trait-associated variants lying within the non-coding genome (Maurano et al., 2012), there have been widespread efforts to integrate GWAS data with expression quantitative trait loci (eQTL) mapping. eQTL mapping functions to identify associations between individual-level genotypes and quantitative molecular impacts on gene expression (reviewed in Nica and Dermitzakis, 2013, Rockman and Kruglyak, 2006). Integrating eQTL

data with GWAS-associated variants could therefore capture trait-associated mutations that make significant contributions at the level of gene regulation (Gamazon et al., 2015, Giambartolomei et al., 2014, Gusev et al., 2016, Hormozdiari et al., 2016, Zhu et al., 2016). However, to date, most GWAS hits cannot be explained by currently identified eQTLs (Chun et al., 2017, Connally et al., 2022), with only 43% of GWAS hits colocalising with eQTLs (GTEx Consortium, 2020). This may be explained by the GWAS hits only colocalising with eQTLs in specific cell type-specific or developmental time point-specific contexts, or in response to physiological stimuli (Jerber et al., 2021, Kim-Hellmuth et al., 2020, Mu et al., 2021, Strober et al., 2019), and equally do not consider potential contributions of processes such as alternative splicing and polyadenylation to complex traits (Li et al., 2021, 2016). Such effects are challenging to capture in eQTL assays using postmortem adult whole tissues. Additionally, even if the phenotypic effects of complex traits were entirely mediated by alterations in gene expression, many GWAS hits would not be discovered as significant eQTLs, and vice versa, due to the inherent differences in GWAS and eQTL mapping strategies (Figure 1.14) (Connally et al., 2022, Mostafavi et al., 2023). Therefore, the disconnect between GWAS and eQTL data further contributes to the "missing heritability" of complex traits identified by GWAS, with lack of disease-associated eQTLs recently being referred to as "missing regulation" (Connally et al., 2022). Integrating GWAS data with regulatory associations has potential to enhance searches for unidentified disease loci, with the colocalisation of GWAS and eQTL data providing power to identify variants that causally impact both disease phenotype and gene expression, ultimately providing evidence for the regulatory mechanisms that likely underpin many complex diseases.

### 1.9.2 Enhancers in disease

*The following subsection is adapted from Harrison and Bose (2022)*

The dynamic nature of enhancer elements allows for the specific modulation of the quantitative, temporal and spatial expression of associated genes. This modularity implies that mutations within enhancer elements often have lower burden on fitness than coding mutations, hence allowing such mutations to reach higher frequency in the population. Variants within *cis*-regulatory elements are predicted to be of small effect size, with the quantitative contribution of many variants resulting in complex disease by the cumulation of small phenotypic effects, for example via the convergent modulations of TF binding characteristics (Maurano et al., 2012).

Although gene expression patterns and TF binding preferences are largely conserved across mammals (King and Wilson, 1975, Nitta et al., 2015), enhancers show changes in

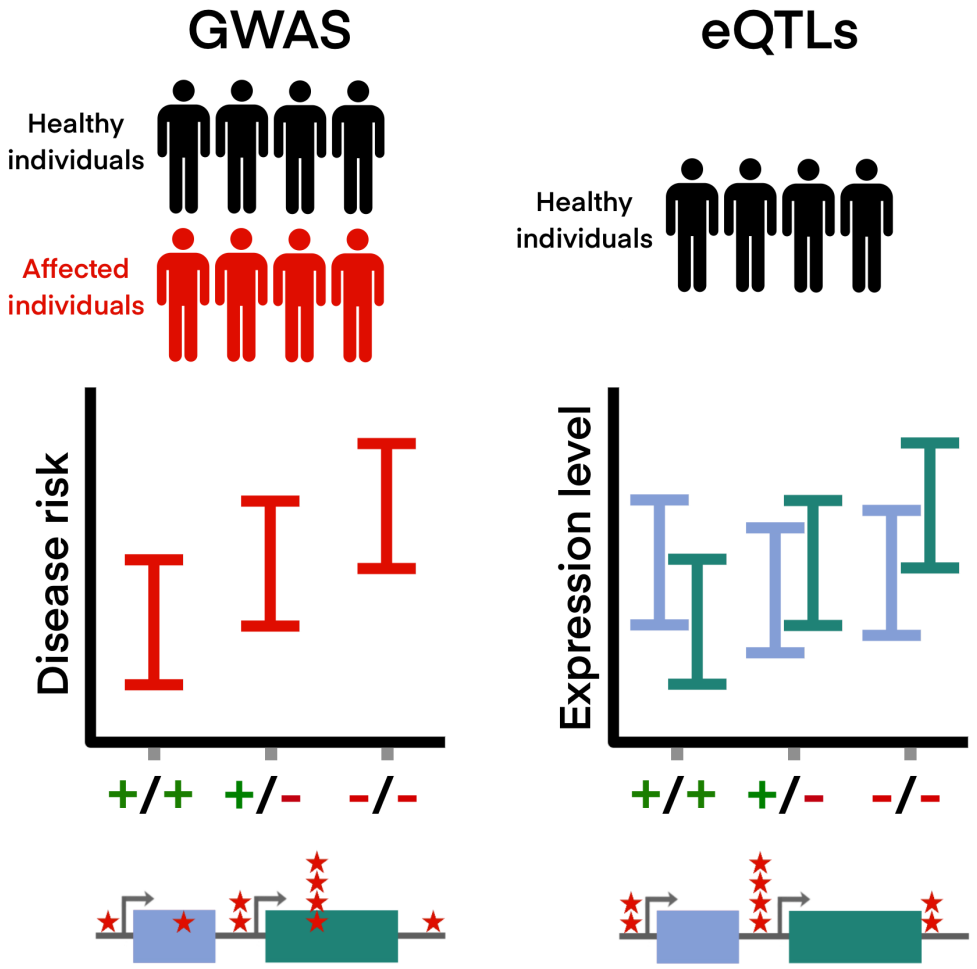



FIGURE 1.14: Integrating GWAS associations with eQTLs may provide power to understanding regulatory genotypes underpinning disease
GWAS signals identify association between variants and complex traits, including disease, whereas eQTLs are identified by examining the correlation between genetic variants and gene expression levels. GWAS hits are intentionally selected for their measurable impact on organismal traits or diseases, while in eQTL mapping, variants must account for a significant percentage of the variance in gene expression. Therefore GWAS hits that are not explicitly linked to disease via alterations in gene expression would not be captured by eQTL mapping, and changes in gene expression caused by eQTLs that are not explicitly linked to disease phenotypes would not be captured by GWAS.

function and sequence over evolutionary time. In contrast to the longstanding belief that sequence conservation implies functionality, many enhancers appear to maintain functions that are relatively independent of precise sequence conservation. Such a dichotomy raises questions about the relationship between enhancer sequence and enhancer function, especially when considering eRNAs. For example, how do you reconcile a functionally important eRNA with a lowly conserved sequence? In addition, SNPs associated with human complex diseases are markedly enriched within *cis*-regulatory elements, particularly enhancers (Bal et al., 2022, ENCODE Project Consortium, 2012, Farh et al., 2015, Hindorff et al., 2009, van Arensbergen et al., 2019). Such localisation of disease

variants within enhancers supports the importance of sequence for regulating enhancer function. However, it is also known that enhancers exhibit promiscuous behaviours – a single enhancer can govern different effects in different cell types (Fukaya et al., 2016) – suggesting that enhancer function is unlikely to be driven by intrinsic sequence features alone.

If enhancer sequence has little relevance to functionality, why do SNPs associated with human complex disease so frequently localise to enhancers (ENCODE Project Consortium, 2012, Farh et al., 2015, Hindorff et al., 2009)?

#### 1.9.2.1 Alterations in chromatin architecture

Pathogenic phenotypes can arise following disruption of chromatin organisation, namely disruption of TAD architecture and rewiring of long-range regulatory interactions. Alterations in higher-order chromatin topology have potential to result in ectopic gene expression. Loss of insulation between topological domains, via loss of CTCF binding or structural variants spanning TAD boundaries, is associated with aberrant gene activation by enhancers, and implicated in disease pathogenesis, as particularly evidenced in cancer (Flavahan et al., 2016, Hnisz et al., 2016, Katainen et al., 2015, Lupiáñez et al., 2015). In fact, disease-associated SNPs, particularly those implicated in cancer, are preferentially located at TAD borders, implying that genetic variation proximal to TAD boundaries may weaken insulation of adjacent domains, promoting spurious interactions between enhancers and genes in neighbouring TADs (Jablonski et al., 2022).

With CTCF playing a key role in the anchoring of chromatin loops, complex disease-associated SNPs that modify CTCF binding sites, and in turn, nucleosome positioning and 3D chromatin architecture, can result in ectopic gene expression, as documented for asthma risk. SNPs within the 17q12-q21 asthma risk locus are linked to changes in CTCF binding patterns, with a domain-wide *cis*-regulatory impact to long-range chromatin interactions (Berlivet et al., 2012, Schmiedel et al., 2016, Verlaan et al., 2009). The presence of two linked asthma risk alleles within an enhancer proximal to the *ORMDL3* gene results in CTCF binding site switching, resulting in genotype-dependent changes in gene expression. One asthma risk SNP (rs4065275) introduces a *de novo* binding site for CTCF, wheras a linked downstream SNP (rs129362) disrupts a CTCF binding site, together resulting in the interchanging of two CTCF sites in the vicinity of the *ORMDL3* gene locus. The 3D organisation of the asthma risk haploblock is modified in these asthma risk alleles to favour recruitment of a distal *cis*-regulatory element to the *ORMDL3* promoter region, resulting in misexpresssion of the *ORMDL3* gene (Figure 1.15a) (Schmiedel et al., 2016).

Further, SNPs within putative enhancer regions have been linked to neurodegenerative disease. The *PM20D1* gene locus has been implicated in Alzheimer's disease (AD) risk, whereby the *PM20D1* promoter is consistently hypermethylated in AD patients, compared to healthy individuals, and thus increased *PM20D1* expression has been proposed to have neuroprotective properties (Sanchez-Mut et al., 2018). In agreement, increased *PM20D1* expression has been observed following induction of AD-associated neurotoxic insults, such as reactive oxygen species (ROS) and amyloid-$\beta$ (A$\beta$) *in vitro* and *PM20D1* expression is increased in AD mouse models and in AD post-mortem human frontal cortex brain samples (Sanchez-Mut et al., 2018). Additionally, *PM20D1* overexpression *in vitro* and *in vivo* was demonstrated to reduce ROS-induced cell death and A$\beta$ levels, respectively. The *PM20D1* locus is coupled to an AD-associated haplotype enriched for enhancer-associated epigenetic marks, containing the AD risk SNPs rs708727 and rs960603, whereby the putative enhancer contacts the *PM20D1* promoter via a CTCF-mediated chromatin loop in haplotype-dependent manner (Sanchez-Mut et al., 2018). Differential chromatin loop-dependent *PM20D1* expression and promoter methylation are therefore determined by this AD risk haplotype, with the CTCF-mediated looping interaction weakened in *PM20D1* transcriptionally silent, methylated risk haplotype carrier cells. AD phenotype may therefore manifest in risk haplotype carriers due to impaired cellular defence to neurotoxicity as caused by decreased *PM20D1* expression and promoter hypermethylation, demonstrating the potential impact of enhancer function in AD predisposition (Figure 1.15b) (Sanchez-Mut et al., 2018).

#### 1.9.2.2 Alterations in transcription factor binding

Through sequence-specific TF binding to DNA, TFs can directly interpret regulatory elements, initially by recognition of binding motifs within regulatory DNA (reviewed in Spitz and Furlong, 2012). Therefore, disruption, or indeed creation, of TF binding sites within enhancers may play a central role in manifestation of disease phenotype, and in complex disease, likely via the contribution to overall disease risk.

##### 1.9.2.2.1 Disruption of TF binding sites

The $\alpha$-synuclein gene *SNCA* is among the strongest risk loci associated with sporadic Parkinson's disease (PD) (Nalls et al., 2014). Further, genomic duplications of *SNCA* in familial PD cases demonstrate that a 50% increase in *SNCA* expression is sufficient for development of the autosomal dominant PD (Chartier-Harlin et al., 2004, Ibáñez et al., 2009). A common PD-associated SNP (rs356168) that lies in an intronic enhancer in

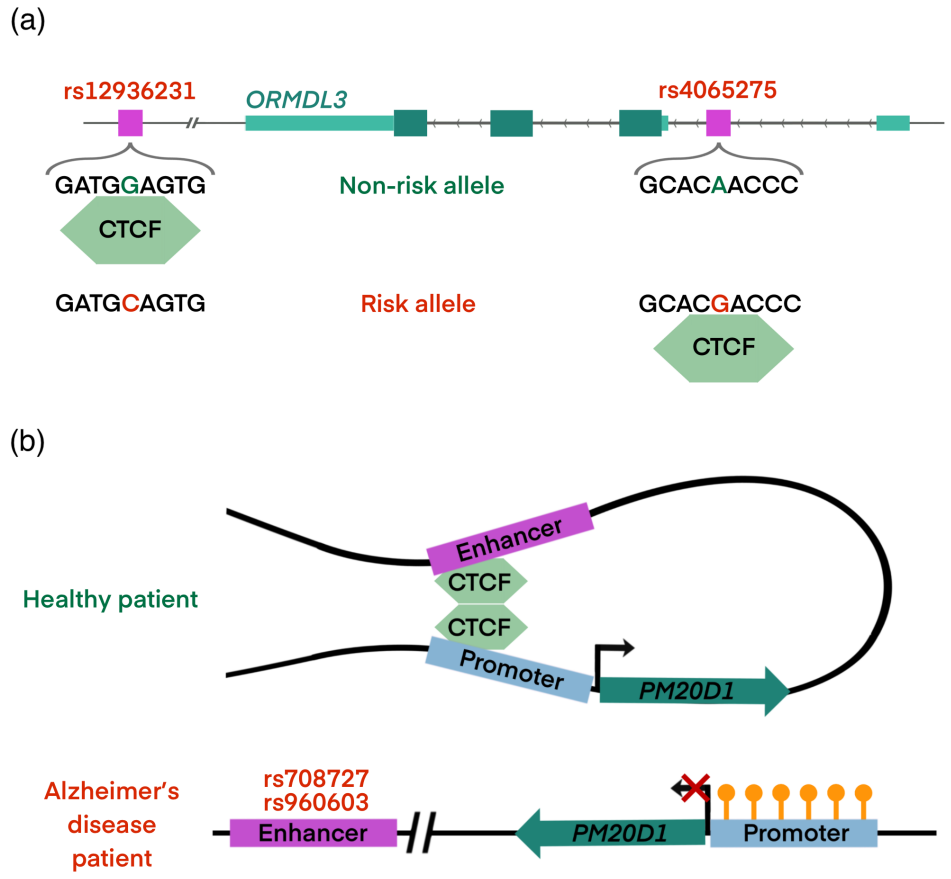



FIGURE 1.15: Enhancer variants can alter chromatin architecture
(a) The presence of linked risk SNPs within an enhancer in the 17q21-q21 asthma risk locus cause a switch in CTCF binding sites that alter 3D enhancer-promoter interactions (b) CTCF-mediated enhancer-promoter interactions are weakened by the presence of a AD risk SNP within a *PM20D1* enhancer

the *SNCA* gene causes a significant increase in total *SNCA* expression, likely via disruption TF-mediated repression of the enhancer. Carriers of the PD protective A-allele at rs356168 exhibit efficient binding of the brain-specific TFs EMX2 and NKX6-1 at the enhancer, leading to suppression of enhancer activity and therefore lower expression of *SNCA*. Conversely, carriers of the PD risk G-allele disrupt the sequence-specific binding of EMX2 and NKX6-1 at the enhancer, and consequently the active enhancer causes PD risk-associated upregulation of *SNCA* expression (Figure 1.16a) (Soldner et al., 2016).

At the genomic level, variation within the clusterin gene (*CLU*) locus has been associated with age-related pathologies, including pseudoexofoliation and AD (Chen et al., 2012b, Krumbiegel et al., 2009, Lambert et al., 2009, Padhy et al., 2014). Clusterin is an extracellular chaperone that influences the formation and the disease-associated toxicity of amyloid deposits via interactions with prefibrallar structures (Yerbury et al., 2007). Both pseudoexofoliation and AD share pathological alterations in the deposition of fibrallar protein aggregates (Karch et al., 2012, Padhy et al., 2014, Schrijvers et al.,

2011). A common disease variant at the *CLU* gene locus, rs2279590, associated with pseudoexofoliation and AD risk (Krumbiegel et al., 2009, Lambert et al., 2009, Padhy et al., 2014) is located within a putative intronic enhancer that differentially regulates *CLU* expression (Padhy et al., 2017). Individuals homozygous for the disease risk G-allele at rs2279590 exhibit a two-fold increase in clusterin expression in lens capsule tissue compared to homozygous carriers of the protective A-allele (Padhy et al., 2014), indicative of a regulatory role of the rs2279590-containing locus over *CLU* expression. The rs2279590 A-allele, but not the G-allele, constitutes a TF binding site for heat shock factor-1 (HSF1), with HSF1 binding abrogating the enhancer effect of the locus, reducing not only *CLU* expression, but also expression of the neighbouring *PTK2B* and *EPHX3* genes (Padhy et al., 2017), both of which are linked to AD risk by GWAS (Chapuis et al., 2017, Lambert et al., 2013). Thereby, in carriers of the G-allele, without the protective suppression of the intronic enhancer, the cytotoxic upregulation of the disease risk genes *CLU*, *PTK2B* and *EPHX3* may promote the pathogenesis of psuedoexofoliation and AD (Figure 1.16b) (Padhy et al., 2017).

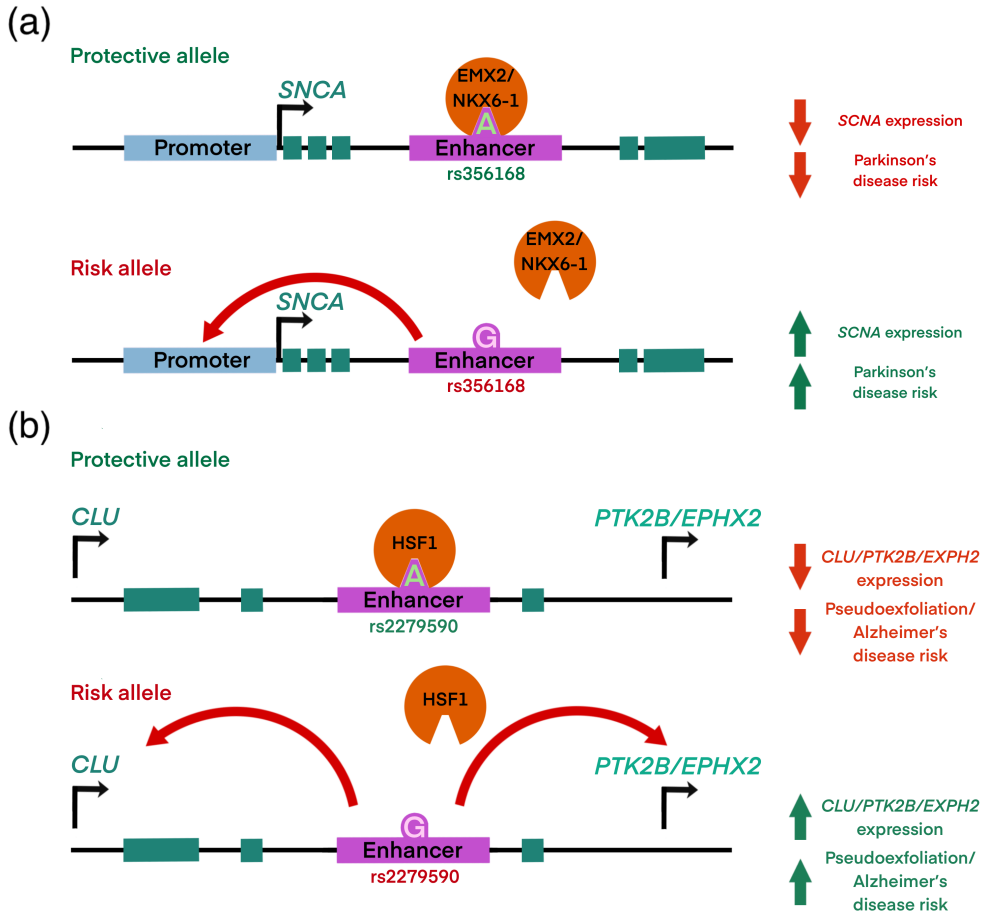



FIGURE 1.16: Enhancer variants can disrupt TF binding sites
Presence of disease risk SNPs within enhancers can disrupt the binding sites of transcriptional repressors, resulting in pathogenic gene upregulation, as evidenced in neurodegenerative diseases.

#### 1.9.2.2.2 Introduction of *de novo* TF binding sites

Enhancer variants can have clinical implications in cancer. For example, the rs11672691 enhancer variant within the 19q13 prostate cancer risk locus underpins the oncogenic regulatory circuit of aggressive prostate cancer, synergising with the oncogenic TF HOXA2, to drive expression of its eQTL genes, *CEACAM21* and *PCAT19* (Gao et al., 2018, Xia and Wei, 2018). The rs11672691 prostate cancer risk SNP, located intronically within the long non-coding RNA *PCAT19*, is predictive of prostate cancer aggressiveness, progression and mortality (Amin Al Olama et al., 2013, Shui et al., 2014). Mechanistically, the presence of the prostate cancer risk G-allele at rs11672691 generates a preferential binding site for HOXA2, with HOXA2 binding at the *PCAT19* intronic enhancer contributing to enhanced *PCAT19* and *CECAM21* expression, promoting prostate cancer cell proliferation, aggressiveness and metastatic capacity (Figure 1.17a) (Gao et al., 2018, Hua et al., 2018, Xia and Wei, 2018). The homozygous G-allele genotype is associated with higher *PCAT19* and *CEACAM21* transcriptional levels compared to heterozygous and homozygous reference A-allele genotypes. Thus, clinically, the rs11672691 genotype, coupled with *PCAT19* and *CECAM21* expression, have potential as biomarkers for predicting prostate cancer prognosis, relapse and survival (Gao et al., 2018, Xia and Wei, 2018).

Toll-like receptor 4 (TLR4) functions as pattern-recognition receptor in innate immunity, with its aberrant activation implicated in perpetuation of chronic inflammation in various autoimmune pathologies, including rheumatoid arthritis (RA) (Ospelt et al., 2008, Radstake et al., 2004) and type-2 diabetes mellitus (T2DM) (Kim et al., 2007, Shi et al., 2006). The 3' UTR of *TLR4* exhibits enhancer activity - in a manner mediated by the presence of the disease-associated SNP, rs7873784 - it is enriched for histone modifications characteristic of active enhancers and strongly enhances the activity of the *TLR4* promoter in human monocytic cell lines (Korneev et al., 2020). Whereby, the RA- and T2DM-linked risk rs7873784 C-allele (Guo et al., 2016, Huang et al., 2015, Kolz et al., 2008, Yang et al., 2013) within the 3' UTR increases enhancer activity by 1.5-fold in comparison to the reference G-allele (Korneev et al., 2020). The observed amplification of enhancer activity, and consequent increase in *TLR4* expression, is due to the introduction of a strong functional PU.1 TF binding site by the C-allele (Figure 1.17b) (Korneev et al., 2020). PU.1 is an established positive regulator of *TLR4* expression (Pedchenko et al., 2005, Rehli et al., 2000), thus enhanced sequence-specific PU.1 binding at the rs7873784 C-allele may augment *TLR4* transcription in monocytes, resulting in pathological inflammation predisposing carriers to the development of RA and T2DM (Korneev et al., 2020).

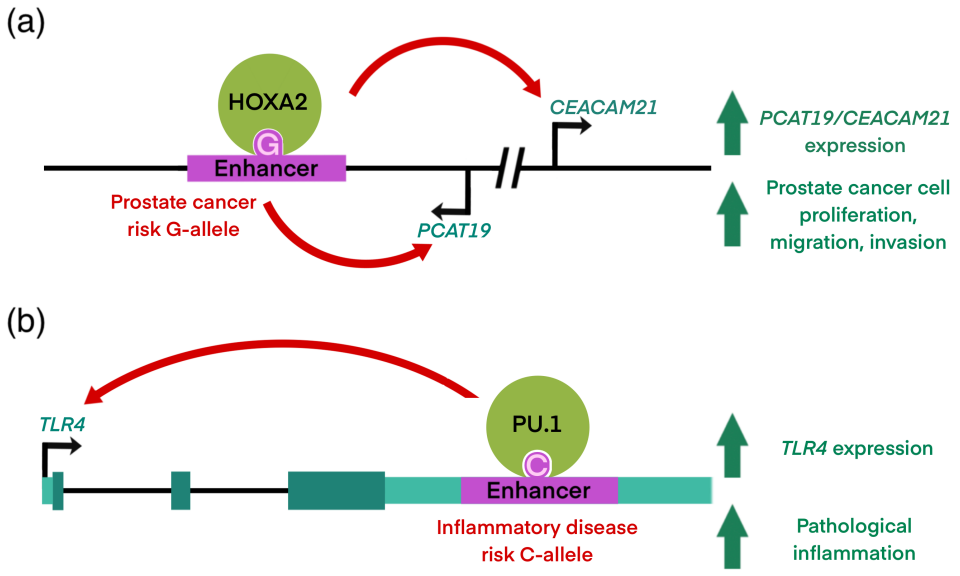



FIGURE 1.17: Enhancer variants introduce *de novo* TF binding sites
Presence of disease risk SNPs within enhancers can create *de novo* binding sites of transcriptional activators, resulting in pathogenic gene upregulation, as evidenced in (a) cancer and (b) autoimmune disease.

### 1.9.3 eRNAs in disease

*The following subsection is abridged from Harrison and Bose (2022)*

eRNAs may act as sequence-independent determinants of enhancer activity that can promote specific transcriptional control mechanisms. For eRNAs, as for many RNAs, lack of sequence conservation does not mean lack of functional conservation. In fact, the relative lack of sequence constraint on eRNA function appears to bestow eRNAs with the functional plasticity necessary to impact a myriad of transcriptional control mechanisms. In disease contexts, mutant enhancers potentially encode aberrant eRNAs that may disrupt this normal eRNA-mediated transcriptional regulation.

#### 1.9.3.1 Transcription factor binding impacts eRNA-mediated pathogenesis

eRNAs can also bind to the transcriptional and epigenetic regulator BRD4, which plays a well-established role in cancer development (reviewed in Donati et al., 2018). In human colorectal cancer, BRD4 is recruited to enhancers that are linked to proinflammatory gene expression patterns and co-occupied by mutant p53 upon chronic TNF-$\alpha$ signalling (Rahnamoun et al., 2018, 2017). Together, BRD4 and p53 cooperate to support synthesis of eRNAs and expression of tumour-promoting genes (Rahnamoun et al., 2018). eRNAs synthesised from these enhancers bind to BRD4 via its tandem bromodomains,

enhancing and stabilising BRD4 binding to acetylated histones and active enhancers. Here, eRNAs support interactions between BRD4 and acetylated histones to maintain enhancer and gene activation (Rahnamoun et al., 2018). BRD4 demonstrates largely promiscuous RNA binding activity but, similar to CBP, binds eRNAs in a locus-specific manner such that it selectively associates with eRNAs produced from specific BRD4-bound enhancers (Bose et al., 2017, Carullo et al., 2020, Rahnamoun et al., 2018).

#### 1.9.3.2 Neurotoxic protein aggregate formation as mediated by RNA structure

Multiple examples exist of condensate-associated diseases caused by alterations in primary RNA sequence, the best-studied being the hexanucleotide repeat expansion within the C9orf72 gene in familial Amyotrophic Lateral Sclerosis (ALS) and Frontotemporal Dementia (FTD) (DeJesus-Hernandez et al., 2011, Renton et al., 2011). In this example, the disease-associated GGGGCC (G4C2) repeat expansion templates multivalent interactions, driving the formation of large gel-like aberrant foci via phase separation (Jain and Vale, 2017). Interestingly, it has been observed that the disease-associated $G_4C_2$ repeat found in the *C9orf72* locus can also readily form G-quadruplexes (Conlon et al., 2016, Reddy et al., 2013). Such stable G-quadruplex structures sequester the splicing regulator protein, hnRNPH, into insoluble aggregates within the brain, resulting in splicing defects in hnRNPH-regulated transcripts, contributing to neurodegeneration (Conlon et al., 2016).

### 1.9.4 3' UTRs in disease

APA is of clinical significance in human disease, coupled with disease-associated global alterations in 3' UTR length and genetic aberrations within 3' UTR disrupting regulatory sequences (Bettens et al., 2009, Mayr and Bartel, 2009, Rademakers et al., 2008, Stacey et al., 2011, Wang et al., 2008a, Wiestner et al., 2007). Tumour tissues demonstrate a global and recurrent preference for short 3' UTR isoforms (Mayr and Bartel, 2009, Xia et al., 2014), preferential expression of shortened 3' UTR isoforms leads to the activation of proto-oncogenes that are typically repressed by 3' UTR-mediated regulation (Lee and Dutta, 2007, Mayr and Bartel, 2009, Mayr et al., 2007). Shortening of 3' UTRs enables mRNAs to escape from the stronger regulatory potential associated with their longer counterparts through loss of miRNA binding sites and regulatory sequences that determine mRNA stability, cytoplasmic localisation and translation efficiency (reviewed in Moore, 2005). APA-mediated 3' UTR shortening has significant functional consequences in tumorigenesis. For example, the shorter mRNA isoform of

the *IMP-1* oncogene promotes more efficient oncogenic transformation than its longer counterpart, through escape from miRNA-mediated repression (Mayr and Bartel, 2009). Moreover, alterations in relative abundance of 3' UTR isoforms may contribute to the deposition of pathological protein aggregates in neurodegenerative disease. Huntington's disease (HD), an inherited autosomal dominant neurodegenerative disorder, is pathologically underpinned by an unstable CAG repeat expansion in the huntingtin (*HTT*) gene (MacDonald et al., 1993). The *HTT* gene produces two alternative 3' UTR isoforms (Lin et al., 1994, 1993), impacting mRNA localisation and protein synthesis (Romo et al., 2017, Xu et al., 2017). The short 3' UTR isoform of *HTT* is more efficiently translated, resulting in more protein aggregates and higher cellular apoptosis rates, when compared to the long 3' UTR isoform (Xu et al., 2017).

Additionally, polymorphisms in 3' UTRs may influence predisposition to several neurodegenerative disorders, whereby these mutations may impact expression of the host gene by disrupting regulatory elements. SNPs located within the 3' UTR that either disrupt or promote miRNA binding have been implicated in PD (Wang et al., 2008a), FTD (Rademakers et al., 2008) and AD (Bettens et al., 2009, Delay et al., 2011). Pathologically, AD is characterised by the extracellular deposition of A$\beta$ peptides, derived from the proteolytic processing of amyloid-beta precursor protein (APP). In familial AD, mutations in the *APP* gene can result in dysregulated APP processing, promoting A$\beta$ overproduction, accumulation and deposition (reviewed in De Strooper, 2010). APP expression is thus intimately linked to AD development and a SNP 171nt into the *APP* 3' UTR was shown to decrease the inhibitory effect of miR-147 on APP expression in *in vitro* reporter assays (Delay et al., 2011). This highlights the potential for mutations within the *APP* 3' UTR to abrogate miRNA-mediated downregulation of APP and A$\beta$ production, with possible pathogenic outcomes in AD patients.

## 1.10 Amyotrophic Lateral Sclerosis: A complex disease underpinned by non-coding genetic variants

### 1.10.1 ALS as a complex disease with little known heritability

A large number of sequence variants have been linked to neurological diseases via GWAS (ENCODE Project Consortium, 2012). ALS is a late-onset, progressive neurodegenerative disorder characterised by the degeneration of the upper and lower motor neurons, presenting with muscular atrophy due to lower motor neuron denervation and upper motor neuron-related gliosis and sclerosis of the corticospinal tract and its cortical origins (reviewed in Bäumer et al., 2014). Clinically, most patients exhibit a combination

of lower motor neuron-related muscle weakness and upper motor neuron involvement, manifesting as spasticity and hyperreflexia. Loss of motor neuron function ultimately arises from the combined degeneration of both upper and lower motor neurons. The risk of ALS onset increases after the age of 40 years, with peak incidence occurring in the early seventies. The site of onset varies, most commonly presenting as limb weakness (limb-onset ALS), although a substantial proportion of patients present with bulbar dysfunction (bulbar-onset ALS), manifesting as dysarthria. Despite the prognostic heterogeneity of ALS, the disease has a median survival of approximately three to four years from symptom onset, with patient death typically resulting from progressive ventilatory failure (reviewed in Bäumer et al., 2014).

Approximately 10% of ALS cases have familial origin, related to mutations in established causative genes (Abhinav et al., 2007, Chiò et al., 2012, Greenway et al., 2006). The vast majority of ALS cases are therefore considered to be sporadic, likely resulting from interactions between genetic and environmental factors in predisposed individuals, indicative of ALS as a complex disease (reviewed in Bäumer et al., 2014). ALS-associated SNPs are increasingly linked to disruption in enhancer function. Sequence polymorphisms within enhancer elements may contribute to ALS aetiology via alterations in TF binding, three-dimensional genome organisation, and eRNA structure and function (reviewed in Carullo and Day, 2019).

Despite genetic factors being estimated to account for up to 52% of the variance in ALS development risk in a population-based parent-offspring heritability study, the majority of sporadic ALS cases have no associated genetic risk factor identified. Additionally, in a population devoid of known ALS-associated mutations, the estimate for ALS heritability remains high at ∼37%, suggesting that the risk variants identified thus far account for only a small fraction of the inherited component of disease risk (Ryan et al., 2019). The bulk of ALS heritability is accounted for by SNPs with associations below genome-wide significance, potentially implying that many genetic risk variants of weak effect size are yet to be discovered and may account for the "missing heritability" associated with ALS as a complex disease (Manolio et al., 2009, Speed et al., 2012, van Rheenen et al., 2016).

#### 1.10.1.1 Functional annotation of GWAS datasets

In order to prioritise association signals from GWAS, genetic variants can be aggregated into groups of functionally related genes for subsequent disease association analysis. Integrating additional biological information allows for improved statistical power for detecting disease-associated gene sets by reduction of the genomic search space (reviewed in Wang et al., 2011b). Functional annotation of the GWAS-associated regions may

enable variants that were previously only statistically associated with a trait to become established risk genes, providing key insights into disease aetiology (reviewed in Spain and Barrett, 2015).

Many GWAS studies neglect to consider the cell type-specificity of genomic function, which is particularly of relevance in noncoding regulatory elements (Schaub et al., 2012, Trynka et al., 2013). This may further explain the "missing heritability" in complex diseases such as ALS (Ryan et al., 2019, Trabjerg et al., 2020). The cell type-specificity of these non-coding regulatory regions creates an opportunity to drastically reduce search space within the genome. This specificity can be exploited, boosting the power to discover ALS risk variants by focusing on genomic regions active within the cell type of interest (Pickrell, 2014). Using cell type-specific epigenetic profiling to estimate disease association probability has been demonstrated to significantly reduce the genomic search space by more than 90% (Zhang et al., 2022).

Integrating fine-mapping methods with cell type-specific biology provides a promising framework for uncovering the genetic basis of complex disease based on GWAS data. A regional fine-mapping approach based on ALS GWAS has recently identified 690 disease-associated genetic variants within neurologically active genomic regions. This approach, referred to as RefMap, employs epigenetic profiling of motor neurons derived from neurologically normal fibroblasts to capture functionally significant genomic variation and combines such data with allele effect sizes estimated from ALS GWAS (Zhang et al., 2022).

Population genetics demonstrates the functional significance of genes identified by RefMap, whereby their conservation, and assumed lack of redundancy, suggests that ALS-associated modulation of RefMap gene expression is likely to be pathogenic. In agreement, RefMap genes are significantly haploinsufficient and intolerant to loss-of-function mutations within the Exome Aggregation Consortium (ExAC) dataset. Transcriptomic data from iPSC-derived motor neurons, human central nervous system (CNS) tissues and ALS mouse models substantiated the functional role of these genes in ALS pathogenesis. In comparison to the overall transcriptome, RefMap genes were upregulated in iPSC-derived neurologically normal motor neurons, substantiating the functional significance of these genes in motor neuron function. As expected, comparatively to the control motor neurons, RefMap genes were downregulated in iPSC-derived motor neurons from ALS patients. To further corroborate a role for RefMap genes in ALS aetiology, it was demonstrated that RefMap gene expression was significantly downregulated in CNS tissues derived from ALS patients in comparison to the overall transcriptome. Utilising the SOD1G93A mouse model of ALS to understand the role these genes may play in ALS prognosis, it

was observed that genes homologous to those identified by RefMap were progressively downregulated throughout disease progression (Zhang et al., 2022).

#### 1.10.1.2 Rare variant burden analysis

Rare variant burden analysis has been recently utilised for the identification of ALS risk variants specifically located within enhancers (Cooper-Knock et al., 2021). Following aggregation of enhancers into sets that regulate common coding genes, enhancer variants were filtered for further association analysis, dependent on minor allele frequencies and evolutionary conservation, prioritising rare variants and those under strong evolutionary conservation. Variant frequencies for each set of enhancers were then collapsed into single variant burden tests (Figure 1.18). Application of this pipeline resulted in the discovery of ALS-associated variants within the enhancers for the genes encoding the membrane lipid raft scaffold proteins, Caveolin-1 and Caveolin-2 (*CAV1* and *CAV2*) (Cooper-Knock et al., 2021). These enhancer variants are located on chromosome 7, spanning the region q31.1 to q31.31 (Figure 1.19a), with the *CAV1* and *CAV2* genes located at 7q31.2 (Figure 1.19b). Such pathogenic enhancer variants are likely to result in reduced transcription of their target genes. The associated reduction in *CAV1*/*CAV2* expression is proposed to induce neurotoxicity via the disruption of the heterooligomeric caveolin complex within membrane lipid rafts of motor neurons, resulting in impaired cell signalling (Figure 1.20) (Cooper-Knock et al., 2021, Head et al., 2008, 2010, Scherer et al., 1997). Fluorescently labelling GM1 gangliosides in lymphoblastoid cell lines derived from ALS patients carrying *CAV1*/*CAV2* enhancer variants visually demonstrates the potential of such enhancer variants to impair membrane lipid raft formation. Fluorescence is significantly reduced in ALS patient-derived lymphoblastoid cells in comparison to controls, consistent with a direct role for CAV1 in membrane lipid raft integrity, and hence promotion of neuronal survival (Cooper-Knock et al., 2021, Head et al., 2011, 2010).

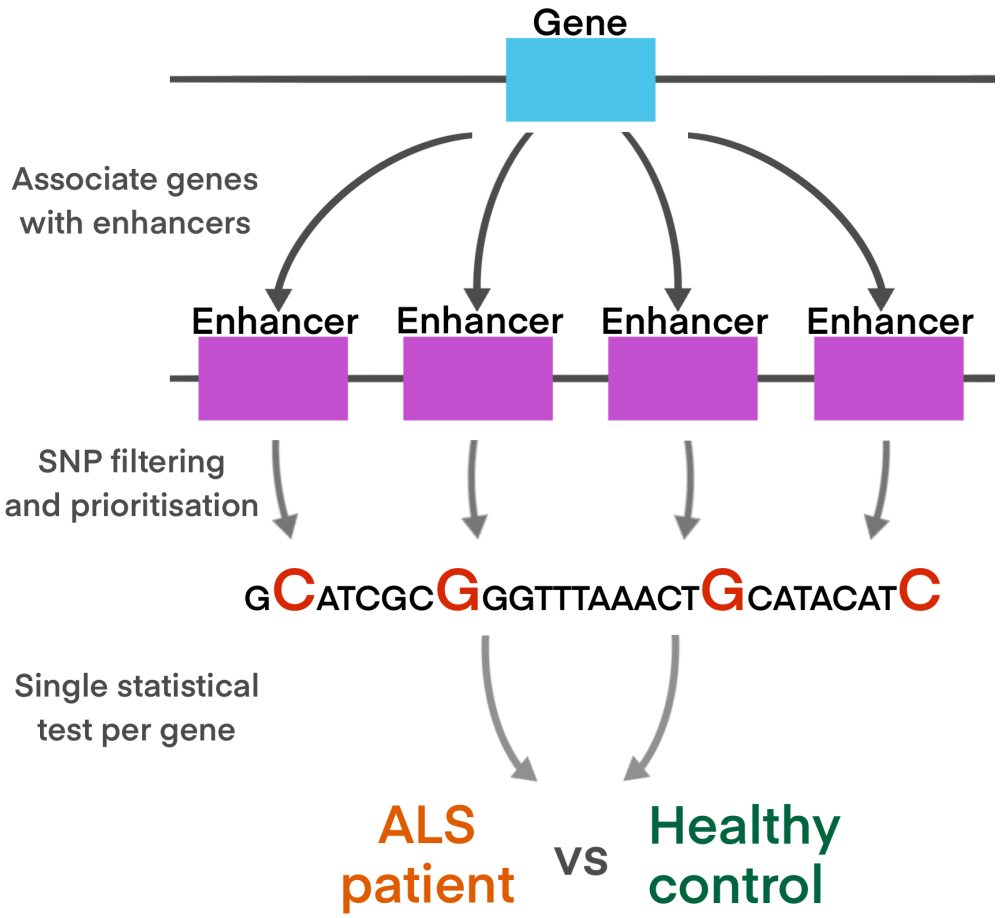



FIGURE 1.18: Rare variant burden analysis pipeline
Pipeline for ALS rare variant filtering and burden analysis. Enhancers are associated with candidate genes, then further filtered for association analysis. Pathogenic variants in enhancer regions are are prioritised for burden testing if they are rare, based on a minor allele frequency (MAF) (MAF <0.01) and if they are evolutionary conserved, based on LINSIGHT scores consistent with strong evolutionary selection (LINSIGHT score >0.8). Following filtering, case and control variant frequencies for each set of enhancers are collapsed into single variant burden tests. Figure modified from Cooper-Knock et al. (2021).

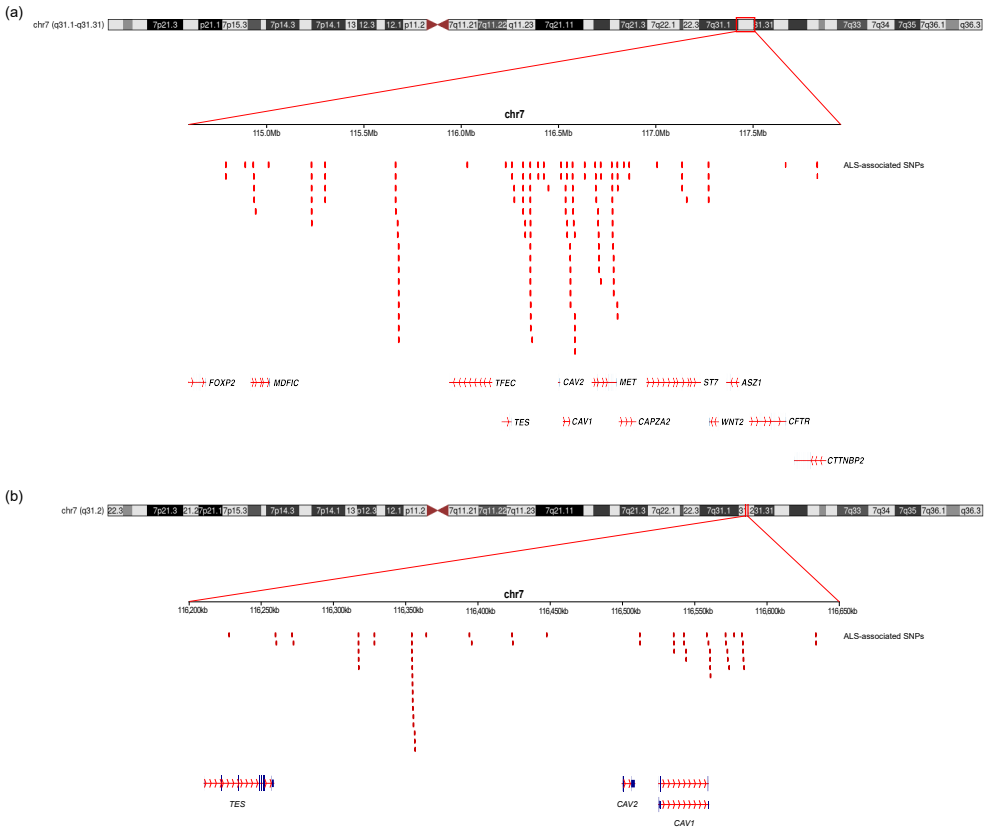



FIGURE 1.19: Chromosomal location of ALS-associated enhancer variants in relation to the *CAV1/CAV2* locus

(a) Schematic representation of chromosome 7, highlighting the q31.1-q31.31 region containing the ALS-associated SNPs within candidate *CAV1/CAV2* enhancers. Location of ALS-associated enhancer variants shown in red. (b) Schematic representation of chromosome 7, highlighting the 7q31.2 cytogenetic band containing the *CAV1* and *CAV2* genes. Location of ALS-associated enhancer variants shown in red.

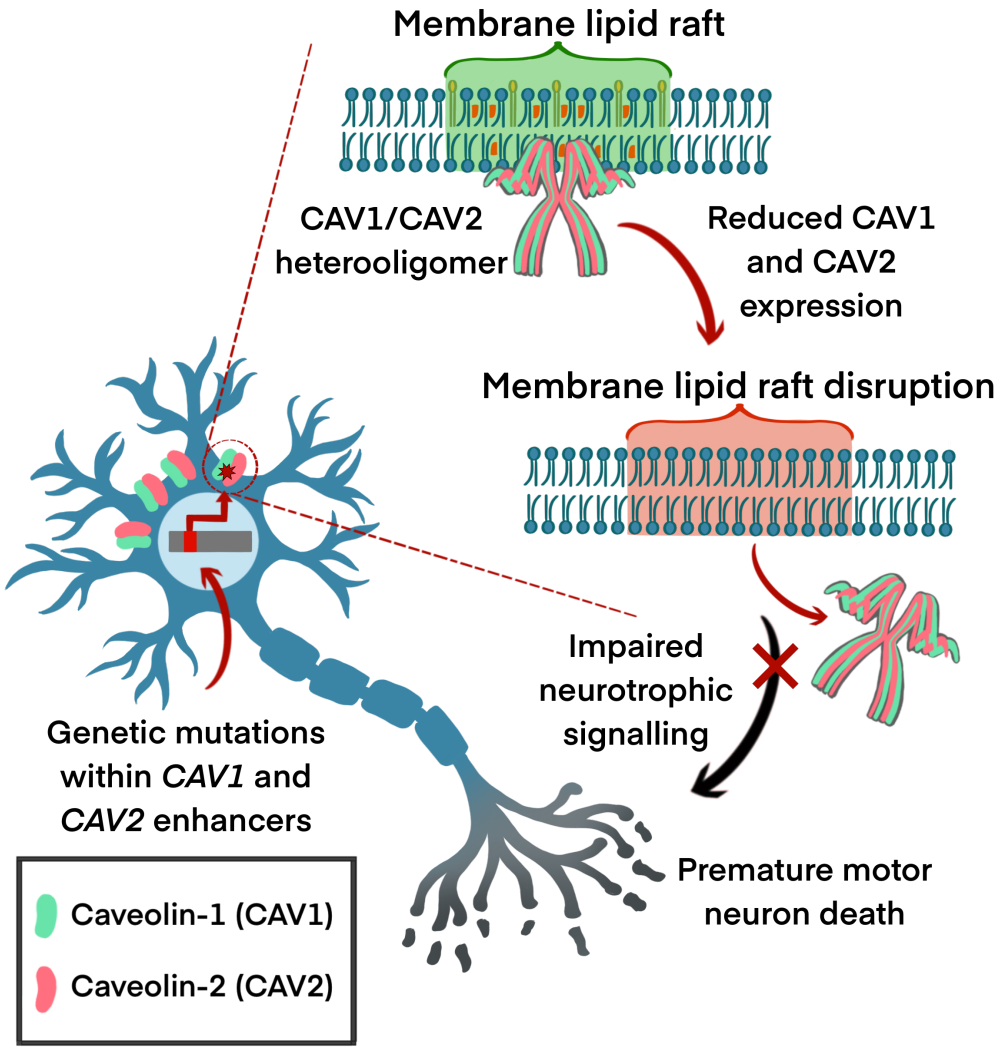



FIGURE 1.20: ALS-associated rare variants are enriched in *CAV1/CAV2* enhancers Pathogenic *CAV1/CAV2* enhancer variants may result in reduced transcription of their target genes. This reduction in expression proposed to induce neurotoxicity via disruption of the caveolin complex within membrane lipid rafts of motor neurons, resulting in impaired neurotrophic signalling and ultimately neurodegeneration

The application of CRISPR/Cas9 genome editing has allowed the recapitulation of the aforementioned ALS-associated enhancer mutations in neurologically-relevant cell lines to investigate their potential roles in ALS aetiology. Introduction of insertion-deletion (indel) mutations proximal to the site of an ALS-associated *CAV1/CAV2* enhancer mutation in neuroblastoma SH-SY5Y cells differentiated to a neuronal phenotype robustly reduced *CAV1* mRNA expression (Cooper-Knock et al., 2021). Such observations may support a mechanism whereby enhancer variants contribute to disease predisposition via altered expression of their target genes (Cooper-Knock et al., 2021, Zhang et al., 2022).

With significant enrichment of functional genetic variants associated with ALS identified within the *CAV1*—but not *CAV2*—coding sequence, *CAV1* has been established as an

ALS risk gene (Cooper-Knock et al., 2021). In this work, we focused our efforts on understanding the impact that ALS-associated variants within candidate *CAV1/CAV2* enhancers have on *CAV1* expression, given the well-established link between *CAV1* genetic variation and ALS.

### 1.10.2 Caveolin-1: Function and disease relevance

The pleiotropic functions of CAV1 may evince a multitude of mechanisms by which dysregulation of *CAV1* expression, as caused by polymorphisms in *CAV1*-associated enhancers, underpin the pathogensis of ALS. CAV1 expression is intimately linked to blood-brain barrier (BBB) integrity, neuroinflammation, neutrophic signalling and autophagy (Andreone et al., 2017, Chow and Gu, 2017, Egawa et al., 2018, Head et al., 2011, Le Lay et al., 2010, Martinez-Outschoorn et al., 2011, Pandit et al., 2020, Shi et al., 2015, Trevino et al., 2023).

Tight control over CAV1 protein expression in brain vascular endothelial cells provide the basis for BBB maintenance, transcellular transcytosis and neurovascular coupling (Andreone et al., 2017, Chow and Gu, 2017, Pandit et al., 2020, Yang et al., 2020). Caveolae-mediated vesicular transport contributes to BBB permeability (Andreone et al., 2017), with disease-associated *CAV1* upregulation contributing to BBB leakage and neuroinflammation (Errede et al., 2012, Knowland et al., 2014, Lutz et al., 2017, Trevino et al., 2023, Wu et al., 2016, Yang et al., 2020). Constitutive transport of plasma circulatory macromolecules by ligand-specific receptor-mediated transcytosis maintains a healthy BBB (reviewed in Chow and Gu, 2015). In the ageing brain, there is a decline in this specific receptor-mediated transport in favour of non-specific caveolar transcytosis (Yang et al., 2020). Indeed, with age, CAV1 expression increases in the brain (Ha et al., 2021).

Beyond its functions in endothelial transport, CAV1 is implicated in neurotransmission, dendrite outgrowth and synaptic plasticity via its membrane lipid raft scaffolding function (Egawa et al., 2018, Head et al., 2011). CAV1 overexpression in primary neurons and hippocampal brain tissue enhance both membrane lipid raft formation and the localisation of neurotransmitter and neurotrophin receptors to membrane lipid rafts, promoting neuronal survival (Egawa et al., 2018, Head et al., 2011). Along with the altered neurotransmitter signalling, akin to that seen in CNS pathologies, observed in CAV1 knockout mice (Francesconi et al., 2009, Head et al., 2008, Takayasu et al., 2010), these studies implicate *CAV1* downregulation in impaired neurotrophic signalling. Further, CAV1, and its associated lipid rafts, are implicated in the regulation of autophagy, whereby, deficiency in CAV1 expression induces autophagy through disruption of membrane lipid rafts (Le Lay et al., 2010, Martinez-Outschoorn et al., 2011, Shi et al., 2015), suggesting

a suppressive role for CAV1. With autophagy functioning as a major pathway for the clearance of neurotoxic protein aggregates (reviewed in Amin et al., 2020), suppression of the autophagic response by the downregulation of *CAV1* may contribute to ALS. Moreover, in the absence of active proteasomal degradation, caveolins themselves tend to form protein aggregates with trafficking defects (Morales-Paytuví et al., 2023), with mutant caveolins capable of seeding further aggregation by sequestering WT caveolins, thus reducing caveolae formation (Copeland et al., 2017, Lee et al., 2002, Pol et al., 2001, Roy et al., 1999).

The multiplicity of CAV1 functions in the CNS therefore provide numerous potential pathways by which dysregulated *CAV1* expression may result in neurodegeneration. Upregulation in *CAV1* may impair the neuroprotective properties of the BBB, whereas downregulation of *CAV1* may impair neurotrophic signalling and the autophagic removal of neurotoxic protein aggregates.

#### 1.10.2.1 Caveolin-1 in the DNA damage response

The pleiotropy of CAV1 is also observed in the DNA damage response, whereby dysregulation of *CAV1*, either by overexpression or underexpression, may result in accumulation of cellular oxidative stress and ultimately senescence. CAV1 overexpression, phosphorylation and improper localisation is linked to the inability to repair oxidative stress-induced DNA insults and impaired mitochondrial autophagy (Goutas et al., 2021, Jiang et al., 2022), whereas CAV1 deficiency can suppress autophagic responses via disruption of membrane lipid rafts and impede mitochondria quality control via dysregulated mitochondrial fusion and fission dynamics (Jiang et al., 2022, Le Lay et al., 2010, Martinez-Outschoorn et al., 2011, Shi et al., 2015).

Phosphorylation and nuclear localisation of CAV1 is implicated in response to oxidative stress. For example, CAV1 is upregulated, phosphorylated and translocated to the nucleus in chondrocytes under stress conditions, prior to returning to basal expression and phosphorylation levels and export from the nucleus upon recovery. However in osteoarthritic disease states, CAV1 is constitutively overexpressed and mislocalised away from the membrane and cytoplasm, into the nucleus, resulting in failure of osteoarthritic chondrocytes to appropriately respond to oxidative stress-induced DNA insults (Goutas et al., 2021).

Mitochondrial fusion and fission ensure mitochondrial quality, with fusion enabling exchange of material between mitochondria and fission promoting the segregation and isolation of damaged mitochondrial components for their targeted removal by autophagy (reviewed in Ni et al., 2015). CAV1 is a pivotal regulator of mitochondrial morphology

and dynamics. CAV1 interacts with the mitochondrial fusion/fission machinery in a phosphorylation-dependent manner, negatively regulating mitochondrial fusion and fission dynamics and mitochondrial autophagy (Jiang et al., 2022). Phosphorylated CAV1 impairs the initiation of mitochondrial autophagy, resulting in accumulation of damaged mitochondria, and thus, mitochondrial ROS (Jiang et al., 2022). Thus, aberrant overexpression of CAV1 or dysregulation of phosphorylation may promote the pathological accumulation of ROS. Conversely, CAV1 depletion enhances mitochondrial fusion and fission dynamics and thus alters mitochondrial morphology, producing short, fragmented, but highly mobile mitochondria that exhibit higher levels of ROS (Jiang et al., 2022, Stephen et al., 2023). This posits a mechanism whereby either upregulation or downregulation of CAV1 have pathogenic potential, via accumulation of damaged mitochondria by escape from mitochondrial autophagy or increase in ROS by alterations in mitochondrial morphology and dynamics, respectively (Jiang et al., 2022). Both of these aforementioned mechanisms have been implicated in ALS. As a result of the limited regenerative capacity and large metabolic demands of post-mitotic neurons, neurons are particularly sensitive and susceptible to damage from oxidative damage (reviewed in Genova et al., 2004). As such, both increased levels of oxidative damage (Calingasan et al., 2005, Ferrante et al., 1997, Shaw et al., 1995), and mitochondrial fragmentation and morphological changes have been described in the CNS of ALS patients (Sasaki et al., 2007). In support, mutations in $Cu^2+/Zn^2+$ superoxide dismutase (SOD1), a key antioxidant enzyme, account for ∼20% of familial and ∼3% of sporadic ALS cases (reviewed in Saccon et al., 2013).

## 1.11 Hypotheses

Enhancer function is governed by a complex interplay of factors at multiple levels: at the chromatin level, through compartmentalisation (Dixon et al., 2012, Nora et al., 2012, Rao et al., 2014), histone modifications (Creyghton et al., 2010, Heintzman et al., 2007, Rada-Iglesias et al., 2011), and accessibility (Ernst et al., 2011); at the DNA sequence level, through linear proximity to target promoters (Zuin et al., 2022) and the presence of TF binding motifs (reviewed in Spitz and Furlong, 2012); and at the transcriptional level, through the production and regulatory roles of eRNAs (De Santa et al., 2010, Hah et al., 2011, Kaikkonen et al., 2013, Kim et al., 2010, Li et al., 2013). This multi-dimensional complexity presents a major challenge in deciphering how enhancer variants contribute to complex disease. We hypothesise that ALS-associated SNPs within enhancer regions may disrupt any one - or multiple - of these molecular features, thereby contributing to disease phenotype.

We propose that active *CAV1/CAV2* enhancers - identified via epigenetic and transcriptomic profiling - located in close linear or 3D proximity to the *CAV1* gene are more likely to harbour ALS-associated SNPs. Based on the assumption that active enhancers in physical contact with the *CAV1* promoter are more likely to influence *CAV1* gene transcription, we expect that disruption of such enhancers by disease-associated SNPs will have a greater impact on *CAV1* expression, potentially contributing to ALS pathogenesis. Furthermore, we hypothesise that there is unlikely to be a single unifying mechanism by which ALS-associated enhancer variants drive disease. Even within a single enhancer element, different SNPs within its sequence may impair enhancer function through distinct mechanisms. We expect to encounter examples of ALS-associated enhancer variants that impact *CAV1/CAV2* enhancer function at the DNA level, for example via disrupted TF binding to the enhancer, and at the RNA level, for example through the transcription of aberrant eRNAs that have altered structures, stability and potentially function. Identifying ALS-associated SNPs in enhancers that exert pathogenic effects via distinct molecular mechanisms could therefore underscore the complexity of enhancer-mediated gene regulation in complex disease.

## 1.12 Aims and motivations of this research

*CAV1* is an established risk gene in ALS (Cooper-Knock et al., 2021), with risk factors associated with both up- and downregulation of *CAV1* expression. As enhancers are essential for fine-tuning overall gene expression levels, we sought to investigate the impact of ALS-associated mutations in *CAV1*-associated enhancer elements. By doing so, we aim to illuminate the pathways by which dysregulation of *CAV1* expression can underpin neurodegenerative disease.

Firstly, we aimed to examine the genome-wide distribution of ALS-associated *CAV1* enhancers and the enrichment of ALS-associated variants within these enhancers, in the context of higher-order chromosome organisation. By linking ALS-associated mutations to enhancer activity, we will examine the activity of ALS-associated *CAV1* enhancers by assessing chromatin accessibility, enhancer-associated chromatin modifications and eRNA transcription.

We will then examine the role of eRNAs at *CAV1* enhancers using nascent RNA sequencing approaches. The transcription of eRNAs is a key factor contributing to enhancer activity. Therefore, not only may enhancer transcription data provide insight into how enhancer activity changes upon mutation, but enhancers containing disease-associated variants may yield aberrant eRNAs capable of disrupting transcriptional regulation to culminate in disease phenotype. By combining nascent RNA sequencing data

with computational RNA structure prediction, we will ask whether enhancers containing ALS-associated variants may generate eRNAs with altered structures, potentially modifying their regulatory functions. Understanding whether eRNA expression levels and structural changes are central to their roles in transcriptional regulation could help to establish eRNAs as key drivers in enhancer-associated disease mechanisms.

Finally, through the identification of putative ALS-associated enhancers and functional assessments using luciferase reporter assays, the regulatory consequences of disease-associated SNPs on enhancer activity can be examined. In disease contexts, we will explore how both distal intergenic and proximal intragenic enhancers influence not only *CAV1* gene expression but also the expression of alternative isoforms. Finally, we will consider the potential implications of functional ALS-associated enhancers localising within the 3' UTR of their target gene.

# Chapter 2

# Materials and Methods

## 2.1 Molecular cloning

### 2.1.1 Buffers and media

**50X Tris-acetate-EDTA (TAE)**: Stock (ThermoFisher), diluted to 1X working concentration with Milli-Q® water

**1X Lithium Borate (10mM lithium acetate dihydrate, 10mM boric acid)**: 1g Lithium acetate dihydrate and 0.62g boric acid dissolved in 1L Milli-Q® water

**Luria Broth (LB) media**: each 25g of LB granules (Melford) was dissolved in 1L Milli-Q® water and sterilized by autoclaving for 20 min at 121°C

**LB agar**: each 32g of LB Agar granules, Low Salt Formula (Melford) was dissolved in 1L Milli-Q® water and sterilized by autoclaving for 20 min at 121°C

### 2.1.2 Molecular biology kits

**PCR product purification**: Monarch® PCR DNA Cleanup Kit (NEB)

**Gel extraction**: QIAquick® Gel Extraction Kit (QIAGEN), Monarch® DNA Gel Extraction Kit (NEB)

**Plasmid DNA extraction and purification (small scale)**: Monarch® Plasmid Miniprep Kit (NEB)

**Plasmid DNA extraction and purification (medium scale)**: QIAGEN® Plasmid Plus Midi Kit (QIAGEN)

**Genomic DNA extraction and purification**: Monarch® Genomic DNA Purification Kit (NEB)

### 2.1.3 Plasmids

Table 2.1: Plasmids utilised for doctoral work

| **Plasmid** | **Bacterial Resistance** | **Mammalian Selection Marker** | **Source** |
|---|---|---|---|
| CAV1_enhancer_WT | Kanamycin | N/A | GeneArt G-block vector (ThermoFisher) |
| pHR-UCOE-SFFV-Zim3-dCas9-P2A-Hygro | Ampicillin | Hygromycin | Addgene (#188768) |
| pSpCas9(BB)-2A-Puro (PX459) V2.0 | Ampicillin | Puromycin | Addgene (#62988) |
| lentiGuide-Puro | Ampicillin | Puromycin | Addgene (#52963) |
| pMD2.G (VSV-G) | Ampicillin | N/A | Addgene (#12259) |
| psPAX2 | Ampicillin | N/A | Addgene (#12260) |
| pGL3-Basic | Ampicillin | N/A | Addgene (#212936) |
| pGL3-Control | Ampicillin | N/A | Addgene (#212937) |
| pGL3-Promoter | Ampicillin | N/A | Promega (#E1761) |
| 759_Renilla | Ampicillin | N/A | Gifted by Cristina Alexandru-Crivac, derived from Renilla luciferase reporter vector pGL4.70[hRluc] from Promega (#E6881) |

### 2.1.4 Polymerase chain reaction (PCR)

Polymerase chain reaction was performed in a 50$\mu$l volume (Table 2.2)

ProFlex PCR thermocycler (Applied Biosystems) was used to run below PCR reactions with the following cycling conditions (Table 2.3)

Table 2.2: Composition of 50$\mu$l PCR reaction

| **Component** | **Volume ($\mu$l)** |
|---|---|
| 10X PfuUltra II reaction buffer (Agilent) | 5 |
| DMSO (Sigma) | 2.5 |
| 100$\mu$M forward primer (IDT) | 0.25 |
| 100$\mu$M reverse primer (IDT) | 0.25 |
| 10mM dNTP mix (Roche) | 5 |
| PfuUltra II Fusion High-Fidelity Polymerase (Agilent) | 0.5 |
| $H_2O$ | Up to 50 |
| DNA | 1 |

Table 2.3: Cycling conditions for PCR reaction

| **Cycle step** | **Temperature** | **Time** | **Cycles** |
|---|---|---|---|
| Initial denaturation | 95°C | 5 minutes | 1 |
| Denaturation | 95°C | 45 seconds | 30 |
| Annealing | 50 – 59°C (primer dependent) | 45 seconds | |
| Extension | 72°C | 30 seconds/kb | |
| Final extension | 72°C | 15 minutes | 1 |

### 2.1.5 Agarose gel electrophoresis

UltraPure$^{TM}$ (Thermo Fisher) and 1X TAE or 1X lithium borate buffer were used to prepare a 1% agarose gel, which was supplemented with SYBR$^{TM}$ Safe DNA Gel Stain (Thermo Fisher) to a 1:10,000 dilution. Gel tank was filled with 1X TAE or 1X lithium borate (dependent on buffer used to make gel). DNA samples were mixed with Gel Loading Dye (6X) (NEB), to give a final six-fold dilution of loading dye, and loaded into wells along with an appropriate DNA ladder (Quick-Load® Purple 1 kb Plus DNA Ladder, Quick-Load® 1kb DNA Ladder or Quick-Load® 100bp DNA Ladder; all NEB). Gel was run at 100V for 0.5 - 1.5 hours and visualised and photographed using G-BOX Chem-XRQ (Syngene).

### 2.1.6 Transformation of bacterial cells

Competent *E. coli* cells of the *Stbl3* strain (genotype: FmcrB mrrhsdS20($r_B^-$, $m_B^-$) *rec*A13 *sup*E44 *ara*-14 *gal*K2 *lac*Y1 *pro*A2 *rps*L20($Str^R$) *xyl*-5 $\lambda$ *leumtl*-1; ThermoFisher) were thawed on ice for 10 minutes. 1pg-100ng (1-10$\mu$l) of plasmid DNA was added to 50$\mu$l of *Stbl3* cells and placed on ice for 30 minutes, then heat-shocked at 42°C for 30 seconds before being returned to ice for 2 minutes. 950$\mu$l of SOC media (Melford) was added to cell suspension, and cells were incubated at 37°C for 1 hour with vigorous shaking

(300rpm). Cell suspensions were centrifuged at 5000rcf for 1 minute, then 900$\mu$l of media supernatant was removed and cell pellet resuspended in residual SOC media. Cell suspension was then plated onto LB agar plates containing either 100$\mu$g/ml ampicillin or 100$\mu$g/ml kanamycin, dependent on plasmid, and incubated at 37°C overnight.

### 2.1.7 Plasmid purification and Sanger sequencing

For small-scale DNA preparation, selected colonies from plated transformed *E. coli* were transferred into a 15ml tube containing 5ml LB media supplemented with 100$\mu$g/ml ampicillin or 50$\mu$g/ml kanamycin, dependent on plasmid. For medium-scale DNA preparation, selected colonies from plated transformed *E. coli* were transferred into 100ml of LB media in a 250ml conical flask containing 100$\mu$g/ml ampicillin or 50$\mu$g/ml kanamycin, dependent on plasmid. Suspensions were incubated overnight at 37°C with vigorous shaking (180rpm on a shaking platform). Plasmid DNA was isolated and purified from cell culture using the Monarch® Plasmid Miniprep Kit (NEB) for small-scale DNA preparation and QIAGEN® Plasmid Plus Midi Kit (QIAGEN) for medium-scale DNA preparation. Purified plasmid DNA was sequenced by Eurofins Genomics with pre-mixed primers (Table 2.4).

Table 2.4: Sanger sequencing primers

| **Primer name** | **Sequence** | **Comments** |
|---|---|---|
| U6_promoter_fwd | GAGGGCCTATTTCCCATGATT | General plasmid sequencing |
| CAV1_116199_val | CAGCAATGTCCGCATCAACTTGC | Checking for integration of chr7:116,199,522:T>C mutation into *CAV1* enhancer G-block vector |
| CAV1_116200_val | TGTGAGCCTATCAGAGTTGCTGC | Checking for integration of chr7:116,200,589:T>A, chr7:116,200,705:C>T, chr7:116,200,719:C>A and chr7:116,300,953:A>T mutations into *CAV1* enhancer G-block vector |
| SV40_CAV1_gRNA1_valF | TTGCTGTCTGCCCTCTTTG | Checking for successful mutation of SV40 gRNA 1 PAM (CCC >TCC) |
| SV40_CAV1_gRNA3_valF | GCACTCATCTGGGGCATTTAC | Checking for successful mutation of SV40 gRNA 3 PAM (CCG >ACG) |
| SV40_CAV1_gRNA4_valF | GCATCAGCCGTGTCTATTCC | Checking for successful reversion of SV40 gRNA 4 PAM to WT (CAT >CCT) and validation of gRNA efficiency |

### 2.1.8 Preparation of plasmids for luciferase reporter assay

#### 2.1.8.1 Site-directed mutagenesis

To generate ALS-associated mutations in *CAV1* enhancer G-block vector (CAV1_enhancer_WT, ThermoFisher GeneArt, Appendix A.1), PCR was performed with the following primers, annealing temperatures and extension temperatures (Table 2.5).

Table 2.5: Site-directed mutagenesis primers and PCR conditions
Primers designed using NEBaseChanger

| **Mutation** | **Forward primer** | **Reverse primer** | **Annealing temperature** | **Extension temperature** |
|---|---|---|---|---|
| chr7:116,199,522:T>C | CTATTTC ATCcATT TTTGGCA GTC | CTCAGAA GAGACAT AATACTT ATG | 51℃ | 68℃ |
| chr7:116,200,589:T>A | CAAGCAT CCCaTTG CCCAGAA | GATTAGG TCCAAAG CTCC | 56℃ | 68℃ |
| chr7:116,200,705:C>T | AAACCTG AGTtGTA CAGAAAG | ACAGGTG AAATTCC TCAG | 52℃ | 68℃ |
| chr7:116,200,719:C>A | CAGAAAG CTGaCTG GTATATC | TACGACT CAGGTTT ACAG | 54℃ | 68℃ |
| chr7:116,200,953:A>T | AACTGTG TACtTAA CTGAAAA TGTG | TAAAAGT AAAACTA TATCCAT GC | 53℃ | 68℃ |

PCR products were gel extracted using either QIAquick® Gel Extraction Kit (QIAGEN) or Monarch® DNA Gel Extraction Kit (NEB). Template DNA was removed from PCR product via *Dpn*I (NEB) digest, via incubation at 37°C for 30 minutes followed by heat inactivation at 80°C for 20 minutes. Linearised vector was then phosphorylated by T4 PNK (ThermoFisher) treatment, using T4 ligase buffer (10X) (ThermoFisher), and then purified using Monarch® PCR DNA Cleanup Kit (NEB). Vector was then recircularised using T4 DNA ligase (ThermoFisher) and incubated at room temperature for 2 hours or at 16℃ overnight, followed by heat inactivation at 65°C for 10 minutes. Plasmid DNA was purified using Monarch® PCR DNA Cleanup Kit (NEB) and then transformed into *Stbl3* E.coli, plated and incubated at 37℃ overnight. Up to 4 colonies were picked per construct and cultures were grown at 37℃ overnight, before isolating plasmid DNA using Monarch® Plasmid Miniprep Kit (NEB). Samples were sent for Sanger sequencing using the appropriate primers detailed in Table 2.4, to confirm successful base substitution.

#### 2.1.8.2 Cloning enhancer DNA into luciferase reporter plasmid

WT *CAV1* enhancer vector (*CAV1*_enhancer_WT, ThermoFisher GeneArt), and *CAV1* enhancer vectors containing ALS-associated SNPs (*CAV1*_enhancer_522, *CAV1*_enhancer_589, *CAV1*_enhancer_705, *CAV1*_enhancer_719 and CAV1_enhancer_953) were double digested using *Mlu*I (NEB) and *Xho*I (NEB) in NEBuffer r3.1 by 15 minute incubation at 37℃ followed by heat inactivation at 80℃ for 20 minutes. Due to supercoiling of the pGL3-Promoter vector (Promega), sequential single digests with *Mlu*I followed by *Xho*I were performed to linearise the vector, as above. Following digestion, *CAV1* enhancer insert DNA and linearised pGL3-Promoter vector DNA were purified using Monarch® PCR DNA Cleanup Kit (NEB), and the inserts ligated into the pGL3-Promoter backbone using T4 DNA ligase (ThermoFisher) with incubation at 16℃ overnight , followed by heat inactivation at 65°C for 10 minutes. Plasmid DNA was purified using Monarch® PCR DNA Cleanup Kit (NEB) and then transformed into *Stbl3 E.coli*, plated and incubated at 37℃ overnight. A colony was picked per construct and cultures were grown at 37℃ overnight, before isolating plasmid DNA using Monarch® Plasmid Miniprep Kit (NEB). Samples were sent for Sanger sequencing using the U6 promoter sequencing primer (Table 2.4). Vectors containing insert were then re-transformed, plated and subsequently cultured before isolating plasmid DNA using QIAGEN® Plasmid Plus Midi Kit (QIAGEN) to generate appropriate amounts of DNA for Neon® transfection.

### 2.1.9 Preparation of plasmids for CRISPR interference

Plasmids used for CRISPR interference (CRISPRi) were pHR-UCOE-SFFV-Zim3-dCas9-P2A-Hygro (Zim3-dCas9) (Addgene #188768) and modified from lentiGuide-Puro (Addgene #52963). Single guide RNAs (sgRNAs) were cloned into lentiGuide-Puro using a modified protocol from the Zhang lab (Sanjana et al., 2014, Shalem et al., 2014).

Two oligonucleotides were designed for each sgRNA to be cloned in the following format:

Oligo 1 (up): 5' – CACCGNNNNNNNNNNNNNNNNNNN – 3'

Oligo 2 (down): 3' – CNNNNNNNNNNNNNNNNNNNCAAA – 5',

where $N_{20}$ is the unique sequence of the sgRNA. The overhangs included in the oligonuleotides are compatible with *BsmB*I digest of the lentiGuide-Puro vector.

For CRISPRi, the following oligonucleotide pairs were annealed and inserted into the lentiGuide-Puro vectors, respectively (Table 2.6)

Table 2.6: Oligonucleotides for cloning sgRNAs into lentiGuide-Puro vectors for generation of CRISPRi SH-SY5Y cell lines

| **sgRNA target** | **Target locus** | **gRNA sequence** | **Oligonucleotide name** | **Oligonucleotide sequence** |
|---|---|---|---|---|
| Proximal *CAV1* candidate enhancer | chr7:116200361-116200383 (+) | TCCAAACTAATCCATCACCG | *CAV1*_CE_gRNA1_up | CACCGTCCAAACTAATCCATCACCG |
| | | | *CAV1*_CE_gRNA1_down | AAACCGGTGATGGATTAGTTTGGAC |
| Proximal *CAV1* candidate enhancer | chr7:116200523-116200545 (+) | GTCTGTTATGCTGTGACACA | *CAV1*_CE_gRNA2_up | CACCGGTCTGTTATGCTGTGACAC A |
| | | | *CAV1*_CE_gRNA2_down | AAACTGTGTCACAGCATAACAGAC C |
| Proximal *CAV1* candidate enhancer | chr7:116200701-116200723 (+) | GAGTCGTACAGAAAGCTGCC | *CAV1*_CE_gRNA3_up | CACCGGAGTCGTACAGAAAGCTGC C |
| | | | *CAV1*_CE_gRNA3_down | AAACGGCAGCTTTCTGTACGACTC C |
| *CAV1* promoter | chr7:116164416-116164438 (-) | TCGTTTACATCTAGTCGGTG | *CAV1*_prom_gRNA_up | CACCGTCGTTTACATCTAGTCGGT G |
| | | | *CAV1*_prom_gRNA_down | AAACCACCGACTAGATGTAAACGA C |

Continued on next page

Table 2.6 – continued from previous page

| **sgRNA target** | **Target locus** | **sgRNA sequence** | **Oligo-nucleotide name** | **Oligo-nucleotide sequence** |
|---|---|---|---|---|
| Distal *CAV1* candidate enhancer | chr7:116223121-116223143 (-) | TTGTGGTGACCAAGCAGCAT | *CAV1*_DE_gRNA_up | CACCGTTGTGGTGACCAAGCAGCAT |
| | | | *CAV1*_DE_gRNA_down | AAACATGCTGCTTGGTCACCACAAC |
| Non-targeting scrambled (negative control) | N/A | GCTGATCTATCGCGGTCGTC | Scramble_gRNA_fwd | CACCGCTGATCTATCGCGGTCGTC |
| | | | Scramble_gRNA_rev | AAACACGACCGCGATAGATCAGCC |

The lentiGuide-Puro vector (5$\mu$g) was digested with *BsmB*I (25U; NEB) by incubation at 37℃ for 15 minutes, followed by heat inactivation at 65℃ for 20 minutes. Linearised vectors were then purified using Monarch® PCR DNA Cleanup Kit (NEB) and then dephosphorylated by incubation with Shrimp Alkaline Phosphatase (rSAP; 2 U; NEB) at 37℃ for 30 mins, followed by heat inactivation at 65℃ for 5 mins. Dephosphorylated DNA was then run on a 1% agarose gel, gel extracted and quantified using Qubit dsDNA Broad Range (ThermoFisher). Meanwhile, each oligonucleotide pair was phosphorylated and annealed at a 10 $\mu$M concentration, in a 10$\mu$l reaction with T4 Polynucleotide Kinase (PNK; 5U; NEB) in 1X T4 Ligation Buffer in a thermocycler using the following conditions: 37℃ for 30 minutes; 95℃ for 5 minutes; then ramped down to 25℃ at 5℃/minute. The annealed oligonucleotides were then ligated into the linearised vector using T4 DNA ligase (400U/50 ng DNA; NEB) by incubation for 15 minutes at room temperature followed by heat inactivation at 65℃ for 10 minutes, and subsequently transformed into *Stbl3 E. coli*.

### 2.1.10 Preparation of plasmids for CRISPR gRNA validation and generation of genetically modified cell line

Plasmids for insertion of premature polyadenylation signal (PAS) in *CAV1* were modified from pSpCas9(BB)-2A-Puro (pX459) V2.0 (Addgene #62988). The cloning strategy for lentiGuide-Puro was modified for pX459, using a *Bbs*I digest to linearise the vector as opposed to *BsmB*I. As for *BsmB*I, the overhangs included in oligonucleotides are compatible *Bbs*I digest of pX459, with the following oligonucleotide pairs annealed and inserted into pX459 (Table 2.7).

Table 2.7: Oligonucleotides for cloning sgRNAs into pX459 for generation of genetically modified SH-SY5Y cell lines

| **sgRNA target** | **sgRNA sequence** | **Oligonucleotide name** | **Oligonucleotide sequence** |
|---|---|---|---|
| *CAV1* final exon | TCCCAACAGCT TCAAAGAGT | SV40_CAV1_ gRNA1_up | CACCGTCCCAA CAGCTTCAAAG AGT |
| | | SV40_CAV1_ gRNA1_down | AAACACTCTTT GAAGCTGTTGG GAC |
| *CAV1* final exon | TTTCTTTCTGC AAGTTGATG | SV40_CAV1_ gRNA3_up | CACCGTTTCTT TCTGCAAGTTG ATG |
| | | SV40_CAV1_ gRNA3_down | AAACCATCAAC TTGCAGAAAGA AAC |
| *CAV1* final exon | TGGAACTTGAA ATTGGCACC | SV40_CAV1_ gRNA4_up | CACCGTGGAAC TTGAAATTGGC ACC |
| | | SV40_CAV1_ gRNA4_down | AAACGGTGCCA ATTTCAAGTTC CAC |
| Non-targeting scrambled (negative control) | GCTGATCTATC GCGGTCGTC | Scramble_ gRNA_fwd | CACCGCTGATC TATCGCGGTCG TC |
| | | Scramble_ gRNA_rev | AAACACGACCG CGATAGATCAG CC |

### 2.1.11 Preparation of donor template for generation of genetically modified cell line

Donor template for insertion of the premature polyadenylation signal (PAS) into *CAV1* were modified from the WT *CAV1* enhancer vector (*CAV1*_enhancer_WT, ThermoFisher

GeneArt). A gene fragment (IDT gBlock) containing SV40 PAS flanked by complementarity to *CAV1* and containing mutation of the SV40_CAV1_gRNA4 PAM site was designed to allow insertion of the SV40 PAS into CAV1_enhancer_WT by restriction digest (Appendix A.2). Sequential single digests with *Mlu*I (NEB) followed by *EcoR*I-HF (NEB) were performed on both the WT *CAV1* enhancer vector and SV40_gBlock_RD insert. Following digestion, DNA purification, ligation, transformation and DNA isolation were performed as for cloning of the pGL3-Promoter vector. To determine successful generation of CAV1_SV40PAS_gRNA4 (Appendix A.3), samples were sent for Sanger sequencing using the CAV1_116199_val sequencing primer (Table 2.4).

#### 2.1.11.1 Generation of donor template for induction of ALS-associated SNPs

To generate donor templates appropriate for use with pX459 plasmids containing SV40_CAV1_gRNA1 and SV40_CAV1_gRNA3, site-directed mutagenesis was performed on the *CAV1* enhancer vector containing the SV40 PAS, as described for preparation of plasmids for luciferase reporter assay. For SV40_CAV1 gRNAs 1 and 3, firstly, the SV40_CAV1_gRNA4 PAM site mutation was restored to the WT sequence prior to induction of the appropriate PAM site mutation for each corresponding sgRNA. For conversion of SV40_CAV1_gRNA4 PAM site mutation to WT sequence, and subsequent mutation of SV40_CAV1_gRNA1 and SV40_CAV1_gRNA3 PAM sites, PCR was performed with the following primers, annealing temperatures and extension temperatures (Table 2.8).

Table 2.8: Site-directed mutagenesis primers and PCR conditions

| **Mutation** | **Forward primer** | **Reverse primer** | **Annealing temperature** | **Extension temperature** |
|---|---|---|---|---|
| Restoration of SV40_CAV1_gRNA4 PAM site to WT | TTTAATTTTCcTGGTGCCAATTTC | AGGAAAAAAAATCAGGTATAC | 52℃ | 68℃ |
| Induction of PAM site mutation for SV40_CAV1_gRNA1 | CCGTCTGTGAtCCACTCTTTG | TGTGGACGTAGATGGAATAG | 57℃ | 68℃ |
| Induction of PAM site mutation for SV40_CAV1_gRNA3 | TCAGCAATGTaCGCATCAACT | ATATTTTCCCAACAGCTTCAAAG | 57℃ | 68℃ |

PCR products were extracted and purified as for plasmids for luciferase reporter assay, and sample were sent for Sanger sequencing using the corresponding Sanger sequencing

primer to confirm successful base substitution (Table 2.4). For further confirmation that donor templates contained the complete SV40 PAS and *CAV1* homology arms, purified DNA underwent whole plasmid sequencing by Eurofins Genomics.

## 2.2 Mammalian Cell culture

### 2.2.1 Cell lines

Table 2.9: Mammalian cell lines utilised for doctoral work

| **Name** | **Description** | **Source** |
|---|---|---|
| SH-SY5Y | Wildtype SH-SY5Y cells, derived from human neuroblastoma | Dr. Johnathan Cooper-Knock, Sheffield Institute of Translational Neuroscience |
| *CAV1* exon 2 CRISPR SH-SY5Y | SH-SY5Y cell line with *CAV1* exon 2 mutated by CRISPR/Cas9. Approximately 60% of the cells contain the indel within exon 2. | Dr. Tobias Moll, Sheffield Institute of Translational Neuroscience |
| *CAV1* distal enhancer KO SH-SY5Y | SH-SY5Y cell line with an enhancer mutation 16bp downstream of the ALS-associated chr7:116,222,625T>C mutation site. Approximately 72% of cells carry this enhancer mutation. | Dr. Tobias Moll, Sheffield Institute of Translational Neuroscience |
| CRISPRi SH-SY5Y | SH-SY5Y cells transduced with Zim3-dCas9 for further transduction with gRNA-containing constructs lentiGuide-Puro | In house |

Continued on next page

Table 2.9 – continued from previous page

| **Name** | **Description** | **Source** |
|---|---|---|
| CRISPRi *CAV1* proximal enhancer SH-SY5Y (gRNA 1) | SH-SY5Y cells transduced with Zim3-dCas9 further transduced with lentiGuide-Puro containing *CAV1* proximal enhancer-targeting gRNA 1 | In house |
| CRISPRi *CAV1* proximal enhancer SH-SY5Y (gRNA 2) | SH-SY5Y cells transduced with Zim3-dCas9 further transduced with lentiGuide-Puro containing *CAV1* proximal enhancer-targeting gRNA 2 | In house |
| CRISPRi *CAV1* proximal enhancer SH-SY5Y (gRNA 3) | SH-SY5Y cells transduced with Zim3-dCas9 further transduced with lentiGuide-Puro containing *CAV1* proximal enhancer-targeting gRNA 3 | In house |
| CRISPRi *CAV1* distal enhancer SH-SY5Y | SH-SY5Y cells transduced with Zim3-dCas9 further transduced with lentiGuide-Puro containing *CAV1* distal enhancer-targeting gRNA | In house |
| CRISPRi *CAV1* promoter SH-SY5Y | SH-SY5Y cells transduced with Zim3-dCas9 further transduced with lentiGuide-Puro containing *CAV1* promoter-targeting gRNA | In house |

Continued on next page

Table 2.9 – continued from previous page

| Name | Description | Source |
|---|---|---|
| CRISPRi negative control SH-SY5Y (scrambled gRNA) | SH-SY5Y cells transduced with Zim3-dCas9 further transduced with lentiGuide-Puro containing scrambled gRNA | In house |
| HEK293FT | HEK293-based cell line modified to produce high viral titres (fast-growing and stably expressing SV40 large T antigen) | ThermoFisher |
| SC3504 | Healthy control-derived lymphoblastoid cell line. | UK Motor Neurone Disease Association DNA Bank (via Dr. Jonathan Cooper-Knock, Sheffield Institute of Translational Neuroscience) |
| SP3363 | ALS patient-derived lymphoblastoid cell line, carrying the ALS-associated SNP chr7:116,222,625T>C. | UK Motor Neurone Disease Association DNA Bank (via Dr. Jonathan Cooper-Knock, Sheffield Institute of Translational Neuroscience) |
| SP3388 | ALS patient-derived lymphoblastoid cell line, carrying the ALS-associated SNP chr7:115,994,269:C>T. | UK Motor Neurone Disease Association DNA Bank (via Dr. Jonathan Cooper-Knock, Sheffield Institute of Translational Neuroscience) |

### 2.2.2 Sources of healthy control- and patient-derived cell lines

Clinical information of both healthy control subjects and ALS patients carrying SNPs within *CAV1/CAV2* enhancers, from whom lymphoblastoid cell lines were derived, is detailed in Table 2.10

Table 2.10: Clinical information of healthy control subjects and ALS patients from whom lymphoblastoid cell lines were derived

| Status | Cell line ID | Project MinE ID | Gender | Sampling age |
|---|---|---|---|---|
| Control | SC3504 | LP6008235-DNA_G11 | Female | 48 |
| Sporadic ALS | SP3363 | LP6008122-DNA_A08 | Male | 72 |
| Sporadic ALS | SP3388 | LP6008122-DNA_A03 | Female | 53 |

#### 2.2.2.1 Ethical permissions

Human lymphoblastoid cell lines derived from two Caucasian ALS patients and one neurologically normal control, all of Northern European descent, were obtained from the UK Motor Neurone Disease Association (MNDA) DNA Bank. Informed consent was obtained from all participants, including patients, family members, and controls. Ethical approval for the collection of samples and the establishment of the MNDA DNA Bank was granted by the Trent Research Ethics Committee (Ref: MREC/02/4/107, February 2003; Ref: 09/H0405/32, July 2009). Participants were provided with comprehensive information and contact details, and were informed of their right to withdraw from the study at any time. Samples were pseudo-anonymised, and clinical data were collected via a bespoke online database designed to support data entry by research nurses and to track clinical parameters such as symptom onset and presentation. Data storage and access are compliant with the UK Data Protection Act (1998). As custodians of the DNA Bank, the MNDA manages all enquiries related to data access and use.

### 2.2.3 Maintenance of mammalian cells

#### 2.2.3.1 SH-SY5Y cell culture

SH-SY5Y cell cultures were maintained in Dulbecco's Modified Eagle's Medium (DMEM) (Cytiva) supplemented with 10% (v/v) heat-inactivated foetal bovine serum (FBS) (Thermofisher) and 50U/ml Penicillin/Streptomycin (ThermoFisher) at 37°C, 5% $CO_2$.

#### 2.2.3.2 HEK293FT cell culture

HEK293FT cell cultures were maintained in DMEM (Cytiva) supplemented with 10% (v/v) FBS (Thermofisher) and 50U/ml Penicillin/Streptomycin (ThermoFisher) at 37°C, 5% $CO_2$.

#### 2.2.3.3 Lymphoblastoid cell culture

Human-derived lymphoblastoid cell cultures were maintained in Roswell Park Memorial Institute (RPMI) 1640 Medium (ThermoFisher) supplemented with 10% (v/v) FBS (Thermofisher) and 50U/ml Penicillin/Streptomycin (ThermoFisher) at 37°C, 5% $CO_2$.

#### 2.2.3.4 CRISPRi SH-SY5Y cell culture

SH-SY5Y cell cultures transduced with Zim3-dCas9 were maintained in DMEM (Cytiva) supplemented with 10% (v/v) foetal bovine serum (FBS) (Thermofisher) and 500$\mu$g/ml Hygromycin B (ThermoFisher) at 37°C, 5% $CO_2$.

### 2.2.4 Routine culturing conditions

Cell concentrations were determined by mixing cell suspension in equal parts with 0.4% Trypan Blue (Gibco) and counted either manually using an Improved Neubauer Haemocytometer (Hawksley) or through use of a Countess II Cell Counter (ThermoFisher). Cryopreservation medium for SH-SY5Y cell lines consisted of 90% FBS and 10% DMSO (v/v). For HEK293FT and lymphoblastoid cell lines, crypreservation medium consisted of their respective full medium supplemented with 10% DMSO (v/v). 1ml aliquots of 1-10x$10^6$ cells/ml in cryopreservation medium were frozen at 80°C in Mr. Frosty Freezing Container (ThermoFisher) overnight and transferred into liquid nitrogen for storage at -196°C. Frozen cell aliquots were thawed rapidly in a water bath at 37°C.

#### 2.2.4.1 Subculturing SH-SY5Ys and their derivative cell lines

SH-SY5Y cells were subcultured every 3-5 days. For passage, media containing suspension cells was removed and retained in a 15ml falcon tube, and adherent cells were washed with PBS (ThermoFisher). Adherent cells were lifted from the flask or dish by addition of 1X Trypsin (ThermoFisher) and incubation for 3 minutes at 37°C. In circumstances where Trypsin incubation would be detrimental to downstream application (CUT&RUN and ATAC-seq), suspension cells were lifted from the flask or dish by repeated PBS washes. Trypsinised cells were dislodged from the flask or dish by addition of the retained media. Cells were then transferred to a 15ml tube and pelleted by centrifugation at 400rcf for 4 minutes at room temperature. The supernatant was then removed from the cell pellet, and the cells resupended in fresh complete DMEM and transferred to a new flask or dish. Flask or dish size was dependent on application, but for standard culturing, SH-SY5Y cells were maintained in T75 flasks (Sarstedt).

#### 2.2.4.2 Subculturing HEK293FT cells

HEK293FT cells were subcultured every 3 days. For passage, media was aspirated and discarded prior to washing with PBS. Following aspiration of PBS, 0.25% Trypsin was added to the dish or flask to evenly cover the surface of the dish or flask and incubated at 37°C for 4 minutes. Trypsinised cells were dislodged from the surface by addition of complete DMEM and further diluted to achieve required cell density and plated in new dish or flask. Flask or dish size was dependent on application, with 100mm dishes (Sarstedt) used for standard culturing, 6-well plates (Sarstedt) used for test transfections, and T25 flasks (Sarstedt) used for lentiviral packaging.

#### 2.2.4.3 Subculturing lymphoblastoid cell lines

Lymphoblastoid cells were subcultured every 3-5 days. For passage, cell suspension was transferred to a 15ml tube and centrifuged at 100rcf for 5 minutes at room temperature. Following removal of the supernatant, cell pellet was resuspended in complete RPMI media to achieve the required cell density and transferred into a new flask. Flask size was dependent on application, but for standard culturing lymphoblastoid cells were maintained in T75 flasks (Sarstedt).

### 2.2.5 Small molecule treatment

#### 2.2.5.1 CBP/p300 inhibition

The small molecule A-485 (Stratech) was used to inhibit CBP/p300 histone acetyltransferase activity in SH-SY5Y cells. WT SH-SY5Y cells were plated into a 6-well plate 24 hours prior to A-485 treatment at a density to achieve $4x10^5$ cells per well. For treatment, media was replaced with DMEM containing a final concentration of 5$\mu$M A-485, with untreated cells receiving the corresponding volume of PBS as a negative control. Cells were incubated at 37°C for 48 hours and then harvested for RNA extraction and downstream analyses.

## 2.3 Transfection and transduction of plasmid DNA

### 2.3.1 Electroporation-based transfection

Electroporation-based transfection of SH-SY5Y cells was utilised for transfections of the pGL3-Promoter, pGL3-Basic, pGL3-Control, 759_Renilla and pX459 plasmids and their

derivatives. SH-SY5Y cells were passaged 1-2 days prior to electroporation to ensure 70-90% confluency on day of transfection. Transfections using the Neon® Transfection system (ThermoFisher) were carried out according to manufacturer's guidelines.

#### 2.3.1.1 Transfection of SH-SY5Y cells for luciferase reporter assays

Prior to transfection, for each condition, 90$\mu$l of DMEM without phenol red (Cytiva, SH30284.01) supplemented with 10% FBS was pre-warmed in a 96 well plate to 37°C. For Dual-Glo luminescence experiments, a 1:1 ratio of modified pGL3-Promoter, pGL3-Basic or pGL3-Control vector to 759_Renilla vector was transfected, with a equal concentrations of plasmid used for each condition. 3ml of Buffer E was transferred to Neon® Tube and placed into Neon® Pipette Station. SH-SY5Y cells were pelleted by centrifugation for 4 minutes at 400rcf, and washed with PBS. Cell pellet was resuspended in Resuspension Buffer R to a final density of $2.0 \times 10^5$ cells in 10$\mu$l of buffer R per electroporation reaction. Plasmid DNA was added to cell suspension, mixed, and then aspirated into the Neon® Pipette. The Neon® Pipette was inserted into the Neon® Tube placed in the Neon® Pipette Station and electroporated. For SH-SY5Ys the pulse conditions were set to 1200V, at a 20ms pulse width with 3 pulses. Following electroporation, the the Neon® Pipette was removed from the Neon® Pipette Station and samples were immediately transferred from the Neon® Tip into the pre-warmed culture plate.

#### 2.3.1.2 Transfection of SH-SY5Y cells for generation of a modified cell line

Transfection was performed as for luciferase reporter assays, but with the following modifications. Prior to transfection, for each condition, 240$\mu$l of DMEM (Cytiva) supplemented with 10% FBS was pre-warmed in a 48 well plate to 37°C. 5-30$\mu$g of modified pX459 plasmids were used for transfection, with equal concentrations of plasmid used for each condition.

### 2.3.2 Lentiviral packaging

For generation of polyclonal cell lines stably expressing Zim3-dCas9 for CRISPRi and subsequent delivery of gRNA constructs (modified from lentiGuide-Puro), lentiviral packaging of transfer plasmids was performed in HEK293FT cells for transduction of SH-SY5Ys. $2.4\text{x}10^6$ HEK293FT cells were plated per transfer plasmid in T25 flasks in complete DMEM 24 hours prior to transfection. 1 hour prior to transfection, HEK293FT media was replaced with 4ml of fresh complete DMEM. After approximately 30 minutes, 4$\mu$g of transfer plasmid, 2$\mu$g of pMD2-G (VSVG) envelope plasmid and 2$\mu$g of psPAX2

packaging plasmid were diluted in 400$\mu$l of serum-free DMEM. For each transfer plasmid, 24$\mu$l of LipoD293 (SignaGen) was diluted in 400$\mu$l of serum-free DMEM. Diluted LipoD293 was added to diluted DNA, mixed and incubated at room temperature for 10 minutes. The LipoD293 DNA complex was added dropwise to cell medium and homogenised. Transfected cells were incubated at 37°C and 5% $CO_2$ and culture medium was replaced after 16 hours.

### 2.3.3 Lentiviral concentration

24 hours post transfection, virus-containing media was collected from HEK293T packaging cells and centrifuged at 500rcf for 10 minutes. 3 volumes of clarified lentiviral supernatant was then combined with 1 volume of lenti-X concentrator (Takara) and incubated at 4°C for 30 minutes. Lentivirus was then centrifuged at 1,500rcf for 45 minutes at 4°C and pellet was resuspended in 50$\mu$l of ice-cold PBS.

### 2.3.4 Transduction of SH-SY5Y cells

0.4x$10^6$ SH-SY5Y cells were plated per sample in a well of a 6 well plate 24 hours prior to transduction, to yield ∼50% confluency for transduction. For transduction, polybrene transfection reagent (Merck Millipore) was added to concentrated lentivirus at a final concentration of 8$\mu$g/ml, combined with 1ml of culture medium and added to SH-SY5Ys. Transduced cells were incubated at 37°C and 5% $CO_2$ for 24 hours prior to addition of antibiotics for selection.

### 2.3.5 Antibiotic selection

For CRISPRi, SH-SY5Ys transduced with Zim3-dCas9, antibiotic selection was performed using DMEM supplemented with 10% FBS and 200$\mu$g/ml hygromycin at 37°C and 5% $CO_2$ for two weeks. Cultures were subsequently maintained in DMEM supplemented with 10% FBS and 500$\mu$g/ml hygromycin at 37°C and 5% $CO_2$. Following transduction of Zim3-dCas9 SH-SY5Y with gRNA constructs modified from lentiGuide-Puro, antibiotic selection was performed by addition of DMEM supplemented with 10% FBS, 500$\mu$g/ml hygromycin and 1mg/ml puromycin at 37°C and 5% $CO_2$ for one week followed by selection with 500$\mu$g/ml hygromycin and 2mg/ml puromycin at 37°C and 5% $CO_2$ for one week.

## 2.4 Validation of stable expression of Zim3-dCas9 in SH-SY5Y cells

### 2.4.1 Western blot

To validate stable expression of Zim3-dCas9 in transduced SH-SY5Y cells, a western blot probing for $\alpha$-Cas9 was performed.

#### 2.4.1.1 Buffers

**RIPA lysis buffer**: 50mM Tris HCl (Sigma) pH 8 at 4°C, 100mM NaCl (Melford), 2mM $MgCl_2$ (ThermoFisher), 1% Triton X-100 (Sigma), 0.1% Sodium deoxycholate (Sigma), 0.1% Sodium Dodecyl Sulfate (SDS, Sigma) made up to volume with Milli-Q water. On the day of use, 1mM Dithiothreitol (DTT, Melford), 1X Halt phosphatase inhibitor cocktail (ThermoFisher), 10mM sodium butyrate (Sigma) was added to the buffer. Benzonase 500U/100$\mu$l (Insight Biotech) was added to the pellet during cell lysis.

**1X Tris-Acetate SDS running buffer**: For 1L, 50ml of 20X NuPAGE Tris-Acetate SDS Running Buffer (ThermoFisher) was diluted into 950ml of $dH_2O$.

**10X TBS**: For 1L, tris buffered saline (TBS) was made up of 24g of Tris, 88g of NaCl at a pH of 7.6 made up to a final volume of 1L.

**1X TBS-T**: For 1L, 100ml of 10X TBS was added to 900ml of $dH_2O$. 1ml of Tween-20 was added to give a final concentration of 0.1%.

**Blocking solution**: For 40ml, 2g of dried skimmed milk (Marvel) was dissolved in 40ml of TBS-T.

#### 2.4.1.2 Antibodies

Antibodies used for western blotting and their corresponding working concentrations are listed below (Table 2.11).

Table 2.11: Antibodies used for Western Blotting

| Antibody | Source | Species | Type | Dilution |
|---|---|---|---|---|
| $\alpha$-GAPDH | Proteintech (60004-1-Ig) | Mouse | Monoclonal | 1:10,000 |
| $\alpha$-Cas9 | Abcam (ab191468) | Mouse | Monoclonal | 1:500 |

**Table 2.11 continued from previous page**

| Antibody | Source | Species | Type | Dilution |
|---|---|---|---|---|
| HRP-conjugated $\alpha$-IgG (Heavy and light chain) | Proteintech (SA00001-1) | Goat | Polyclonal | 1:10,000 |

#### 2.4.1.3 Cell harvest and bicinchoninic acid (BCA) protein assay

SH-SY5Y cells were washed with PBS and subsequently harvested by centrifugation at 400rcf for 4 minutes. Cell pellet was resuspended in 50-100$\mu$l RIPA Cell Lysis Buffer and incubated at 37°C for 20 minutes. Cell debris was then pelleted by centrifugation at 20,000rcf for 10 minutes at 4°C. Supernatant was then transferred into a low adhesion microcentrifuge tube on ice and protein content of the cell lysate was calculated by Pierce Bicinchoninic Acid (BCA) Protein Assay Kit (ThermoFisher), according to the manufacturer's instructions.

#### 2.4.1.4 SDS-PAGE electrophoresis

The desired concentration of protein lysate, typically between 25-50$\mu$g was prepared, with addition of 1X Invitrogen NuPAGE LDS Sample Buffer (ThermoFisher) to each cell lysate sample, followed by protein denaturation at 70°C for 10 minutes. XCell SureLock Mini-Cell Electrophoresis System (ThermoFisher) gel tank was filled 1X Tris-Acetate SDS running buffer, with 500$\mu$l Invitrogen NuPAGE Antioxidant (ThermoFisher) added to the upper buffer chamber. Cell lysates of equal quantity were loaded onto a NuPAGE 3-8% Tris-Acetate gel (ThermoFisher), with 5$\mu$l of sample buffer loaded into empty wells and SeeBlue Plus2 Pre-stained Protein Standard standard (ThermoFisher). Electrophoresis was carried out at 100V for 2 hours.

#### 2.4.1.5 Protein transfer

Proteins were transferred from the SDS-PAGE gel to a nitrocellulose membrane using the semi-dry Trans-Blot Turbo Transfer System (Bio-Rad) and the Trans-Blot Turbo Mino 0.2$\mu$m nitrocellulose transfer packs (Bio-Rad) at 1.3A, 25V for 25 minutes.

#### 2.4.1.6 Membrane probing and visualisation

Following protein transfer, membranes were washed twice with TBS-T before being blocked in milk blocking solution for a minimum of 40 minutes. Primary antibodies

were made up to the corresponding working concentration in milk blocking solution (Table 2.11), and the membrane incubated in primary antibody solution overnight on an agitator. The membrane was then washed thrice in TBS-T, each with 5 minute incubations at room temperature on an agitator. Horseradish peroxidase (HRP)-secondary antibody was prepared in milk blocking solution, and the membrane incubated in secondary antibody solution at room temperature for 30 minutes on an agitator. Membrane was washed a further 3 times in TBS-T, as it was following primary antibody incubation. HRP signal was visualised using SuperSignal™ West Pico PLUS Chemiluminescent Substrate and G-BOX Chem-XRQ (Syngene) and GeneSys software.

## 2.5 Validation of gRNAs for generation of genetically modified cell lines

### 2.5.1 Liposome-based transfection

For validation of pX459 derived gRNA constructs, plasmids were transfected into HEK293FT cells using liposome-based transfection. HEK293FT were seeded to achieve 30-50% confluency on the day of transfection. Approximately 30-60 minutes prior to transfection fresh complete media was added to the cells. For each reaction in a 6-well plate, 1$\mu$g of plasmid DNA was diluted in 50$\mu$l of serum-free DMEM and 3$\mu$l of LipoD293 reagent (SignaGen) was diluted in 50$\mu$l serum-free DMEM. Diluted LipoD293 reagent was immediately added to diluted DNA solution and incubated for 10 minutes at room temperature. Following incubation, the LipoD293-DNA solution was added dropwise to the cell medium for each reaction and gently mixed to homogenise the mixture. LipoD293-DNA solution containing media was removed from cells after 12-18 hours and replaced with complete DMEM.

### 2.5.2 gRNA validation

48 hours post transfection with LipoD293, cells were collected for genomic DNA (gDNA) extraction using the Monarch Spin gDNA extraction kit (NEB), as per the manufacturer's instructions for cultured cells.

#### 2.5.2.1 T7 endonuclease I assay

PCR primers flanking the Cas9 cut-site for each pX459 derivative plasmid were used to amplify gDNA surround the Cas9-mediated double strand break (Table 2.12). PCR

reactions were set up as in Table 2.2 with cycling conditions as in Table 2.3. Following amplification, PCR cleanup was performed using Monarch® PCR DNA Cleanup Kit (NEB), with resultant DNA concentration measured using Qubit dsDNA High Sensitivity quantitation assay (ThermoFisher) on the Qubit 4 fluorimeter (ThermoFisher). T7 endonuclease I (T7E1) assay was performed by adding 2$\mu$l of T7E1 reaction buffer (Vazyme) to 200ng DNA, in a final reaction volume of 19$\mu$l. PCR products were denatured to generate single-stranded DNA by heating at 95°C for 5 minutes. DNA stands were hybridised by decreasing the temperature from 95°C to 85°C, at a ramp rate of -2°C/second, then from 85°C to 25°C, at a ramp rate of -0.1°C/second. Following reannealing of DNA strands, 1$\mu$l of T7E1 enzyme (Vazyme) was added per reaction and incubated at 37°C for 15 minutes. To quench the reaction, 1.5$\mu$l of 0.25M EDTA was then added. The reaction volume was then visualised on a 1% agarose gel.

Table 2.12: Primers used for gRNA validation

| **Primer name** | **Sequence** | **Comment** |
|---|---|---|
| SV40_CAV1_gRNA1_valF | TTGCTGTCTGCCCTCTTTG | Validation of pX459_gRNA1 |
| SV40_CAV1_gRNA1_valR | GTTCATATCTCAGCCAATAAAGCGA | |
| SV40_CAV1_gRNA3_valF | GCACTCATCTGGGGCATTTAC | Validation of pX459_gRNA3 |
| SV40_CAV1_gRNA3_valR | TTTGGAGTAGGCAGTTGAGGTT | |
| SV40_CAV1_gRNA4_valF | GCATCAGCCGTGTCTATTCC | Validation of pX459_gRNA4 |
| SV40_CAV1_gRNA4_valR | TGCTGAAAATCTGCCCACATAC | |
| 601_PCR_fwd | CATCACTTGGCTCTCCTGACTT | Validation of positive control construct |
| 601_PCR_rev | GGGTCTCCACAACTAGAATCTGG | |

## 2.6 Luciferase reporter assays

24 hours following transfection of luciferase reporter constructs into SH-SY5Y cells, firefly and Renilla luciferase activity was measured using the Promega Dual-Glo luciferase assay system (Promega). Following equilibration of cell cultures to room temperature, 45$\mu$l of Dual-Glo luciferase reagent was mixed with each culture, followed by 15 minute incubation at room temperature. Firefly luciferase activity was subsequently measured using Hidex Sense plate reader and luminescence settings (orbital shake with 5 second

duration, IR cutoff filter and 5 second counting time). 45$\mu$l of Dual-Glo Stop & Glo reagent was then mixed with each culture and incubated at room temperature for 15 minutes. Renilla luciferase activity was then measured using the Hidex Sense plate reader with the same settings. To determine luciferase activity for each sample, the ratio of luminescence from the experimental firefly reporter to luminescence from the control Renilla reporter was calculated and normalised to the PBS control.

### 2.6.1 Visualisation of data

Error bars were calculated using standard error. Significance of results were determined by one-way ANOVA test using the `aov` function in `R`, followed by computation of Tukey Honest Significant Differences using the `TukeyHSD` function.

## 2.7 Quantitative reverse transcription polymerase chain reaction (RT-qPCR)

### 2.7.1 RNA extraction

For RNA extraction, cell lysates were transferred to Phasemaker tubes (ThermoFisher), with 200$\mu$l of chloroform (Sigma) added per 1ml of TRIzol lysis reagent. Samples were vigorously shaken for 15 seconds, left to rest at room temperature for 2 mins, then shaken for a further 15 seconds. Samples were centrifuged at 12,000rcf for 5 mins at 4℃ and the upper aqueous phase was transferred to a new microcentrifuge tube. RNA was precipitated by addition of 1/10 volume (of aqueous phase) 5M NaCl, an equal volume of isopropanol (Sigma) and 1$\mu$l of GlycoBlue coprecipitant (ThermoFisher), followed by gentle vortexing and incubation at room temperature for 10 minutes. Samples were centrifuged at 20,000rcf for 20 mins at 4℃ to pellet the RNA. The RNA pellet was then washed twice via centrifugation at 20,000rcf for 5 minutes at 4℃, first using an equal volume then 1ml of ice cold 75% ethanol. Ethanol was completely removed and pellets were air dried to up to 5 minutes at room temperature to remove residual ethanol. RNA pellets were resuspended in 20$\mu$l of RNase-free $H_2O$ (Sigma) and the concentration of RNA was quantified using Qubit RNA Broad Range assay kit (ThermoFisher) and a Qubit 4 fluorimeter (ThermoFisher).

### 2.7.2 Reverse transcription of RNA

Following RNA extraction, 2$\mu$g of RNA per sample was taken to generate cDNA by reverse transcription. Reverse transcription was performed using either the High-Capacity cDNA kit (ThermoFisher) or the High-Capacity RNA-to-cDNA kit (ThermoFisher), according to manufacturer's guidelines.

### 2.7.3 RT-qPCR

Samples were prepared, at least in triplicate, with primers targeting the region of interest (Table 2.13) with Power SYBR Green PCR master mix (ThermoFisher).

Table 2.13: Primers used for RT-qPCR

| **Target** | **Primers** | **Sequence** | **Working concentration** |
|---|---|---|---|
| 18S rRNA | 18S_fwd | GTAACCCGTTGAACCCCATT | 100nM |
| | 18S_rev | CCATCCAATCGGTAGTAGCG | 100nM |
| *CAV1* CDS | CAV1_exon3_CDS_fwd | GCTTGCTGTCTGCCCTCTTTG | 200nM |
| | CAV1_exon3_CDS_rev | AGACACGGCTGATGCACTGAA | 600nM |
| *CAV1* alpha isoform | CAV1_alpha_fwd | CGTAGACTCGGAGGGACATCTC | 600nM |
| | CAV1_alpha_rev | CCACGTCATCGTTGAGGTGTTTAGG | 600nM |
| *CAV1* sense eRNA | CAV1_sense_eRNA_fwd | CCTGAGTCGTACAGAAAGCTGCC | 100nM |
| | CAV1_sense_eRNA_rev | GACTGGAAGAGGCAAAAATCAAGCAG | 100nM |
| *CAV1* antisense eRNA | CAV1_antisense_eRNA_fwd | TAGGTCCAAAGCTCCTGGCAG | 600nM |
| | CAV1_antisense_eRNA_rev | TGGAGGCCATTGTGTGAGCC | 200nM |

Reverse transcription quantitative real-time PCR (RT-qPCR) was performed on QuantStudio 12K Flex Real-Time PCR System with the following cycling conditions (Table 2.14).

Table 2.14: Cycling conditions for RT-qPCR reaction

| Stage | Temperature | Time | Cycles |
|---|---|---|---|
| Hold Stage | 50°C | 2 minutes | 1 |
| | 95°C | 10 minutes | 1 |
| PCR Stage | 95°C | 15 seconds | 40 |
| | 60°C | 1 minute | |
| Melt curve | 95°C | 15 seconds | 1 |
| | 60°C | 1 minute | 1 |
| | 95°C | 15 seconds | 1 |

### 2.7.4 Visualisation of data

Error bars were calculated using standard error. Signifance of results were determined using `R`, with unpaired moderated t-tests computed using the `limma` function `eBayes` and paired t-tests computed using the `ttest` function from base `R` stats package.

## 2.8 Hi-C

### 2.8.1 Bioinformatic analysis

*In situ* Hi-C contact matrices for *in vitro* differentiated motor neurons (Zhang et al., 2022) were downloaded from the ENCODE portal with the identifiers ENCSR379CII, ENCSR550JLK, ENCSR350NJV, ENCSR794RDS, ENCSR444BAR and ENCSR094EIC (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020). The `HiCExplorer` command line suite was utilised to convert file formats using the tool `hicConvertFormat` with the flag `--resolutions 10000`. Contact domains and TAD boundaries were identified using the tool `hicFindTADs` and the flag `--correctForMultipleTesting fdr`.

To generate heatmaps demonstrating CTCF binding at TAD boundaries, CTCF ChIP-seq data from SK-N-SH neuroblastoma cells was downloaded from the ENCODE portal with the identifier ENCSR000EIC (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020). Matrices of CTCF signal around TAD boundaries were computed using the `deepTools` suite with the tool `computeMatrix` with the `scale-regions` parameter and the flags `-a 500 -b 500`. Subsequently, heatmaps were plotted using the `deepTools` tool `plotHeatmap`.

For downstream analysis of chromosome topology in ALS patients, Hi-C contact matrices from ALS patient derived motor neurons were combined using the `HiCExplorer` tool `hicSumMatrices`.

### 2.8.2 Liftover of genomic annotations

Due to the complexity of lifting over Hi-C contact matrices, UCSC `liftOver` was utilised to lift over enhancer and ALS-associated SNP locations from hg19 to hg38. For this reason, all analyses related to chromosome topology were performed in hg38.

### 2.8.3 Visualisation of data

Configuration files for plotting Hi-C track lines were generated using the `make_tracks_file` function from the `pyGenomeTracks` package. Subsequently track lines were plotted using `hicPlotTADs` tool from `HiCExplorer` with the hg38 plotting region defined using the `--region` flag.

## 2.9 Assay for transposase-accessible chromatin with sequencing (ATAC-seq)

### 2.9.1 Generation of ATAC-seq libraries

We generated ATAC-seq libraries using the Active Motif ATAC-seq kit (Active Motif), according to manufacturer's guidelines, utilising 100,000 WT SH-SY5Y cells as input. Following library preparation, library preparation size distribution was examined using Tapestation 4150 system (Agilent) and High Sensitivity DNA ScreenTape (Agilent). Paired-end (PE150) sequencing of libraries was performed by Novogene using Illumina V1.5 reagents and the Illumina NovaSeq6000 platform, generating 9GB raw data.

### 2.9.2 Bioinformatic analysis

#### 2.9.2.1 Quality control and adapter trimming

Quality control and adapter trimming of raw fastq files was performed using `trim_galore` with the flags `--phred33 --fastqc --adapter CTGTCTCTTATACACATCT --adapter2 CTGTCTCTTATACACATCT --length 20`.

#### 2.9.2.2 Alignment

Reads were aligned to the hg19 reference genome using `bowtie2` using the flags `--very-sensitive -q --phred33 -X 2000`.

#### 2.9.2.3 Post-alignment quality control

Aligned reads were filtered using `samtools view` with the flags `-bq20 -h`, and subsequently sorted using `samtools sort`. Subsequently, the aligned reads were further filtered to remove reads that aligned to the mitochondrial genome.

#### 2.9.2.4 Peak calling

ATAC-seq peaks were called from the alignment BAM file using `MACS3 callpeak` with the flags `-f BAMPE -B -g hs --nomodel --extsize 73 --shift 37 --keep-dup all --SPMR`

### 2.9.3 Validation of data

To validate ATAC-seq datasets, called peaks were compared to a publicly available SK-N-SH ATAC-seq dataset, downloaded from the ENCODE portal with the identifier ENCSR587TRP (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020), using the `bedtools` command `reldist`.

### 2.9.4 Comparison to published datasets

ATAC-seq datasets from iPSC-derived motor neurons from healthy controls and ALS patients (Zhang et al., 2022) were downloaded from the ENCODE portal with the identifiers ENCSR704VZY, ENCSR709QRD, ENCSR913OWV, ENCSR516YAD, ENCSR634WYX and ENCSR065CER (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020).

## 2.10 Cleavage under targets and release using nuclease (CUT&RUN) sequencing

### 2.10.1 Generation of CUT&RUN libraries

Cleavage under targets and release using nuclease (CUT&RUN) sequencing libraries were generated using the CUT&RUN kit from Active motif, as per manufacturers guidelines. Isolated nuclei from 500,000 SH-SY5Y cells were utilised as input, using 1$\mu$g histone H3 as the control antibody and 1$\mu$g H3K27ac as the target antibody. Antibodies used for CUT&RUN and their corresponding concentrations are listed below (Table 2.15).

Table 2.15: Antibodies used for CUT&RUN

| **Antibody** | **Source** | **Species** | **Type** | **Concentration** |
|---|---|---|---|---|
| $\alpha$-Histone H3 | Abcam (ab1791) | Rabbit | Polyclonal | 0.5mg/ml |
| $\alpha$-Histone H3 (acetyl K27) | Abcam (ab4729) | Rabbit | Polyclonal | 0.5mg/ml |

Library preparation was performed using the NEBNext® Ultra™ II kit. Following library preparation, an additional single-sided cleanup was performed to remove sub-nucleosomal peaks. Prior to sequencing, library preparation size distribution was examined using Tapestation 4150 system (Agilent) and High Sensitivity DNA ScreenTape (Agilent). Paired-end (PE150) sequencing of libraries was performed by Novogene using Illumina V1.5 reagents and the Illumina NovaSeq6000 platform, generating 9GB raw data.

### 2.10.2 Bioinformatic analysis

#### 2.10.2.1 Quality control and adapter trimming

Quality control and adapter trimming of raw fastq files was performed using `trim_galore` with the flags `--phred33 --fastqc --adapter AGATCGGAAGAGCACACGTCTGAACTCCAGTCA --adapter2 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --length 20`.

#### 2.10.2.2 Alignment

Reads were aligned to the hg19 reference genome using `bowtie2` using the flags `--local --very-sensitive-local --no-unal --no-mixed --no-discordant -q --phred33 -I 10 -X 700 --dovetail`.

#### 2.10.2.3 Post-alignment quality control

Aligned reads were filtered using `samtools view` with the flags `-bq20 -h`, and sorted using `samtools sort`.

#### 2.10.2.4 Peak calling

CUT&RUN peaks were called from the alignment BAM files using `MACS3 callpeak` using histone H3 as the control file and H3K27ac as the treatment file, with the flags `-f BAMPE -B -g hs --nolambda --keep-dup all --SPMR -q 1e-1`. Subsequently H3K27ac fold enrichment over H3 control was calculated using `MACS3 bdgcmp` using the flag `-m FE`, to allow for comparison to published ChIP-seq datasets.

#### 2.10.2.5 Validation of data

To validate CUT&RUN datasets, called peaks were compared to a publicly available SK-N-SH H3K27ac ChIP-seq dataset (Zhang et al., 2020a), downloaded from the ENCODE portal with the identifier ENCSR000FCU (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020), using the `bedtools` command `reldist`

## 2.11 Chromatin immunoprecipitation sequencing (ChIP-seq)

### 2.11.1 Use of published datasets

H3K27ac and H3K4me1 datasets, including pseudoreplicated peak and fold change over control files, from iPSC-derived motor neurons from healthy controls and ALS patients (Zhang et al., 2022) were downloaded from the ENCODE portal with the identifiers ENCSR207VLY, ENCSR778FKK, ENCSR698HPK, ENCSR564EFE, ENCSR682BFG, ENCSR358AOC, ENCSR312HLG, ENCSR489LNU, ENCSR680IWU, ENCSR754DRC, ENCSR540KQC and ENCSR425FUS (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020).

## 2.12 Transient transcriptome sequencing (TT-seq)

### 2.12.1 Generation of TT-seq libraries

Tranisent transcriptome sequencing libraries were generated using a protocol modified from Gregersen et al. (2020).

#### 2.12.1.1 4sU labelling and extraction of 4sU-labelled RNA

For SH-SY5Y cells, cells were seeded into two 150mm dishes containing 35ml of supplemented DMEM 48 hours prior to 4-thiouridine (4sU) labelling to yield ∼70% confluency at the time of labelling. For labelling, DMEM was supplemented with 4sU (Scientific Laboratory Supplies) to a final concentration of 500$\mu$M and adherent cell population was incubated in 4sU-supplemented media at 37°C, 5% $CO_2$ for 5 minutes. Following labelling, media was removed and cells were lysed by addition of 10ml TRIzol reagent (ThermoFisher) and cells removed from the dish surface using a cell scraper. For lymphoblastoid cells, cells were seeded in two T175 flasks (Starstedt) 48 hours prior to 4sU-labelling, to yield ∼70% confluency on day of labelling, with RPMI media exchanged 24 hours prior to labelling. For labelling, RPMI was supplemented with 4sU to a final concentration of 500$\mu$M and incubated at 37°C, 5% $CO_2$ for 5 minutes. Following labelling, cell suspensions were centrifuged at 600rcf for 4 mins, the supernatant removed and cells lysed by resuspension of the cell pellet in of 5ml TRIzol reagent.

For RNA extraction, cell lysates were transferred to Phasemaker tubes (ThermoFisher), with 200$\mu$l of chloroform (Sigma) added per 1ml of TRIzol lysis reagent. Samples were vigorously shaken for 15 seconds, left to rest at room temperature for 2 mins, then shaken for a further 15 seconds. Samples were centrifuged at 13,000rcf for 15 mins at 4℃ and the upper aqueous phase was transferred to a new microcentrifuge tube. An volume of isopropanol (Sigma) equal to the upper aqueous phase and 1$\mu$l of GlycoBlue coprecipitant (ThermoFisher) was added to the sample to precipitate the RNA, followed by gentle vortexing and incubation at room temperature for 10 minutes. Samples were centrifuged at 13,000rcf for 15 mins at 4℃ to pellet the RNA. The RNA pellet was then washed twice via centrifugation at 13,000rcf for 10 minutes at 4℃, first using an equal volume then 1ml of ice cold 75% ethanol. Ethanol was completely removed and pellets were air dried to up to 5 minutes at room temperature to remove residual ethanol. RNA pellets were resuspended in 100$\mu$l of RNase-free $H_2O$ (Sigma) and the concentration of 4sU-labelled RNA was quantified using Qubit RNA Broad Range assay kit (ThermoFisher) and a Qubit 4 fluorimeter (ThermoFisher).

#### 2.12.1.2 RNA fragmentation

For normalisation purposes, 500ng of *S. cerevisiae* 4TU-labelled RNA (gifted by Ang Li) was added to 100$\mu$g of 4sU-labelled RNA. To chemically fragment the RNA, 20$\mu$l of 1M NaOH (ThermoFisher) was added to RNA samples and incubated on ice for 20 minutes. To stop the fragmentation reaction, 80$\mu$l of 1M Tris-HCl, pH 6.8 was added to reaction and immediately cleaned up by isopropanol extraction. In brief, an equal volume of isopropanol and 1$\mu$l of GlycoBlue was added to the sample and incubated at room temperature for 10 minutes. Samples were centrifuged at 13,000rcf for 10 minutes at 4℃ then RNA pellets were washed as for extraction of 4sU-labelled RNA. RNA pellets were resuspended in 200$\mu$l of RNase-free $H_2O$ and efficacy of RNA fragmentation was assessed using Tapestation 4150 system (Agilent) and High Sensitivity RNA ScreenTape (Agilent).

#### 2.12.1.3 Biotinylation of 4sU-labelled RNA

3$\mu$l of biotinylation buffer, composed of 833mM Tris-HCl pH 7.4 and 83.3mM EDTA, and 50$\mu$l of 0.1mg/ml MTSEA biotin-XX linker (dissolved in Dimethylformamide (DMF)/DMSO) (Biotum) was combined with 200$\mu$l of fragmented RNA. Biotinylation reaction was then incubated at room temperature in the dark for 30 minutes. To generate control samples for dot blot analysis, biotinylation of non-4SU-labelled RNA was also performed. Biotinylated RNA was purified from excess biotin linker by acidic phenol chloroform isoamyl alcohol-based RNA extraction. Biotinylated RNA was transferred to Phasemaker tubes with an equal volume of acid-phenol:chloroform, pH 4.5 (ThermoFisher, AM9722), and samples were shaken as for initial RNA extraction. Samples were centrifuged at 12,000rcf for 10 minutes at 4℃, then to remove excess biotin, 200$\mu$l of chloroform was added to the upper aqueous phase and shaken and centrifuged, as before. Upper aqueous phase was transferred to a new microcentrifuge tube and RNA was precipitated by addition of 1/10 volume of 5M NaCl () and an equal volume of isopropanol. Samples were mixed by inversion and incubated at room temperature for 10 minutes prior to centrifugation at 20,000rcf for 20 minutes at 4℃. RNA pellets were washed with ethanol and dried as for initial RNA extraction and RNA was reconstituted in 50$\mu$l RNase-free $H_2O$.

#### 2.12.1.4 Diagnostic dot blot

Optionally, a diagnostic dot blot was performed on biotinylated 4sU-labelled RNA to test for 4sU incorporation. Biotinylated non-4sU-labelled RNA (negative control) and

biotinylated 4sU-labelled RNA were denatured by incubation 65°C for 10 mins then placed on ice for 5 minutes. Hybond-N membrane and 2-3 sheets of Whatman paper were soaked in RNase-free $H_2O$, and the membrane was placed atop the Whatman paper in SLS Flowgen Dot blot manifold (Scientific Laboratory Supplies, CSLD96) and attached to a vacuum. 10-fold dilutions of denatured RNA (1$\mu$g, 100ng, 10ng and 1ng) were dropped onto the membrane. Following removal from the manifold, the membrane was UV crosslinked at 0.2J/cm$^2$ in a UV crosslinker. Membrane was blocked for 20 minutes at room temperature in dot blot blocking solution, composed of 10% (wt/vol) sodium dodecyl sulfate (SDS) and 1 mM EDTA in PBS. Membrane was probed using a 1:1000 dilution of 1mg/ml streptavidin-HRP conjugate (Epigentek Group Inc) in PBS for 15 minutes at room temperature. Membrane was washed twice with 10% (wt/vol) SDS in PBS, then twice with 1% SDS (wt/vol) in PBS, then twice with 0.1% SDS in PBS, each with 5 minute incubations at RT on an agitator. HRP signal was visualised using SuperSignal™ West Pico PLUS Chemiluminescent Substrate and G-BOX Chem-XRQ (Syngene) and GeneSys software.

#### 2.12.1.5 Streptavidin pull-down of biotinylated 4sU-labelled RNA

Biotinylated 4sU-labelled RNA was denatured at 65℃ for 10 minutes, followed by rapid cooling on ice for 5 minutes. Streptavidin pull-down of 4sU-labelled RNA was performed using the $\mu$MACS Streptavidin kit (Miltenyi Biotec). 200$\mu$l of $\mu$MACS Streptavidin MicroBeads were added to each biotinylated 4sU-labelled RNA sample and incubated on a rotating wheel for 15 minutes at room temperature. Per sample, a $\mu$MACS $\mu$Column was placed under the magnetic field of the $\mu$MACS magnetic separator. $\mu$Columns were equilibrated by rinsing with 100$\mu$l nucleic acid equilibration buffer. RNA-bound beads were then applied to the column matrix, allowing non-4sU-labelled RNA to flow through the column. The column was then washed twice with 500$\mu$l pull-down wash buffer, composed of 100 mM Tris-HCl, pH 7.4, 10mM EDTA, 1M NaCl and 0.1% (vol/vol) Tween 20, pre-warmed to 55℃. 4sU-RNA was then eluted from the column by addition of 100$\mu$l 100mM DTT, followed by a 5 minute incubation and an additional elution with a further 100$\mu$l 100mM DTT. Following elution, 4sU-RNA was purified and concentrated using acidic phenol chloroform isoamyl alcohol-based RNA extraction, as for clean-up following biotinylation, with RNA pellet reconstitued in 15µl RNase-free $H_2O$. The size distribution of 4sU-RNA was examined using Tapestation 4150 system (Agilent) and High Sensitivity RNA ScreenTape (Agilent), and RNA concentration was measured using Qubit RNA HS Assay kit (ThermoFisher) and a Qubit 4 fluorimeter (ThermoFisher).

#### 2.12.1.6 Library preparation and next generation sequencing (NGS)

Library preparation was performed using Lexogen CORALL Total RNA-seq Library Prep Kit (V1 and V2 with UDI 2nt Set A1) with UDIs (Lexogen) as per manufacturer's guidelines. In brief, reverse transcription and adaptor ligation were performed. Prior to library amplification with unique dual indices, a qPCR assay using the Lexogen PCR Add-on Kit for Illumina (Lexogen) was performed, as per manufacturer's guidelines, to determine the optimal number of PCR cycles for endpoint PCR. The optimal cycle number was determined by identifying the cycle number that corresponds to 50% maximum fluorescence (MF) and subtracting 3 cycles. Following library amplification and purification, library preparation size distribution was examined using Tapestation 4150 system (Agilent) and High Sensitivity DNA ScreenTape (Agilent). Paired-end (PE150) sequencing of libraries was performed by Novogene using Illumina V1.5 reagents and the Illumina NovaSeq6000 platform, generating 30GB of raw data.

### 2.12.2 Bioinformatic analysis

#### 2.12.2.1 Quality control

Quality control on raw fastq files was performed using `fastqc`

#### 2.12.2.2 Unique molecular identifier (UMI) extraction

Unique molecular identifier (UMI) sequences were extracted and appended to header using `UMI-tools extract` with the flags `--extract-method=string --bc-pattern= NNNNNNNNNNNN`

#### 2.12.2.3 Adapter trimming

Adapters were trimmed and post-trimming quality control was performed using `Trim Galore` with the flags `--phred33 --fastqc --adapter AGATCGGAAGAGCACACGTCTGAAC TCCAGTCAC --adapter2 AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA --length 18`.

#### 2.12.2.4 Alignment

Trimmed reads were aligned to the hg19 human reference genome and the sacCer3 yeast reference genome using `STAR` aligner using the flags `--outSAMtype BAM SortedByCoordinate`

```
--quantMode GeneCounts --outFilterType BySJout
--peOverlapNbasesMin 40 --peOverlapMMp 0.8
```

#### 2.12.2.5 Post-alignment quality control and normalisation

Post-alignment quality control was performed using `samtools view` with the flags `-h` and `-q 7`. BAM files were then sorted and indexed using `samtools`. Reads were grouped based on their UMI and mapping coordinates using `UMI-tools group` with the flags `--chimeric-pairs output --paired` and then deduplicated using `UMI-tools dedup` with the flags `--method=unique --multimapping-detection-method=NH --paired`. Normalisation was performed by calculating scaling factors based on exogenous yeast spike-in read count. Scaling factors were calculated using `DESeq2` and the function `estimateSizeFactors` to generate size factors using `ReadsPerGene.out.tab` output files from the sacCer3 alignment generated by `STAR` aligner.

To identify correlation between replicates, a `DESeq2` dataset was generated using the `ReadsPerGene.out.tab` output files from the hg19 alignment generated by `STAR` aligner. The `DESeq2` dataset was normalised using spike-in scaling factors using the `sizeFactors` function, and normalised gene counts were generated using the `counts` function with the argument `normalized = TRUE`. Correlation matrices were then calculated using the base `R` function `cor` with the argument `method = "pearson"`

Stranded reads files were generated using the `samtools` function `view`, whereby forward reads were identified using the flags `-f 99` and `-f 147` to identify proper paired reads where the read is on the forward strand and first in the pair and where the read is on the reverse strand and second in the pair, respectively. Reverse reads were identified similarly, except using the flags `-f 83` and `-f 163` to identify proper paired reads where the read is on the reverse strand and first in the pair and where the read is on the forward strand and second in the pair, respectively. The output files were then merged using `samtools merge` to generate separate BAM files of forward and reverse reads. Stranded BAM files were scaled based on exogenous spike-in using the `bamCoverage` function from `deepTools` using the `--scaleFactor` flag and scaling factors calculated using `DESeq2`.

#### 2.12.2.6 Transcript assembly

To assemble RNA transcripts from TT-seq data, *de novo* transcript assembly was performed using `StringTie` with the flags `--fr -m 50 -s 1 -c 1 -j 1` (Pertea et al., 2015). Novel eRNA transcripts were then identified using the package textttgffcompare (Pertea and Pertea, 2020), comparing the output transcript assembly GTF file to the

reference hg19 GENCODE annotation GTF (Frankish et al., 2019). For downstream analyses, transcripts that were transcribed from within, and restricted to candidate *CAV1/CAV2* enhancer regions with the class codes 'u', 'e', 'x' and 'o' were considered to be potential eRNAs.

#### 2.12.2.7 RNA structure prediction

Minimum free energy (MFE) eRNA structures were predicted using `RNAfold` with the flags `-p --noLP --salt 0.1` (Lorenz et al., 2011). RNA secondary structures were visualised using forna from ViennaRNA (Lorenz et al., 2011). Thermodynamic z-scores of RNA sequences were calculated using `ScanFold` (Andrews et al., 2018), with a window size of 10nt and a step size of 1nt for 100 randomisations. Input sequences for eRNA structure prediction are provided in Appendix B.2.

To calculate the mean MFEden of sequences with a given GC%, MFE of shuffled RNA sequences for sequence lengths ranging from 50nt to 600nt (in 50nt intervals), and GC% of 20%, 40%, 50%, 60%, 70% and 80%, was calculated using a custom `bash` shell script (see Appendix B.3.1). To calculate MFE density, shuffled RNA sequences with equimolar ratios of each nucleotide for a given RNA sequence length were generated using a custom `bash` shell script (see Appendix B.3.2), and output RNA sequence was then input into `RNAfold`. For the generation of shuffled RNA sequences of the same nucleotide composition as the native RNA sequence, the `bash` command `shuf` was utilised, again with output then input into `RNAfold`.

## 2.13 RNA-sequencing (RNA-seq)

Total SH-SY5Y RNA-seq raw fastq files (Liu et al., 2022) were downloaded from the NCBI GEO database, with the accession GSM6217322 (Edgar et al., 2002) and aligned to the hg19 reference genome using the `STAR` aligner with the flags `--outSAMtype BAM SortedByCoordinate --quantMode TranscriptomeSAM --outFilterType BySJout --peOverlapNbasesMin 40 --peOverlapMMp 0.8`.

Total RNA-seq datasets, including unique read signal and gene quantifications generated by `RSEM`, from iPSC-derived motor neurons from healthy controls and ALS patients (Zhang et al., 2022), were downloaded from the ENCODE portal with the identifiers ENCSR215KUO, ENCSR350OKU, ENCSR319GRO, ENCSR373XVN, ENCSR967DSJ and ENCSR575MQY (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020).

## 2.14 Cap analysis of gene expression sequencing (CAGE-seq)

Cap analysis of gene expression sequencing (CAGE-seq) dataset from SK-N-SH cells was downloaded from the ENCODE portal with the identifier ENCSR000CLC (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020).

## 2.15 Visualisation of next-generation sequencing data

For visualisation of output bigWig files and transcript GTF files, tracklines were generated using `pyGenomeTracks` (Lopez-Delisle et al., 2021, Ramírez et al., 2018), with gene locations provided by `TxDb.Hsapiens.UCSC.hg19.knownGene` (Carlson and Maintainer, 2015).

## 2.16 Computational genomics

### 2.16.1 Published datasets

#### 2.16.1.1 Clinical information

We utilised published Hi-C, ATAC-seq, RNA-seq and ChIP-seq data from iPSC-derived motor neurons of healthy control and ALS patient origin. Clinical information of both healthy control subjects and ALS patients is detailed in Table 2.16.

Table 2.16: Clinical information on both healthy control subjects and ALS patients from whom iPSCs were derived

| **Sample name** | **iPSC cell line and source** | **Clinical remarks** | **Mutation** | **Race** | **Gender** | **Sampling age** |
|---|---|---|---|---|---|---|
| SMA | CS00iCTR-nxx, Cedars Sinai Medical Centre | Clinically normal and healthy volunteer | N/A | AA | Male | 6 |
| CS14 | CS14iCTR-nxx, Cedars Sinai Medical Centre | Clinically normal and healthy volunteer | N/A | CAUC | Female | 30-35 |
| PGP | GM23338, Coriell Institute for Medical Research | Clinically normal and healthy volunteer | N/A | CAUC | Female | 55 |
| ALS28 | CS28iALS-C9nxx, Cedars Sinai Medical Centre | Age of onset: 46; Site of onset: left upper extremity; Disease duration: 51 months* | *C9orf72* hexa-nucleotide repeat expansion (6-8kb expanded allele) | CAUC | Male | 47 |

Continued on next page

Table 2.16 – continued from previous page

| Sample name | iPSC cell line and source | Clinical remarks | Mutation | Race | Gender | Sampling age |
|---|---|---|---|---|---|---|
| ALS29 | CS29iALS-C9nxx, Cedars Sinai Medical Centre | Age of onset: 46; Site of onset: left upper extremity; Disease duration: 60 months* | *C9orf72* hexa-nucleotide repeat expansion (6-8kb expanded allele) | CAUC | Male | 47 |
| ALS52 | CS52iALS-C9nxx, Cedars Sinai Medical Centre | Age of onset: 45; Site of onset: left upper extremity; Disease duration: 57 months | *C9orf72* hexa-nucleotide repeat expansion (6-8kb expanded allele) | CAUC | Male | 48 |

**Clinical information on both healthy control subjects and ALS patients**

Clinical information on subjects with *C9ORF72* hexanucleotide expansions and healthy controls used for iPSC lines (Zhang et al., 2022).

* indicates that at time of sampling, subject was still living

Abbreviations: CAUC – Caucasian; AA – African American

#### 2.16.1.2 Published dataset sources

In addition to iPSC-derived motor neuron datasets, we also utilised published datasets from WT SK-N-SH and SH-SY-5Y neuroblastoma cells. A summary of the sources of published datasets, including acessions is provided in Table 2.17.

Table 2.17: Sources of published datasets

| Sample name | Dataset | Source | Accession |
|---|---|---|---|
| iPSC-derived motor neurons (SMA) | Hi-C | ENCODE | ENCSR379CII |
| | ATAC-seq | ENCODE | ENCSR704VZY |
| | RNA-seq | ENCODE | ENCSR215KUO |
| | H3K4me1 ChIP-seq | ENCODE | ENCSR207VLY |
| | H3K27ac ChIP-seq | ENCODE | ENCSR312HLG |
| iPSC-derived motor neurons (CS14) | Hi-C | ENCODE | ENCSR550JLK |
| | ATAC-seq | ENCODE | ENCSR709QRD |
| | RNA-seq | ENCODE | ENCSR350OKU |
| | H3K4me1 ChIP-seq | ENCODE | ENCSR778FKK |
| | H3K27ac ChIP-seq | ENCODE | ENCSR489LNU |
| iPSC-derived motor neurons (PGP) | Hi-C | ENCODE | ENCSR350NJV |
| | ATAC-seq | ENCODE | ENCSR913OWV |
| | RNA-seq | ENCODE | ENCSR319GRO |
| | H3K4me1 ChIP-seq | ENCODE | ENCSR698HPK |
| | H3K27ac ChIP-seq | ENCODE | ENCSR680IWU |
| iPSC-derived motor neurons (ALS28) | Hi-C | ENCODE | ENCSR794RDS |
| | ATAC-seq | ENCODE | ENCSR516YAD |
| | RNA-seq | ENCODE | ENCSR373XVN |
| | H3K4me1 ChIP-seq | ENCODE | ENCSR564EFE |
| | H3K27ac ChIP-seq | ENCODE | ENCSR754DRC |
| iPSC-derived motor neurons (ALS29) | Hi-C | ENCODE | ENCSR444BAR |
| | ATAC-seq | ENCODE | ENCSR634WYX |
| | RNA-seq | ENCODE | ENCSR967DSJ |
| | H3K4me1 ChIP-seq | ENCODE | ENCSR682BFG |
| | H3K27ac ChIP-seq | ENCODE | ENCSR540KQC |
| iPSC-derived motor neurons (ALS52) | Hi-C | ENCODE | ENCSR094EIC |
| | ATAC-seq | ENCODE | ENCSR215KUO |
| | RNA-seq | ENCODE | ENCSR575MQY |
| | H3K4me1 ChIP-seq | ENCODE | ENCSR358AOC |
| | H3K27ac ChIP-seq | ENCODE | ENCSR425FUS |
| WT SK-N-SH | CTCF ChIP-seq | ENCODE | ENCSR000EIC |
| | ATAC-seq | ENCODE | ENCSR587TRP |
| | H3K27ac ChIP-seq | ENCODE | ENCSR000FCU |
| | CAGE-seq | ENCODE | ENCSR000CLC |
| WT SH-SY-5Y | RNA-seq | NCBI GEO | GSM6217322 |

### 2.16.2 Calculating read coverage over genomic features

A number of genomic feature BED files were generated or obtained for downstream analyses - the coding genome; the non-coding genome; coding exons; introns; intergenic regions; 5' UTRs; 3' UTRs; ABC enhancers; shuffled control regions; ALS-associated SNPs; quantile bins across TADs; 50kb bins up- and downstream from *CAV1* transcription start site (TSS). A BED file of the coding genome was generated by filtering the GENCODE V19 annotation GTF file for exons of protein coding genes and translated pseudogenes using the `awk` command. The resultant BED file was then used as input into `bedtools complement` to generate a BED file of the non-coding genome. To generate bed files containing genomic ranges of coding exons, introns, intergenic regions and UTRs, intronic regions were initially added to the GENCODE V45 annotation GFF3 file using `GenomeTools` function `gff3` with the flags `-retainids` and `-addintrons`. BED files of exonic and intronic regions and UTRs were then retrieved from the resultant GFF file using the `awk` command. These BED files were then merged as input to `bedtools complement` to generate a BED file of intergenic regions. Candidate enhancer regions identified using the ABC model (Fulco et al., 2019) were provided by Dr Johnathan Cooper-Knock. Shuffled control regions for comparison to enhancers were generated using the ABC enhancer predictions as input for `bedtools shuffle` using the flags `-chrom` and `-excl`, providing the locations of exons from GENCODE V19. A BED file of ALS-associated SNPs within *CAV1/CAV2* enhancers was generated using information retrieved from Cooper-Knock et al. (2021). To create quantile bins within the TADs across the *CAV1/CAV2* enhancer landscape, a custom `R` script was utilised (Appendix B.4), generating equal sized bins within each TAD, dependent on the size of a given TAD. To generate 50kb bins up- and downstream from the *CAV1* TSS, the `bedtools` command `makewindows` was used, with the flag `-w 50000`.

#### 2.16.2.1 Visualisation of data

Downstream, the `deepTools` function `multibigWigSummary` or `bedtools coverage` were used interchangeably dependent on whether the output required mean coverage over features or number of overlaps with features in the BED file. Plots were subsequently generated using `ggplot2`.

### 2.16.3 Per gene exonic versus intronic read coverage

For downstream comparison of coverage over intronic versus exonic regions for TT-seq and total RNA-seq, we generated General Transfer Format (GTF) files of introns

present in hg19. To generate a feature file containing hg19 introns, we utilised the R package `GenomicFeatures` (Lawrence et al., 2013) to convert the GENCODE V19 comprehensive gene annotation GTF file (Frankish et al., 2019) into a TxDb object using the `makeTxDbFromGFF` function. We extracted the disjoint intronic parts from the TxDb object using `intronicParts` and resultant object was subsetted to remove any spurious introns of larger than 5kb. The dataframe was then exported as a GTF file for further use using `rtracklayer` (Lawrence et al., 2009). We utilised `featureCounts` (Liao et al., 2014), a read summarisation programme to count TT-seq and RNA-seq reads over exons and introns. To count reads over exons we utilised `featureCounts` with the flags `-p -F GTF -t exon` with the GENCODE V19 comprehensive gene annotation GTF file (Frankish et al., 2019) as the input annotation file. To count reads over introns, we utilized `featureCounts` with the flags `-p -F GTF -t sequence_feature` and the intron annotation GTF file generated using `GenomicFeatures` as the input annotation file. To count the ratio of exonic reads to intronic reads for each given gene, for both TT-seq and total RNA-seq, the output of `featureCounts` was subsetted to generate dataframes containing read counts for exons and introns for each given gene. The ratio of exonic to intronic reads were then calculated and plotted using `ggplot2` (Wickham, 2016).

## 2.17 Identification of polyadenylation sites

*CAV1* polyadenylation sites were identified using the polyA_DB (Wang et al., 2018), with polyadenylation (polyA) sites specific to the human brain further identified by the using the polyA-sequencing track on the UCSC genome browser (Raney et al., 2024) (data accessible at NCBI GEO database (Edgar et al., 2002), accession GSM747470).

### 2.17.1 Transcription factor binding motifs

Transcription factor binding motifs within *CAV1/CAV2* candidate enhancers were identified using the JASPAR 2024 Transcription Factor Binding Site database (JASPAR CORE 2024) (Rauluseviciute et al., 2024) trackline on the UCSC genome browser (Raney et al., 2024).

### 2.17.2 Identification of AU-rich motifs

AU-rich elements (AREs) within the *CAV1* 3' UTR were identified using ARESite2 (Fallmann et al., 2016).

### 2.17.3 GNF atlas

Tissue specific relative expression of the *CAV1* coding sequence (CDS) and 3' UTR were retrieved from the GNF Gene Expression Atlas 2 (Su et al., 2004) trackline on the UCSC genome browser (Raney et al., 2024).

# Chapter 3

# Genome architecture of the *CAV1/CAV2* enhancer landscape: Structural insights into disease

## 3.1 Introduction

Despite the predicted prevalence of ALS-associated SNPs and structural variants within the non-coding genome (Cooper-Knock et al., 2021, Roses, 2021, Theunissen et al., 2020, van Rheenen et al., 2016), the non-coding genome remains a largely understudied component in ALS heritability. Along with the *CAV1* gene itself being enriched for ALS-associated variants, enhancers linked to *CAV1* and *CAV2* expression are significantly enriched for ALS-associated SNPs, identified as part of the whole genome sequencing consortium, Project MinE (Cooper-Knock et al., 2021).

### 3.1.1 Mapping enhancer-promoter interactions

Enhancers interact with their target promoters in 3D space (Dixon et al., 2012, Nora et al., 2012). Therefore, interrogating the 3D chromatin architecture of the genome may provide useful insights into the interactions between enhancers and their target genes. Chromosome conformation capture (3C) and its variants can be used to probe the spatial organisation of the genome at high resolution, providing means of quantifying the physical interactions between distal enhancers and their target promoters (Kagey et al., 2010, Rao et al., 2014). Hi-C, a high-throughput derivative of 3C, integrates

proximity-based ligation with high-throughput sequencing to comprehensively explore the 3D architecture of chromatin across the entire genome, revealing genome-wide chromatin interactions. This method utilises crosslinking of interacting chromatin regions, followed by fragmentation, ligation and generation of a genomic library comprising chimaeric DNA fragments. The relative abundance of these chimaeras correlates with the likelihood of their corresponding chromatin fragments interacting in 3D space across the cell population (Lieberman-Aiden et al., 2009). Consequently, Hi-C maps have been used to discern local chromatin domain structure, compartmentalisation of chromatin, and identify chromatin loops (Dixon et al., 2012, Lieberman-Aiden et al., 2009, Rao et al., 2014). Thus, the high-throughput annotation of DNA interactions via Hi-C enables the identification of promoter-enhancer interactions across the genome (Rao et al., 2014).

### 3.1.2 The Activity-By-Contact model

The Activity-By-Contact (ABC) model predicts enhancer-gene relationships based on epigenomic inputs representing enhancer activity and enhancer-promoter contact frequency. Enhancer activity is inferred from ATAC-seq, DNAse-seq and H3K27ac ChIP-seq read counts at the enhancer; the enhancer-promoter contact frequency is defined by Hi-C contact matrices, relative to the activity and contact frequencies of all proximal elements. As such, the quantitative effect of an enhancer is a function of enhancer strength and the frequency with which it contacts its cognate promoter in 3D space (Fulco et al., 2019, Nasser et al., 2021).

The regulatory input of an enhancer element $E$ on gene $G$ is given by its ABC score (Fulco et al., 2019):

$$\mathrm{ABC}_{\mathrm{Score},E,G} = \frac{A_E \times C_{E,G}}{\sum_{e \text{ within 5 Mb of } G} A_e \times C_{e,G}}$$

Where the ABC score of a given enhancer is dependent on the activity of the enhancer ($A_E$) and the contact frequency of the enhancer with its target gene ($C_{E,G}$), relative to the summation of the activity and the enhancer-gene contact frequency of all enhancers ($e$) within 5Mb of the target gene.

This model, consistent with *in vitro* perturbation data, highlights a number of key insights into enhancer functionality and gene regulatory networks. Enhancers can often regulate more than one target gene; most enhancers with detectable regulatory activity are located within 100kb of their cognate promoters; many enhancers may have quantitative effects on gene expression (Fulco et al., 2019).

### 3.1.3 Enhancer cooperativity

Recent predictive models of enhancer regulation, congruent with the ABC model, describe the impacts of enhancer perturbation as approximately proportional with intrinsic enhancer strength and enhancer-promoter contact frequency (Fulco et al., 2019, Gschwind et al., 2023). The ENCODE-rE2G model further predicts additional factors underpinning regulatory networks, including the potential super-additive cooperative activity of enhancers in close 3D proximity, exhibiting activities greater than the sum of their parts (Gschwind et al., 2023). Mammalian genomes house many more putative enhancers than genes, implying that regulatory control of a single gene may often involve the activities of multiple enhancers (reviewed in Long et al., 2016). Indeed, the study of the mammalian $\alpha$- and $\beta$-globin locus control regions has provided evidence for cooperation among several enhancers for the control of gene expression (Bender et al., 2012, Blayney et al., 2022, Hay et al., 2016). However, the exact nature of enhancer cooperation, whether it be by additive, redundant or synergistic effects on cognate gene expression, has largely been elusive. The ENCODE-rE2G model suggests that cooperative enhancer function is dependent on genomic distance between enhancers, whereby enhancers in frequent contact have super-additive regulatory effects and enhancers that contact less frequently have additive regulatory effects (Gschwind et al., 2023). In fact, in reporter assays, weak enhancers, which cannot activate gene expression in isolation, can strongly increase activation of the promoter, when flanked by the target promoter and a strong enhancer (Thomas et al., 2023).

While enhancers coexisting in the same locus are unlikely to be functionally equivalent, they may provide transcriptional robustness and accuracy (Frankel et al., 2010, Fulco et al., 2016, Montavon et al., 2011, Osterwalder et al., 2018). This degree of enhancer cooperativity, and indeed potential functional redundancy, may suggest that individual enhancers can be inactivated without severe phenotypic consequence, as redundant enhancers could compensate for loss of regulatory control. Enhancer redundancy would therefore be evolutionarily advantageous, dampening potential phenotypic consequences of disease-associated mutation in enhancers (Osterwalder et al., 2018). Despite the apparent prevalence of enhancer redundancy, single enhancer deletions have been demonstrated have severe phenotypic consequences, such as the resultant limb malformations following loss of the distal *Shh* enhancer (Lettice et al., 2003, Sagai et al., 2005).

As established, increasing genomic distance between the enhancer and the promoter decreases the capacity of the enhancer to activate the promoter (Fulco et al., 2019, Gschwind et al., 2023, Thomas et al., 2023). Many long-range enhancers may operate at distances where, if functioning independently, they would be unable to activate the promoter effectively. Therefore, integration of multiple weak synergistic enhancers,

in addition to mitigating the effects of enhancer loss, could partially counterbalance the decline in enhancer activity due to increased distance from the target promoter, thereby increasing the distance at which an enhancer can communicate with its promoter (Thomas et al., 2023).

The inherent complexity of endogenous enhancer regulatory networks can indeed hinder the dissection of the impacts of disease-associated mutations in enhancer regions. The impact of a disease-associated SNP is not only dependent on the strength of the enhancer and genomic distance from its cognate promoter, but also any cooperative functions that it may have. As demonstrated in reporter assays and for the $\alpha$-globin super-enhancer, weak enhancers can function synergistically with strong enhancers to facilitate target gene expression (Blayney et al., 2022, Thomas et al., 2023). Therefore, the presence of a disease-associated SNP in a weak enhancer, while having minimal impact on cognate gene expression when studied in isolation, may impact overall activity through disruption of synergistic cooperation with additional enhancer elements. Composite enhancer function thereby highlights the requirement to study the enhancer landscape as a whole when considering the impact of disease-associated SNPs. This raises the question - with some *CAV1* enhancers likely having minimal independent impact on *CAV1* gene expression, would the impact of mutation on these enhancers be analogously minimal when considering the enhancer landscape as a whole?

### 3.1.4 Chapter aims

To model the *CAV1/CAV2* epigenomic landscape, in our work we utilised the neuronal-like neuroblastoma-derived SH-SY5Y cell line, which is commonly used as a model system in neurobiology research (Biedler et al., 1978, Kovalevich and Langford, 2013). Utilising candidate *CAV1/CAV2*-associated enhancers identified by the ABC model (Fulco et al., 2019), a number of which have been identified to contain rare ALS-associated variants via GWAS and burden testing (Cooper-Knock et al., 2021), we sought to characterise the genomic landscape of *CAV1/CAV2* enhancers in the context of ALS.

This chapter focuses on the genomics of the *CAV1/CAV2* enhancer landscape, considering the genome-wide identification of enhancers and exploring the impact of chromatin structure and genomic positioning on enhancer activity and mutation susceptibility. Firstly, we identify the chromosome topology of the *CAV1/CAV2* enhancer landscape, defining TAD architecture in iPSC-derived motor neurons. We draw comparisons between the chromatin structure in iPSC-derived motor neurons and neuroblastoma cells,

via the enrichment of architectural proteins associated with TAD boundaries. Furthermore, we evaluate the implications of TAD structure, intra- and inter-TAD chromatin interactions, and enhancer location, exploring the influence these features have on enhancer-gene interactions and their potential contribution to the propensity for enhancers to gain ALS-associated SNPs.

## 3.2 Chromosome topology of the *CAV1/CAV2* regulatory network

TAD organisation is largely conserved between cell types (Dixon et al., 2015, 2012, Nora et al., 2012), but can be altered in disease states (Flavahan et al., 2016, Hnisz et al., 2016, Katainen et al., 2015, Lupiáñez et al., 2015). While altered TAD architecture underpinning disease is largely evidenced in cancer (Flavahan et al., 2016, Hnisz et al., 2016, Katainen et al., 2015, Lupiáñez et al., 2015), recent reports may suggest remodelling of large-scale chromatin organisation can occur in neurodegenerative disease, with evidence of diminished CTCF binding in the post-mortem prefrontal cortex tissue of AD patients (Patel et al., 2023).

### 3.2.1 *CAV1* and *CAV2* are likely co-regulated

In order to assess the impact of TADs on ALS-associated enhancers in neuronal-like cell lines, we utilised published Hi-C count matrices from human fibroblast-derived iPSCs differentiated *in vitro* into motor neurons to identify the chromosome topology around the *CAV1/CAV2* gene locus (Zhang et al., 2022), downloaded from the ENCODE portal with the identifiers ENCSR379CII, ENCSR550JLK, ENCSR350NJV, ENCSR794RDS, ENCSR444BAR and ENCSR094EIC (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020). Clinical information of both healthy control subjects and ALS patients is detailed in Table 2.16. To provide insight into which genes may be co-regulated in these samples, we identified the TAD organisation around the *CAV1/CAV2* gene locus in healthy control-derived *in vitro* differentiated motor neurons compared to motor neurons derived from ALS patients carrying *C9ORF72* hexanucleotide repeat expansions (Figure 3.1). In all samples, *CAV1* and *CAV2* were contained within the same TAD, along with the *TES* gene. While there was some variability in overall TAD organisation between samples, the organisation of the TAD containing *CAV1* and *CAV2* was largely maintained across samples. With exception, in the youngest control sample (SMA), the *CAV1/CAV2/TES* TAD extended upstream to also contain the *MET* gene, whereas in all other samples *MET* was contained in its own sub-Mb-sized TAD.

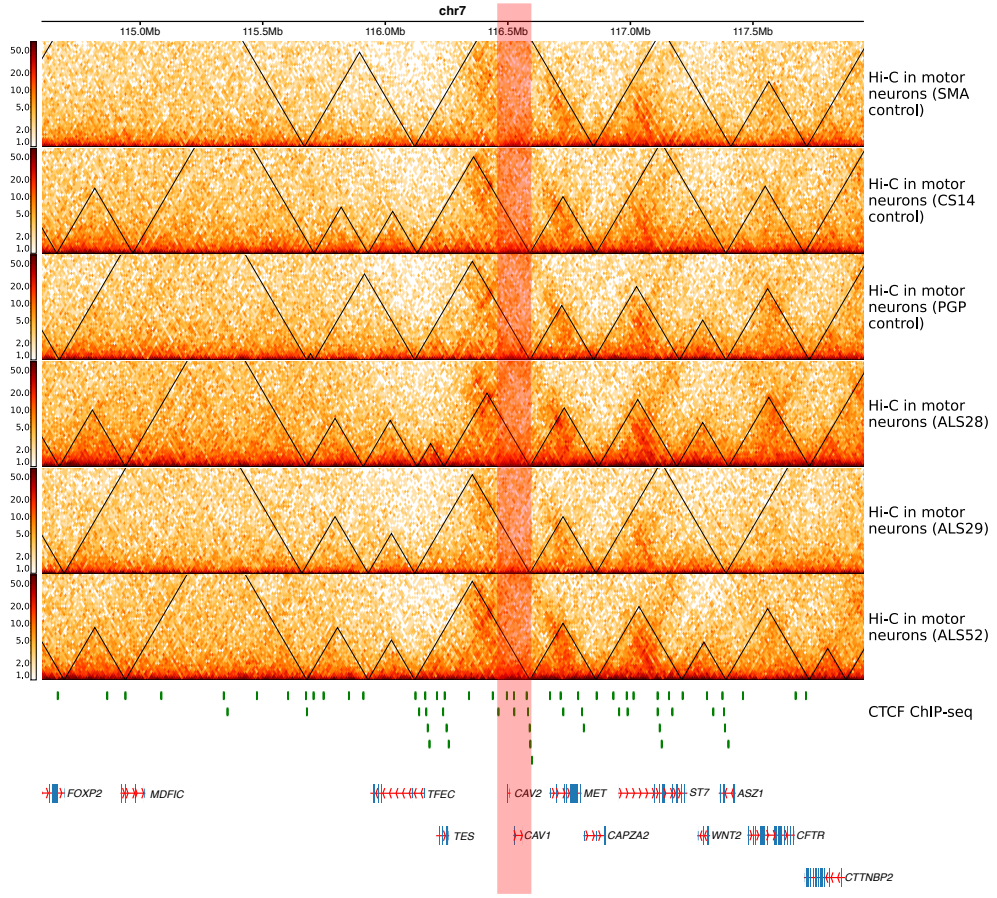



FIGURE 3.1: Chromosome topology around the *CAV1/CAV2* gene locus in *in vitro* differentiated motor neurons

Chromosome topology, as determined by Hi-C contact frequencies at chr7:114,600,000-117,950,000 (hg38) in fibroblast-derived iPSCs differentiated *in vitro* into motor neurons (Zhang et al., 2022), and CTCF binding as determined by CTCF ChIP-seq in SK-N-SH neuroblastoma cells (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020). The genomic location of *CAV1* and *CAV2* highlighted by red box.

### 3.2.2 CTCF binds at TAD boundaries

TAD boundaries are enriched for the insulator protein CTCF (Dixon et al., 2012). With TAD organisation considered to be largely invariant regardless of cell type (Dixon et al., 2015, 2012, Nora et al., 2012), identifying whether CTCF in neuroblastoma model cell lines binds at regions identified as TAD boundaries in *in vitro* differentiated motor neurons may support invariance in TAD structure between these cell types. Due to the unavailability of published CTCF ChIP-seq data from SH-SY5Y cells, we opted to utilise data from the SK-N-SH cell line, from which the SH-SY5Y cell line was derived as a subclone. Using published CTCF ChIP-seq data from SK-N-SH neuroblastoma cell lines, downloaded from the ENCODE portal with the identifier ENCSR000EIC (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020), we identified the genome-wide binding affinity of CTCF at TAD boundaries, as measured by ChIP-seq

signal (Figure 3.2). SK-N-SH CTCF ChIP-seq signal was enriched at TAD boundaries in all samples, as expected. For downstream analysis of chromosome topology in ALS patients, we used combined Hi-C contact matrices from ALS patient-derived motor neurons.

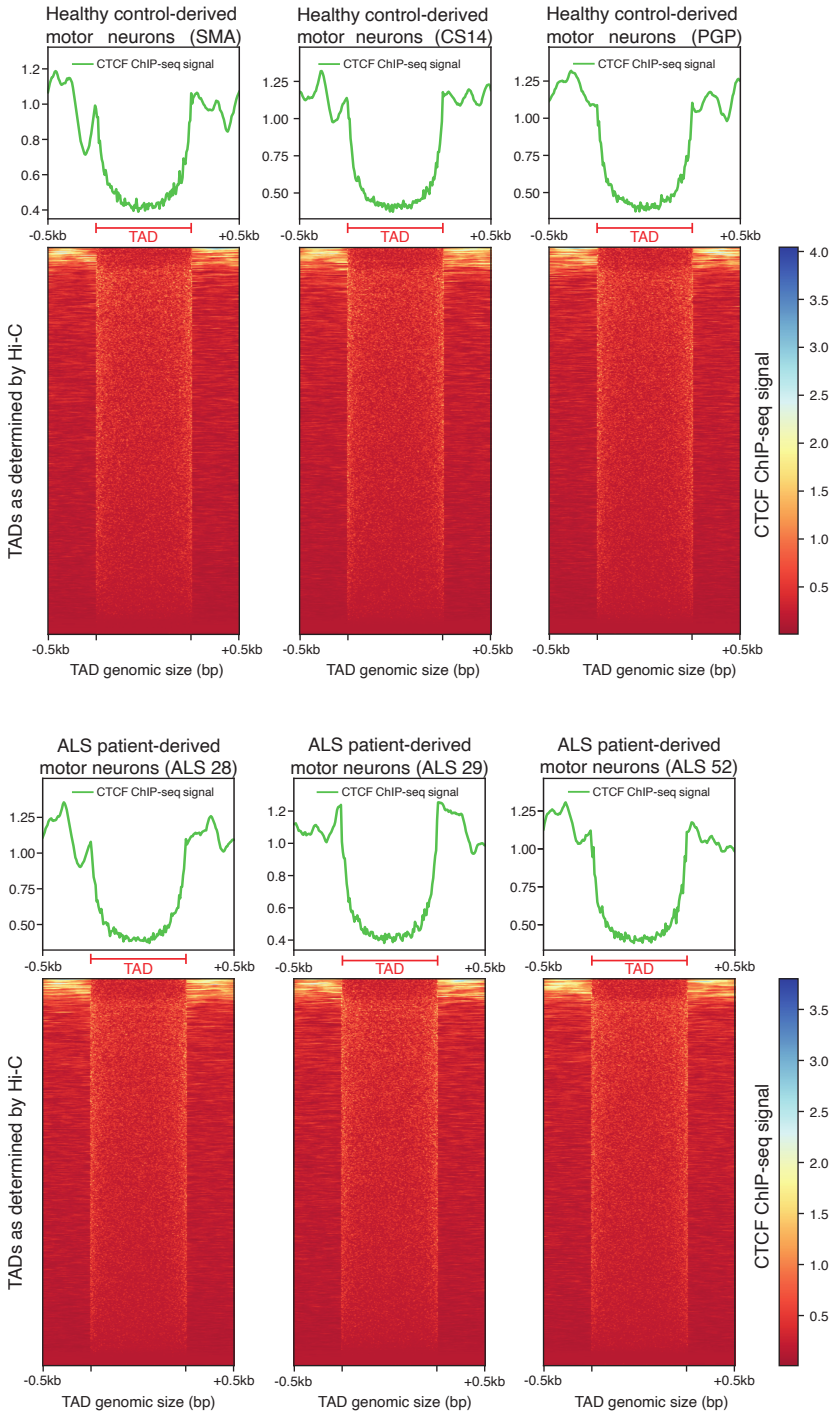



FIGURE 3.2: CTCF binds at TAD boundaries
CTCF binding as determined by ChIP-seq signal in SK-N-SH cells across all TADs genome-wide. TAD boundaries are determined by Hi-C contact frequencies in fibroblast-derived iPSCs differentiated *in vitro* into motor neurons (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020, Zhang et al., 2022).

### 3.2.3 *CAV1/CAV2* candidate enhancers localise within the *CAV1/CAV2* TAD

Higher order chromatin organisation can impact gene expression by compartmentalisation into TADs. TADs function as self-interacting domains that are largely structurally and functionally insulated from adjacent domains, with enhancer interactions typically limited to loci co-occurring within the same TAD (Dixon et al., 2012, Nora et al., 2012, Sexton et al., 2012, Shen et al., 2012). Therefore, we investigated whether *CAV1/CAV2* enhancers localised to the same TAD as *CAV1* and *CAV2*.

Utilising combined Hi-C contact matrices from ALS patient-derived motor neurons, we identified which of the candidate enhancers predicted to be associated with *CAV1/CAV2* by the ABC model lie within the *CAV1/CAV2* TAD in ALS patients. Across the *CAV1/CAV2* enhancer landscape, we identified nine TADs containing *CAV1/CAV2* candidate enhancers, hereafter referred to as TADs A-I, where, of note, *CAV1* and *CAV2* reside in TAD D (Figure 3.3). 55 of the 114 total candidate *CAV1/CAV2* enhancers localised within TAD D (Figure 3.4a). As expected, candidate *CAV1/CAV2* enhancers predominantly localised within TAD D, with 48% of all candidate *CAV1/CAV2* enhancers identified within the same TAD as *CAV1* and *CAV2* (Figure 3.4b).

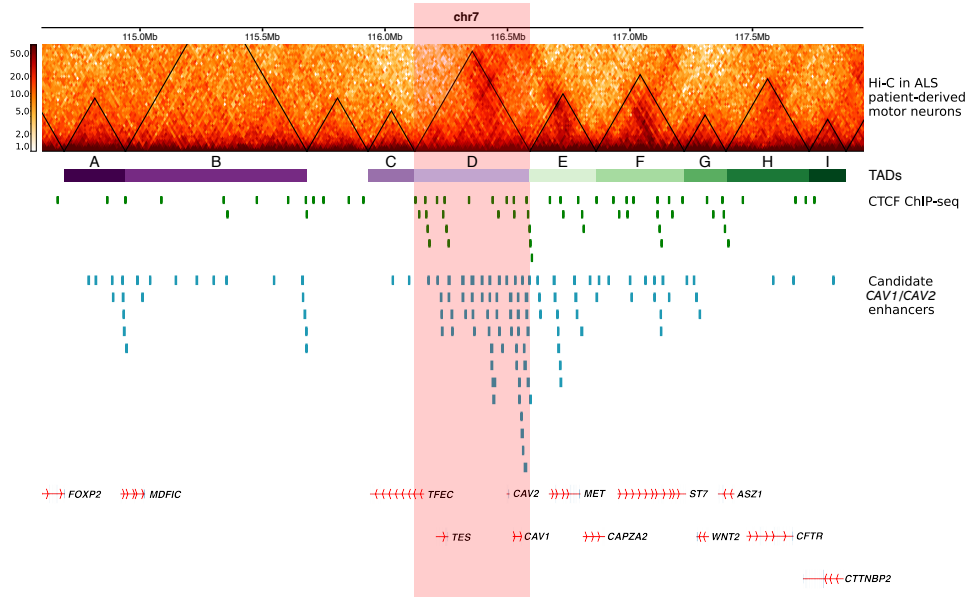



FIGURE 3.3: Chromosome topology of the *CAV1/CAV2* regulatory network in *in vitro* differentiated motor neurons

Chromosome topology, as determined by Hi-C contact frequencies at chr7:114,600,000-117,950,000 (hg38) in patient fibroblast-derived iPSCs differentiated *in vitro* into motor neurons (Zhang et al., 2022), CTCF binding as determined by CTCF ChIP-seq in SK-N-SH neuroblastoma cells (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020) and candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019). The TAD containing *CAV1* and *CAV2* highlighted by red box.

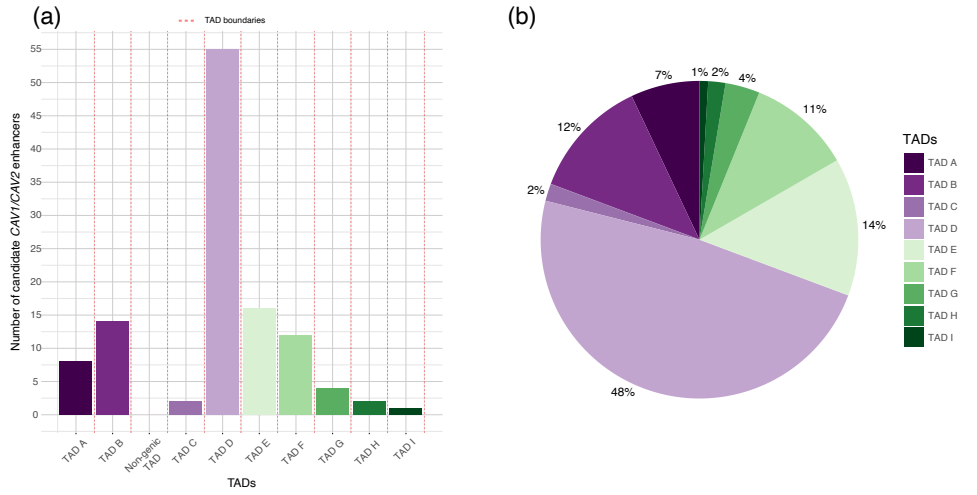



FIGURE 3.4: Candidate *CAV1/CAV2* enhancers largely localise to the *CAV1/CAV2* TAD
(a) Number of candidate *CAV1/CAV2* enhancers that localise within a given TAD.
(b) Percentage of total candidate *CAV1/CAV2* enhancers that localise within a given TAD.

### 3.2.4 Intra-TAD enhancer localisation

Alterations in TAD architecture can result in pathological ectopic gene expression through aberrant modulations in enhancer activity (Flavahan et al., 2016, Hnisz et al., 2016, Katainen et al., 2015, Lupiáñez et al., 2015). Thereby, enhancers that localise near TAD boundaries are more likely to be impacted by large-scale changes in chromatin structure. To identify whether candidate *CAV1/CAV2* enhancers localise at specific intra-TAD locations, we generated 10 quantile bins within each of the TADs and measured the number of enhancers located within each given bin (Figure 3.5a). When investigating the intra-TAD location of all enhancers at chr7:114,690,000-117,880,000, boundary-proximal bins contained an average of 3 enhancers, to the nearest integer, compared to all other bins that contained an average of 2 enhancers, to the nearest integer (Figure 3.5b). There appeared to be some preference for the bins most proximal to the TAD boundaries. When investigating the localisation of enhancers in these quantile bins for each TAD independently, three TADs appear demonstrate this preference for enhancers to localise proximal to TAD boundaries (Figure 3.5c). For TAD A, TAD B and TAD D, the boundary-proximal bins contained the most enhancers.

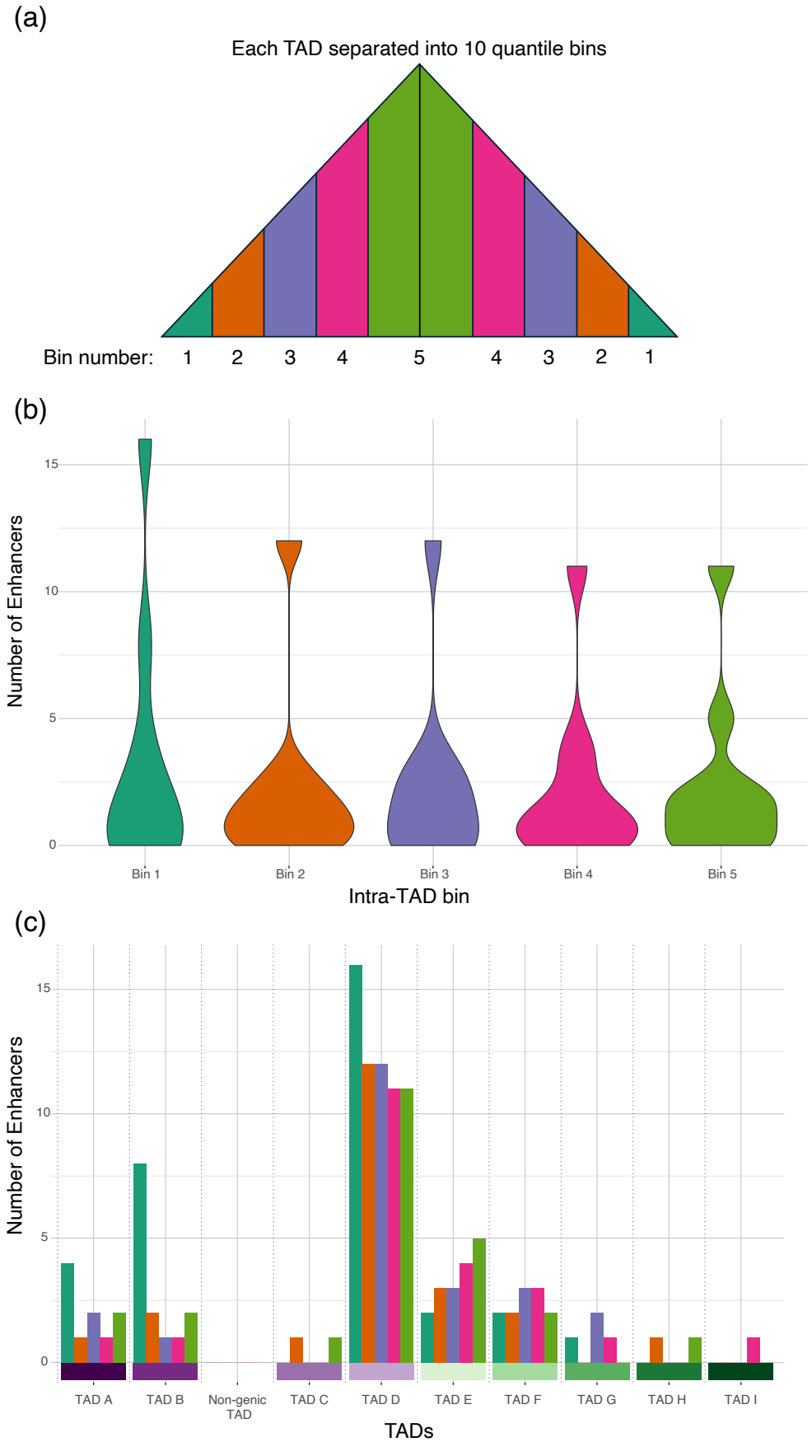



FIGURE 3.5: Candidate *CAV1/CAV2* enhancers localise proximal to TAD boundaries (a) Schematic representing the generation of quantile intra-TAD bins. (b) Distribution of candidate *CAV1/CAV2* enhancers in intra-TAD bins for all TADs at chr7:114,690,000-117,880,000 (hg38). (c) Distribution of candidate *CAV1/CAV2* enhancers in intra-TAD bins for each TAD at chr7:114,690,000-117,880,000 (hg38).

### 3.2.5 Enhancer location and disease relevance

Next, we investigated if enhancer location is related to the propensity to gain disease-associated SNPs. The inter- and intra-TAD locations of ALS-associated SNPs within candidate *CAV1/CAV2* enhancers, identified by GWAS as part of Project MiNE (Table B.2) (Cooper-Knock et al., 2021), were visualised (Figure 3.6). Quantification of ALS-associated SNPs within TADs identified that 89 of the 185 total ALS-associated SNPs localised within TAD D (Figure 3.7a). In line with the high proportion of candidate *CAV1* enhancers localising to same TAD as *CAV1* and *CAV2*, 48% of the ALS-associated SNPs within *CAV1/CAV2* enhancers also resided within TAD D (Figure 3.7b). To explore the distribution of SNPs across enhancers within each TAD, we calculated the average number of SNPs per enhancer for each TAD (Figure 3.7c). Although SNPs are discrete and cannot be fractional, the average number of SNPs per enhancer serves as a useful summary statistic to understand the overall trend. TAD I contained the highest average number of SNPs per enhancer; however, this result was influenced by the presence of a single enhancer with a high SNP count in this TAD. Additionally, TAD D exhibited an average of 2.8 SNPs per enhancer, indicating a moderately high SNP density across its enhancers.

Further, as numbers of enhancers and ALS-associated SNPs across the investigated TADs appeared to be somewhat proportional, we quantified the number of candidate *CAV1/CAV2* enhancers in each TAD that contained $\geq 1$ SNP(s), in comparison to enhancers that were unaffected by SNPs (Figure 3.7d). TAD D contained 26 enhancers containing SNPs and 29 enhancers containing no SNPs. The most noteworthy difference is that of enhancers within TAD E, where 12 enhancers contained $\geq 1$ SNP(s) in comparison to the 4 enhancers unaffected by SNPs.

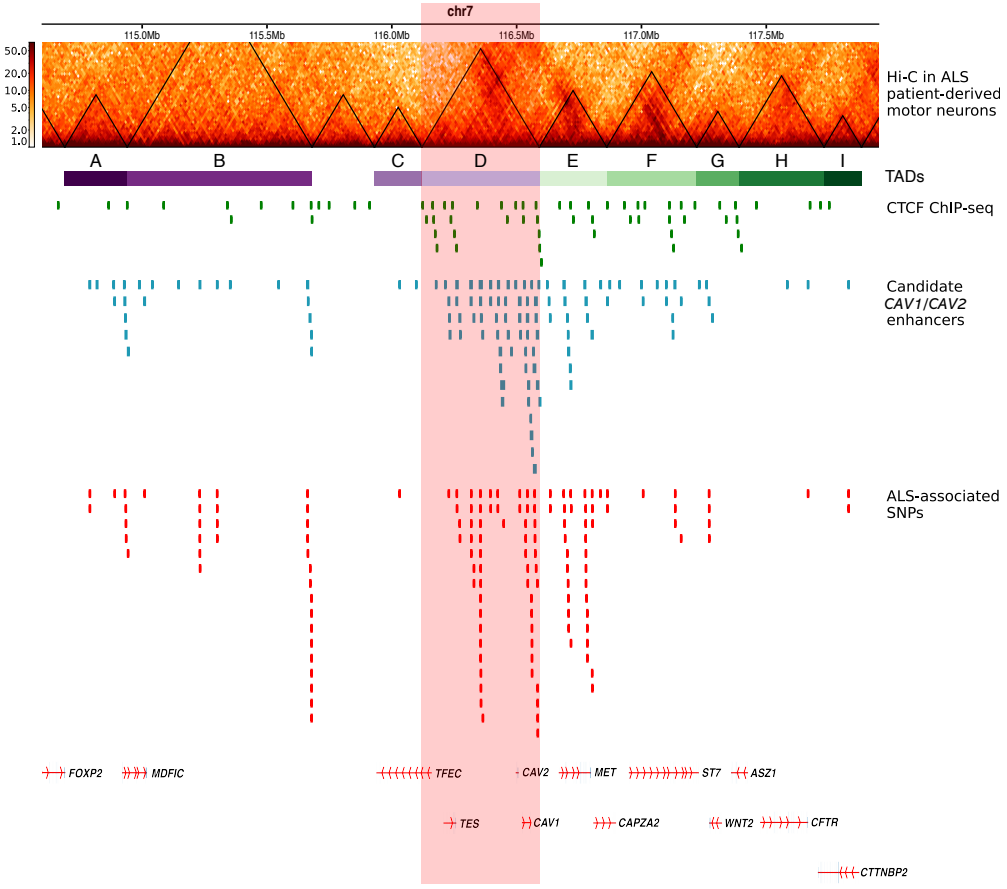



FIGURE 3.6: ALS-associated SNPs in *CAV1/CAV2* candidate enhancers cluster proximal TAD boundaries

Chromosome topology, as determined by Hi-C contact frequencies at chr7:114,600,000-117,950,000 (hg38) in patient-derived iPSCs differentiated *in vitro* into motor neurons (Zhang et al., 2022), CTCF binding as determined by CTCF ChIP-seq in SK-N-SH neuroblastoma cells (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020), candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019) and ALS-associated SNPs within candidate *CAV1/CAV2* enhancers identified by GWAS (Cooper-Knock et al., 2021).

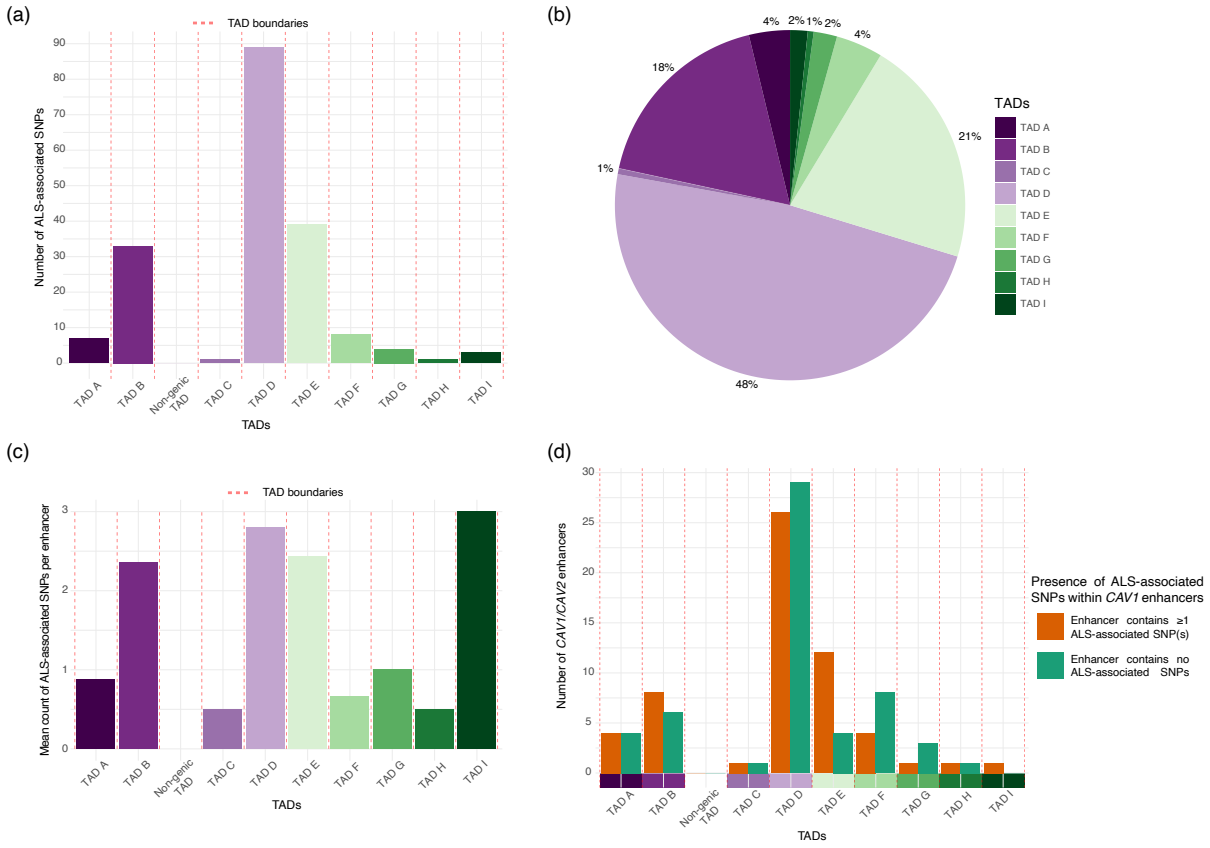



FIGURE 3.7: ALS-associated SNPs within *CAV1/CAV2* enhancers largely localise to the *CAV1/CAV2* TAD
(a) Number of ALS-associated SNPs that localised within a given TAD. (b) Percentage of all ALS-associated SNPs that localise within a given TAD. (c) Mean count of ALS-associated SNPs per enhancer in each TAD. (d) Comparison of the number of candidate *CAV1/CAV2* enhancers that do and do not contain ALS-associated SNPs for each given TAD.

#### 3.2.5.1 Enhancer density and SNP localisation

To investigate if the apparent clustering of ALS-associated SNPs in TAD D was an artefact of enhancer density, observed SNP frequency was compared to expected SNP frequency. For simplicity, expected SNP frequency was calculated based on the assumption that each candidate *CAV1/CAV2* enhancer should contain an equal number of SNPs. Given that there were 185 SNPs distributed across 114 candidate enhancers, expected SNP frequency per TAD was calculated by multiplying the number of enhancers within a given TAD by 1.62.

The expected SNP frequency within enhancers in a given TAD was calculated by:

$$\text{Expected SNP frequency} = \text{Enhancers}_{\text{TAD}} \times \left( \frac{\text{SNPs}_n}{\text{Enhancers}_n} \right)$$

where $\text{Enhancers}_{\text{TAD}}$ represents the number of enhancers in a given TAD, $\text{SNPs}_n$ represents the total number of ALS-associated SNPs, and $\text{Enhancers}_n$ represents the total number of *CAV1/CAV2* candidate enhancers.

The observed number of SNPs in TAD D was equal to the expected frequency, suggesting that the clustering of SNPs within this TAD was a result of the high number of enhancers in this TAD. TAD B, TAD E and TAD I, however, contained more observed SNPs than expected based on enhancer number alone (Figure 3.8a). For further insight into the relationship between SNP frequency and enhancer density, we investigated the observed versus expected SNP frequency in quantile intra-TAD bins (Figure 3.8b). Interestingly, while TAD D, when taken as a whole, had the expected number of SNPs, when looking at intra-TAD bins this is only true of the enhancers located within the same bin as the *CAV1* TSS, with 40 observed SNPs versus 39 expected. However for the central TAD D bin, the number of observed SNPs greatly exceeds that of the expected based on enhancer density, with 39 observed SNPs versus 15 expected.

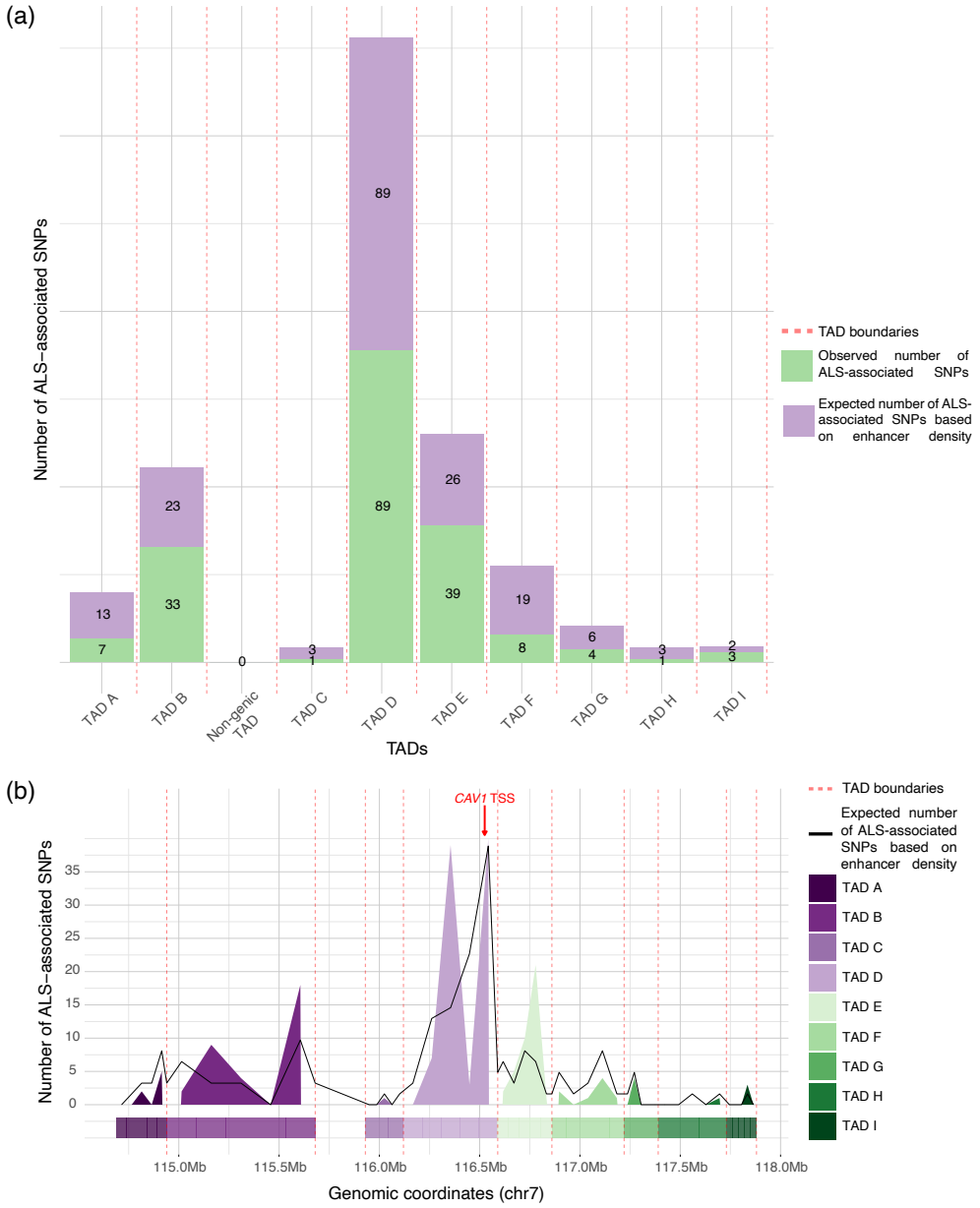



FIGURE 3.8: The relationship between inter- and intra-TAD enhancer location and number of ALS-associated SNPs in enhancers

(a) Number of observed ALS-associated SNPs in TADs at chr7:114,690,000-117,880,000 (hg38) compared to number of expected SNPs based on intra-TAD enhancer density. (b) Number of observed ALS-associated SNPs in TADs within quantile bins at chr7:114,690,000-117,880,000 (hg38) to identify intra-TAD SNP density. TAD boundary locations are demonstrated by red dashed lines. Expected number of ALS-associated SNPs within intra-TAD bins is demonstrated as the black line.

Taken together, we have demonstrated a preference for both *CAV1/CAV2* enhancers, identified via the ABC model (Fulco et al., 2019), and ALS-associated SNPs within these enhancers, identified via GWAS and burden testing (Cooper-Knock et al., 2021), to reside within the same TAD as the *CAV1* and *CAV2* genes. Furthermore, when considering intra-TAD clustering of enhancers, for some TADs in the *CAV1/CAV2* enhancer

landscape, there was some preference for enhancer localisation and SNP enrichment proximal to TAD boundaries.

## 3.3 Genomic location of candidate *CAV1/CAV2* enhancers and ALS-associated SNPs

### 3.3.1 Candidate *CAV1/CAV2* enhancers localise proximal to the *CAV1* transcription start site

Enhancer strength tends to increase with decreasing distance from its cognate target gene (Zuin et al., 2022), potentially corroborated by TAD organisation. Hence, we can utilise gene-enhancer distance to again predict whether ALS-associated enhancers are linked to *CAV1*. By generating equal-frequency 50kb bins both up- and downstream from the *CAV1* TSS (chr7:116,524,994), we identified the number of candidate enhancers in these bins (Figure 3.9a). The bin containing the most enhancers was the bin directly downstream of the bin containing the *CAV1* TSS, containing 13 enhancers.

### 3.3.2 ALS-associated SNPs localise proximal to the *CAV1* transcription start site

We additionally considered the location of ALS-associated SNPs relative to the *CAV1* TSS to identify whether enhancer genomic location has any influence over the propensity for an enhancer to gain disease-associated SNPs. Using equal-frequency 50kb bins both up- and downstream from the *CAV1* TSS
(chr7:116,524,994), we identified the number of ALS-associated SNPs in these bins (Figure 3.9b). As for the number of enhancers, the bin directly downstream of the bin containing the *CAV1* TSS contained the most SNPs, containing 18 ALS-associated SNPs. Given the assumption that enhancers proximal to the *CAV1* TSS are stronger enhancers, we hypothesise that enhancers proximal to the TSS would be more likely to gain ALS-associated SNPs, as they would potentially have a larger impact on enhancer-regulated *CAV1* expression.

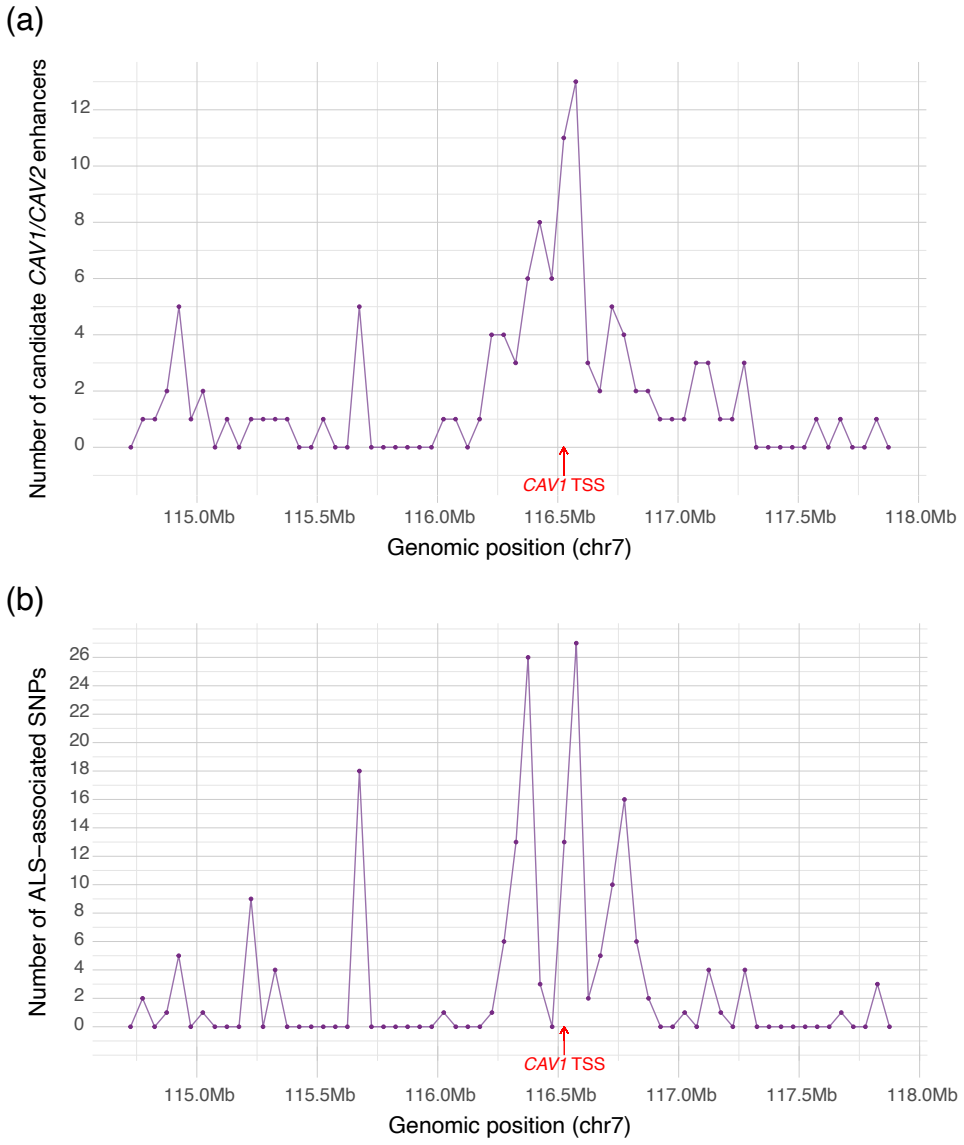



FIGURE 3.9: Localisation of *CAV1/CAV2* enhancers relative to the *CAV1* transcription start site

(a) Number of enhancers located in 50kb bins up- and downstream from the *CAV1* TSS at chr7:116,524,994 (hg38). (b) Number of ALS-associated SNPs located in 50kb bins up- and downstream from the *CAV1* transcription start site.

### 3.3.3 Tissue-specificity of enhancers and genomic location

We secondly considered the location of enhancers relative to genomic features, including introns, coding exons, UTRs and intergenic regions. The vast majority of candidate *CAV1/CAV2* enhancers localised to intronic regions, with 94 out of the 114 total enhancers located within introns (Figure 3.10a). In Cooper-Knock et al. (2021) they subset the candidate *CAV1/CAV2* enhancers into those predicted to be tissue-agnostic and those predicted to be specific to the CNS. We identified that tissue-agnostic enhancers only localised to intronic and intergenic regions, and enhancers that localised to exonic regions (coding exons and UTRs) were specific to the CNS (Figure 3.10b).

### 3.3.4 Genomic location of ALS-associated SNPs

As candidate *CAV1/CAV2* enhancers localised primarily within introns we investigated whether the location of ALS-associated SNPs demonstrated a similar pattern. As for enhancers, ALS-associated SNPs primarily localised within intronic regions followed by intergenic regions, with 148 and 32 SNPs of the total 185 SNPs, respectively (Figure 3.10c). Of note, no SNPs fell within coding exons, however while only 1 enhancer fell within a UTR region, 5 ALS-associated SNPs were identified within this enhancer. Therefore, we calculated the expected SNP frequency as before, multiplying the number of enhancers within a given genomic feature by 1.62. We observed that intronic, intergenic and coding exonic regions contained less SNPs than expected based on enhancer density, whereas UTRs contained more SNPs than expected (Figure 3.10d).

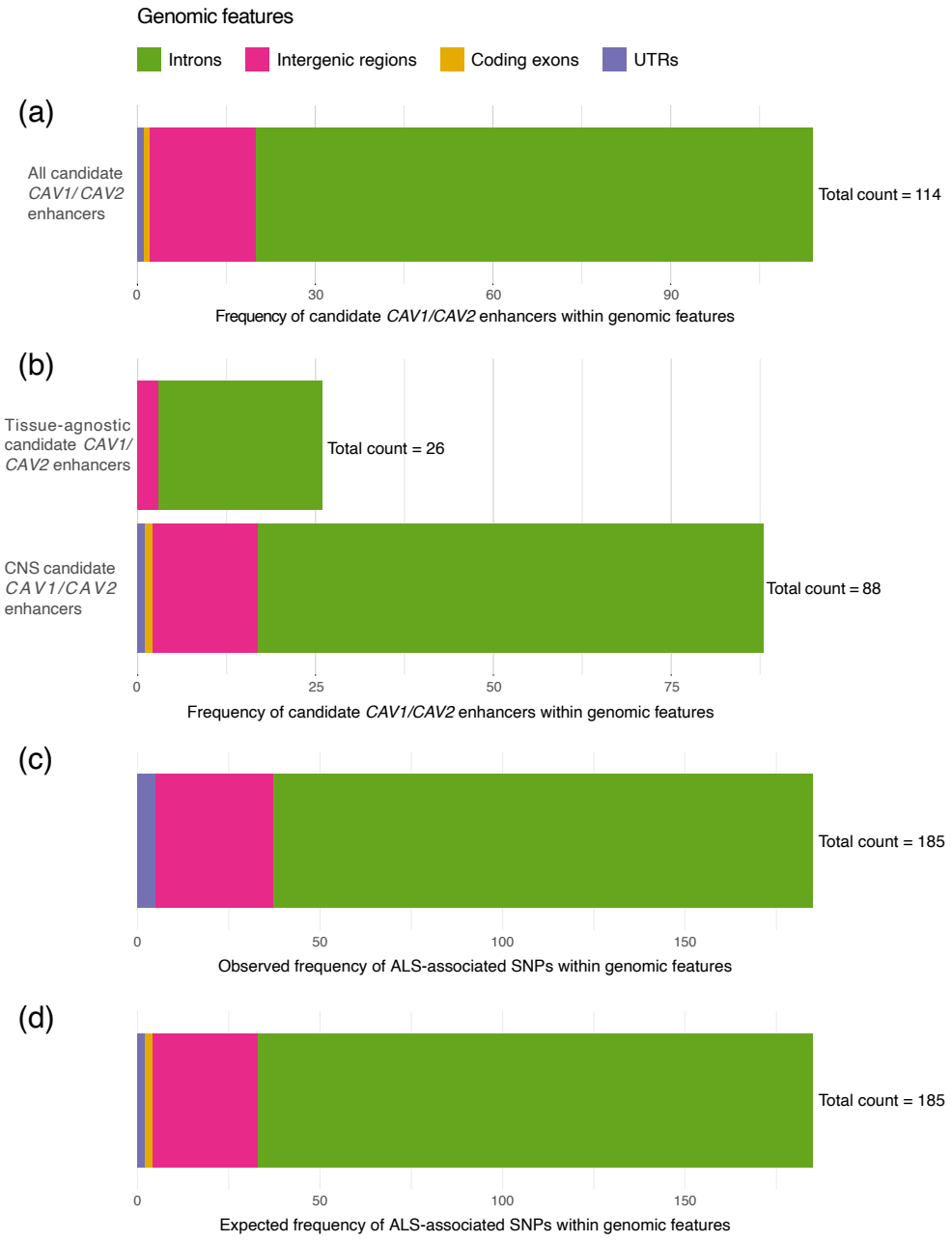



FIGURE 3.10: Frequency of *CAV1/CAV2* enhancers and ALS-associated SNPs relative to genomic features

(a) Frequency of all candidate *CAV1/CAV2* enhancers within introns, intergenic regions, coding exons and untranslated regions. (b) Frequency of tissue-agnostic and CNS-specific enhancers within introns, intergenic regions, coding exons and untranslated regions. For enhancers that overlap multiple features, enhancers were assigned to the feature that accounts for the majority of the enhancer sequence. (c) Frequency of all observed ALS-associated SNPs within *CAV1/CAV2* candidate enhancers within introns, intergenic regions and untranslated regions. (d) Expected frequency of ALS-associated SNPs within introns, intergenic regions, coding exons and untranslated regions, based on the frequency of candidate *CAV1/CAV2* enhancers within these genomic features.

Taken together, when considering linear enhancer-gene distance, we have identified a preference for both candidate *CAV1/CAV2* enhancers and ALS-associated SNPs to localise proximal to *CAV1*. Additionally, while *CAV1/CAV2* enhancers predominantly reside within intronic regions, ALS-associated SNPs show a greater than expected enrichment in enhancers located within UTRs, based on enhancer density.

## 3.4 Discussion

### 3.4.1 Chromatin architecture of the *CAV1/CAV2* enhancer landscape

#### 3.4.1.1 Co-regulation of *CAV1* and *CAV2*

*CAV1* and *CAV2* are thought to be co-regulated, as supported by ABC enhancer-gene relationship predictions, whereby a number of candidate enhancers associated with *CAV1* are also linked to *CAV2* expression (Fulco et al., 2019). Genes located within the same TAD are more likely to be co-regulated, and consequently co-expressed, in a manner that goes beyond what can be explained by purely linear proximity between genes in the genome (Long et al., 2022). We identified that in iPSC-derived motor neurons, regardless of disease state, *CAV1* and *CAV2* are located within the same TAD (Figure 3.1), supporting their co-regulation. Further, paralogous genes are more likely to lie within the same TAD, whereby common regulatory mechanisms constrained to the shared TAD can enable their coordinated expression (Ibn-Salem et al., 2017, Long et al., 2022). When we consider the frequent co-expression of CAV1 and CAV2, and their capacity to form the stable heterooligomeric caveolin complex (Scherer et al., 1997), the close proximity of the *CAV1* and *CAV2* genes both linearly and in 3D chromatin, and thus their plausible co-regulation, likely have functional significance.

#### 3.4.1.2 TADs and the implications for long-range enhancer-promoter interactions

In support of the co-regulation of *CAV1* and *CAV2*, 48% of the candidate enhancers predicted to be associated with *CAV1/CAV2* by the ABC model fell within the same TAD as *CAV1* and *CAV2* (TAD D) (Figure 3.4). Further, this supports the notion that enhancers are largely located within the same TAD as their target gene, with insulation by TAD boundaries providing a functional barrier to enhancer-gene interactions (Dixon et al., 2012, Nora et al., 2012, Sexton et al., 2012, Shen et al., 2012). TAD E, directly downstream of TAD D, contained the second highest proportion of enhancers, at 14% (Figure 3.4). Contact frequencies, as determined by Hi-C, demonstrated cross-domain

contacts between TAD D and TAD E, as can be seen in the contact pattern (Figure 3.3). Further, Hi-C in iPSC-derived motor neurons from the healthy control individual, SMA, identified a singular TAD across this region containing the *TES*, *CAV1*, *CAV2* and *MET* genes (Figure 3.1). Together, this suggests that inter-TAD interactions between enhancers in TAD E and the *CAV1* and *CAV2* gene promoters may contribute to the regulation of *CAV1* and *CAV2*.

Moreover, the apparent clustering of CTCF around the *CAV1* gene and at the upstream TAD boundary may provide a mechanism for interactions with the neighboring TADs (Figure 3.3). While TAD boundaries typically function as insulators, they may have potential to facilitate long-range enhancer-promoter interactions through the preferential alignment or "stacking" of the boundaries themselves. In this model, neighbouring TAD boundaries are stacked as a result of CTCF-guided loop extrusion, bringing boundary-proximal *cis*-regulatory elements into contact (Hung et al., 2024). The candidate *CAV1* promoter at chr7:116,520,393-116,531,041 (hg38) both contains CTCF binding sites, as determined by SK-N-SH CTCF ChIP-seq (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020), and is in relatively close proximity to the TAD boundary, approximately 54kb away from the closest TAD boundary in a TAD that totals 0.47Mb in size. Therefore, there is a possibility that this promoter could be involved in such boundary stacking phenomena, enabling enhancers within adjacent TADs to form long-range interactions with the *CAV1* promoter, despite insulation by TAD boundaries.

Beyond TAD D, we also observed a generalised preference for enhancers to localise proximal to TAD boundaries (Figure 3.5), with TAD A, TAD B and TAD D demonstrating enhancer clustering proximal to TAD boundaries. While such enhancer clustering may promote long-range enhancer-promoter interactions, this apparent preference may be further influenced by additional factors, including distance from the target gene, and the surrounding genomic landscape. As mentioned, the *CAV1* gene is located relatively close to the upstream TAD boundary. Thus, the apparent clustering of enhancers here may, in fact, be an artefact of proximity to the target gene. While enhancers can function independently of distance from the target gene (Banerji et al., 1981), there is a relationship between enhancer strength and enhancer-gene distance, with proximal enhancers typically considered stronger enhancers that those at long-range from the cognate gene (Zuin et al., 2022). We have established that candidate *CAV1/CAV2* enhancers largely localise proximal to the *CAV1* TSS (Figure 3.9a), highlighting the difficulty in distinguishing whether the clustering of enhancers at this locus is a function of TAD boundary proximity or target gene proximity. Further, we identified that the upstream boundary of TAD B lies largely across the *MDFIC* intronic region. Enhancers, especially those with tissue-specific activities, localise to intronic regions (Borsari et al., 2021). As the majority of the candidate *CAV1/CAV2* enhancers in our catalogue are CNS-specific,

as expected, the vast majority of our candidate enhancers are located within intronic regions (Figure 3.10a,b). However, the downstream TAD B boundary is neither proximal to the target genes nor located in an intronic region, therefore the enrichment of enhancers at this boundary may in fact be due to a preference for enhancers to localise at TAD boundaries.

#### 3.4.1.3 Disease relevance of chromatin architecture

When investigating the propensity of *CAV1/CAV2*-associated enhancers to gain ALS-associated SNPs, we must consider the following assumptions: disease-associated SNPs are more likely to occur in strong enhancers (Ernst et al., 2011), as SNPs within these enhancers are more likely to yield phenotypic effects; disease-associated SNPs largely localise to enhancers active in pathologically relevant cell types (Andersson et al., 2014, ENCODE Project Consortium, 2012, Nott et al., 2019). Therefore, we may expect both enhancers proximal to *CAV1* and those active in the CNS to have an increased propensity to gain ALS-associated SNPs.

With enhancer strength linked to enhancer-gene distance (Fulco et al., 2019, Gschwind et al., 2023, Zuin et al., 2022), it could be hypothesised that enhancers in closer proximity to *CAV1*, be that in close linear distance or 3D distance through the preferential chromatin interactions within TAD D, may be stronger enhancers. If a given enhancer is stronger, that is, the expression of *CAV1* is more dependent on the function of this enhancer, then SNPs in this enhancer may be more likely to result in a pathological phenotype. Moreover, in terminally differentiated tissues, such as motor neurons, the enhancer-promoter distance is more closely linked to enhancer activity, when compared to cells undergoing cell fate specification (Beagan et al., 2020, Pollex et al., 2024). In these mature cell states, enhancers may only interact with their target promoters when needed in specific spatial or temporal contexts, forming what is known as an instructive enhancer loop (Bonev et al., 2017, Pollex et al., 2024). In the context of motor neurons, this suggests that enhancer-promoter distance may play a role in the activity of the *CAV1* enhancers, suggestive that SNPs in proximal enhancers may have more impact of overall *CAV1* expression.

Using this rationale, we characterised if enhancer location is of relevance to the propensity to gain ALS-associated SNPs, investigating both the intra- and inter-TAD locations of SNPs and their distance from the *CAV1* TSS. The pattern of SNP localisation largely follows the trend observed for enhancers, whereby 48% of ALS-associated SNPs were located within candidate enhancers in TAD D (Figure 3.7a,b). As observed for enhancers, the TAD containing the second most SNPs was TAD E, containing 21% of the

total ALS-associated SNPs (Figure 3.7a,b), in congruence with the cross-domain contacts between TAD D and TAD E, as observed by Hi-C contact patterns (Figure 3.3). When we consider the average number of SNPs per enhancer in a given TAD, TAD D had a relatively high SNP density within its enhancers, with on average 3 SNPs, to the closest integer (Figure 3.7c). This may reflect that a large effect size, by mutation of strong enhancers, is required for functional downstream impact on *CAV1* expression to ultimately culminate in ALS phenotype.

Further, there was some enrichment for ALS-associated SNPs around the *CAV1* TSS, specifically, in the 50kb bin surrounding the TSS we identified 13 SNPs, and the bin directly downstream contained 27 SNPs (Figure 3.9b). The clustering of enhancers and ALS-associated SNPs proximal to the *CAV1* gene may underlie the pathological mechanisms by which SNPs in these enhancers impact *CAV1* expression. With cooperative enhancer function dependent on genomic distance between enhancers, enhancers in frequent contact have super-additive regulatory effects on their target gene expression (Gschwind et al., 2023), thereby the impact of an ALS-associated SNP within one of these enhancers may impact *CAV1* expression more so than if the enhancers were equally spaced within the TAD, likely forming additive effects.

The enrichment of ALS-associated SNPs in the 50kb bin downstream of the *CAV1* TSS may reflect the preference for SNPs to accumulate proximal to TAD boundaries, with this bin lying over the TAD D boundary. This apparent clustering of ALS-associated SNPs proximal to TAD boundaries is not unexpected in the context of disease - as observed in other diseases including cancer (Flavahan et al., 2016, Hnisz et al., 2016, Katainen et al., 2015, Lupiáñez et al., 2015). In addition to TAD D, this clustering of SNPs is apparent across the majority of TADs investigated, with TAD B and TAD E demonstrating more SNPs proximal to TAD boundaries than expected based on enhancer density in these regions (Figure 3.8).

Enhancer-promoter 3D contacts exhibit dynamism across different tissues, reflecting tissue-specific activity patterns. Typically, enhancer loci engage in tissue-specific 3D interactions with their corresponding promoters, and these interactions are notably more frequent in the relevant tissue types (Chen et al., 2024). Promoters of ubiquitously expressed genes are generally less responsive to the activities of distal enhancers, compared to genes that demonstrate variable expression across cell types (Bergman et al., 2022). In particular, more distal enhancers often regulate tissue-specific genes (Li et al., 2012a), and upon terminal tissue differentiation, many new distal enhancer-promoter interactions emerge in addition to the pre-existing enhancer landscape (Pollex et al., 2024). Therefore, the insulation provided by TADs may facilitate the enhancer-mediated regulation of genes within the same TAD, in the absence of the further specificity that may

exist for enhancer-promoter interactions beyond TADs (Long et al., 2022, Symmons et al., 2014). In concordance, all of the tissue-agnostic *CAV1/CAV2* enhancers included in our candidate enhancer dataset are localised to the the same TAD as *CAV1* and *CAV2* (Table B.2). Interestingly, while TAD D contained more enhancers unaffected by SNPs than those containing SNPs, TAD B, TAD E and TAD I contained more enhancers with at least 1 SNP than those unaffected by SNPs (Figure 3.7d). Further, this is corroborated by investigation of the number of expected SNPs within a TAD, based on enhancer density alone. While TAD D contained the same number of ALS-associated SNPs as expected based on enhancer density, TAD B, TAD E and TAD I contained more SNPs than expected, based on enhancer density alone (Figure 3.8a). This may reflect the propensity of long-range CNS-specific enhancers to obtain ALS-associated SNPs.

It is established that tissue-specific enhancers are largely situated in intronic regions (Borsari et al., 2021), with some additional evidence of enhancers localising within exons in a tissue-specific manner (Birnbaum et al., 2012). This mirrors what we observed when comparing the location of tissue-agnostic versus CNS-specific candidate *CAV1/CAV2* enhancers (Figure 3.10b). Whilst the majority of enhancers localised to introns in both tissue-specific and CNS-specific enhancers, we only identified enhancers in coding exons and UTRs in the CNS-specific subset of our candidate enhancers. Further, we observed more ALS-associated SNPs in UTRs than expected based on the number of enhancers identified in UTRs (Figure 3.10c,d), whereby we identified 1 enhancer located in the 3' UTR of the *CAV1* gene that contains 5 ALS-associated SNPs.

### 3.4.2 Conclusion

By characterisation of the *CAV1/CAV2* enhancer landscape in iPSC-derived motor neurons, we have uncovered the intricate chromatin architecture and regulatory mechanisms governing the expression of the *CAV1* and *CAV2* genes. We have demonstrated the invariance in TAD structure between patient-derived *in vitro* differentiated motor neuron samples, regardless of disease state. We further highlighted the maintenance of chromatin structure between iPSC-derived motor neurons and neuroblastoma model cell lines by identifying the enrichment of SK-N-SH CTCF ChIP-seq enrichment at TAD boundaries.

Consistent with the ABC model, which emphasizes Hi-C contact frequencies (Fulco et al., 2019, Nasser et al., 2021), we observed that a significant number of our candidate enhancers reside within the same TAD as *CAV1* and *CAV2*. Through examination of Hi-C contact frequencies and linear enhancer-gene distance, we have established which of our candidate enhancers are in close proximity *CAV1*, establishing a relationship between

enhancer-gene proximity - both along the linear chromosome and in 3D space - and the likelihood of an enhancer harbouring ALS-associated SNPs. Our findings underscore the significance that 3D chromatin structure and enhancer-gene distance may have in influencing ALS progression.

Having mapped the genome architecture of the *CAV1/CAV2* enhancer landscape and its influence on the susceptibility of enhancers to gain ALS-associated SNPs, we next aim to characterise the chromatin signatures of candidate *CAV1/CAV2* enhancers, both in the SH-SY5Y model cell line and in iPSC-derived motor neurons. This comparison will allow us to identify characteristic epigenetic signatures of enhancers in both our model system and ALS patient-derived *in vitro* differentiated motor neurons, facilitating the selection of a disease-relevant subset of *CAV1/CAV2* enhancers for further study in SH-SY5Y cells.

# Chapter 4

# Chromatin signatures of *CAV1/CAV2* enhancer landscape: Epigenetic regulation and disease implications

## 4.1 Introduction

Given the prevalence of ALS-associated SNPs in enhancers linked to *CAV1* and *CAV2* gene expression (Cooper-Knock et al., 2021), we require methodologies for not only the detection of active ALS-associated *CAV1/CAV2* enhancers, but also means for studying the impact these SNPs may have on linked gene expression and various enhancer-associated characteristics.

### 4.1.1 Methodologies for enhancer identification

In the age of next-generation sequencing, the application of high-throughput technologies has facilitated the genome-wide identification of enhancer elements across numerous cell types, resulting in the generation of extensive candidate enhancer catalogues (Fishilevich et al., 2017, Fulco et al., 2019, Gschwind et al., 2023, Visel et al., 2007). Such approaches typically leverage the distinctive chromatin signatures and binding patterns of TFs and cofactors associated with enhancer regions. Integrating these methods offers a comprehensive overview of enhancer locations and activities across various cell types, developmental stages, and disease states, providing insight into enhancer-mediated gene regulation in complex biological systems.

#### 4.1.1.1 Mapping of accessible chromatin

Enhancers typically reside in regions of accessible chromatin, indicative of a transcriptionally active state (Heintzman et al., 2007, Thurman et al., 2012). Methods that exploit the vulnerability of open chromatin regions, unprotected by nucleosomes, allow for generation of high-resolution maps of chromatin accessibility genome-wide. Nucleosome-free, accessible regions of chromatin are hypersensitive to digestion by nucleases, such as DNase I (reviewed in Gross and Garrard, 1988). Therefore, coupling digestion with DNase I with high-throughput sequencing (DNase-seq) allows for the identification of these open regions of chromatin (Boyle et al., 2008, Song and Crawford, 2010). A more recent and higher-resolution technique, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), employs a hyperactive Tn5 transposase that simultaneously fragments genomic DNA and inserts sequencing adapters into regions of open chromatin (Figure 4.1) (Buenrostro et al., 2013). Such techniques allow for an improved understanding of chromatin dynamics and for identification of regions of open chromatin, indicative of active regulatory elements.

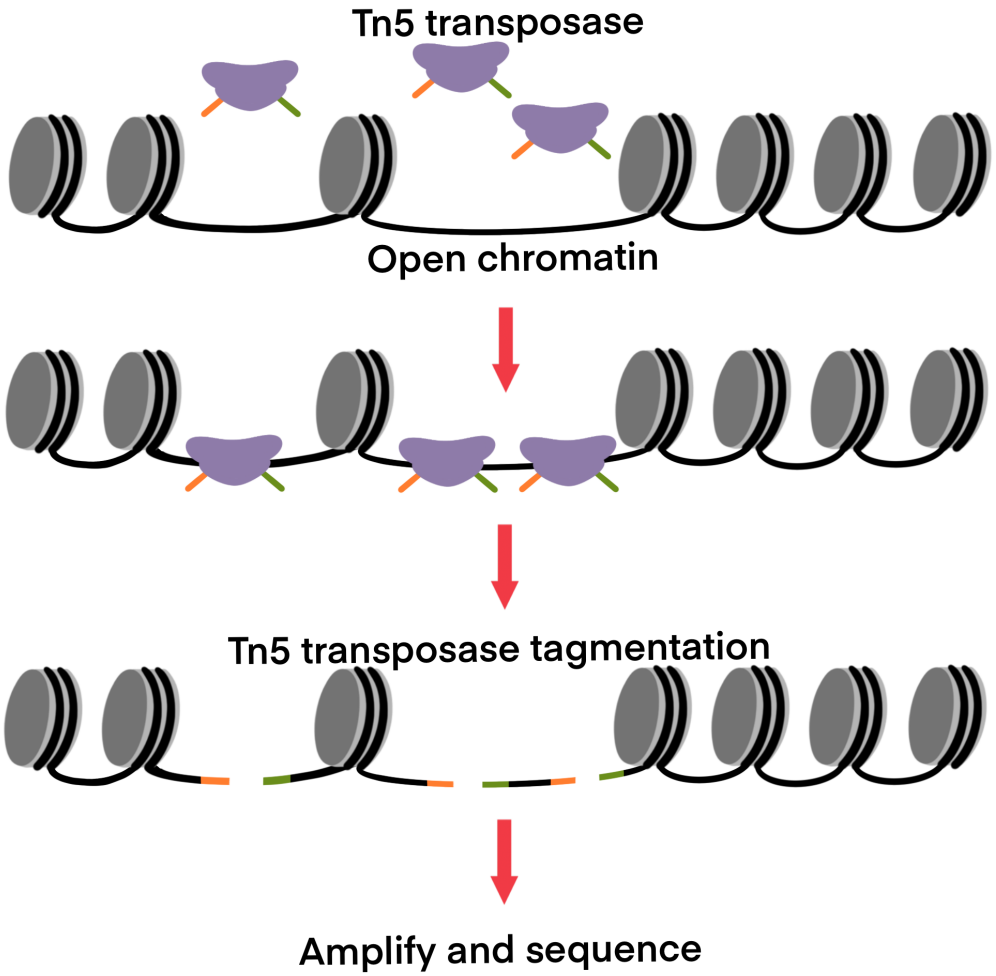



FIGURE 4.1: ATAC-seq workflow
Schematic representing the workflow of ATAC-seq. ATAC-seq probes regions of open chromatin using hyperactive Tn5 transposase loaded with adapters for high-throughput DNA sequencing. In a process known as tagmentation, Tn5 simultaneously fragments and tags regions of accessible chromatin with sequencing adapters, allowing for downstream identification of chromatin accessibility genome-wide (Buenrostro et al., 2013).

#### 4.1.1.2 Mapping of enhancer characteristics

Enhancer potential is greatly influenced by characteristic TF and cofactor binding, and the histone PTMs decorating their flanking histones. Mapping of such enhancer-associated characteristics can be informative for their identification and characterisation genome-wide (reviewed in Calo and Wysocka, 2013, Spitz and Furlong, 2012). Immunoprecipitation-based techniques, such as ChIP-seq and CUT&RUN allow for the high-throughput biochemical annotation of enhancers based on their chromatin state and protein occupancy, through employment of antibodies specific to enhancer-associated TFs, cofactors and histone modifications. In ChIP-seq, the most widely-used approach for chromatin profiling, targets are crosslinked to their bound DNA, chromatin

is sheared, solubilised and immunoprecipitated, with antibody-bound chromatin recovered for downstream sequencing of the associated DNA fragments (Barski et al., 2007, Johnson et al., 2007, Landt et al., 2012). However, the solubilisation of chromatin is an inherent limitation of ChIP-seq, requiring large input cell numbers and deep sequencing for resolution of target binding above background (Landt et al., 2012). Unlike ChIP-seq, CUT&RUN is performed *in situ*, utilising a protein A-protein G-micrococcal nuclease (pAG-MNase) fusion protein to cleave and release target-bound DNA fragments within intact nuclei for subsequent sequencing (Figure 4.2) (Skene and Henikoff, 2017). As a result, required input cell numbers and sequencing depths are dramatically reduced for CUT&RUN relative to ChIP-seq (Skene and Henikoff, 2017). Overall, chromatin profiling methods allow for the identification of enhancer epigenomic signatures genome-wide, enabling the comprehensive cataloging and characterisation of putative enhancers.

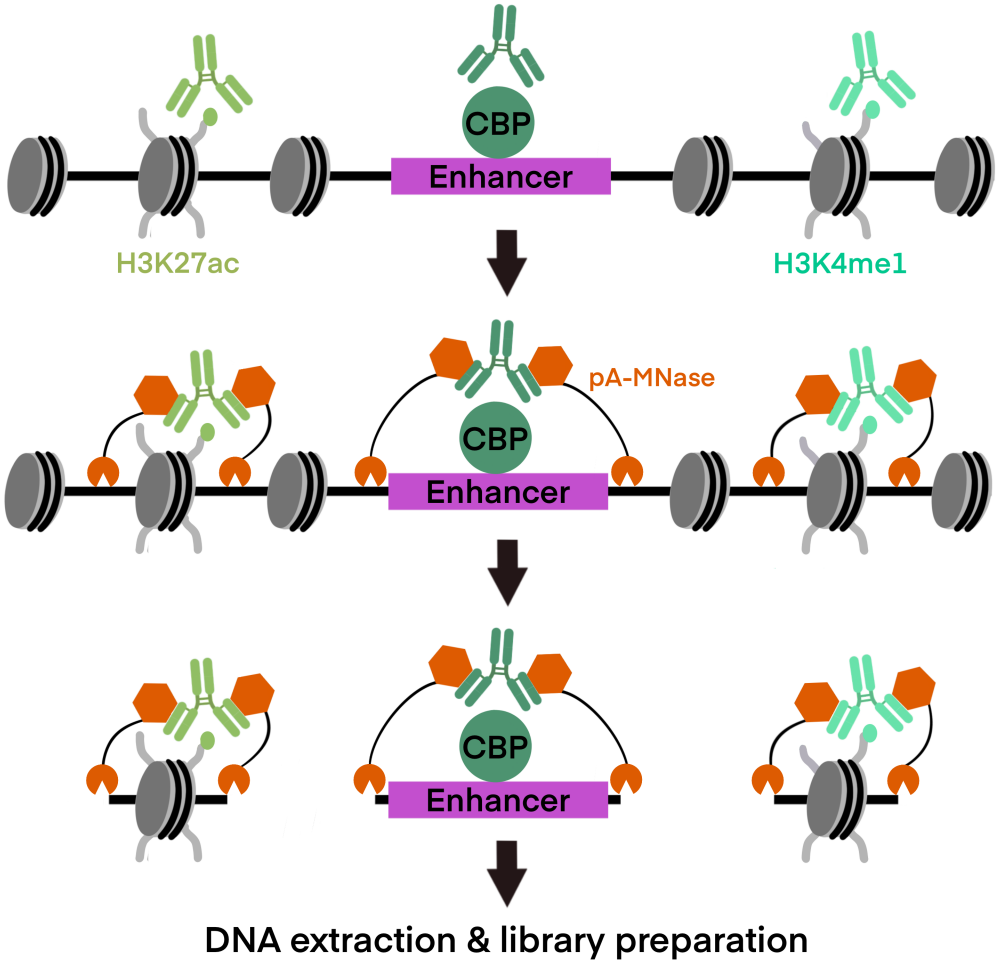



FIGURE 4.2: CUT&RUN workflow
Schematic representing the workflow of CUT&RUN. Cells are firstly immobilised and permeabilised, followed by addition of an antibody specific to the target of interest. Shown here are antibodies specific to the histone modifications H3K27ac and H3K4me1, and the transcriptional coactivator CBP, all of which are known to localise to enhancer regions. pAG-MNase is added, binding to the antibody and selectively cleaving the DNA at the bound site of interest. Cleaved DNA fragments are subsequently released into the supernatant, extracted and prepared for sequencing (Skene and Henikoff, 2017).

### 4.1.2 Chapter aims

To model the *CAV1/CAV2* epigenomic landscape in neuronally-relevant cells, here we define the chromatin signatures of *CAV1/CAV2* enhancers in both the SH-SY5Y model cell line and in iPSC-derived motor neurons (Zhang et al., 2022). Despite the neuronal-like cell identity of SH-SY5Y cells (Guvenek et al., 2022) and their expression of neuronal-specific markers upon differentiation (Kovalevich and Langford, 2013), it is imperative that we identify the aspects of the *CAV1/CAV2* epigenomic landscape in SH-SY5Y cells that resemble that of other neuronal cell types. This will allow the selection of an

appropriate subset of enhancers from the large repertoire of potential *CAV1* enhancers for study in our model system.

To characterise the epigenomic landscape of *CAV1/CAV2* enhancers in the context of ALS, we leverage candidate *CAV1/CAV2*-associated enhancers identified by the ABC model (Fulco et al., 2019), with a particular emphasis on those which have been identified to contain rare ALS-associated SNPs via GWAS and burden testing (Cooper-Knock et al., 2021). This chapter focuses on the epigenetic signatures of *CAV1/CAV2* enhancers, examining chromatin accessibility and enhancer-associated chromatin modifications to define enhancer activity in both the SH-SY5Y model cell line and in iPSC-derived motor neurons. We further explore the epigenomic landscape of *CAV1/CAV2* enhancers in the context of ALS-associated SNPs, investigating how enhancer accessibility and activity correlate with their susceptibility to acquire disease-associated variants.

## 4.2 Chromatin accessibility at enhancers

### 4.2.1 Chromatin accessibility of enhancers in SH-SY5Y cells

An open chromatin state, as determined by ATAC-seq, is an established characteristic of active enhancers (Heintzman et al., 2007, Song and Crawford, 2010, Thurman et al., 2012). Open chromatin at an enhancer region, in a given cell or tissue type, is indicative that that enhancer may have regulatory activity in that cell or tissue type. To explore chromatin accessibility at candidate *CAV1/CAV2* enhancers in our model cell line, we performed ATAC-seq in WT SH-SY5Y cells.

#### 4.2.1.1 Validation of ATAC-seq in SH-SY5Y cells

To validate our dataset, we compared peaks called from our ATAC-seq data to that of publicly available ATAC-seq in SK-N-SH cells (due to the unavailability of published ATAC-seq data in SH-SY5Y cells), downloaded from the ENCODE portal with the identifier ENCSR587TRP (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020), utilising the relative distance metric (Favorov et al., 2012). Using this metric, we assessed the spatial correlation between the two datasets by calculating the distribution of relative distances between each peak in our SH-SY5Y ATAC-seq and comparing that to the two closest peaks in the SK-N-SH ATAC-seq data. If there is no spatial correlation between the two datasets, we would expect the relative distances to be uniformly distributed. However, if the peaks are closer than expected by chance, the distribution of observed relative distances shifts towards low relative distance values.

When comparing our data with SK-N-SH ATAC-seq data, we observed a high frequency of low relative distance values (Figure 4.3), suggesting spatial correlation between the SH-SY5Y ATAC-seq peaks and those identified in the published SK-N-SH dataset. This approach allowed for the evaluation of how closely our peaks align with those identified in previous studies beyond the proportion of intersecting intervals.

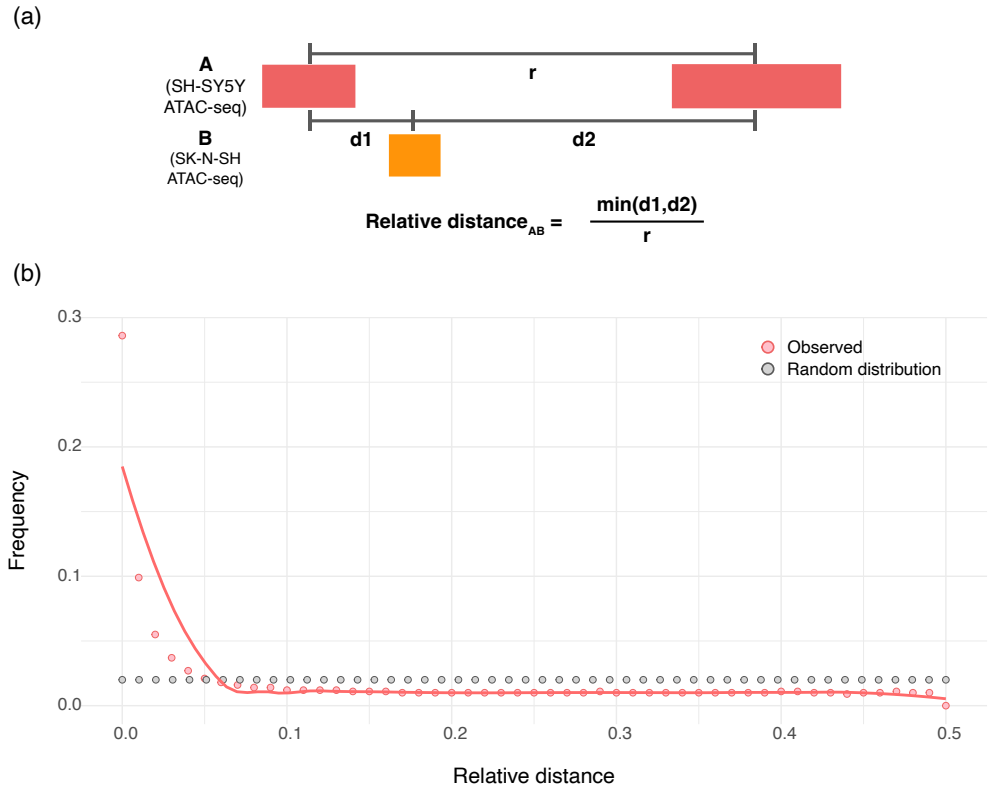



FIGURE 4.3: Relative distance between peaks in SH-SY5Y ATAC-seq publicly available ATAC-seq
(a) Schematic describing the use of the relative distance metric to assess spatial correlation between SH-SY5Y ATAC-seq and SK-N-SH ATAC-seq (Favorov et al., 2012) (b) Distribution of observed relative distances between SH-SY5Y ATAC-seq peaks and SK-N-SH ATAC-seq peaks (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020).

#### 4.2.1.2 Chromatin accessibility of *CAV1/CAV2* enhancers in SH-SY5Y cells

To characterise the chromatin state of candidate *CAV1/CAV2* enhancers in our model cell line, we investigated ATAC-seq signal enrichment over enhancers in WT SH-SY5Y cells (Figure 4.4). Of the 114 candidate *CAV1/CAV2* enhancers, 25 were found to be in an open chromatin state, as indicated by the presence of ATAC-seq peaks.

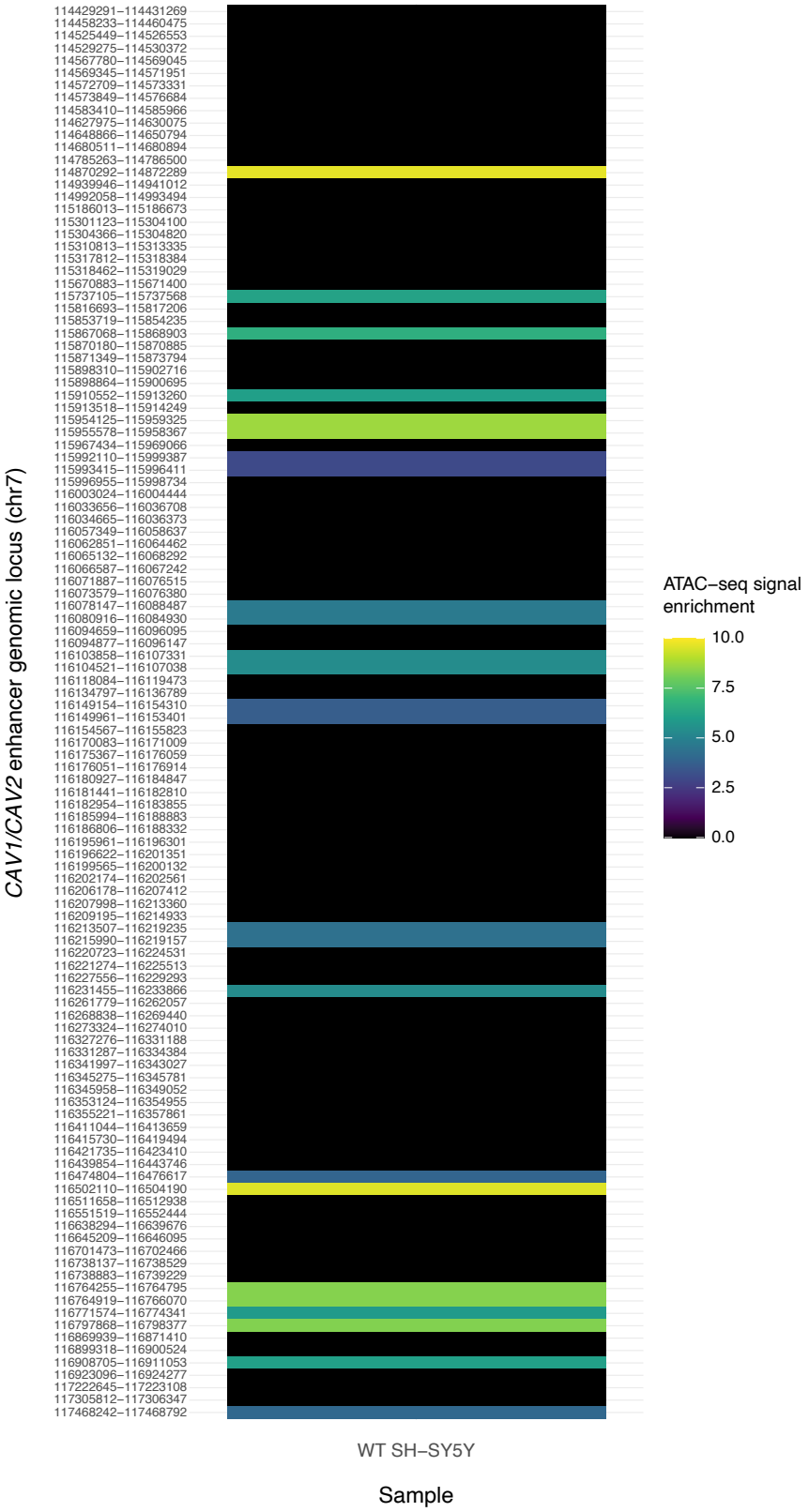



FIGURE 4.4: Chromatin accessibility at candidate *CAV1/CAV2* enhancers in SH-SY5Y cells

ATAC-seq signal (signal per million reads) over candidate *CAV1/CAV2* enhancers in WT SH-SY5Y cells.

### 4.2.2 Chromatin accessibility of enhancers in iPSC-derived motor neurons

To explore how the chromatin state of enhancers may differ between healthy and disease states, we have utilised published ATAC-seq datasets from iPSC-derived motor neurons of healthy controls and ALS patients to identify *CAV1/CAV2* enhancers in an open chromatin state in healthy versus disease states ((Zhang et al., 2022), downloaded from the ENCODE portal with the identifiers ENCSR704VZY, ENCSR709QRD, ENCSR913OWV, ENCSR516YAD, ENCSR634WYX and ENCSR065CER (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020))(Figure 4.5a). The oldest healthy control participant (PGP) exhibited the most *CAV1/CAV2* enhancers in open chromatin, with 48 out of the total 114 enhancers in open chromatin, as determined by the presence of ATAC-seq peaks in these enhancers.

Utilising chromatin accessibility, as determined by ATAC-seq, as a proxy for enhancer activation, we determined which enhancers were active in all healthy control motor neuron samples (Figure 4.5b), and in all ALS-derived motor neuron samples (Figure 4.5c). Further, of the 27 enhancers that were commonly active in healthy control motor neuron samples and the 27 enhancers that were commonly active in ALS-derived motor neuron samples, 22 enhancers were active across all samples, regardless of disease state (Figure 4.5d). Therefore, for both control and ALS samples, 5 enhancers were uniquely active for each condition.

To investigate the correlation between chromatin accessibility at candidate *CAV1/CAV2* enhancers and *CAV1* and *CAV2* mRNA expression, we utilised gene count estimates from RNA-seq in iPSC-derived motor neurons Zhang et al. (2022), downloaded from the ENCODE portal with the identifiers ENCSR215KUO, ENCSR350OKU, ENCSR319GRO, ENCSR373XVN, ENCSR967DSJ and ENCSR575MQY (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020). In congruence with chromatin accessibility, PGP exhibited a markedly higher expression of *CAV1* than all other samples, regardless of disease state (Figure 4.6a). Of note, all samples except PGP demonstrate similar patterns of chromatin accessibility and *CAV1* expression. There appeared to be less of a trend when comparing chromatin accessibility with *CAV2* expression, likely reflecting the lack of enhancers controlling *CAV2* expression alone in our catalogue of candidate enhancers (Figure 4.6b).

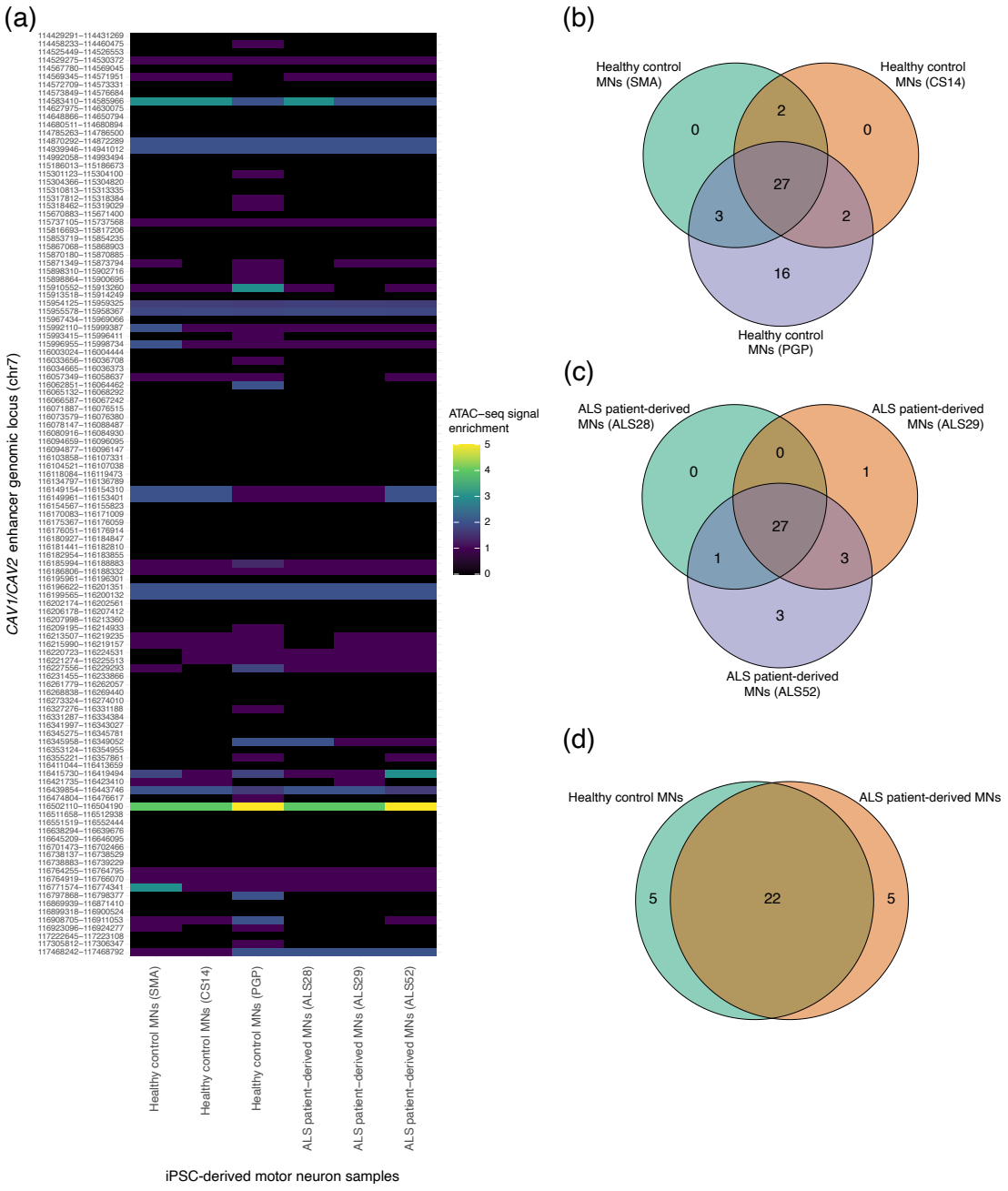



FIGURE 4.5: Chromatin accessibility at candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons

(a) ATAC-seq signal (signal per million reads) over candidate *CAV1/CAV2* enhancers in healthy control- and ALS patient-derived *in vitro* differentiated motor neurons (Zhang et al., 2022) (b) Candidate enhancers in regions of accessible chromatin, as determined by ATAC-seq peaks, in healthy control motor neurons, demonstrating the enhancers in open chromatin common between samples. (c) Candidate enhancers in regions of accessible chromatin, as determined by ATAC-seq peaks, in ALS patient-derived motor neurons, demonstrating the enhancers in open chromatin common between samples. (d) Among the enhancers in open chromatin common to all control motor neurons and all ALS-derived motor neurons, demonstrating the number of enhancers in an open chromatin state in all samples, regardless of disease state.

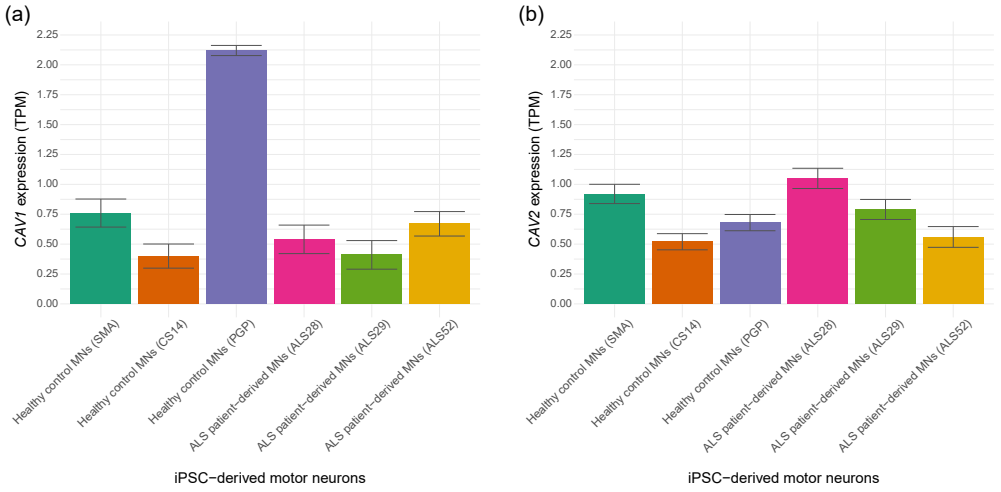



FIGURE 4.6: *CAV1* and *CAV2* gene expression in iPSC-derived motor neurons (a) *CAV1* mRNA expression (transcripts per million) in iPSC-derived motor neurons (Zhang et al., 2022). (b) *CAV2* mRNA expression (transcripts per million) in iPSC-derived motor neurons (Zhang et al., 2022).

### 4.2.3 Comparatative analysis of chromatin accessibility at *CAV1/CAV2* enhancers in SH-SY5Y cells and iPSC-derived motor neurons

To explore how chromatin accessibility — and consequently enhancer activity — differs between our SH-SY5Y model cell line and iPSC-derived *in vitro* differentiated motor neurons, we compared ATAC-seq signal enrichment at *CAV1/CAV2* enhancers across samples (Figure 4.7a). SH-SY5Y cells had fewer enhancers in open chromatin than that of either healthy control or ALS patient-derived motor neurons, with 25 enhancers in an open chromatin state (Figure 4.7b). We further performed pairwise comparisons of enhancers in an open chromatin state between each motor neuron sample and SH-SY5Y cells (Figure 4.7c). Since the PGP motor neuron sample exhibited the highest number of enhancers in an open chromatin state, as expected, SH-SY5Y cells shared the most enhancers in open chromatin with the PGP sample, with 19 enhancers in common. Of the 25 enhancers that are in active chromatin in SH-SY5Y cells, the 27 enhancers that are commonly active in healthy control motor neurons and the 27 enhancers that are commonly active in ALS-derived motor neurons, 12 were in an active chromatin state among all samples (Figure 4.7d). There were 4 enhancers in an active chromatin state that were common between SH-SY5Y cells and healthy control motor neurons, that were not identified as active in disease state motor neurons. However, while there were 5 uniquely active enhancers in disease state patient-derived lines, none of these were found to be active in SH-SY5Y cells. Therefore, no enhancers were identified in common between SH-SY5Y cells and ALS patient-derived motor neurons that were not also active in healthy control motor neurons.

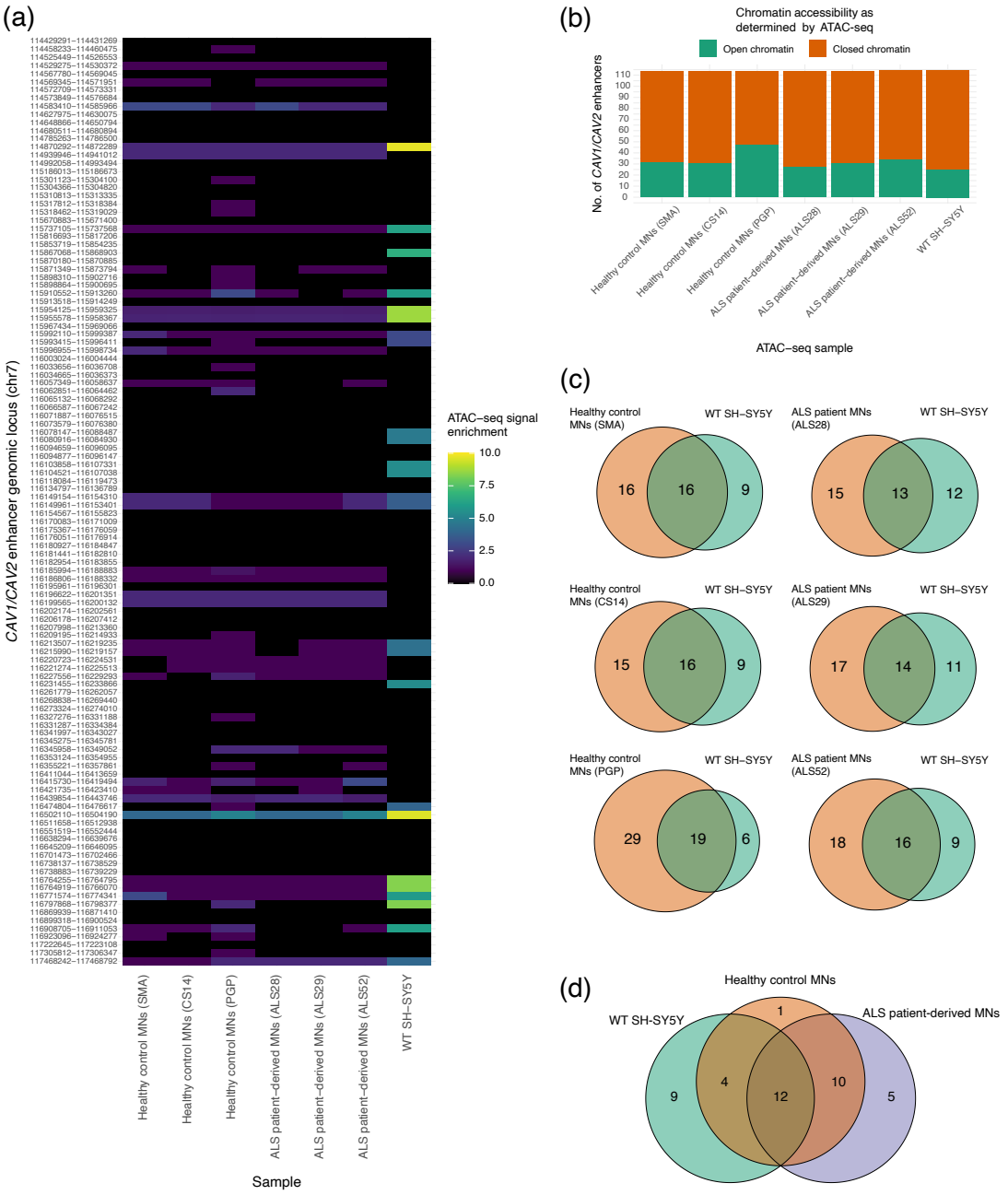



FIGURE 4.7: Comparison of chromatin accessibility at candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons and SH-SY5Y cells

(a) ATAC-seq signal (signal per million reads) over candidate *CAV1/CAV2* enhancers in iPSC-derived *in vitro* differentiated motor neurons and WT SH-SY5Y cells. (b) Chromatin accessibility of *CAV1/CAV2* enhancers in iPSC-derived motor neurons and WT SH-SY5Y cells as determined by called ATAC-seq peaks. (c) Pairwise comparisons of enhancers in an open chromatin state between each iPSC-derived motor neuron sample versus WT SH-SY5Y cells. (d) Among the enhancers in open chromatin common to all control motor neurons and all ALS-derived motor neurons, the number of these enhancers in an open chromatin state in WT SH-SY5Y cells.

### 4.2.4 Chromatin accessibility at *CAV1/CAV2* enhancers containing ALS-associated SNPs

We investigated the relevance of open chromatin at enhancers in disease states by examining the relationship between chromatin accessibility at enhancers and the propensity for enhancers to gain ALS-associated SNPs. Our hypothesis posited that disease-associated SNPs would localise to active enhancers in physiologically relevant cell types, thus potentially inducing pathological changes in gene expression and contributing to the disease phenotype. Consequently, we expected to identify more SNPs in enhancers within accessible chromatin compared to those in closed chromatin.

Therefore, we compared the number of ALS-associated SNPs in *CAV1/CAV2* enhancers between open and closed chromatin states in SH-SY5Y cells, as determined by ATAC-seq peaks (Figure 4.8a). As expected, significantly more ALS-associated SNPs were identified in enhancers within active chromatin states in SH-SY5Y cells compared to closed chromatin states. On average, enhancers in an active chromatin state contained 4 SNPs, to the closest integer, whereas those in a closed chromatin state contained an average of 2 SNPs, to the closest integer.

We further investigated whether similar trends were present in iPSC-derived motor neurons. We again compared the number of ALS-associated SNPs in *CAV1/CAV2* enhancers between open and closed chromatin states (Figure 4.8b). Consistent with our findings in SH-SY5Y cells, significantly more ALS-associated SNPs were identified in enhancers within active chromatin states compared to inactive ones. Across all samples, enhancers in an open chromatin state contained an average of 4 ALS-associated SNPs, to the closest integer, while those in a closed chromatin state contained an average of 1 SNP, to the closest integer.

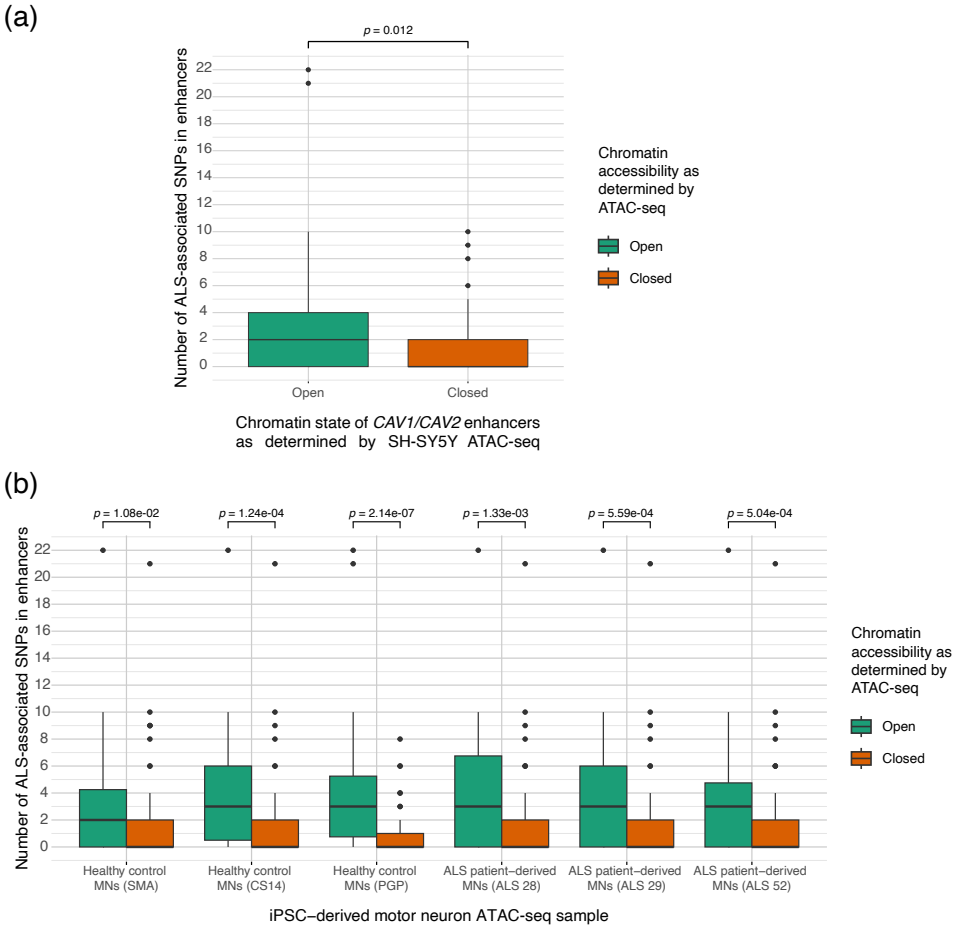



FIGURE 4.8: Number of ALS-associated SNPs in candidate *CAV1/CAV2* enhancers in open and closed chromatin states

(a) Comparison of the number ALS-associated SNPs in enhancers in an open versus closed chromatin state, as determined by ATAC-seq in WT SH-SY5Y cells. Points represent outliers identified based on the interquartile range upper bound. $p$-value determined by Wilcoxon Rank Sum. (b) Comparison of the number ALS-associated SNPs in enhancers in an open versus closed chromatin state, as determined by ATAC-seq in iPSC-derived motor neurons. Points represent outliers identified based on the interquartile range upper bound. $p$-values determined by Wilcoxon Rank Sum.

Taken together, we have characterised how chromatin accessibility varies across distinct neuronal-like cell types, including healthy and disease states. Specifically, we observed increased levels of *CAV1* expression and chromatin accessibility at candidate *CAV1/CAV2* enhancers in the PGP healthy control motor neuron sample, compared to the remaining healthy control-derived samples and all ALS patient-derived samples. While this may suggest a potential link between chromatin accessibility at these enhancers and *CAV1* expression, further analyses are needed to establish a direct relationship between chromatin accessibility at candidate *CAV1/CAV2* enhancers and *CAV1* expression in patient-derived motor neurons. Moreover, in the context of disease, we

identified an enrichment of ALS-associated SNPs within enhancers in an open chromatin state, linking chromatin accessibility at these enhancers to mutation propensity.

## 4.3 Mapping enhancer-associated characteristics

With the functional potential of an enhancer largely influenced by the local chromatin environment, including the histone modifications decorating flanking histones (reviewed in Calo and Wysocka, 2013, Spitz and Furlong, 2012), identifying genomic regions enriched for such PTMs can be informative for not only the identification but the characterisation of enhancers. Enhancers, regardless of activity, can be identified by mapping regions enriched for H3K4me1, with active enhancers further enriched for H3K27ac, the predominant histone modification associated with transcriptional activity. Conversely, promoters are typically identified by mapping regions enriched from H3K4me3, which akin to enhancers, are further enriched for H3K27ac upon activation (Creyghton et al., 2010, Heintzman et al., 2007, Rada-Iglesias et al., 2011).

### 4.3.1 Enhancer-associated characteristics in SH-SY5Y cells

To assess enhancer activity in our SH-SY5Y model cell line, we mapped regions enriched for H3K27ac, a histone modification commonly associated with active enhancers. By performing CUT&RUN, we identified regions enriched for H3K27ac, highlighting enhancers in an active chromatin state. This data provides insight into the regulatory environment surrounding *CAV1/CAV2* candidate enhancers in SH-SY5Y cells, providing a framework for understanding their role in gene regulation.

#### 4.3.1.1 Validation of CUT&RUN

To capture active enhancers, we performed H3K27ac CUT&RUN in WT SH-SY5Y cells. To validate our dataset, we compared peaks called from our H3K27ac CUT&RUN data to that of publicly available H3K27ac ChIP-seq in SK-N-SH cells (due to the unavailability of published H3K27ac CUT&RUN data in SH-SY5Y cells) (Zhang et al., 2020a), downloaded from the ENCODE portal with the identifier ENCSR000FCU (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020). As for our ATAC-seq dataset, we applied the relative distance metric to evaluate the spatial correlation between our H3K27ac CUT&RUN peaks and and those from the SK-N-SH ChIP-seq data (Favorov et al., 2012). The distribution of relative distances between the closest peaks in both datasets revealed a distinct shift towards lower values, indicating spatial correlation

between the datasets (Figure 4.9). This spatial correlation underscores the reliability of our CUT&RUN dataset in comparison to previously published SK-N-SH ChIP-seq data.

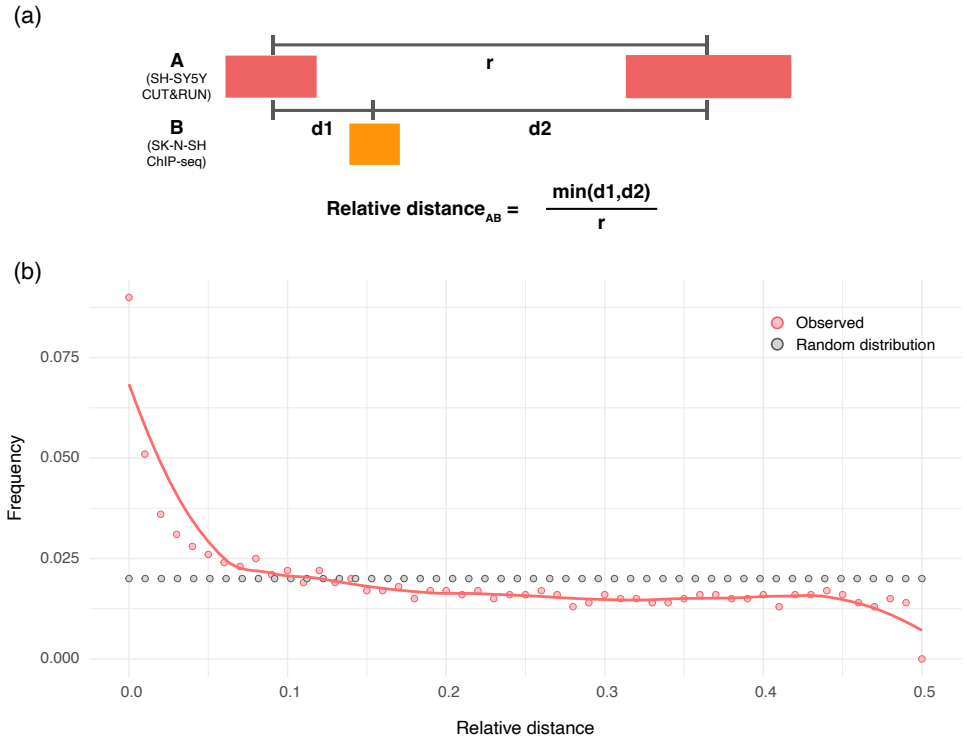



FIGURE 4.9: Relative distance between peaks in H3K27ac CUT&RUN and publicly available ChIP-seq
(a) Schematic describing the use of the relative distance metric to assess spatial correlation between H3K27ac CUT&RUN and H3K27ac ChIP-seq. (b) Distribution of observed relative distances between SH-SY5Y H3K27ac CUT&RUN peaks and SK-N-SH H3K27ac ChIP-seq peaks (Zhang et al., 2020a).

#### 4.3.1.2 Enhancer-associated characteristics of *CAV1/CAV2* enhancers in SH-SY5Y cells

To explore *CAV1/CAV2* enhancer activity in SH-SY5Y cells, we measured H3K27ac CUT&RUN signal enrichment relative to histone H3 control across candidate *CAV1/CAV2* enhancers (Figure 4.10). Considering enhancers with at least a 2-fold higher H3K27ac signal relative to control ($\log_2$-transformed fold enrichment $\geq 1$), we identified 20 active *CAV1/CAV2* enhancers, as determined by H3K27ac enrichment.

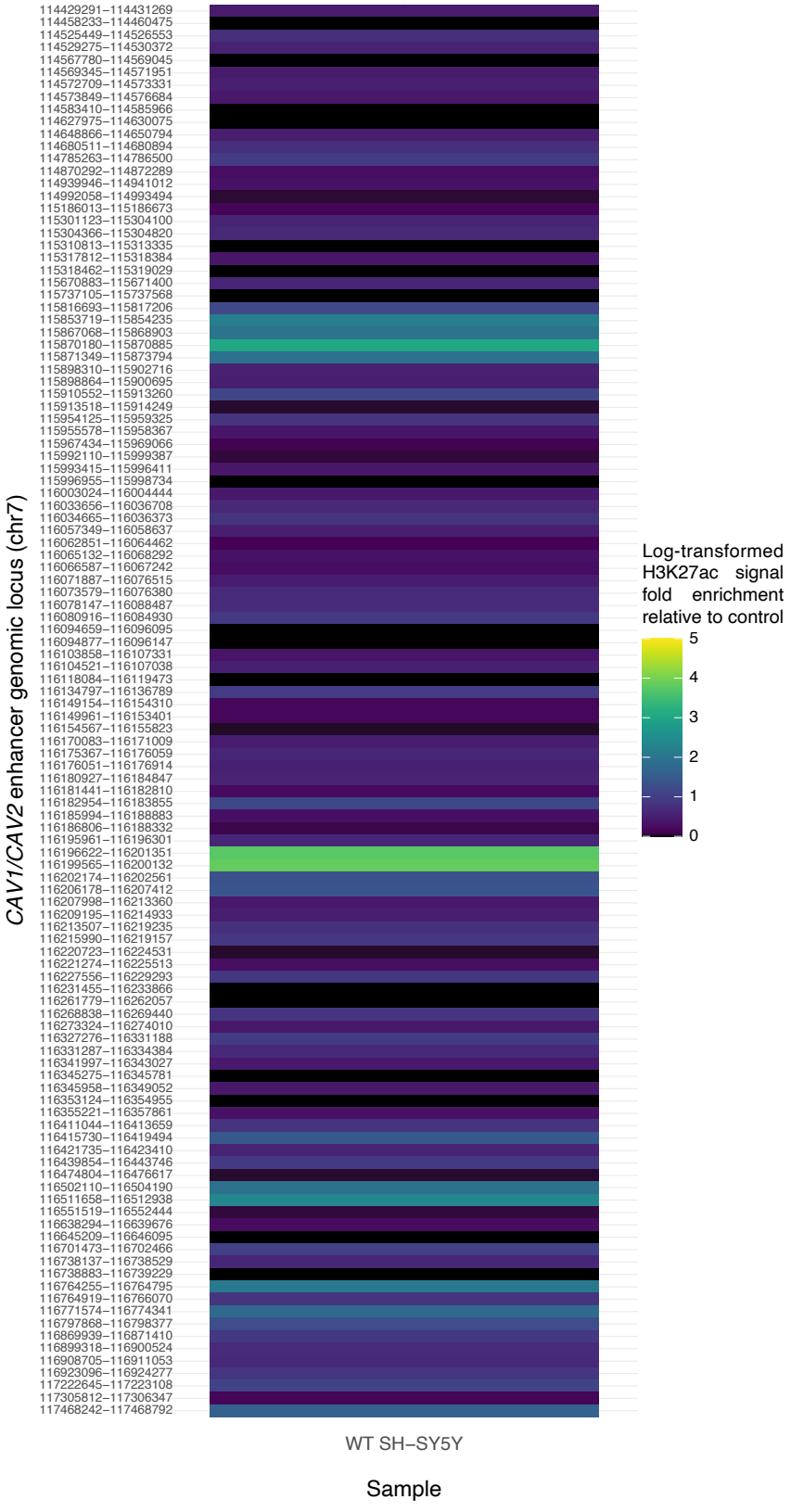



FIGURE 4.10: Enrichment of H3K27ac at candidate *CAV1/CAV2* enhancers in SH-SY5Y cells

(a) Log-transformed H3K27ac H3K27ac signal fold enrichment relative to control for CUT&RUN in WT SH-SY5Y cells over candidate *CAV1/CAV* enhancers

### 4.3.2 Enhancer-associated characteristics in iPSC-derived motor neurons

To explore the distribution of the enhancer-associated histone modifications, H3K4me1 and H3K27ac across the *CAV1/CAV2* enhancer landscape, we have utilised published ChIP-seq datasets from iPSC-derived motor neurons from healthy controls and ALS patients, (Zhang et al., 2022), downloaded from the ENCODE portal with the identifiers ENCSR207VLY, ENCSR778FKK, ENCSR698HPK, ENCSR564EFE, ENCSR682BFG, ENCSR358AOC, ENCSR312HLG, ENCSR489LNU, ENCSR680IWU, ENCSR754DRC, ENCSR540KQC and ENCSR425FUS (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020), to visualise the pattern of histone modification enrichment over *CAV1/CAV2* candidate enhancers

We identified of which candidate enhancers there was enrichment of H3K4me1, positing that these genomic regions may have enhancer functionality in iPSC-derived motor neurons (Figure 4.11a). Consistent with chromatin accessibility, as determined by ATAC-seq, the oldest healthy control, PGP, demonstrated the largest number of candidate enhancers enriched for H3K4me1, with 40 of the 114 total enhancers enriched for H3K4me1 ChIP-seq signal, while the remaining patient sample showed similar numbers of enhancers enriched for H3K4me1 (Figure 4.11b).

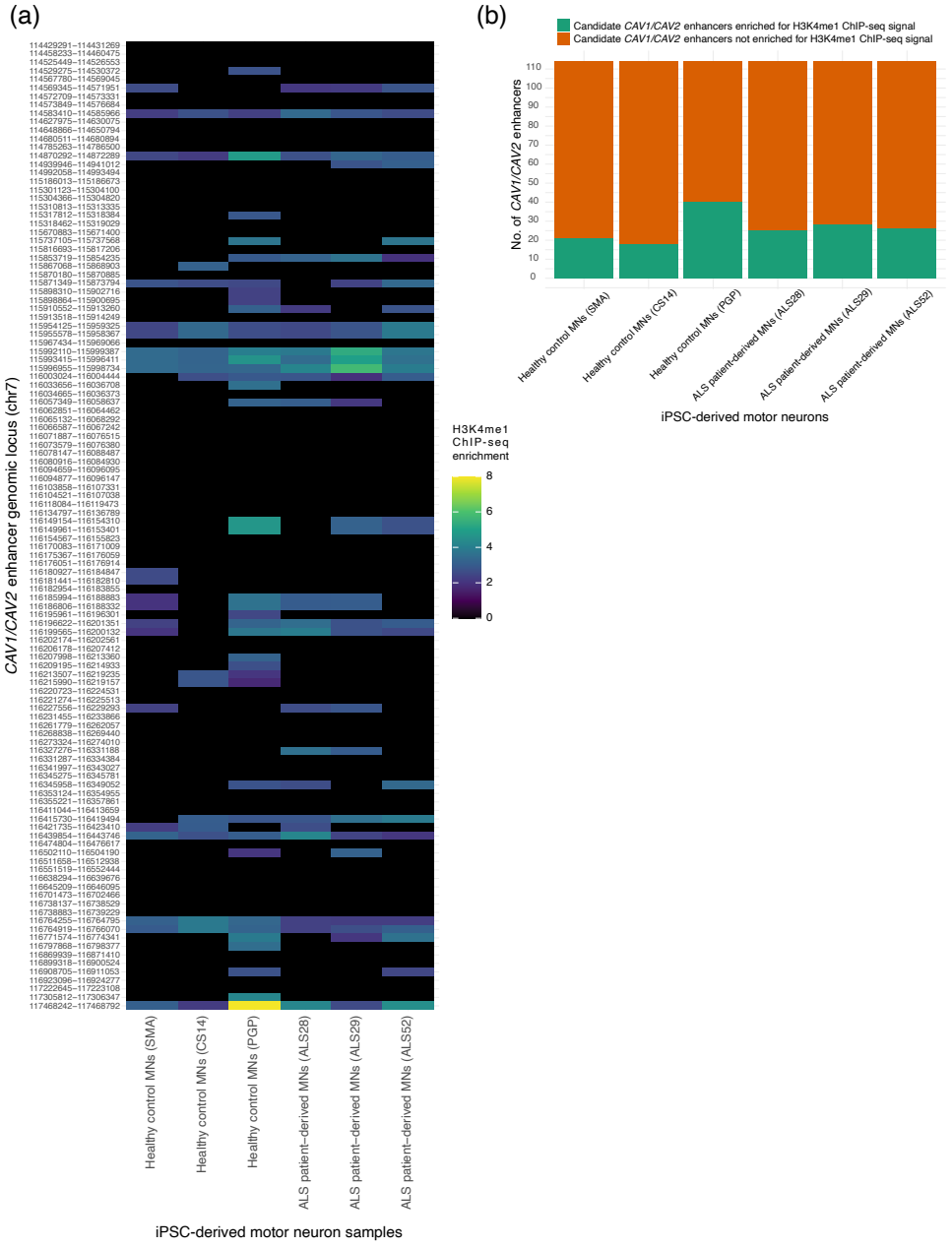



FIGURE 4.11: Enrichment of H3K4me1 at candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons
(a) H3K4me1 ChIP-seq signal (signal per million reads) over candidate *CAV1/CAV2* enhancers in iPSC-derived *in vitro* differentiated motor neurons (Zhang et al., 2022). (b) Number of total candidate *CAV1/CAV2* enhancers enriched for H3K4me1 ChIP-seq signal in healthy control and ALS patient derived motor neurons (Zhang et al., 2022).

Further, we identified of which candidate enhancers there was enrichment of H3K27ac, potentially indicative of active enhancers (Figure 4.12a). Again, PGP was identified as the sample with the most enhancers enriched for H3K27ac, with 32 candidate enhancers enriched for H3K27ac. Contrary to the trend observed for H3K4me1 enrichment, there were notably fewer enhancers enriched for H3K27ac in the youngest healthy control

sample, SMA, and the ALS patient sample, ALS28, with 3 and 7 enhancers enriched for H3K27ac, respectively (Figure 4.12b).

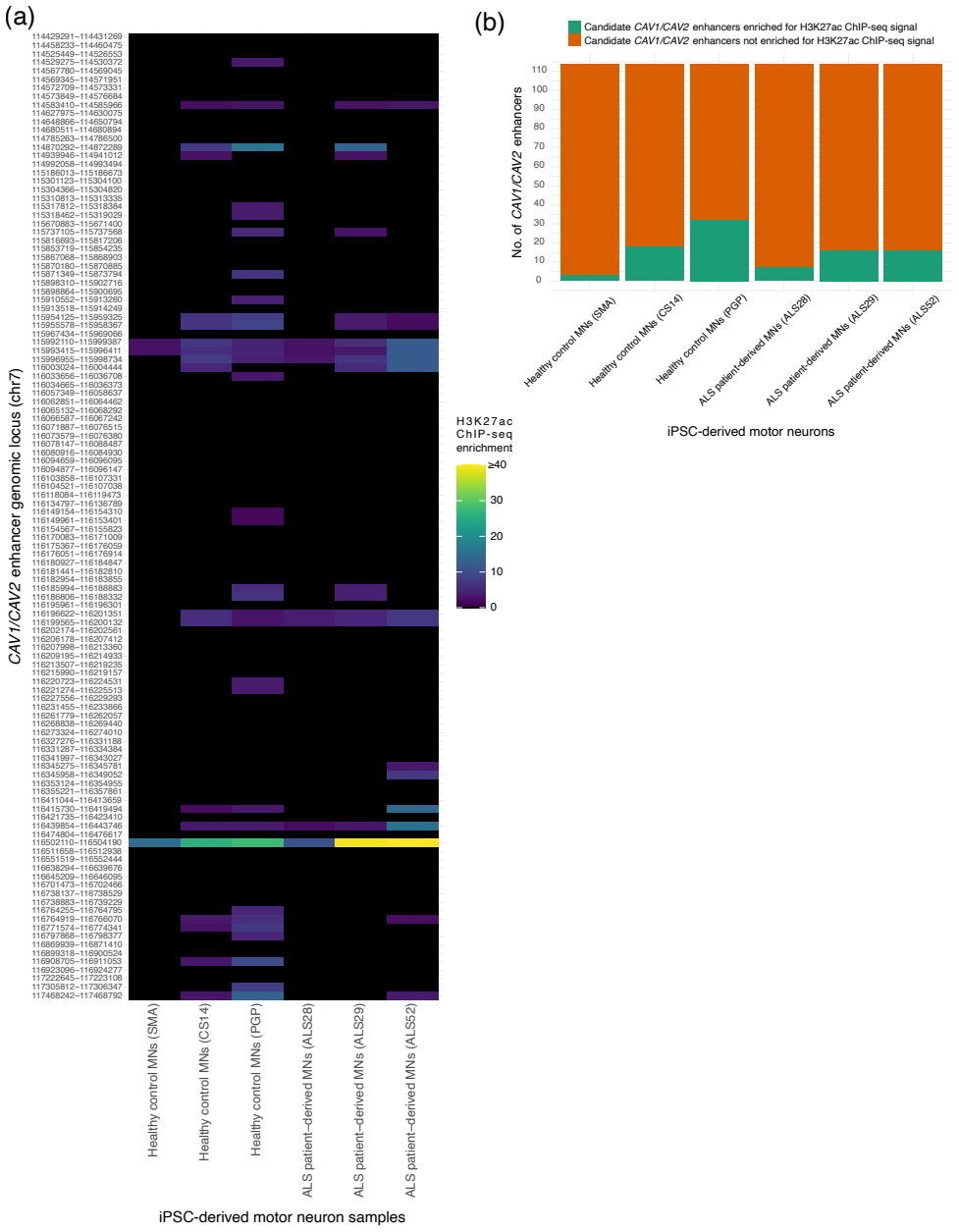



FIGURE 4.12: Enrichment of H3K27ac at candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons
(a) H3K27ac ChIP-seq signal (signal per million reads) over candidate *CAV1/CAV2* enhancers in iPSC-derived *in vitro* differentiated motor neurons (Zhang et al., 2022) (b) Number of total candidate *CAV1/CAV2* enhancers enriched for H3K27ac ChIP-seq signal in healthy control and ALS patient derived motor neurons (Zhang et al., 2022).

Therefore, while we observed a somewhat uniform enrichment for H3K4me1 at *CAV1/CAV2* enhancers, with the exception of the healthy control-derived sample, PGP, there were

distinct differences in H3K27ac enrichment patterns across iPSC-derived motor neuron samples. This highlights the potential variability in enhancer activity across these samples.

#### 4.3.2.1 Characterisation of *CAV1/CAV2* enhancers in iPSC-derived motor neurons

Moreover, to further characterise the activity of candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons, we identified at which enhancers there was a co-occurrence of H3K4me1 and H3K27ac ChIP-seq peaks (Figure 4.13). Enhancers enriched for H3K4me1 alone may be considered inactive or primed enhancers, whereas enhancers enriched for both H3K4me1 and H3K27ac may be considered active enhancers, with H3K27ac indicative of transcriptional activity. As expected based on the number of enhancers independently enriched for H3K4me1 and H3K27ac, PGP had the largest number of active enhancers, as determined by co-occurrence of H3K4me1 and H3K27ac. Also, as expected, based on the limited number of enhancers enriched for H3K27ac in SMA and ALS28, these samples had the fewest active enhancers, with 2 and 6 enhancers, respectively. Hence, we identified the highest number of enhancers enriched for H3K4me1 alone in these samples, potentially suggesting inactive or primed enhancers, with both samples having 19 enhancers enriched for H3K4me1 alone.

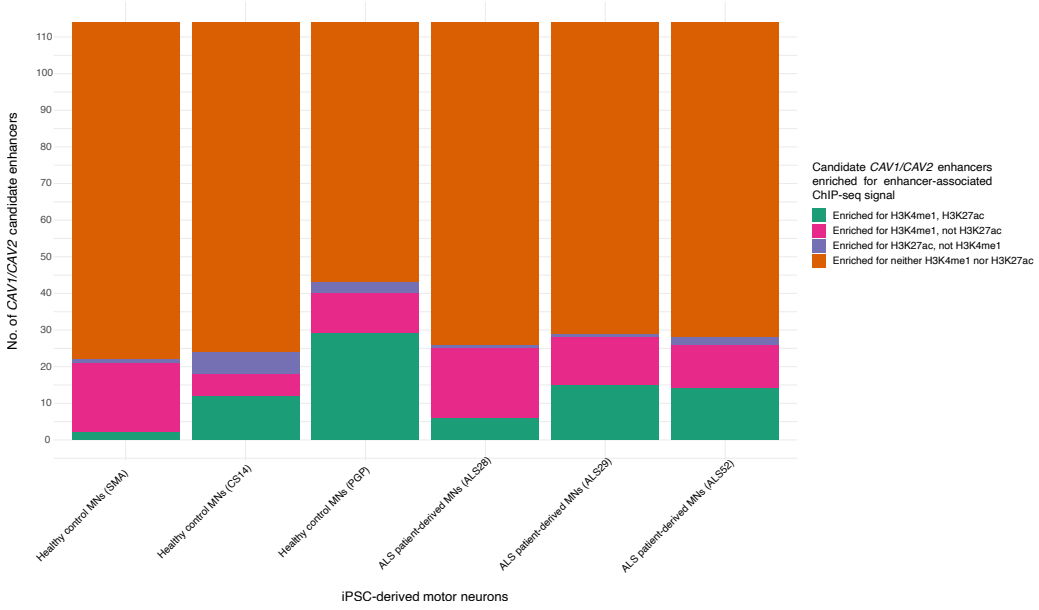



FIGURE 4.13: Characterisation of candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons

Quantification of intersections of ChIP-seq peaks for H3K4me1 and H3K27ac at candidate *CAV1/CAV2* enhancers in iPSC-derived motor neurons (Zhang et al., 2022).

#### 4.3.2.2 Enhancer characteristics in Amyotrophic Lateral Sclerosis

To explore the role of *CAV1/CAV2* enhancers in ALS, we firstly investigated if any of our candidate enhancers were active across all of the ALS patient-derived motor neuron samples. Utilising co-enrichment of H3K4me1 and H3K27ac ChIP-seq signal, we identified 6 enhancers that were active among the ALS samples (Figure 4.14a). Further, we investigated if these enhancers were similarly active in the healthy control samples, using the same methods. Upon quantification, all 6 of these enhancers were active in the PGP sample, 4 were active in the CS14 sample and only 2 were active in the SMA sample (Figure 4.14b). H3K4me1 ChIP-seq signal was observed over all of these enhancers for all samples, except CS14, which lacked signal over 2 enhancers (Figure 4.14c). H3K27ac ChIP-seq signal was observed over all of these enhancers for all samples, except SMA, which lacked signal over 4 enhancers (Figure 4.14d).

Uniquely, the enhancers at chr7:116,196,622-116,201,351 and chr7:116,199,565-116,200,132, considered to be tissue-agnostic and CNS-specific, respectively, hereafter collectively referred to as the *CAV1/CAV2* proximal enhancer, appeared to be in distinct epigenetic states across the healthy control samples (Figure 4.15). Whereby, SMA demonstrated enrichment for H3K4me1 alone, CS14 demonstrated enrichment for H3K27ac alone, and PGP demonstrated co-enrichment for both H3K4me1 and H3K27ac at these enhancers.

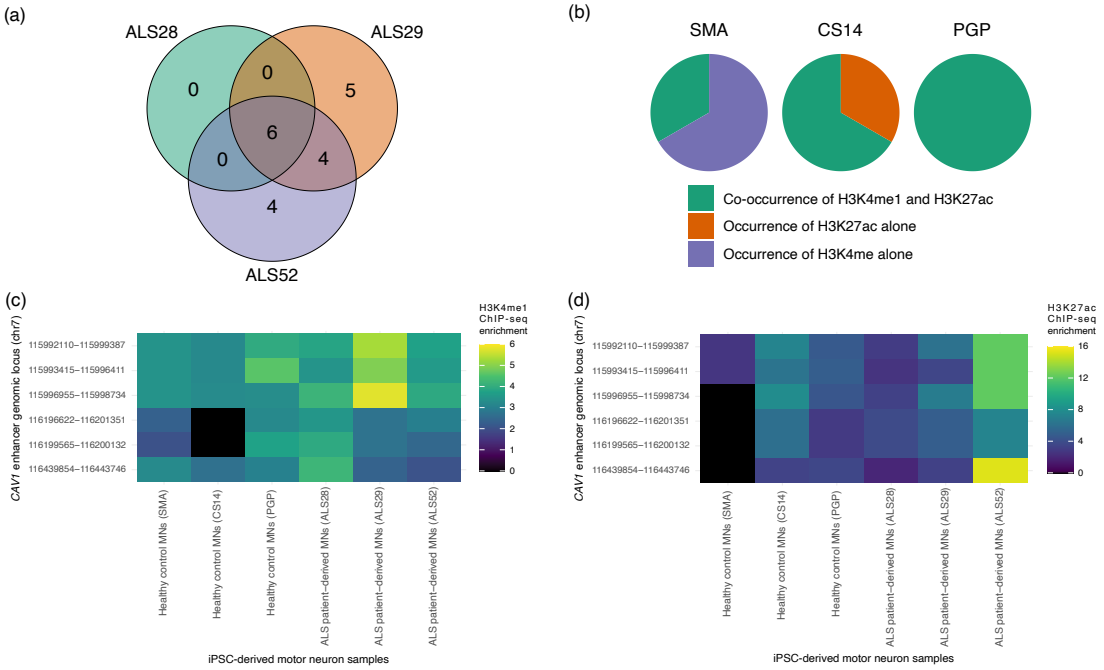



FIGURE 4.14: Characterisation of active *CAV1/CAV2* enhancers in Amyotrophic Lateral Sclerosis

(a) Identification of active enhancers common to all ALS patient-derived motor neuron samples, through intersection of enhancers enriched for both H3K4me1 and H3K27ac in each sample (Zhang et al., 2022) (b) Epigenetic characterisation of the 6 enhancers identified as active in ALS patient samples, in the healthy control-derived motor neurons (Zhang et al., 2022). (c) H3K4me1 ChIP-seq signal (signal per million reads) in iPSC-derived *in vitro* differentiated motor neurons over the enhancers identified as commonly active in ALS patients (Zhang et al., 2022). (d) H3K27ac ChIP-seq signal (signal per million reads) in iPSC-derived *in vitro* differentiated motor neurons over the enhancers identified as commonly active in ALS patients (Zhang et al., 2022).

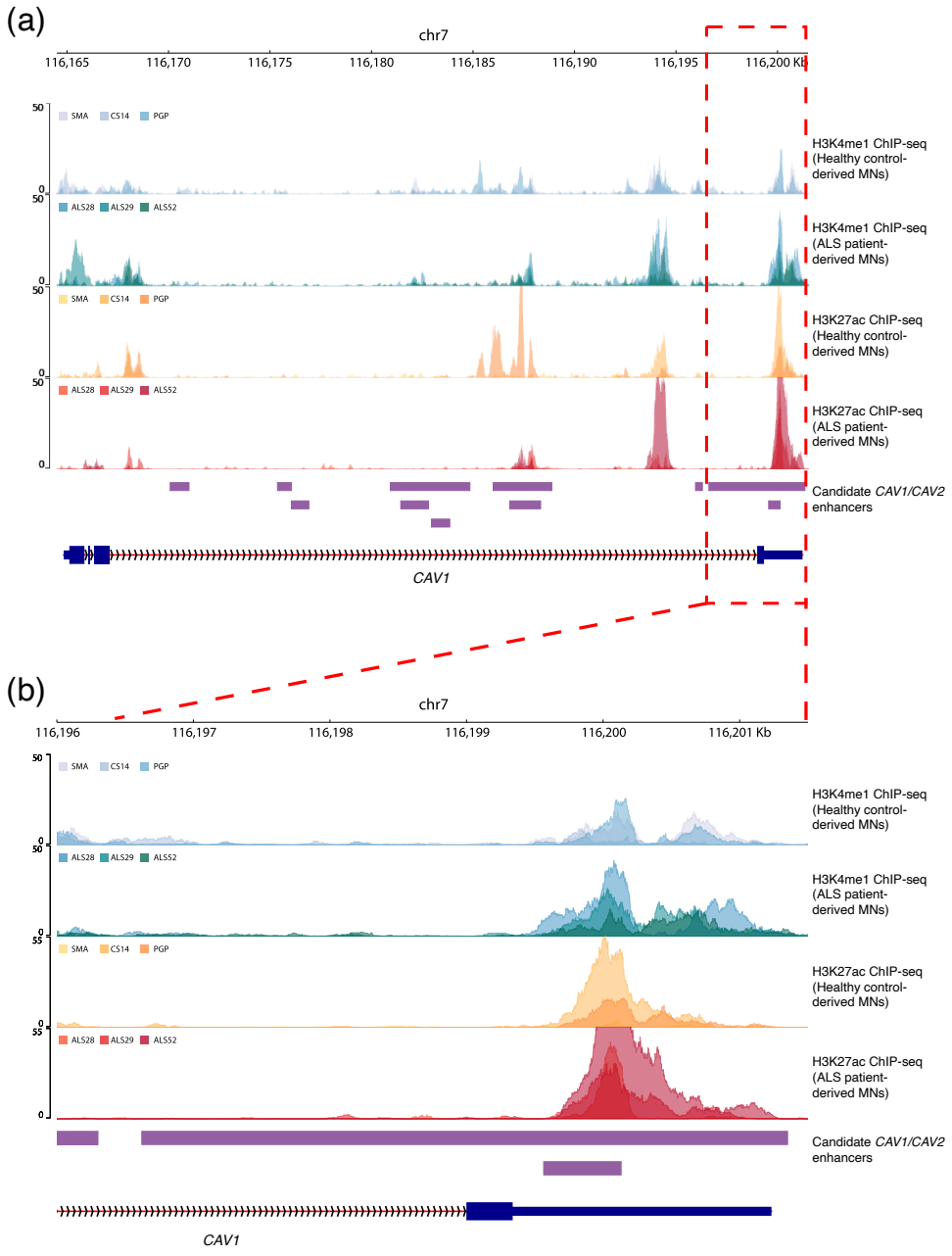



FIGURE 4.15: Characteristic enhancer-associated histone modifications at the *CAV1/CAV2* proximal enhancer in iPSC-derived motor neurons
(a) Trackline at chr7:116,164,500-116,201,500 (hg19) demonstrating H3K4me1 ChIP-seq in iPSC-derived motor neurons, H3K27ac ChIP-seq in iPSC-derived motor neurons (Zhang et al., 2022) and candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019). (b) Magnified view of trackline at chr7:116,196,000-116,201,500 (hg19) demonstrating H3K4me1 ChIP-seq in iPSC-derived motor neurons, H3K27ac ChIP-seq in iPSC-derived motor neurons (Zhang et al., 2022) and candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019). Position of the *CAV1/CAV2* proximal enhancer (chr7:116,196,622-116,201,351) is highlighted by red dashed box.

#### 4.3.2.3 Enhancer characteristics at *CAV1/CAV2* enhancers containing ALS-associated SNPs

While only two candidate *CAV1/CAV2* enhancers demonstrated co-occurrence of H3K27ac and H3K4me1 ChIP-seq signal across all iPSC-derived motor neuron samples, regardless of disease state, these were the enhancers in our catalogue that contained the highest number of ALS-associated SNPs. The enhancer at chr7:115,992,110-115,999,387, considered to be tissue-agnostic, contained 22 SNPs and the enhancer at chr7:115,993,415-115,996,411, considered to be CNS-specific, contained 21 SNPs (Figure 4.16). Hereafter, these enhancers are collectively referred to as the intergenic ALS-associated enhancers. Therefore, we explored how the presence of H3K4me1 and H3K27ac ChIP-seq peaks at enhancers relates to the propensity to gain ALS-associated SNPs. Across all iPSC-derived motor neuron samples, regardless of disease state, enhancers overlapping both H3K4me1 (Figure 4.17a) and H3K27ac (Figure 4.17b) ChIP-seq peaks contained more ALS-associated SNPs than enhancers not enriched for ChIP-seq signal.

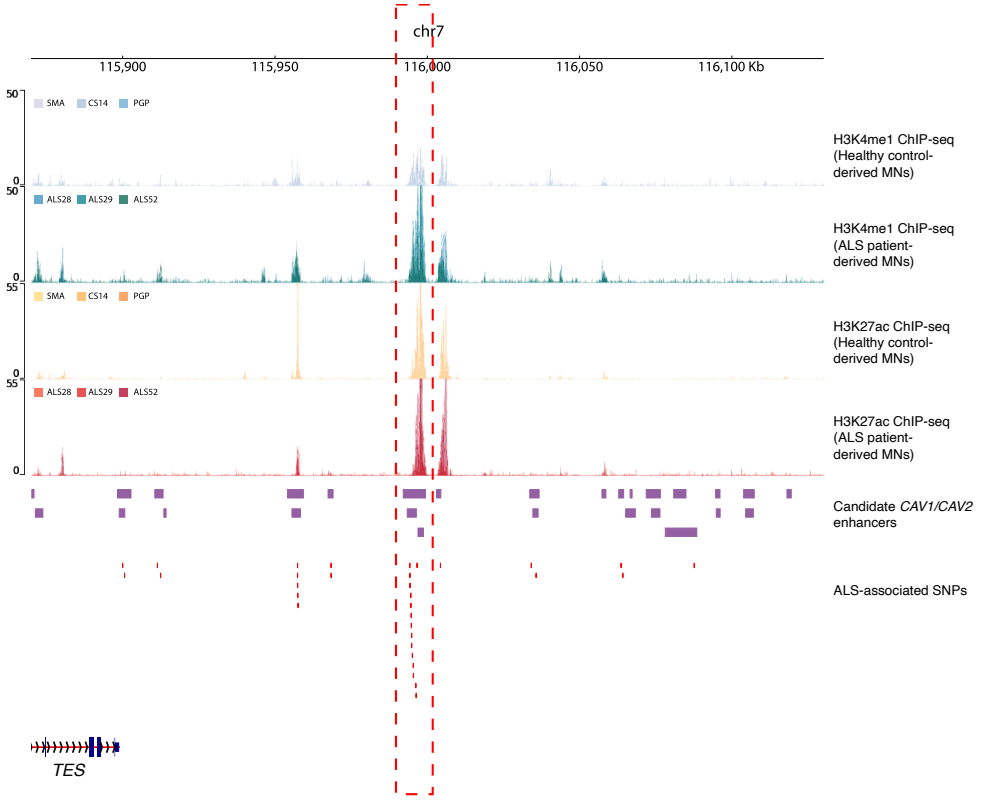



FIGURE 4.16: Enhancer-associated histone modifications are enriched at ALS-associated *CAV1/CAV2* enhancers

Trackline at chr7:115,850,000-115,150,000 (hg19) demonstrating H3K4me1 ChIP-seq in iPSC-derived motor neurons, H3K27ac ChIP-seq in iPSC-derived motor neurons (Zhang et al., 2020a), candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019) and ALS-associated SNPs (Cooper-Knock et al., 2021). Intergenic ALS-associated enhancers highlighted by red dashed box.

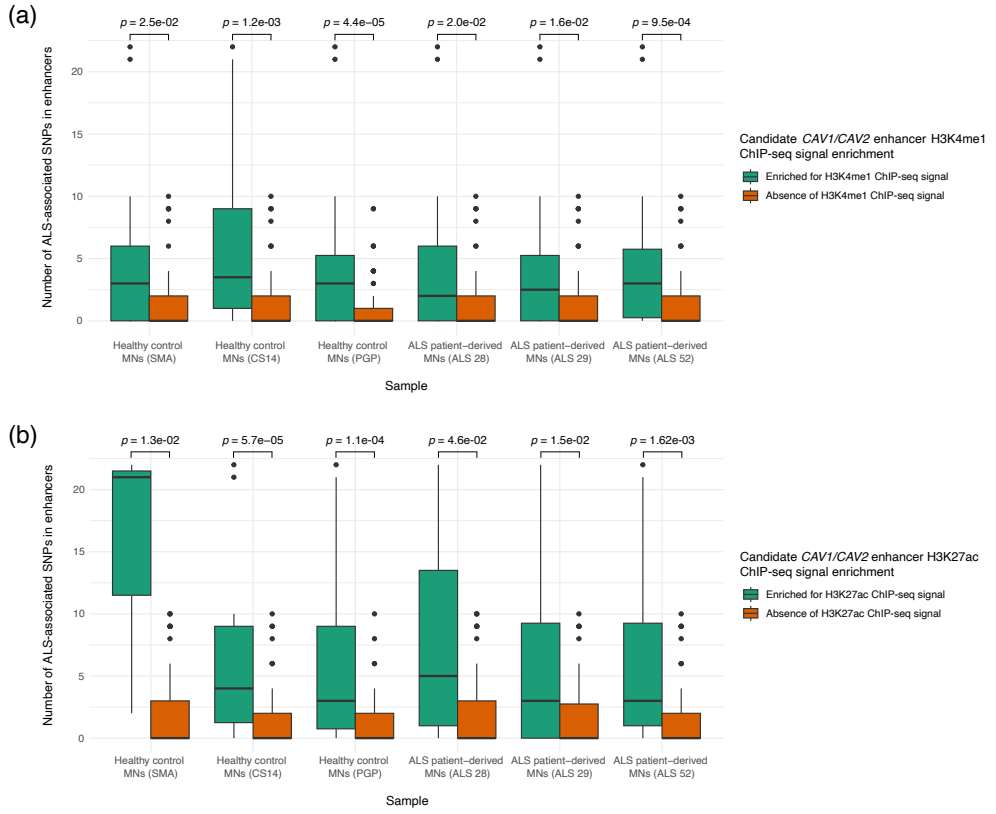



FIGURE 4.17: Number of ALS-associated SNPs in candidate *CAV1/CAV2* enhancers enriched for histone modifications characteristic of enhancers
(a) Comparison of the number of ALS-associated SNPs in enhancers overlapping H3K4me1 ChIP-seq peaks versus those absent of H3K4me1 ChIP-seq signal in iPSC-derived motor neurons (Cooper-Knock et al., 2021, Zhang et al., 2022). Points represent outliers identified based on the interquartile range upper bound. *p*-values determined by Wilcoxon Rank Sum. (b) Comparison of the number of ALS-associated SNPs in enhancers overlapping for H3K27ac ChIP-seq peaks versus those absent of H3K27ac ChIP-seq signal in iPSC-derived motor neurons (Cooper-Knock et al., 2021, Zhang et al., 2022). Points represent outliers identified based on the interquartile range upper bound. *p*-values determined by Wilcoxon Rank Sum.

Taken together, we have established a relationship between the enrichment of characteristic enhancer-associated histone modifications and mutation susceptibility at enhancers. Consistently, candidate enhancers containing the greatest number of ALS-associated SNPs were co-enriched for H3K4me1 and H3K27ac across all iPSC-derived motor neuron samples.

#### 4.3.2.4 Comparison of enhancer characteristics between SH-SY5Y cells and iPSC-derived motor neurons

Having established a positive correlation between H3K27ac deposition and presence of ALS-associated SNPs at *CAV1/CAV2* enhancers in iPSC-derived *in vitro* differentiated

motor neurons, we next aimed to compare enhancer activity between these iPSC-derived motor neurons and our SH-SY5Y model cell line.

To identify parallels in enhancer activity between SH-SY5Y cells and iPSC-derived motor neurons, we compared the H3K27ac signal enrichment in SH-SY5Y cells, to the H3K27ac signal enrichment in iPSC-derived motor neurons, across candidate *CAV1/CAV2* enhancers (Figure 4.18a). As H3K27ac enrichment was measured by CUT&RUN in SH-SY5Y cells and by ChIP-seq in iPSC-derived motor neurons, we utilised $\log_2$-transformed fold enrichment values for both CUT&RUN and ChIP-seq to account for potential differences in signal between the two techniques. Whereby, a $\log_2$-transformed fold enrichment of 0 indicates equal signal abundance between treatment and control. Given that H3K27ac is a deposited only at active enhancer elements, this comparison provides a direct measure of enhancer activity in both iPSC-derived motor neurons and our model cell line.

A fold enrichment cut-off of $\geq$ 2-fold enrichment over control ($\log_2$-transformed fold enrichment of $\geq 1$) is generally accepted as an appropriate threshold for identifying significant ChIP-seq peaks (Zhang et al., 2008). We categorized H3K27ac signal enrichment based on the ratio of treatment to control signal, distinguishing between cases where the treatment signal was less than or equal to the control signal, 2-fold higher, between 2-fold and 4-fold higher, or more than 4-fold higher than the control signal. We identified a greater number of candidate enhancers with at least a 2-fold higher H3K27ac signal relative to control in WT SH-SY5Y cells compared to iPSC-derived motor neurons from patient samples (Figure 4.18b). Specifically, we identified 20 *CAV1/CAV2* enhancers in SH-SY5Y cells with a $\log_2$-transformed H3K27ac fold enrichment $\geq 1$, 6 of which demonstrated a fold enrichment of $> 2$. Similarly, in the oldest healthy control, PGP, we observed 20 *CAV1/CAV2* enhancers with a $\log_2$-transformed H3K27ac ChIP-seq fold enrichment of $\geq 1$, 6 of which also had a fold enrichment of $> 2$. Of the ALS patient-derived samples, sample ALS52 had the most candidate enhancers with with $\log_2$-transformed H3K27ac ChIP-seq fold enrichment $\geq 1$, with 12 enhancers in total, 6 of which had a fold enrichment $> 2$.

Of note, we identified two enhancers with strong H3K27ac enrichment ($\log_2$-transformed fold enrichment of $> 2$) across multiple samples. The *CAV1/CAV2* CNS-specific proximal enhancer (chr7:116,199,565-116,200,132) had a $> 2$ $\log_2$-transformed H3K27ac signal fold enrichment for the iPSC-derived samples CS14, ALS29 and ALS52 and for SH-SY5Y cells, with a $\geq 1$ $\log_2$-transformed H3K27ac signal fold enrichment in iPSC-derived samples PGP and ALS28. Additionally, the intronic enhancer within the *CAPZA2* gene (chr7:116,502,110-116,504,190) had a $> 2$ $\log_2$-transformed H3K27ac signal fold enrichment for the iPSC-derived samples SMA, CS14, PGP, ALS29 and ALS52, with a $\geq 1$

$log_2$-transformed H3K27ac signal fold enrichment in the patient-derived sample ALS28 and in SH-SY5Y cells.

To identify parallel active enhancers in SH-SY5Y cells and iPSC-derived motor neurons, we identified of which candidate *CAV1/CAV2* with a more than 2-fold H3K27ac CUT&RUN signal enrichment in SH-SY5Y cells also had a more than 2-fold H3K27ac ChIP-seq enrichment in iPSC-derived motor neurons, for both healthy control and ALS patient-derived motor neurons (Figure 4.18c). Mirroring our results investigating strong signal enrichment, the intronic enhancer across the *CAPZA2* gene (chr7:116,502,110-116,504,190) was identified as an H3K27ac-enriched enhancer across all samples, and the CNS-specific *CAV1/CAV2* proximal enhancer (chr7:116,199,565-116,200,132) was identified as an H3K27ac-enriched enhancer for all samples, except the youngest healthy control sample, SMA. In addition, the tissue-agnostic *CAV1/CAV2* proximal enhancer (chr7:116,196,622-116,201,351) was identified as H3K27ac-enriched in SH-SY5Y cells, and the patient samples CS14, ALS29 and ALS52, and the intronic enhancer within the *CTTNBP2* (chr7:117,468,242-117,468,792) was identified as H3K27ac-enriched in SH-SY5Y cells and the patient samples, CS14, PGP and ALS52.

By comparing H3K27ac signal enrichment across both SH-SY5Y cells and iPSC-derived motor neurons, we have identified multiple enhancers that consistently exhibited high H3K27ac enrichment, pointing to active regulatory elements shared between these systems. This suggests similarities in the *CAV1/CAV2* enhancer landscape regulating gene expression in both neuronally-relevant model systems.

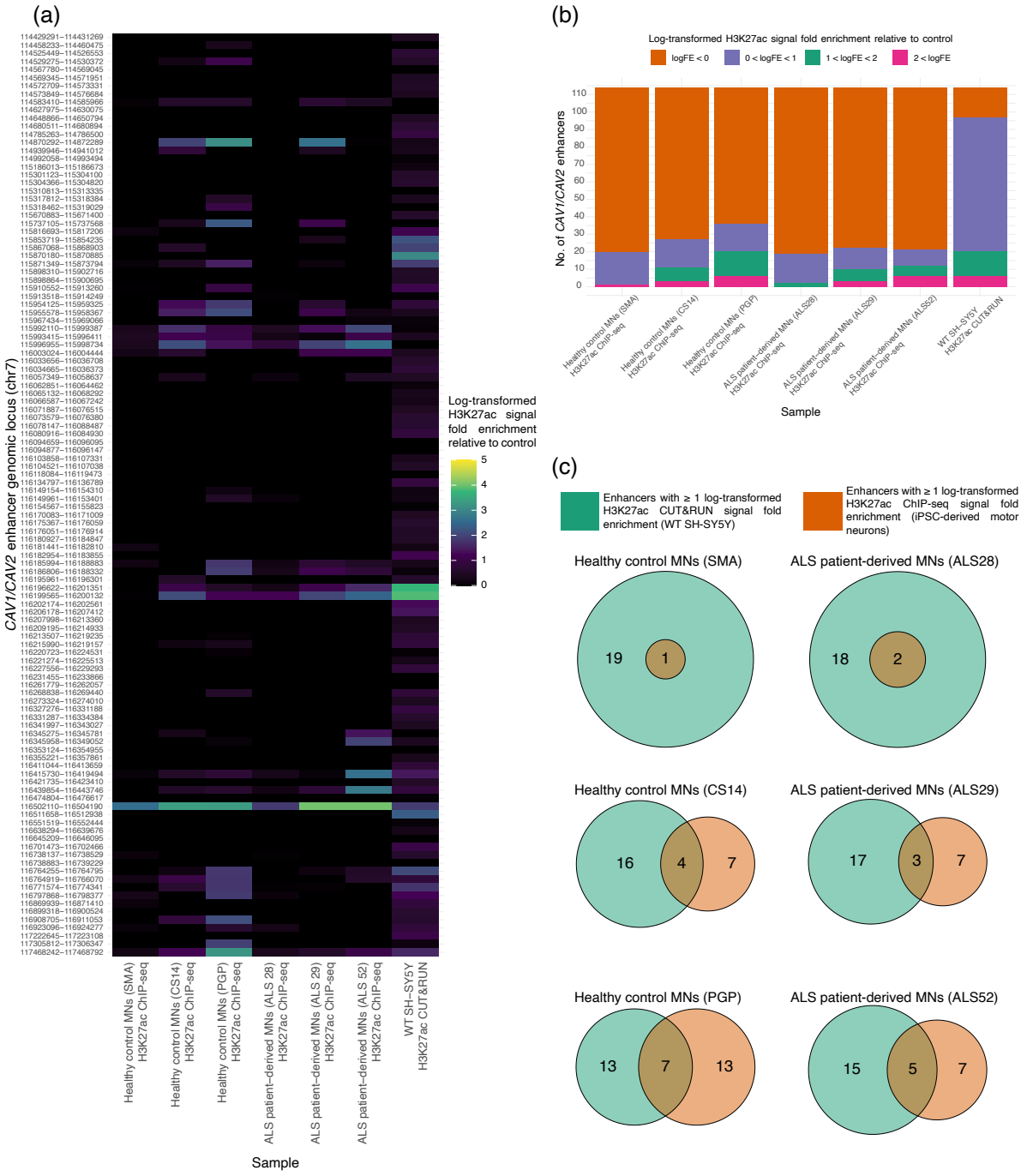



FIGURE 4.18: Comparison of enhancer-associated characteristics at candidate *CAV1/CAV2* enhancers in SH-SY5Y cells and iPSC-derived motor neurons
(a) Log-transformed H3K27ac signal fold enrichment relative to control for ChIP-seq in iPSC-derived motor neurons and CUT&RUN in WT SH-SY5Y cells over candidate *CAV1/CAV2* enhancers in WT SH-SY5Y cells (Zhang et al., 2022). (b) Quantification of total candidate *CAV1/CAV2* enhancers with differential levels of log-transformed H3K27ac signal fold enrichment relative to control, for iPSC-derived motor neurons and WT SH-SY5Y cells (Zhang et al., 2022). (c) Quantification of intersections of candidate *CAV1/CAV2* enhancers with a more than 2-fold H3K27ac CUT&RUN signal enrichment in SH-SY5Y cells and a more than 2-fold H3K27ac ChIP-seq H3K27ac signal enrichment in iPSC-derived motor neurons (Zhang et al., 2022).

## 4.4 Discussion

### 4.4.1 Chromatin accessibility at *CAV1/CAV2* enhancers

A key characteristic of active enhancers is an open chromatin conformation (Heintzman et al., 2007, Song and Crawford, 2010, Thurman et al., 2012). In agreement, the iPSC-derived motor neurons derived from the oldest healthy control demonstrated both the most candidate *CAV1/CAV2* enhancers in open chromatin, as determined by ATAC-seq signal, and also significantly higher *CAV1* expression than all other iPSC-derived motor neuron samples (Figure 4.5, Figure 4.6). This was not a surprising result, given that *CAV1* expression is known to increase with age (Ha et al., 2021). However, our results do not reflect the potentially neuroprotective properties of CAV1, where we expected increased *CAV1* and *CAV2* expression in ALS patients versus healthy controls, as previously observed in post-mortem brain tissues (Adey et al., 2023). In fact, we observed similarities in both chromatin accessibility at enhancers and *CAV1* expression between ALS patients and the younger healthy control individuals.

We observed similarities in the landscape of chromatin accessibility in SH-SY5Y cells compared to iPSC-derived motor neurons, where we observe likeness in ATAC-seq signal enrichment across some of the candidate *CAV1/CAV2* enhancers. Of the 22 enhancers commonly open in all iPSC-derived motor neurons, 12 enhancers are also in open chromatin in WT SH-SY5Y cells (Figure 4.7), suggestive that we may be able to model the activity of these potentially disease-relevant enhancers in our model cell line. Further, we observed a statistically significant increase in the number of ALS-associated SNPs in enhancers in open chromatin when compared to those in closed chromatin, for both iPSC-derived motor neurons and SH-SY5Y cells (Figure 4.8). This observation is in keeping with the suggestion that active enhancers are more likely to obtain ALS-associated SNPs. In addition, disease-associated SNPs have been found to alter chromatin accessibility in regulatory regions, including enhancers (Bysani et al., 2019, Wang et al., 2022), thus we can imagine if a disease-associated SNP in an enhancer typically in regions of open chromatin drives the chromatin into a more closed conformation, this may result in a pathogenic reduction in *CAV1* expression.

### 4.4.2 Enhancer-associated histone modifications at *CAV1/CAV2* enhancers

Open chromatin and enrichment of characteristic histone modifications are used as indicators of enhancer activity. We observed a greater number of candidate enhancers

enriched for both H3K27ac and H3K4me1 in the oldest healthy control (PGP) (Figure 4.13), consistent with the increased *CAV1* expression and chromatin accessibility at candidate *CAV1/CAV2* enhancers in this cell line (Figure 4.5,Figure 4.6).

Interestingly, while the youngest healthy control exhibited a similar number of candidate enhancers enriched for H3K4me1 as seen in other samples (Figure 4.11), we observed much fewer enhancers enriched for H3K27ac (Figure 4.12). This may suggest a higher number of primed enhancers in this sample - enhancers typically marked by H3K4me1 in the absence of H3K27ac, allowing for later activation in response to external stimuli (Creyghton et al., 2010, Zentner et al., 2011). Given that neuronal *CAV1* expression increases with age (Ha et al., 2021), it is unsurprising that *in vitro* differentiated motor neurons derived from a 6 year-old would have fewer active enhancers, as determined by H3K27ac enrichment.

In the pursuit of identifying ALS-specific enhancers, we identified candidate enhancers that demonstrated enrichment for both H3K4me1 and H3K27ac signal in all samples derived from ALS patients (Figure 4.14). However, we did not observe any active enhancers common between all ALS patients that were not active in at least one of the healthy control samples. This is not necessarily a surprise, given that the mutation underlying disease in these patients is the *C9orf72* hexanucleotide repeat expansion, which may not inherently cause any alterations in *CAV1* expression. However, we did observe a significantly lower *CAV1* expression in ALS patient-derived *in vitro* differentiated motor neurons compared to the healthy control individual most similar in age (PGP) (Figure 4.6). This may suggest that alterations in *CAV1* enhancer architecture may not occur uniformly across ALS patients, with 5 and 4 active enhancers unique to the patient samples ALS29 and ALS52, respectively (Figure 4.14a).

Interestingly, the two enhancers that were ubiquitously marked by both H3K27ac and H3K4me1 in all patient samples, regardless of disease state, were the enhancers with the highest number of SNPs (Figure 4.16). This further supports the idea that enhancer activity and strength may be linked to propensity to gain SNPs, as if these enhancers are active regardless of age and disease state, they are likely functional. We observed a supporting trend when identifying the number of ALS-associated SNPs in candidate enhancers enriched for H3K27ac (Figure 4.17b). Whereby, active enhancers, that is those enriched for H3K27ac ChIP-seq peaks, contained significantly more ALS-associated SNPs than enhancers not marked by H3K27ac.

Similarly to the trend observed in all candidate *CAV1/CAV2* enhancers, 4 of the enhancers marked by both H3K27ac and H3K4me1 across all ALS patient samples

lacked enrichment for H3K4me1 in the youngest control sample (SMA). Further, we observed an interesting trend at the *CAV1/CAV2* proximal enhancer(s) (chr7:116,196,622-116,201,351 and chr7:116,199,565-116,200,132), while dually enriched for H3K4me1 and H3K27ac in ALS patients, they exhibited distinct epigenetic states in control samples (Figure 4.14c,d). The oldest control (PGP) exhibited co-enrichment for both H3K4me1 and H3K27, indicative of active enhancers, while the youngest control (SMA) exhibited enrichment for H3K4me1 in the absence of H3K27ac, indicative of primed enhancers. Uniquely, iPSC-derived motor neurons from the healthy control (CS14) demonstrated enrichment for H3K27ac in the absence of H3K4me1 at these enhancers. This may suggest high levels of transcriptional activation in the absence of inherent enhancer activity in these regions. Since these enhancers are located across the 3' UTR of the *CAV1* gene, the transcriptional activity associated with H3K27ac here may signify transcriptional readthrough along the length of the 3' UTR (Caldas et al., 2024), generating the long 3' UTR isoforms associated with neuronal cells (Guvenek and Tian, 2018, Miura et al., 2013, Zhang et al., 2005).

### 4.4.3 Comparative analysis of *CAV1/CAV2* enhancer activity in neuronal-like model systems

To explore similarities in candidate *CAV1/CAV2* enhancer activity between iPSC-derived motor neurons and our model cell line, SH-SY5Y cells, we utilised H3K27ac signal fold enrichment to allow for differences in signal between CUT&RUN and ChIP-seq (Figure 4.18). When comparing the usage of fold enrichment to the presence or absence of peaks, with their associated signal, as used in (Figure 4.11, Figure 4.12), fold enrichment appeared to be a more stringent approach, identifying fewer active enhancers in iPSC-derived motor neurons. Minimising the number of called active enhancers using this approach may explain why we identified so few commonly active enhancers across all samples than we did when comparing chromatin accessibility. To highlight this, the 1 enhancer that we identified significant H3K27ac fold enrichment for in all samples is the intronic enhancer across the *CAPZA2* gene (chr7:116,502,110-116,504,190), which may have strong H3K27ac signal because it intersects the *CAPZA2* TSS, and is located within the *CAPZA2* candidate promoter with the highest gene association score, according to GeneHancer (Fishilevich et al., 2017). Beyond the distinctions in method between ChIP-seq and CUT&RUN that may influence fold enrichment, fold enrichment can be dependent on sequencing depth, since peaks with a higher fold enrichment are saturated more readily than those with a lower fold enrichment (Zhang et al., 2008). As a result,

differences in sequencing depth between our CUT&RUN data and the published ChIP-seq datasets may also contribute to the observed variation in fold enrichment across samples.

### 4.4.4 Conclusion

By epigenomic characterisation of the *CAV1/CAV2* enhancer landscape in both iPSC-derived motor neurons and SH-SY5Y cells, we have revealed the distinct patterns of chromatin accessibility and enhancer-associated histone modifications that underpin the enhancer-mediated regulation of *CAV1* and *CAV2* expression. These patterns vary across SH-SY5Y cells, healthy control- and ALS patient-derived motor neurons, highlighting how enhancer activity may influence gene regulation in both disease and healthy contexts. We have assessed chromatin accessibility at enhancers using ATAC-seq and transcriptional activity using H3K27ac CUT&RUN, revealing that active enhancers in an open chromatin states are more prone to harbouring ALS-associated SNPs. By independently demonstrating that enhancers in an active chromatin state and those associated with activating histone modifications are more likely to contain SNPs, we underscore the potential for enhancers to contribute to ALS aetiology.

Our findings highlight the significance that chromatin accessibility and epigenetic modifications may have in influencing ALS progression. Finally, we have identified a subset of candidate *CAV1/CAV2* enhancers that warrant further investigation, including profiling of eRNA transcription and conducting additional functional analyses, to better understand their role in disease aetiology.

# Chapter 5

# Characterisation of *CAV1* enhancer transcriptional activity: eRNAs and disease

## 5.1 Introduction

### 5.1.1 Mapping enhancer transcription

Active enhancers are transcribed into eRNAs, with the level of transcription often considered a proxy for enhancer activity level (De Santa et al., 2010, Hah et al., 2011, Kaikkonen et al., 2013, Kim et al., 2010, Li et al., 2013). However, the transience of eRNAs, due to their sensitivity to degradation by the RNA exosome (Andersson et al., 2014, De Santa et al., 2010), has long impeded their detection. Nascent RNA constitutes less than 0.5% of the total cellular RNA (Jackson et al., 1998), therefore methods that specifically capture nascent RNA transcripts, such as transient transcriptome sequencing (TT-seq), are particularly valuable for identifying short-lived RNA species that are often missed using conventional RNA-seq technologies (Schwalb et al., 2016). TT-seq pairs 4-thiouridine (4sU)-labelling with RNA fragmentation, to specifically isolate and sequence RNA transcribed within the 5 minute metabolic labelling pulse and remove bias caused by the presence of long pre-existing 5' regions. TT-seq can therefore identify and quantify eRNAs with high sensitivity and temporal resolution, providing a snapshot of the dynamic transcriptional landscape (Figure 5.1) (Schwalb et al., 2016).

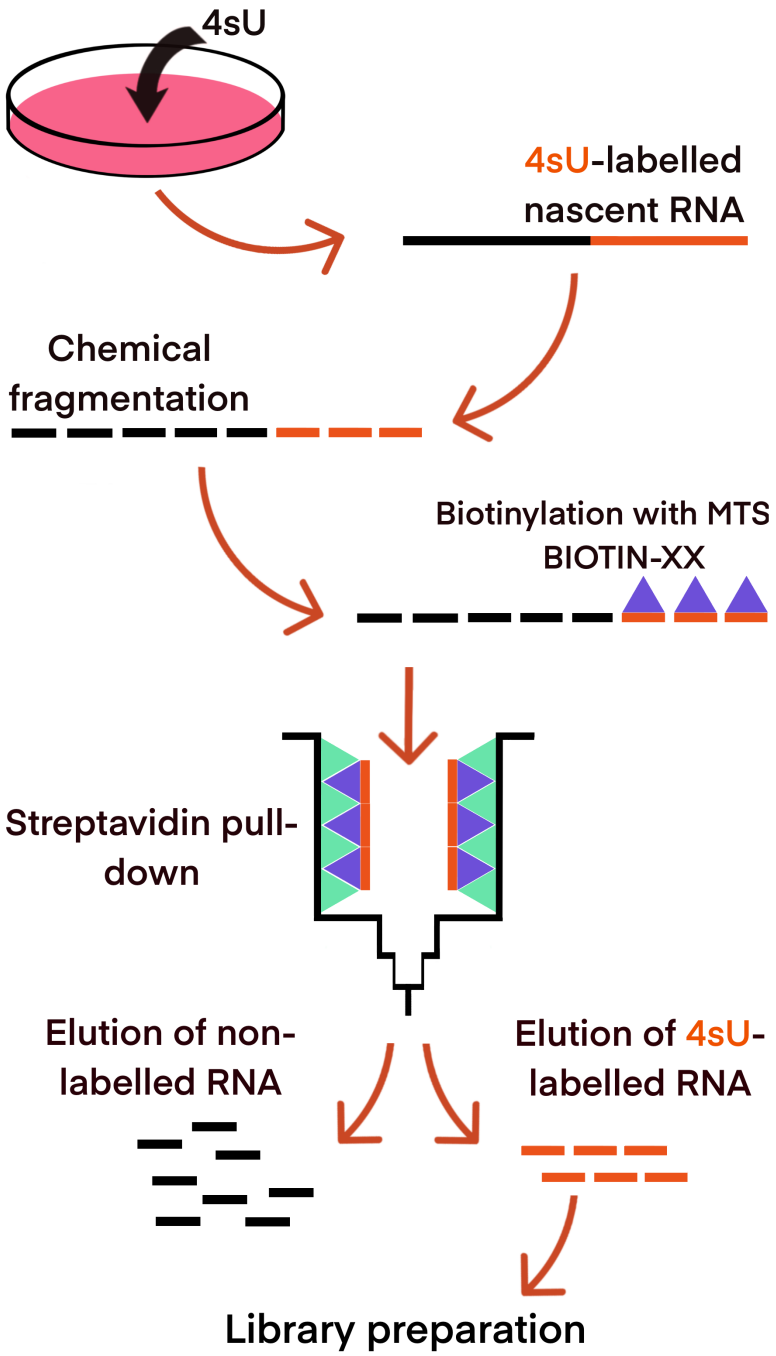



FIGURE 5.1: TT-seq workflow
Schematic representing the workflow of TT-seq. Nascent RNA is metabolically labelled *in vivo* via incubation with the nucleoside analogue, 4-thiouridine (4sU). Total RNA is extracted and fragmented. 4sU-labelled fragmented RNA is biotinylated and captured using steptavidin pull-down prior to library preparation and sequencing (Gregersen et al., 2020, Schwalb et al., 2016)

### 5.1.2 RNA structure

The pairing of the four nucleotide bases that make up an RNA molecule, namely adenine, cytosine, guanine and uracil, culminate in its secondary structure. Akin to DNA, RNA also forms standard Watson-Crick base pairs, but can additionally form non-canonical wobble and Hoogsteen base pairs (Crick, 1966, Leontis and Westhof, 2001, Nagaswamy et al., 2002, Watson and Crick, 1953). Non-canonical base-pairing in RNA molecules not only influences RNA structure, but also function (Chang et al., 1999, Zhong et al., 2006), with the structural and chemical properties of non-canonical base pairs favouring intra- or inter-molecular interactions (reviewed in Ananth et al., 2013, Leontis and Westhof, 2003). The conformation of an RNA molecule is dictated not only by sequence, but also by the thermodynamic landscape and the biological conditions in which it exists. *In vivo*, a given RNA populates an ensemble of potential structures that vary in stability, flexibility and compactness (Ganser et al., 2019). With increasing RNA length, the complexity of this ensemble increases due to increasing numbers of potential unique conformational states dynamically sampled within the thermodynamic landscape. Thereby, formation of functional RNA structures within biologically-relevant timescales is determined by the interplay between polynucleotide interactions in the context of the free-energy of the system in which it exists (reviewed in Ganser et al., 2019, Thirumalai and Woodson, 1996, Vicens and Kieft, 2022).

The folding kinetics of an RNA molecule are determined by an ensemble of unfolded structures navigating the free-energy landscape in search of the native conformation, a process sensitive to mutations and external conditions. Within the free-energy profile of an RNA, the RNA predominantly occupies the most thermostable structure, namely the lowest free-energy state, which is often the native, biologically-relevant state. Higher energy states are less energetically favourable and therefore less populated (Figure 5.2). Some RNA molecules rapidly reach the native state without encountering intermediate states, while others assume misfolded structures, represented as additional metastable free energy minima in the free-energy landscape. Subsequent rearrangement of misfolded RNA structures to the native state requires activation to overcome the free-energy barrier, resulting in slower folding kinetics. However, even the dominant conformation shows minor structural variations under biological conditions, with RNA structure unlikely to be truly static. In all cases, RNA is structured, with the non-structured state highly energetically unfavourable, though the characteristics of the conformational ensemble can differ. Alterations in the biochemical environment, such as binding of proteins, metal ions, and metabolites, or chemical modifications of nucleotides, can further stabilise the native state, while sequence mutations, and some RNA-protein interactions, may destabilise it and potentially stabilise a non-native state (Figure 5.2) (reviewed in Ganser

et al., 2019, Thirumalai and Woodson, 1996, Vicens and Kieft, 2022).

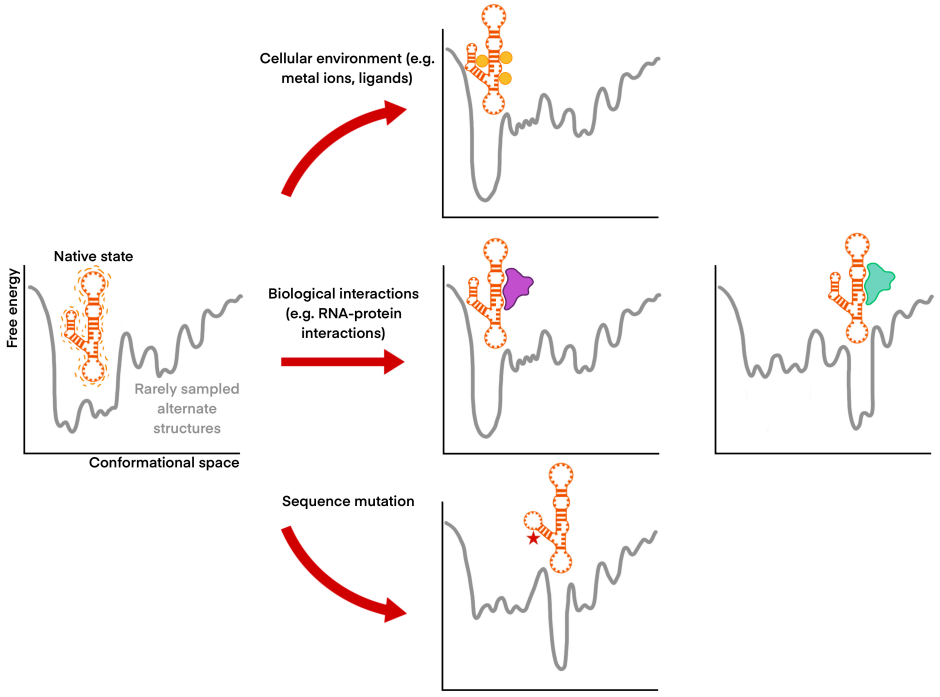



FIGURE 5.2: Dynamics of RNA structure in the free energy landscape
In the free energy landscape of an RNA, an RNA will populate the native state, with some limited structural changes occurring within the conformational space. Less energetically favourable states are less populated, separated from the native state by high energy states. In different cellular environments, for example by binding of metal ions or ligands, the native state can be further stabilised. Upon binding of protein interaction partners, the native state can be further stabilised, or destabilised along with the stabilisation of non-native state conformations. A mutation in the RNA sequence can result in destabilisation of the native state and stabilisation of a non-native state. Figure adapted from Ganser et al. (2019), Vicens and Kieft (2022)

#### 5.1.2.1 RNA structure prediction

The minimum free-energy (MFE) structure of an RNA refers to the most thermodynamically favourable, and therefore most stable, secondary structure that an RNA molecule can adopt under given temperature and ionic conditions (Zuker and Stiegler, 1981). RNA structure prediction software, such as RNA Structure (Reuter and Mathews, 2010) and RNAfold from the ViennaRNA package (Hofacker et al., 1994, Lorenz et al., 2011), utilise dynamic programming algorithms to predict the MFE structure and compute the full equilibrium partition function of the input RNA (Hofacker et al., 1994, Lorenz et al., 2011, McCaskill, 1990, Reuter and Mathews, 2010, Zuker and Stiegler, 1981).

MFE values can serve as a proxy for predicted RNA stability, as they reflect the thermodynamic favourability of a given RNA structure compared to all possible conformations

of the RNA sequence. Simplistically, the MFE structure of an RNA may represent more stable structures, depicting configurations which RNA molecules are more likely to adopt and maintain in the cellular environment (Hofacker et al., 1994, Lorenz et al., 2011, Reuter and Mathews, 2010, Zuker and Stiegler, 1981). While MFE structure prediction can provide valuable insights into RNA stability, biochemical factors, including RNA modifications, RNA-protein interactions and cellular conditions, can influence stability beyond the MFE structures alone (Lorenz et al., 2016).

A limitation to MFE structure prediction is that it assumes the RNA secondary structure is at equilibrium. *In vivo*, an RNA molecule likely samples a number of potential structures, dependent on the cellular environment (reviewed in Ganser et al., 2019, Thirumalai and Woodson, 1996, Vicens and Kieft, 2022). Computing the partition function of an RNA can establish a more comprehensive understanding of the ensemble of secondary structures an RNA molecule may adopt. This approach quantifies the sum of all possible RNA structures, including suboptimal conformations, weighted by their Boltzmann probabilities, providing a measure of the ensemble of structures an RNA can adopt at thermal equilibrium (McCaskill, 1990, Zuker, 1989).

### 5.1.3 Chapter aims

This chapter focuses on the use of TT-seq to quantify and characterise eRNA transcription across the *CAV1/CAV2* enhancer landscape, discussing the use of eRNA transcription data to infer enhancer activity, and further utilising *de novo* transcript assembly to predict the structure and assess the stability of eRNAs, in the context of disease.

## 5.2 Using nascent transcription data to identify active enhancers

For the detection of eRNA transcription we utilised TT-seq. TT-seq is capable of capturing transient ncRNAs prior to degradation by the RNA exosome (Schwalb et al., 2016), thus has great potential to capture eRNAs despite their generalised transience.

### 5.2.1 TT-seq normalisation and replicate correlation

For normalisation purposes, we have utilised a 4-thiouracil (4tU)-labelled *Saccharomyces cerevisiae* RNA spike-in, allowing for control of technical variation between replicates, and allowing for comparison of expression levels between different cell lines.

Following alignment to the *S. cerevisiae* (sacCer3) genome, we utilised gene count data output from `STAR` RNA-seq aligner to perform differential expression analysis, estimating size factors for normalisation using `DESeq2`. As the same amount of 4tU-labelled *S. cerevisiae* RNA was added to each replicate, we were able to use the size factors to control for technical variation between replicates. To demonstrate the efficacy of the spike-in for controlling variation in global gene expression counts between replicates, we compared the gene counts between unnormalised versus spike-in normalised TT-seq data (Figure 5.3). Following alignment to the hg19 genome, we utilised gene count data output from `STAR` RNA-seq aligner to perform differential expression analysis using `DESeq2`, with scaling factors of 1.195229 and 0.836660 used for normalisation of replicate 1 and replicate 2, respectively. Slight distributional shifts were observed due to downsampling of replicate 2. Mean unnormalised TT-seq gene counts were 117.2 and 197.6, for replicates 1 and 2 respectively. Whereas, following normalisation, mean TT-seq gene counts were 141.1 and 164.0, for replicates 1 and 2 respectively. This demonstrated the effectiveness of spike-in normalisation for stabilising gene expression measurements among replicates, crucial for reliable downstream analyses.

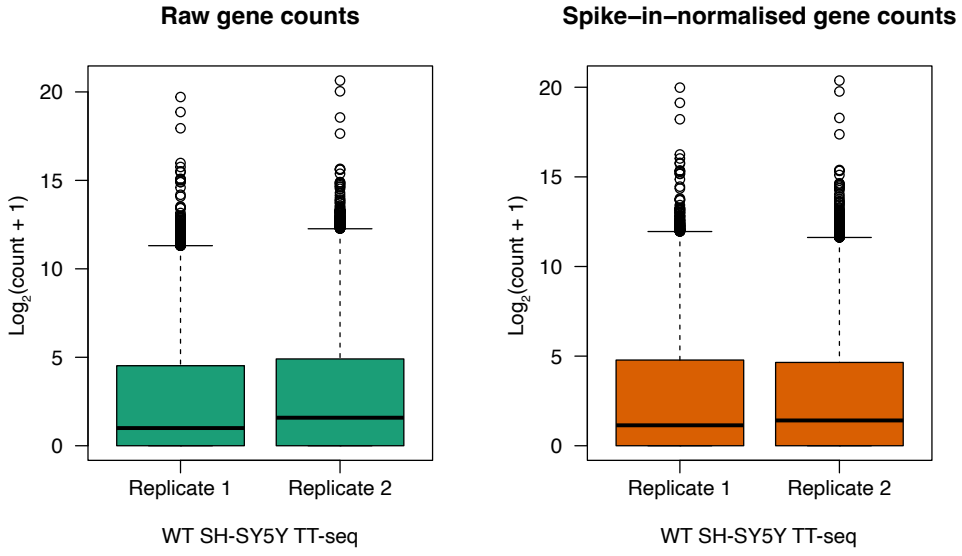



FIGURE 5.3: Gene count differences for TT-seq data before and after spike-in normalisation
Log-transformed gene counts of WT SH-SY5Y TT-seq data from two replicates both prior and post spike-in normalisation. Demonstrates a distributional shift in replicate 2 following downsampling based on spike-in generated scaling factors.

To demonstrate similarities in gene expression counts between the two replicates of WT

SH-SY5Y TT-seq, we considered the correlation in gene counts between replicates (Figure 5.4). The Pearson correlation coefficient, $r$ between the two replicates was 0.9913805, indicating a very strong positive linear relationship between the gene expression counts of the two replicates. This suggests that the gene expression measurements were highly consistent across the two replicates, suggestive of reproducibility between replicates. Further, the high correlation confirms that technical differences between the replicates were successfully minimised by spike-in normalisation

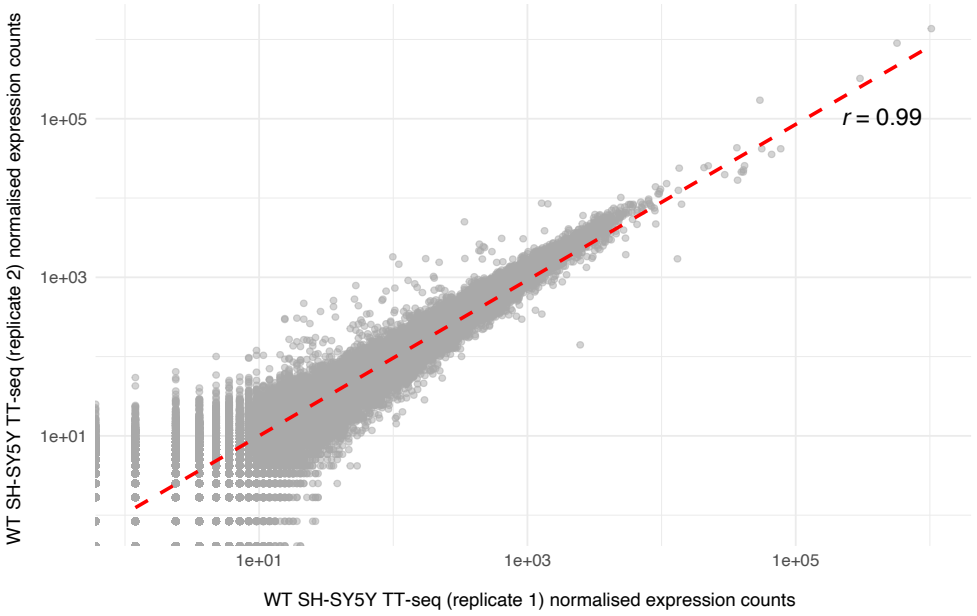



FIGURE 5.4: Correlation between WT SH-SY5Y TT-seq replicates
Correlation of spike-in normalised gene counts between WT SH-SY5Y TT-seq replicates. Correlation strength quantified by the Pearson correlation coefficient.

### 5.2.2 TT-seq is capable of capturing transient non-coding transcripts

The coverage of transient intronic transcripts with respect to exonic transcripts is estimated to be ∼60% for TT-seq, compared to ∼8% for total RNA-seq (Schwalb et al., 2016). To validate the usage of TT-seq in SH-SY5Y cells for the capture of transient non-coding transcripts, we investigated whether we could recapitulate this ratio in our samples. For each given GENCODE gene annotation (Frankish et al., 2019), the ratio of intronic versus exonic transcripts was calculated, for both our TT-seq data and a published total RNA-seq dataset from WT SH-SY5Y cells (Liu et al., 2022), data accessible at NCBI GEO database (Edgar et al., 2002), accession GSM6217322. For TT-seq, we observed a mean percentage intron read count of ∼69% and ∼68% per gene, for replicate 1 and 2, respectively, compared to ∼21% for total RNA-seq (Figure 5.5).

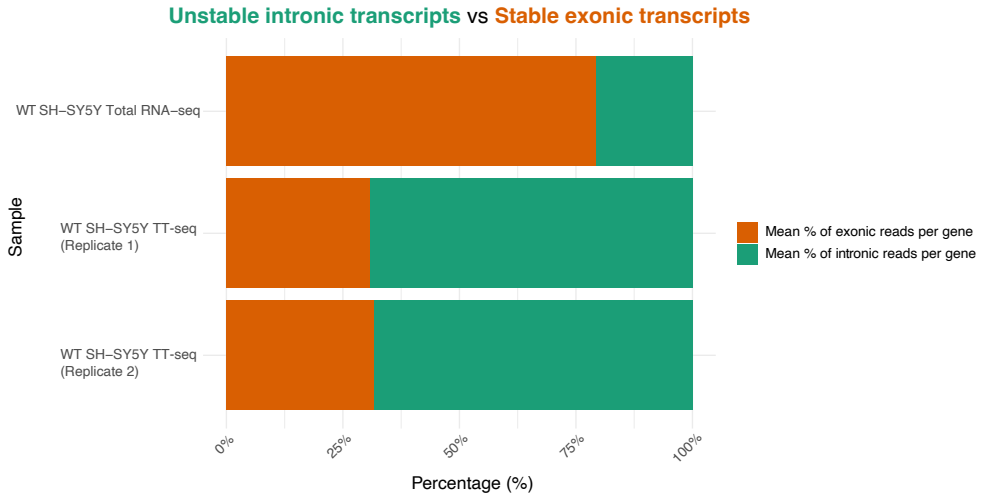



FIGURE 5.5: TT-seq captures more unstable intronic transcripts than total RNA-seq
Exonic to intronic per gene transcript ratios for TT-seq in SH-SY5Y cells, compared to total RNA-seq in SH-SY5Y cells (Liu et al., 2022), to recapitulate the expected values for TT-seq, with the coverage of intronic to exonic transcripts per gene estimated to be ∼60% (Schwalb et al., 2016).

### 5.2.3 TT-seq captures enhancer transcription

As suggested by the exonic to intronic per gene transcript ratios, TT-seq is capable of capturing unstable, specifically intronic, transcripts. In order to identify if the capacity to capture unstable transcripts extended beyond introns, we investigated the TT-seq coverage over the non-coding genome as a whole. In comparison to total RNA-seq, where only 13.2% of the total reads mapped to the non-coding genome, 32.5% and 31.5% of the total reads mapped to non-coding genome for TT-seq replicate 1 and 2, respectively (Figure 5.6).

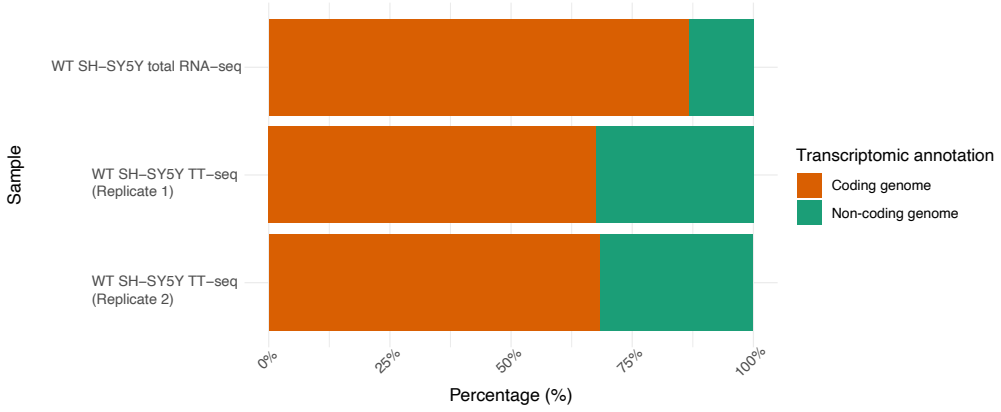



FIGURE 5.6: TT-seq reads map to the non-coding genome
Of the total TT-seq reads, a larger proportion of reads map to the non-coding genome relative to total RNA-seq in the same cell type.

Next, we explored the capacity for TT-seq to capture eRNAs by investigating the TT-seq coverage over enhancer predictions from the ABC model (Figure 5.7) (Fulco et al., 2019). As a control, we generated chromosome-matched, non-exonic shuffled control regions, where we used the ABC enhancer predictions as input and randomly permuted their locations along the same chromosome, outside of coding exons. We then investigated the TT-seq coverage over these shuffled control regions (Figure 5.8). This comparison enabled us to investigate whether random non-coding regions produce a similar number of nascent RNA transcripts relative to enhancers, which are known to produce eRNAs upon activation. As expected, we identified more ABC enhancers with TT-seq coverage than shuffled control regions, with 54.4% of the ABC enhancers exhibiting TT-seq coverage, compared to 45.8% of shuffled control regions (Figure 5.9). Moreover, when considering regions of very high TT-seq read coverage, we identified 17 ABC enhancers with more than 1e+01 CPM TT-seq coverage, while only 4 of the shuffled regions demonstrated such high coverage.

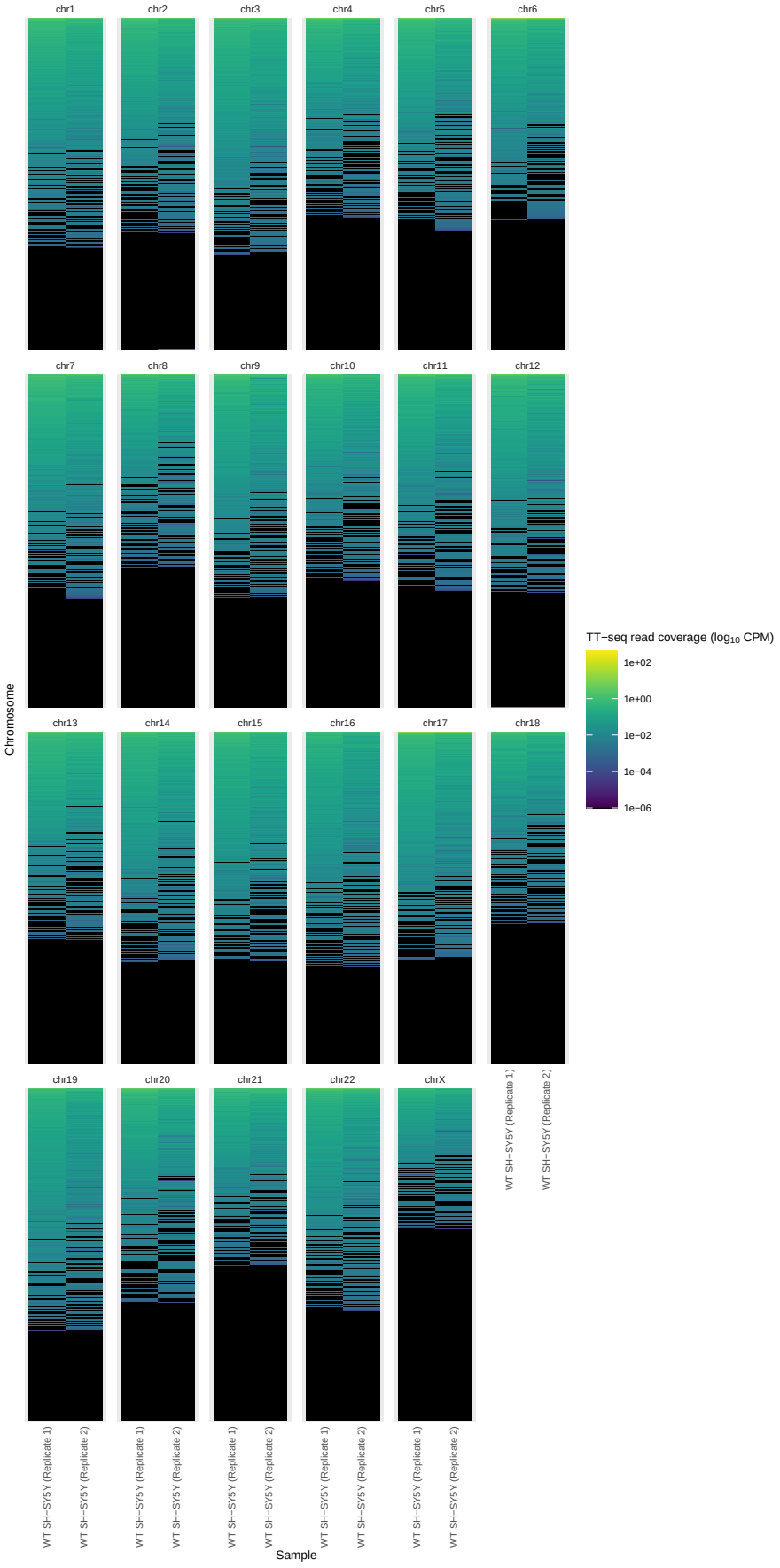



FIGURE 5.7: TT-seq coverage over ABC enhancer predictions
WT SH-SY5Y TT-seq read coverage in $\log_{10}$-transformed CPM over ABC enhancer predictions (Fulco et al., 2019), ordered by descending mean read coverage per chromosome.

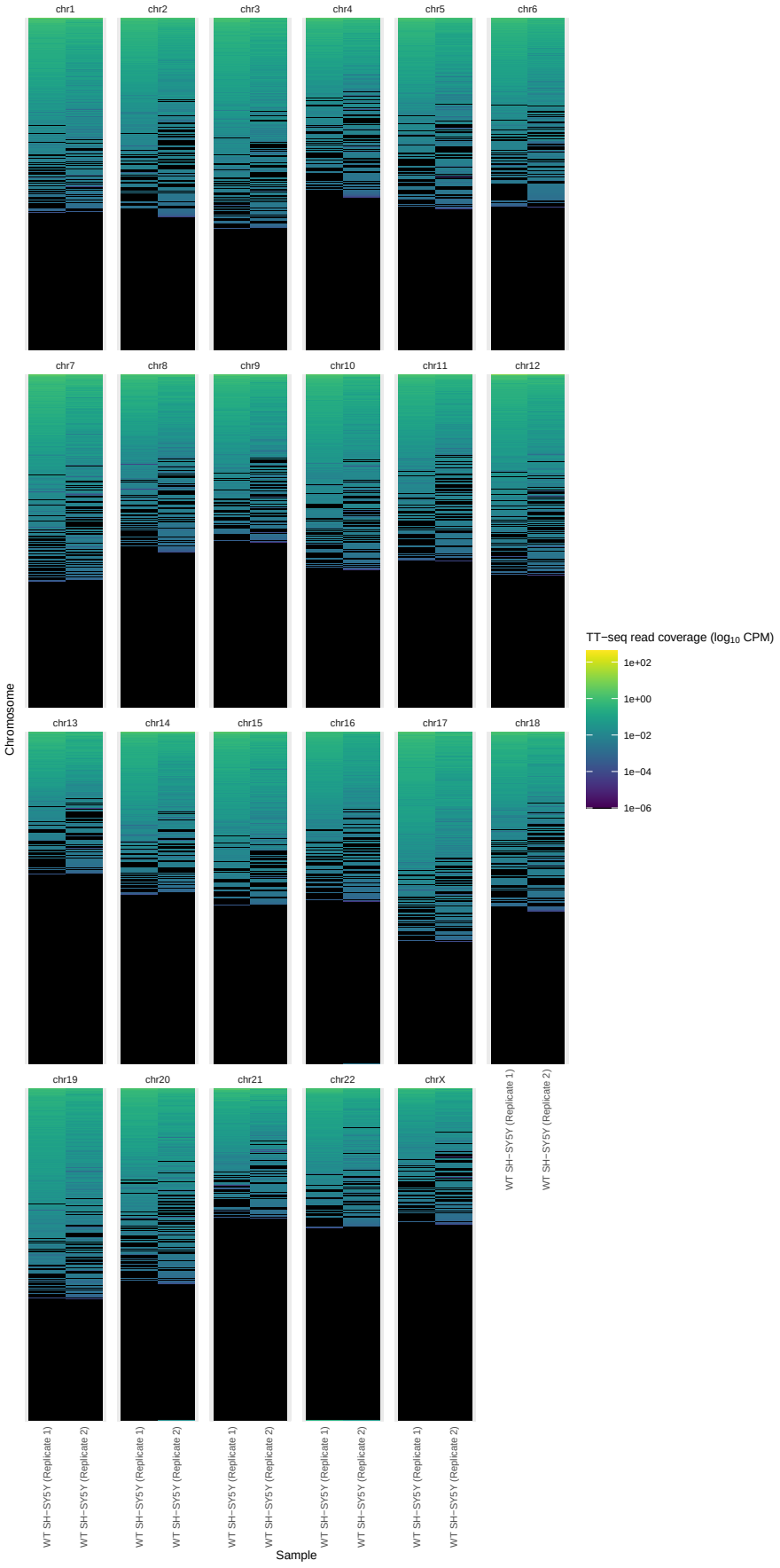



FIGURE 5.8: TT-seq coverage over shuffled control regions
WT SH-SY5Y TT-seq read coverage in $\log_{10}$-transformed CPM over shuffled control regions, ordered by descending mean read coverage per chromosome.

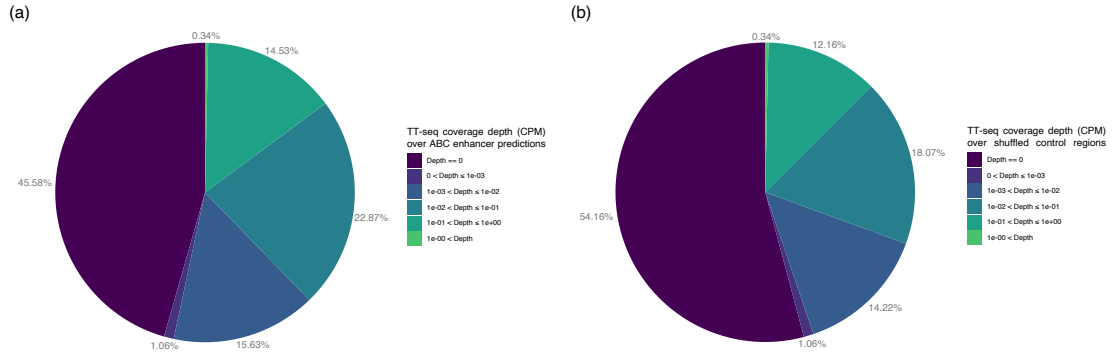



FIGURE 5.9: Comparison of TT-seq coverage over ABC enhancer predictions and shuffled control regions
(a) Quantification and categorisation of mean WT SH-SY5Y TT-seq coverage over ABC candidate enhancers (Fulco et al., 2019). (b) Quantification and categorisation of mean WT SH-SY5Y TT-seq coverage over shuffled control regions.

In a direct comparison of TT-seq read coverage between ABC enhancer regions and shuffled control regions, we observed significantly higher transcriptional activity at ABC enhancers (Wilcoxon Rank Sum test, $p <$ 2.2e-16) (Figure 5.10). This significant difference in coverage distributions underscores the broader range and higher maximum coverage over ABC enhancers compared to shuffled control regions. To further quantify this difference, we calculated the rank-biserial correlation, which measures how transcription coverage ranks differ between enhancer regions and shuffled control regions. By comparing the rank distributions between these groups, we obtained values of 0.0949 and 0.0880 for replicates 1 and 2, respectively. Therefore, while we confirmed that the TT-seq coverage over enhancers tended to be higher than over shuffled regions, the overall magnitude of this effect is modest, likely attributable to the large number of regions with zero TT-seq coverage for both ABC enhancers and shuffled control regions.

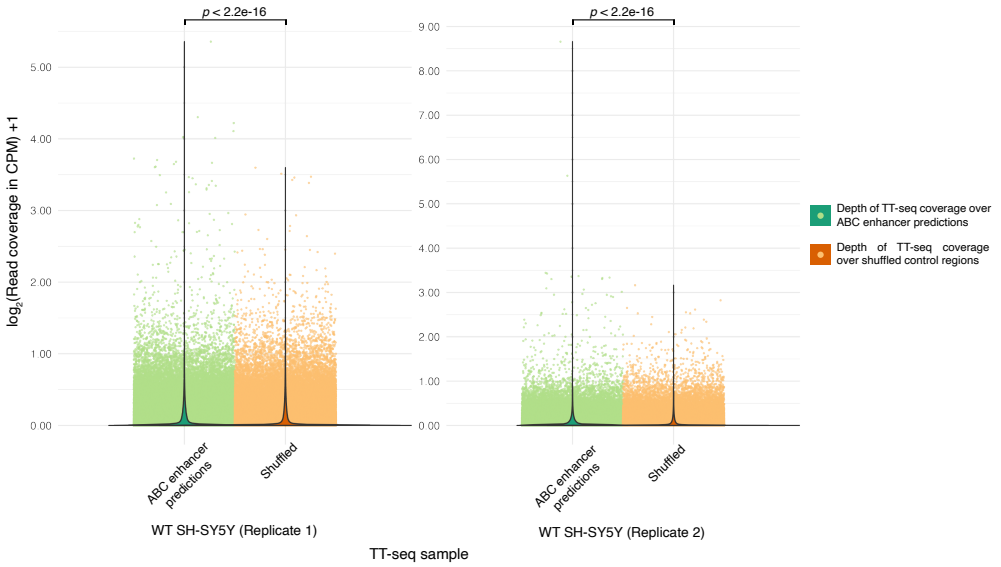



FIGURE 5.10: TT-seq identifies transcription from enhancers
WT SH-SY5Y TT-seq coverage ($\log_2$ CPM+1) over ABC candidate enhancers (Fulco et al., 2019) compared to shuffled control regions. $p$-values determined by Wilcoxon Rank Sum.

### 5.2.4 eRNA transcription at candidate *CAV1/CAV2* enhancers

Given that eRNA transcription is a fundamental characteristic of active enhancers (Kim et al., 2010), we examined the TT-seq coverage over candidate *CAV1/CAV2* enhancers in WT SH-SY5Y cells (Figure 5.11a). Utilising eRNA transcription as a proxy for enhancer activity level, we can identify the generalised activity at our candidate enhancers. This further demonstrated the congruence between our TT-seq replicates. We identified that the majority of candidate *CAV1/CAV2* enhancers were transcribed in WT SH-SY5Y cells, with 51% (58 out of 114) and 61% (69 out of 114) of enhancers demonstrating TT-seq coverage for replicate 1 and replicate 2, respectively (Figure 5.11b)

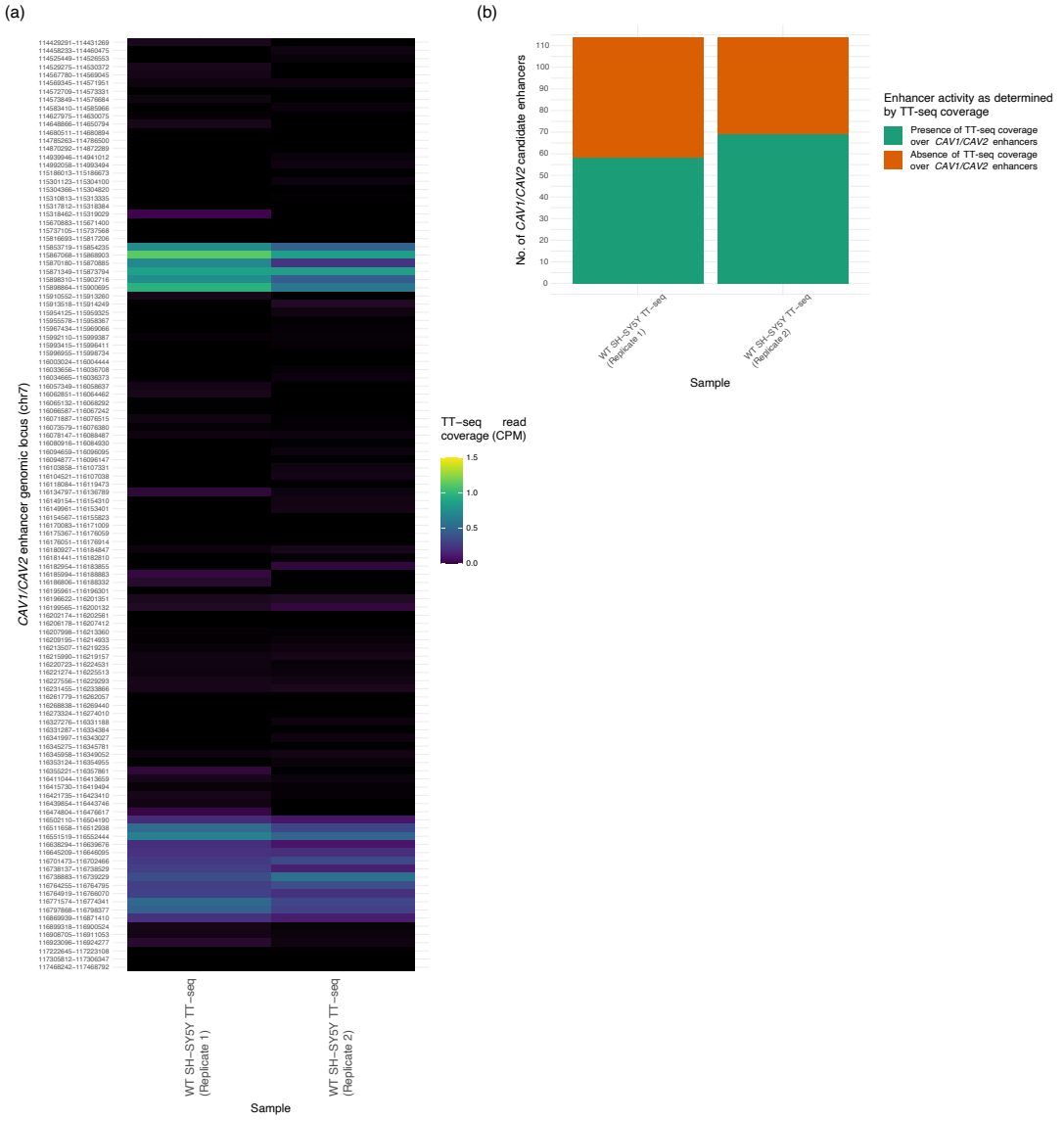



FIGURE 5.11: eRNA transcription at *CAV1/CAV2* candidate enhancers in SH-SY5Y cells

(a) TT-seq read coverage in CPM at the *CAV1/CAV2* candidate enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) in WT SH-SY5Y cells. (b) Number of candidate *CAV1/CAV2* candidate enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) considered to be active, as determined by the presence of TT-seq coverage over enhancer regions.

#### 5.2.4.1 eRNA transcription at candidate *CAV1/CAV2* enhancers containing ALS-associated SNPs

To further test the hypothesis that disease-associated SNPs localise to active enhancers in physiologically relevant cell types, we investigated the number of SNPs in actively

transcribed candidate *CAV1/CAV2* enhancers versus those absent of TT-seq read coverage. As expected, transcribed *CAV1/CAV2* enhancers, as determined by TT-seq coverage, contain significantly more ALS-associated SNPs than untranscribed enhancers (Figure 5.12). Transcribed enhancers contained on average 3 SNPs, whereas untranscribed enhancers contained on average 1 SNP, to the closest integer.

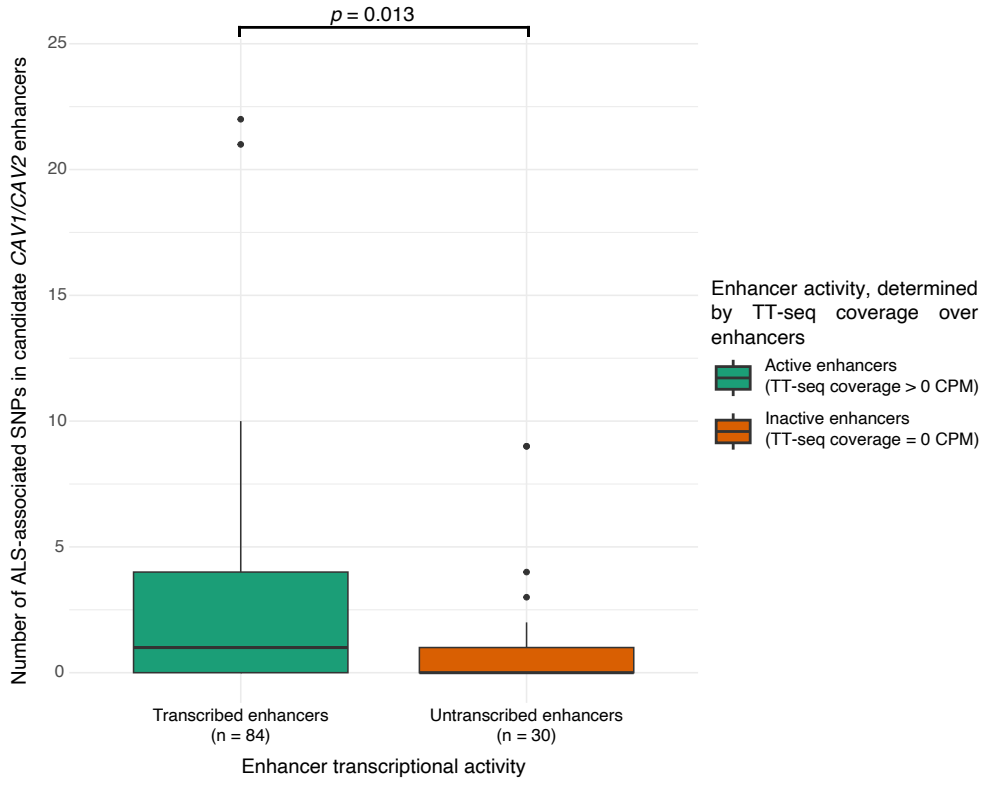



FIGURE 5.12: Transcribed *CAV1/CAV2* candidate enhancers contain ALS-associated SNPs

Number of ALS-associated SNPs (Cooper-Knock et al., 2021) transcribed versus untranscribed candidate *CAV1/CAV2* enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019), as determined by TT-seq in WT SH-SY5Y cells. $p$-value determined by Wilcoxon Rank Sum.

## 5.3 Integration of eRNA transcription data with other enhancer characteristics

Integrating nascent transcription data at enhancers with additional identifiers of enhancer characteristics, such as ATAC-seq and H3K27ac CUT&RUN, may provide improved insight into the functional characteristics of enhancers. Given that an open chromatin state is considered a fundamental requirement for transcription (Georges et al., 2002, Miller and Widom, 2003), and H3K27ac is associated with transcriptional activation at enhancers (Creyghton et al., 2010, Heintzman et al., 2007, Rada-Iglesias et al.,

2011, Stasevich et al., 2014), we would expect enrichment of ATAC-seq and H3K27ac CUT&RUN signal proximal to sites of nascent transcription from enhancers.

### 5.3.1 eRNA transcription and chromatin accessibility

With the well-established repressive effect of chromatin compaction on transcription (Georges et al., 2002, Miller and Widom, 2003), we would expect there to be correlation between regions of open chromatin and nascent transcription. However, we observed no significant difference in transcription levels between the 25 candidate *CAV1/CAV2* enhancers in open chromatin, as indicated by presence of ATAC-seq signal, and the remaining 89 enhancers in a closed chromatin state, as characterised by absence of ATAC-seq signal (Figure 5.13a). We further considered the relationship between the depth of TT-seq read coverage and log-transformed ATAC-seq signal fold-enrichment at candidate enhancers, observing a weak positive correlation ($r = 0.192$) (Figure 5.13b).

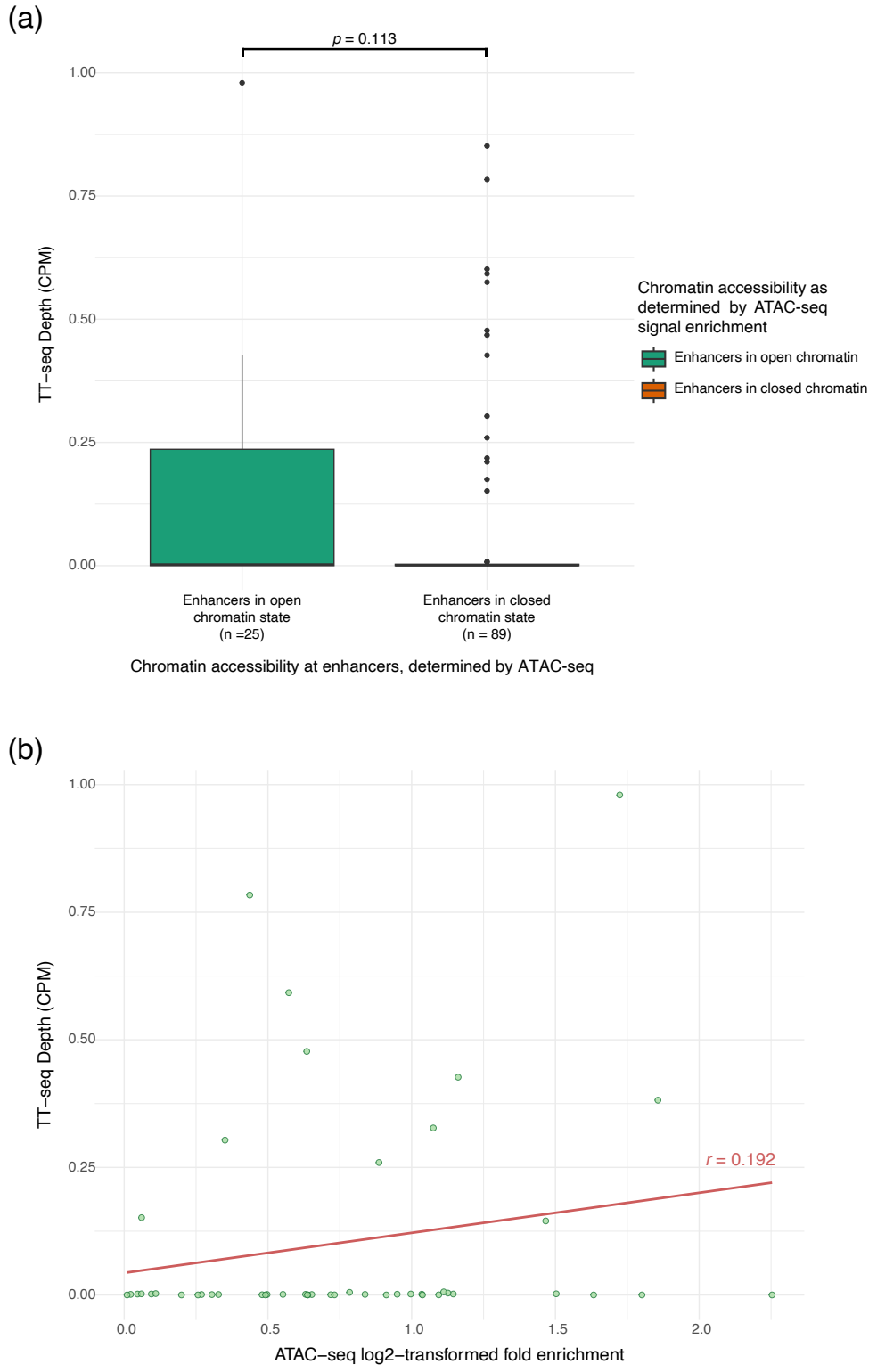



FIGURE 5.13: The relationship between chromatin accessibility at candidate *CAV1/CAV2* enhancers and eRNA transcription

(a) WT SH-SY5Y TT-seq read coverage over *CAV1/CAV2* candidate enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) in an open chromatin state (n = 25) versus a closed chromatin state (n = 89), as determined by presence or absence of ATAC-seq signal, respectively. Points represent outliers identified based on the interquartile range upper bound. *p*-value determined by Wilcoxon Rank Sum. (b) Relationship between levels of nascent transcription at *CAV1/CAV2* enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) and log-transformed ATAC-seq signal fold enrichment. Correlation strength quantified by the Pearson correlation coefficient.

### 5.3.2 eRNA transcription and histone modification

With H3K27ac deposition at enhancers a well-defined signifier of enhancer activity (Creyghton et al., 2010, Heintzman et al., 2007, Rada-Iglesias et al., 2011), we would expect a tight link between presence of H3K27ac at enhancers and eRNA transcription. As expected, we observed a significant increase in the levels of nascent transcription at the 20 candidate *CAV1/CAV2* enhancers enriched for H3K27ac signal compared to the remaining 94 enhancers not enriched for H3K27ac signal (Figure 5.14a). Additionally, we observed a moderate positive correlation ($r = 0.381$) between the depth of TT-seq read coverage and the log-transformed fold enrichment of H3K27ac signal at these candidate enhancers (Figure 5.14b).

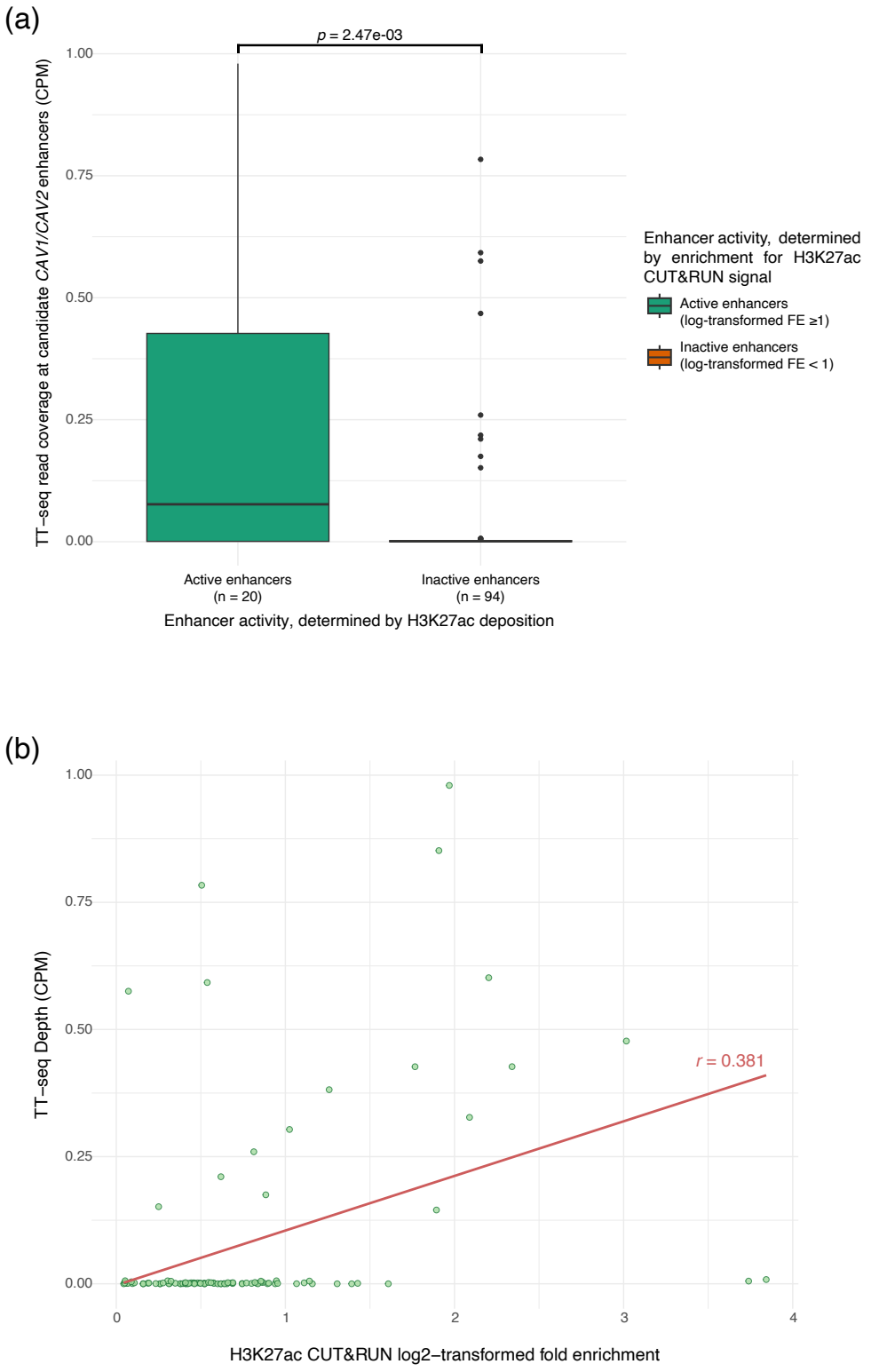



FIGURE 5.14: The relationship between H3K27ac deposition at candidate *CAV1/CAV2* enhancers and eRNA transcription

(a) WT SH-SY5Y TT-seq read coverage over *CAV1/CAV2* active (n = 20) and inactive (n = 94) candidate enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019), as determined by enrichment or absence of H3K27ac CUT&RUN, respectively. Points represent outliers identified based on the interquartile range upper bound. $p$-value determined by Wilcoxon Rank Sum. (b) Relationship between levels of nascent transcription at *CAV1/CAV2* enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) and log-transformed H3K27ac CUT&RUN signal fold enrichment. Correlation strength quantified by the Pearson correlation coefficient.

For further inspection of the relationship between H3K27ac deposition and eRNA transcription, we investigated of which enhancers considered to be active based on enrichment of H3K27ac CUT&RUN signal were also transcribed (Figure 5.15a). We identified that only 4 of the 20 candidate *CAV1/CAV2* enhancers with a log-transformed H3K27ac signal fold enrichment $\geq 1$ were absent of TT-seq coverage. Similarly, we identified of which enhancers considered to be inactive based on absence of H3K27ac CUT&RUN signal enrichment were not transcribed (Figure 5.15b). 26 out of the 30 candidate enhancers that were not transcribed were similarly not enriched for H3K27ac CUT&RUN signal.

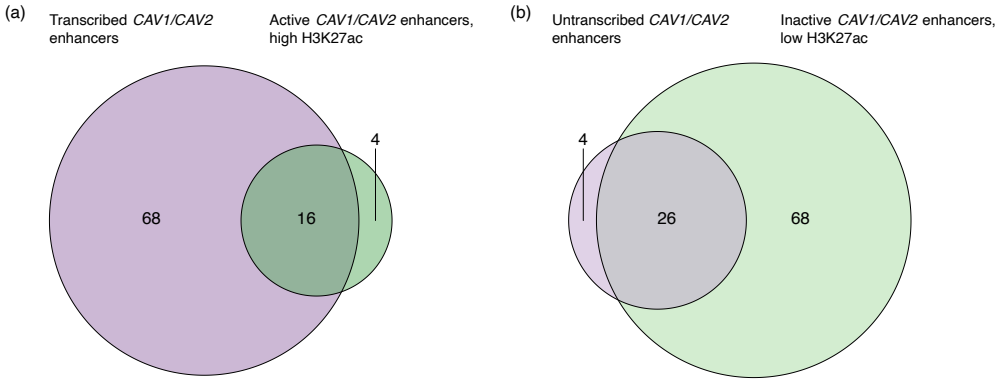



FIGURE 5.15: The relationship between H3K27ac deposition at candidate *CAV1/CAV2* enhancers and eRNA transcription
(a) Quantification of intersections of candidate *CAV1/CAV2* enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) with TT-seq read coverage (TT-seq read depth >0 CPM) with those with a more than 2-fold H3K27ac CUT&RUN signal enrichment in SH-SY5Y cells (log-transformed H3K27ac signal fold enrichment $\geq 1$). (b) Quantification of intersections of candidate *CAV1/CAV2* enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) absent of TT-seq read coverage (TT-seq read depth = 0 CPM) with those with a less than 2-fold H3K27ac CUT&RUN signal enrichment in SH-SY5Y cells (log-transformed H3K27ac signal fold enrichment $\leq 1$).

## 5.4 Identification of novel eRNAs in SH-SY5Y cells

While TT-seq has the capacity to capture eRNA transcription, providing valuable insights into enhancer activity (Schwalb et al., 2016), it also detects nascent transcription of other RNA species such as pre-mRNAs and lncRNAs. This broader transcriptional capture can introduce potential biases when specifically quantifying enhancer transcription, as the signal may not be solely attributed to eRNA production. Therefore, in order to identify novel eRNAs, we performed *de novo* transcript assembly, and subsequently compared the assembled RNA transcripts to the hg19 reference genome using `GffCompare` (Pertea and Pertea, 2020). We disregarded transcripts that were exact matches to or contained in the reference, retaining transcripts classified as unknown intergenic; single

exon transflag partially covering an intron; same strand overlap with reference exons; exonic overlap on the opposite strand. We further filtered for transcripts transcribed from within our candidate *CAV1/CAV2* enhancers, whereby the entire length of the transcript was contained within the coordinates of the candidate enhancer. Upon transcript assembly and filtering, we identified 18 and 13 novel potential eRNA transcripts from the WT SH-SY5Y TT-seq data in replicates 1 and 2, respectively (Figure 5.16). We identified 3 loci at which assembled eRNA transcripts overlapped between replicates.

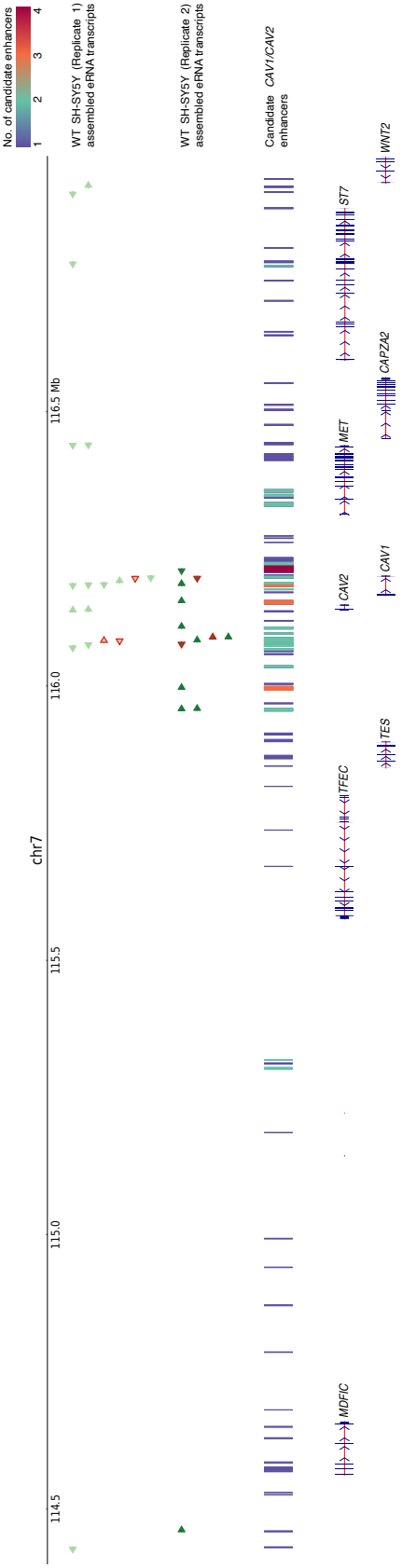



FIGURE 5.16: eRNA transcript assembly in WT SH-SY5Y cells
Trackline view of chr7:114,400,000-116,965,000 (hg19) demonstrating the genomic locations of *de novo* assembled eRNA transcripts from TT-seq in WT SH-SY5Y cells, relative to candidate *CAV1/CAV2* enhancers locations (Cooper-Knock et al., 2021, Fulco et al., 2019) and genic locations. eRNA transcripts overlapping between replicates are highlighted with red border. Candidate enhancers are coloured for clarity due to collapse in the track line.

## 5.5 eRNA structure and stability prediction

### 5.5.1 Minimum free energy of RNA secondary structures

The observation that mRNAs and miRNA precursors have greater negative MFE values than expected based on their nucleotide compositions and ratios (Bonnet et al., 2004, Seffens and Digby, 1999), has resulted in the widespread application of RNA MFE as an indicator of not only potential RNA stability, but also functionality. Therefore, in order to predict the stability, and thereby potential functionality of *CAV1/CAV2* enhancer-associated eRNA sequences, we performed RNA structure prediction using `RNAfold` (Lorenz et al., 2011), using the resultant MFE values as a proxy for predicted RNA stability.

We employed MFE calculations as a direct measure of stability, identifying the lowest free energy state that a given RNA can adopt, where more negative MFE values imply a greater structural stability. MFE encompasses all three structural determinants of stability: sequence length, nucleotide content, and nucleotide order.

#### 5.5.1.1 Normalisation of MFE

For effective normalisation and comparison of predicted free energy values, we employed the MFE density (MFEden) normalisation index, effectively removing the component of sequence length, evaluating the contributions of nucleotide content and order to the overall free energy (Trotta, 2014). Moreover, it is established that pre-miRNAs, a class of ncRNAs, typically have MFEden values lower than expected based on their nucleotide context alone (Trotta, 2014). Therefore, we were interested to identify whether this was true for our eRNAs, under the pretext that a lower than expected MFEden may suggest some degree of stability.

The MFE ($MFE$) of a given RNA sequence with a nucleotide length ($L$) is normalised by calculating the MFEden using the following formula (Trotta, 2014):

$$\text{MFEden} = 100 \times \frac{MFE - MFEref_L}{L - L_0}$$

Where the $MFEref_L$ is the expected MFE for an RNA sequence of $L$ nucleotides with equimolar ratios of each nucleotide (A, C, G, and U) and $L_0$ for MFEs computed by RNAfold is empirically equal to 8nt (Trotta, 2014).

MFEden successfully excludes the component of MFE associated with sequence length, but is still greatly impacted by nucleotide order and composition and thus, may indirectly

provide an estimate of the contributions of the nucleotide order and ratio within the RNA sequence (Trotta, 2014). To account for the component of GC content in MFEden, we further calculated the expected MFEden of sequences based on GC content. Since longer and more GC-rich sequences generally form more stable RNA structures, calculating MFEden and comparing the observed MFEden values to those expected based on GC content helps determine whether a predicted RNA structure is more or less stable than anticipated, based on sequence length and GC content.

As in Trotta (2014), we calculated the mean MFEden of shuffled sequences for sequence lengths ranging from 50nt to 600nt, and GC% of 20%, 40%, 50%, 60%, 70% and 80%, with 10 shuffled sequences for each interval (Figure 5.17a). We confirmed that regardless of RNA length, GC content is a considerable contributor to MFEden value, with MFEden largely uniform for all tested RNA lengths for each given GC percentage. We therefore plotted the mean MFEden for each given GC percentage to predict the expected MFEden for RNA sequences of GC contents between 20% and 80%, fitting a second-degree polynomial regression model to capture the non-linear relationship between GC content and MFEden (Figure 5.17b). This allows us to interpolate the predicted MFEden for eRNAs based on their GC content, enabling the assessment of whether eRNAs are predicted to be more stable than expected given their nucleotide composition.

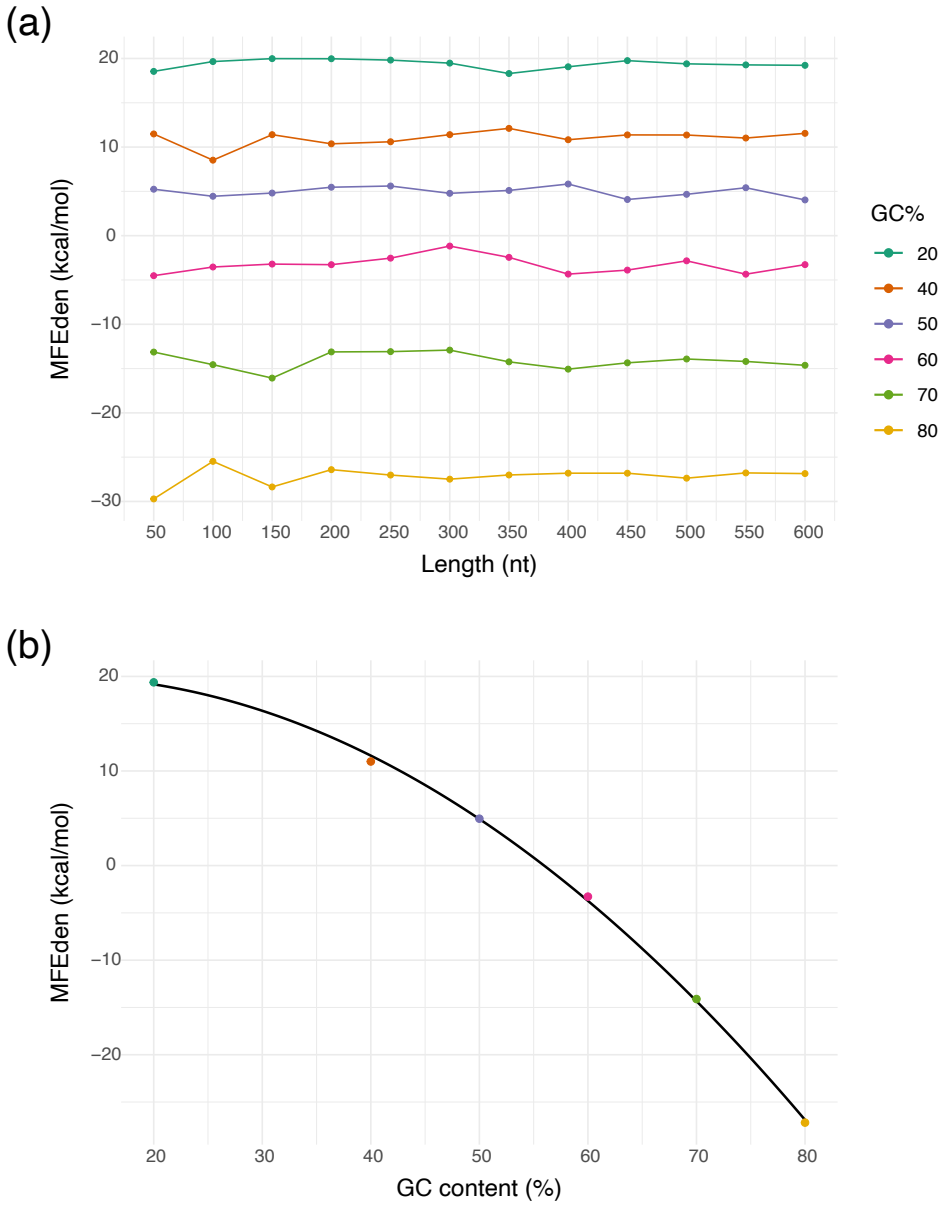



FIGURE 5.17: MFEden of shuffled sequences with given GC-content
(a) Mean MFEden versus RNA sequence length for shuffled sequences with GC content of 20%, 40%, 50%, 60%, 70% and 80%. Each point corresponds to the mean MFEden of 10 shuffled sequences. (b) Mean MFEden for sequences 50 nt to 600 nt versus GC content. Line is fitted using a quadratic linear regression model.

### 5.5.2 Thermodynamic stability prediction

Beyond normalisation, the thermodynamic z-score can be employed to provide a statistical perspective on the contribution of nucleotide order to predicted eRNA stability (Andrews et al., 2018, 2022). This metric is based on the principle that functional ncRNAs have lower MFE values compared to randomised sequences with the same nucleotide composition. Fundamentally, the nucleotide arrangement in functional RNA

sequences is optimised to form specific secondary structures, that are ultimately disrupted in shuffled sequences, resulting in a relatively more positive MFE values (Babak et al., 2007, Clote et al., 2005). Therefore, we used thermodynamic z-scores to compare the predicted stability of the native eRNA sequence to that of shuffled sequences generated by random permutations of the input nucleotides. Since the shuffled sequences share the same nucleotide composition as the native sequence, the z-score index isolates the contribution of nucleotide order to folding energy.

The thermodynamic z-score compares the MFE of the native sequence to the mean MFE of multiple shuffled sequences, normalised by the standard deviation of all MFE values (Andrews et al., 2018, 2022, Clote et al., 2005):

$$\text{z-score} = \frac{MFE_{native} - \overline{MFE_{random}}}{\sigma}$$

Where the $MFE_{native}$ is the MFE of the native RNA sequence, $MFE_{random}$ is the mean MFE of the shuffled RNA sequence and $\sigma$ is the standard deviation of all MFE values.

Thus, negative z-scores indicate that the native RNA MFE is more negative than the MFE values that randomised RNAs with the same length and nucleotide composition would generate by chance.

### 5.5.3 Comparison of predicted MFE values between native and shuffled sequences

With nucleotide order a significant component influencing RNA stability, we also computed the MFE of shuffled RNA sequences with the same length and nucleotide composition as the native eRNA sequence. This provides a benchmark for evaluating the predicted stability of native eRNA structures, assessing whether the observed predicted stability is due to specific structural motifs or is primarily driven by generalised nucleotide composition. This approach assumes that explicitly stable RNA structural motifs contribute more significantly to overall stability than nucleotide content alone. If the native eRNA structures exhibit more negative MFE values than that of the randomised sequences, this would suggest a greater degree of structural stability than expected by chance, attributed to the capacity to form local secondary structures that contribute to overall stability. This comparison helps reveal the extent to which native eRNA sequences are optimised for specific structural features that may contribute to biological function.

Taken together, the discussed metrics provide a comprehensive view of RNA stability by offering distinct perspectives on the structural factors that contribute to the free energy

profile of an RNA. Inconsistencies between these metrics may highlight biases in the structural components shaping the free energy profile. Conversely, consistent patterns across these measures may suggest that all three structural determinants contribute to stability of the RNA, providing confidence in our stability assessments.

### 5.5.4 RNA structure and stability prediction of example eRNAs

For comparison purposes we generated structure predictions for the known eRNAs transcribed from the well-studied *TAL1* super-enhancer (Danko et al., 2015, Lidschreiber et al., 2021), and the $^{DRR}$eRNA which has established functionality, and known to be stable due to its in *trans* functionality (Cichewicz et al., 2018, Przanowska et al., 2022, Tsai et al., 2018). MFE values and RNA structure dot-bracket notations were obtained from `RNAfold` (Lorenz et al., 2011). For normalisation by MFEden, we calculated MFE values for 10 shuffled sequences of equal length to the native RNA sequence with equimolar ratios of each nucleotide. Further, thermodynamic z-scores were calculated using `ScanFold` (Andrews et al., 2018), with a window size of 10nt, a step size of 1nt and 100 randomisations.

To model a stable, spliced and functional eRNA we performed RNA structure and stability predictions for the $^{DRR}$eRNA. We obtained a MFE value of -133.39kcal/mol, suggestive of a stable RNA. Moreover, we obtained a MFEden of -0.28kcal/mol and a mean z-score of -0.01 (Figure 5.18). Based on the 49.9% GC content of the $^{DRR}$eRNA, we would expect a MFEden of 5.00kcal/mol, thus its exhibited MFEden is notably more negative than expected.

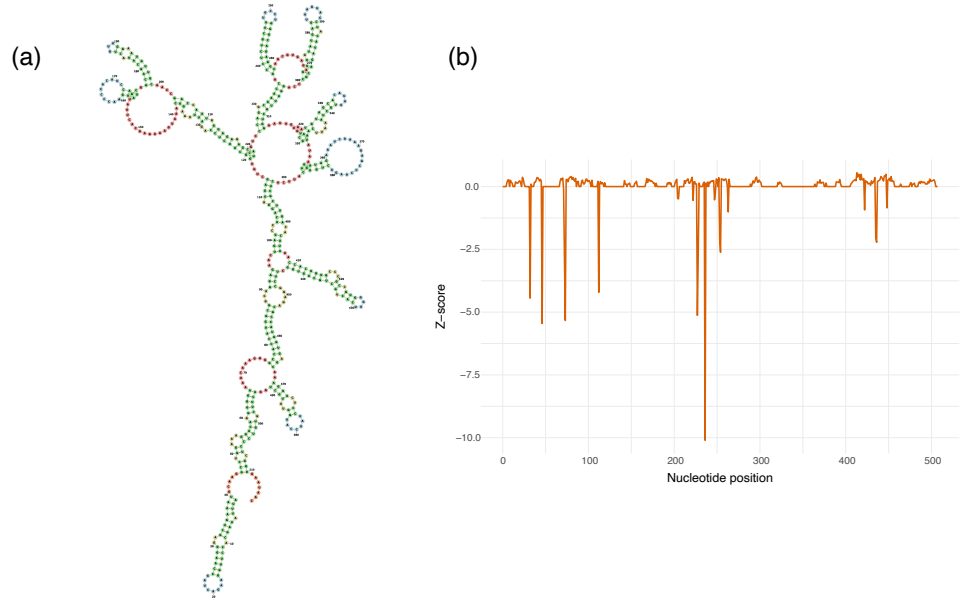



FIGURE 5.18: eRNA structure and stability predictions for the $^{DRR}$eRNA (a) MFE structure prediction for the $^{DRR}$eRNA. (b) Thermodynamic z-score across the length of the $^{DRR}$eRNA.

TABLE 5.1: Predicted stability of $^{DRR}$eRNA

| **RNA** | **MFE (kcal/mol)** | **Observed MFEden (kcal/mol)** | **Expected MFEden (kcal/mol)** | **Mean shuffled MFE (kcal/mol)** | **Mean z-score** | **Min. z-score** |
|---|---|---|---|---|---|---|
| $^{DRR}$eRNA | -133.39 | -0.28 | 5.00 | -126.05 | -0.01 | -10.10 |

The *TAL1* sense eRNA had a predicted MFE of -34.04kcal/mol, with a MFEden of 9.92 kcal/mol and a mean z-score of -0.06 (Figure 5.19a). In contrast, the *TAL1* antisense eRNA had a MFE of -90.52 kcal/mol, an MFEden of 1.54 kcal/mol, and a mean z-score of 0.04 (Figure 5.19b). Based on their respective GC contents, the *TAL1* sense eRNA would be expected to have a MFEden of 5.23kcal/mol, and the antisense eRNA would be expected to have a MFEden of 10.44kcal/mol. Compared to the expected MFEden values, the *TAL1* sense eRNA had a more positive observed MFEden and the antisense eRNA had a more negative observed MFEden. While this may suggest relative thermodynamic instability and stability for the sense and antisense eRNAs, respectively, the antisense eRNA actually demonstrated a more positive z-score when compared to the sense eRNA.

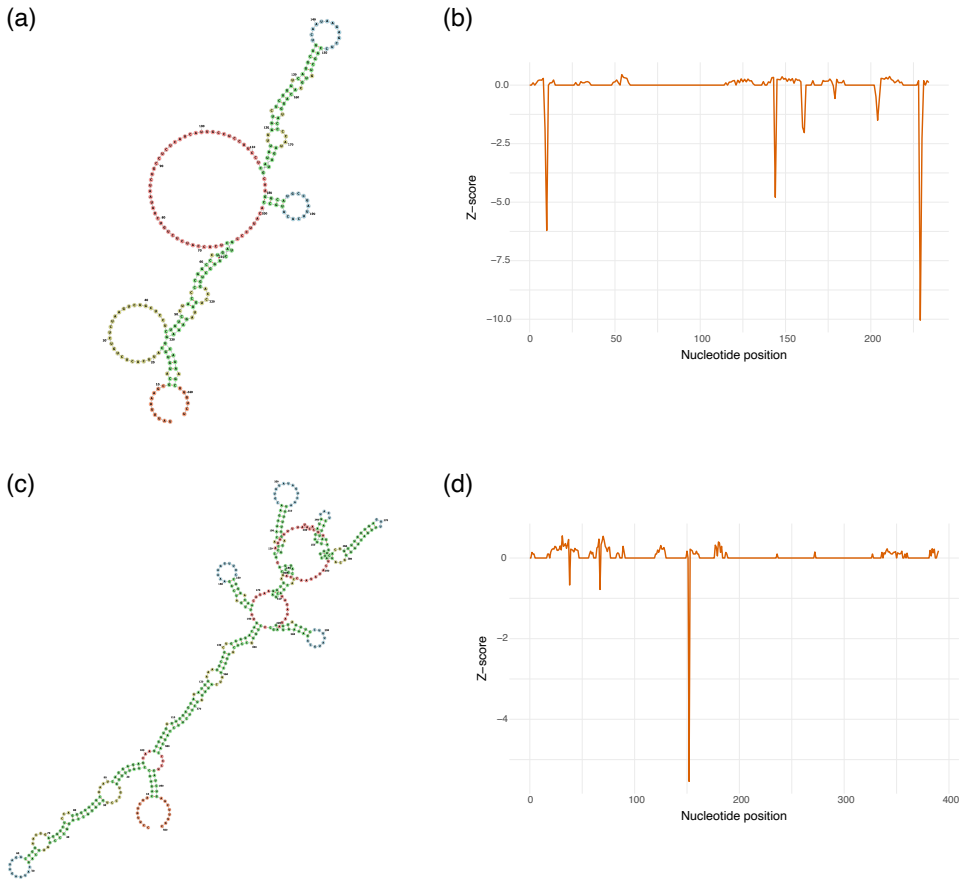



FIGURE 5.19: eRNA structure and stability predictions for the eRNAs produced by the *TAL1* super-enhancer
(a) MFE structure prediction for the sense *TAL1* eRNA. (b) Thermodynamic z-score across the length of the the sense *TAL1* eRNA. (c) MFE structure prediction for the antisense *TAL1* eRNA. (d) Thermodynamic z-score across the length of the the antisense *TAL1* eRNA.

TABLE 5.2: Predicted stability of *TAL1* eRNAs

| **RNA** | **MFE (kcal/mol)** | **Observed MFEden (kcal/mol)** | **Expected MFEden (kcal/mol)** | **Mean shuffled MFE (kcal/mol)** | **Mean z-score** | **Min. z-score** |
|---|---|---|---|---|---|---|
| *TAL1* sense eRNA | -34.04 | 9.92 | 5.23 | -48.95 | -0.06 | -10.04 |
| *TAL1* antisense eRNA | -90.52 | 1.54 | 10.44 | -79.75 | 0.04 | -5.54 |

Visually, we observed more peaks of negative z-score across the length of the $^{DRR}$eRNA compared to the *TAL1* eRNAs, indicative of more stable structural motifs within the

sequence. Further, the $^{DRR}$eRNA had a slightly lower minimum z-score of -10.10, compared to the minimum z-scores of -10.04 and -5.54 for the sense and antisense *TAL1* eRNAs, respectively. To identify if these example eRNAs had more negative predicted MFE values than expected by chance, based on nucleotide composition, for each sequence, we generated 10 random shuffled RNA sequences of the same nucleotide composition and length and calculated the MFE values. We identified that the native sense *TAL1* eRNA sequence had an MFE that was consistently more positive than that of the shuffled sequences. When comparing the native antisense *TAL1* eRNA and $^{DRR}$eRNA MFE values to that of shuffled sequences, we identified that while on average they yielded MFE values more negative than the shuffled sequences, some shuffled sequences did yield more negative MFE values than the native sequences (Figure 5.20).

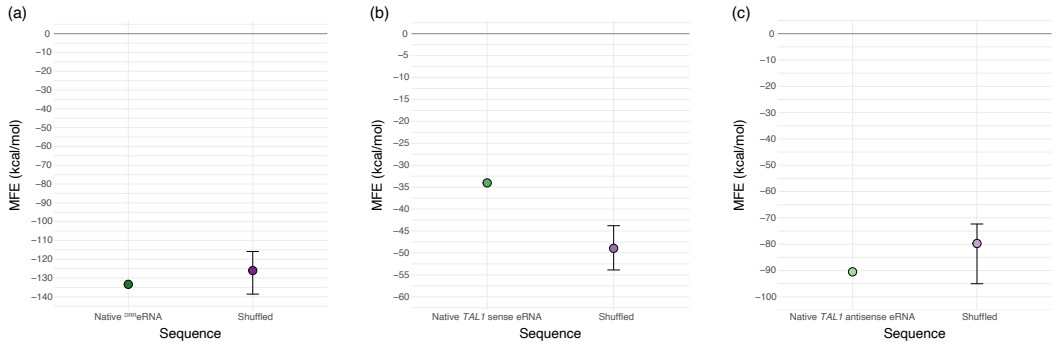



FIGURE 5.20: Native eRNA MFE predictions compared to shuffled randomised control MFE values for example eRNAs
(a) Native $^{DRR}$eRNA MFE compared to MFE values of shuffled RNA sequences (n=10). (b) Native sense *TAL1* eRNA MFE compared to MFE values of shuffled RNA sequences (n=10). (c) Native antisense *TAL1* eRNA MFE compared to MFE values of shuffled RNA sequences (n=10).

### 5.5.5 RNA structure and stability prediction of eRNAs in SH-SY5Y cells to model ALS-associated SNPs

In order to model the potential impact of ALS-associated SNPs on eRNA structure and stability, we identified novel eRNAs assembled from WT SH-SY5Y TT-seq data that overlapped in both replicates and intersected ALS-associated SNPs. We identified the antisense eRNA at chr7:116,199,437-116,199,680 transcribed from the proximal *CAV1/CAV2* candidate enhancer (chr7:116,196,623-116,201,351), that encapsulates the ALS-associated SNP chr7:116,199,522T>C. We predicted the secondary RNA structure of this eRNA, hereafter referred to as *eCAV1_WT* (Figure 5.21a), yielding a MFE of -26.73kcal/mol and an MFEden of 12.69kcal/mol. Moreover, we calculated the thermodynamic z-score across the length of the eRNA sequence, obtaining a mean z-score of -0.032 and a minimum z-score of -6.78 (Figure 5.21b).

To model the impact of the chr7:116,199,522T>C SNP, we computationally induced this SNP, generating an eRNA hereafter referred to as *eCAV1_ALS* and again predicted the RNA structure (Figure 5.21c), yielding a relatively more negative MFE of -29.61kcal/mol and a MFEden of 11.47kcal/mol. In addition to the changes in the MFE, we visually observed changes to the secondary RNA structure upon induction of the ALS-associated SNP. The large multiloop, containing two small hairpin loops and a helix in the WT eRNA sequence, upon induction of the ALS SNP, is replaced with a smaller multiloop containing two helices and an interior loop and a larger hairpin loop branching off the multiloop. Further, there is a larger difference between the predicted MFEden and the exhibited MFEden for the eRNA containing the ALS-associated SNP than that of the WT sequence, with predicted MFEden values of 16.84kcal/mol and 17.00kcal/mol for the mutated and WT sequences, respectively.

In concordance with the reduction in MFE upon induction of the mutation, the mean z-score across the eRNA sequence containing the ALS-associated SNP is -0.066, with a minimum z-score of -10.1 (Figure 5.21d). Of note, the minimum z-score in the WT eRNA sequence, and hence the most stable structure within the eRNA, is attributed to the stem loop 84 nt into the RNA sequence. On the other hand, the minimum z-score in the eRNA sequence containing the ALS-associated SNP can be attributed to the larger stemloop structure that is absent in the WT eRNA structure.

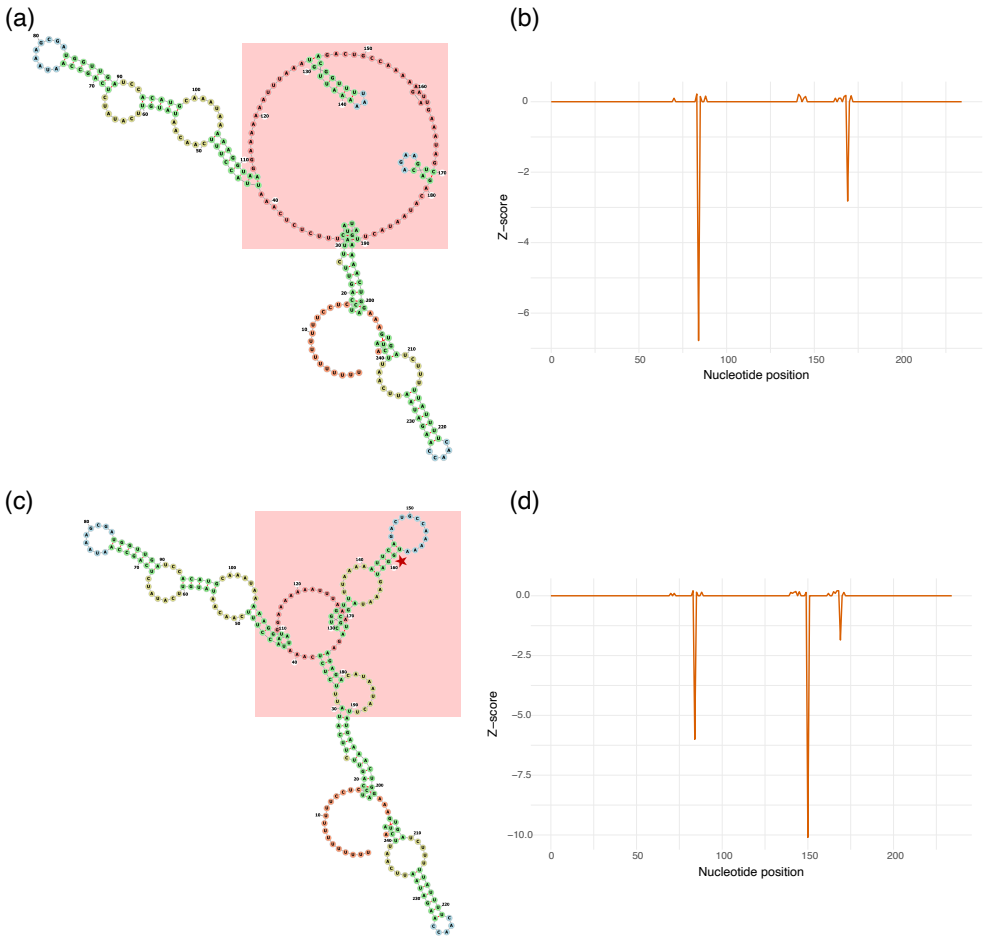



FIGURE 5.21: eRNA structure and stability predictions for the novel ALS-associated antisense eRNA in SH-SY5Y cells
(a) MFE structure prediction for the novel *CAV1* enhancer-associated antisense eRNA *eCAV1_WT* (chr7:116,199,437-116,199,680). (b) Thermodynamic z-score across the length of *eCAV1_WT*. (c) MFE structure prediction for the *CAV1* enhancer-associated antisense eRNA upon computational induction of the chr7:116,199,522T>C ALS-associated SNP, *eCAV1_ALS*. Location of SNP is denoted by the red star. (d) Thermodynamic z-score across the length of *eCAV1_ALS*. SNP-dependent structural changes are highlighted with red boxes.

TABLE 5.3: Predicted stability of ALS-associated eRNAs in SH-SY5Y cells

| **RNA** | **MFE (kcal/mol)** | **Observed MFEden** | **Expected MFEden** | **Mean shuffled MFE** | **Mean z-score** | **Min. z-score** |
|---|---|---|---|---|---|---|
| *eCAV1_WT* | -26.73 | 12.69 | 17.00 | -25.30 | -0.032 | -6.78 |
| *eCAV1_ALS* | -39.61 | 11.47 | 16.84 | -24.44 | -0.066 | -10.10 |

To identify the structural stability of this novel antisense eRNA, and to identify if this stability is altered by the computational induction of the chr7:116,199,522T>C ALS-associated SNP, we compared the MFE values of each RNA to that of 10 random

shuffled RNA sequences of the same nucleotide composition and length. While the native WT eRNA sequence generated a MFE similar to that of the average of the randomised shuffled sequences (Figure 5.22a), the eRNA sequence containing chr7:116,199,522T>C generated an MFE that was consistently lower than that of the randomised shuffled sequences (Figure 5.22b).

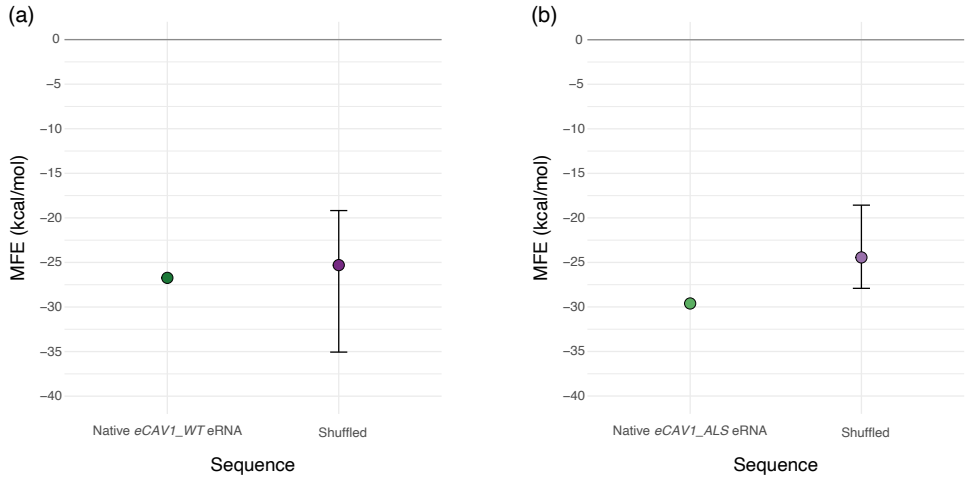



FIGURE 5.22: eRNA MFE predictions compared to shuffled randomised control for native SH-SY5Y antisense eRNA and eRNA with induced ALS-associated SNP
(a) Native WT antisense SH-SY5Y *eCAV1_WT* eRNA (chr7:116,199,437–116,199,680) MFE compared to MFE values of shuffled RNA sequences ($n = 10$).
(b) MFE of antisense SH-SY5Y eRNA (chr7:116,199,437–116,199,680) with computationally induced chr7:116,199,522T→C mutation (*eCAV1_ALS*) compared to MFE values of shuffled RNA sequences ($n = 10$).

Taken together, this provides an example of an ALS-associated SNP that could considerably alter both the structure and stability of an eRNA.

## 5.6 eRNA transcription in human-derived lymphoblastoid cells

To identify alterations in eRNA transcription in ALS patients harbouring *CAV1* enhancer mutations, we performed TT-seq in human-derived lymphoblastoid cells. We compared the healthy control line, SC3504 to two ALS patient-derived cell lines: SP3363, from an ALS patient with the enhancer mutation chr7:116,222,625T>C and SP3388, from an ALS patient with the enhancer mutation chr7:115,994,269:C>T. Normalisation was performed as for SH-SY5Y cells, using scaling factors of 0.9787169, 1.0846587 and 0.9269028, for SC3504, SP3363 and SP3388, respectively. To provide insight into

the *CAV1* enhancer landscape in ALS disease models, we examined the TT-seq coverage over *CAV1/CAV2* enhancers in control- and patient-derived lymphoblastoid cells (Figure 5.23a).

Overall, we observed moderate uniformity in enhancer activity between cell lines, however for some loci we observed a prominent increase in enhancer transcription in the ALS patient-derived lymphoblastoid cell lines compared to the healthy control. Of note, enhancers around the *CAV1* gene locus appeared to be more highly transcribed in SP3363 and SP3388 cells than in the control SC3504 cells. Despite the apparent increase in transcription levels at particular enhancer loci in ALS patient-derived cells, more enhancers demonstrated TT-seq read coverage in SC3504 cells than in SP3363 and SP3388 cells (Figure 5.23b). The healthy control-derived cells, SC3504, demonstrated TT-seq read coverage over 76 enhancers, compared to 68 and 54 enhancers for the ALS patient-derived cell lines, SP3363 and SP3388, respectively.

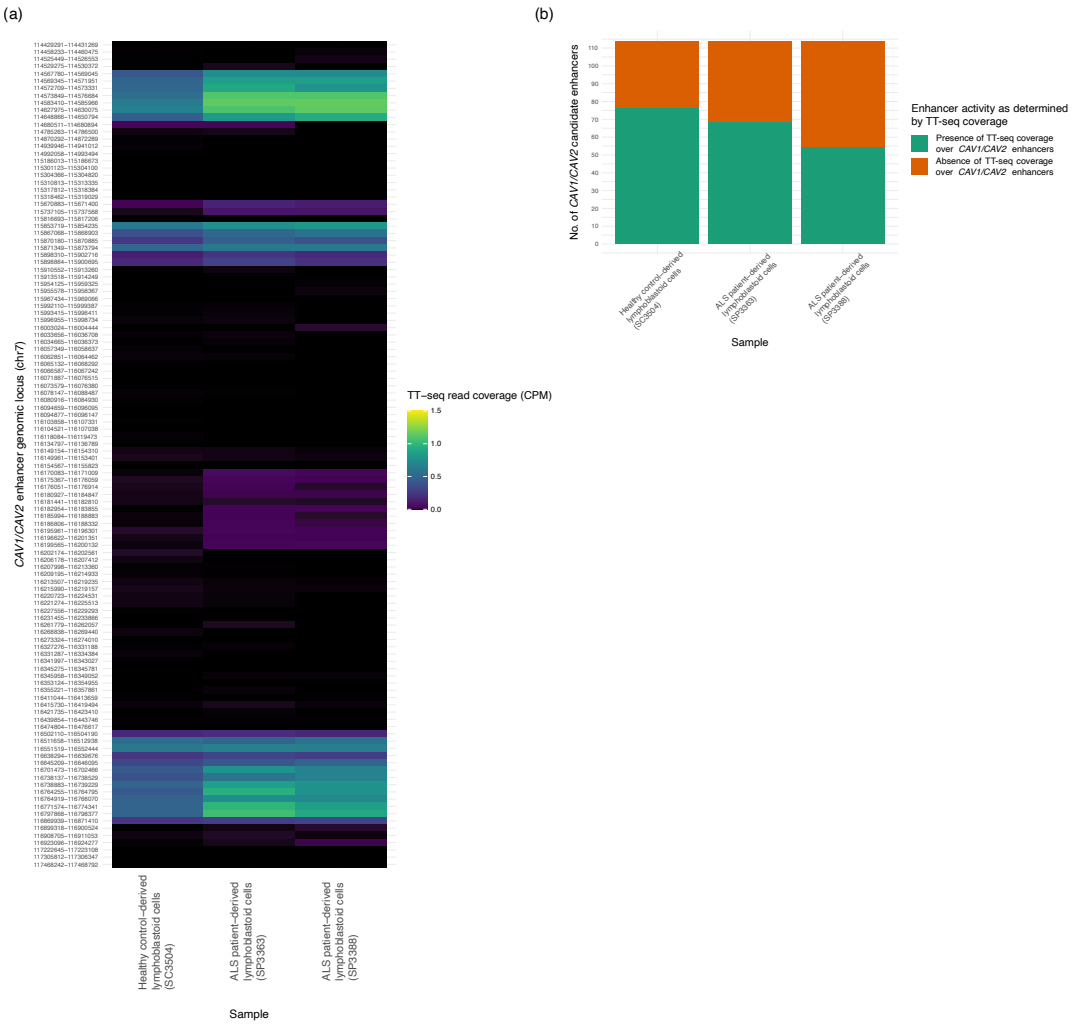



FIGURE 5.23: eRNA transcription at *CAV1/CAV2* enhancers in human-derived lymphoblastoid cells

(a) TT-seq read coverage in CPM at the *CAV1/CAV2* candidate enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) in healthy control- and ALS patient-derived lymphoblastoid cells. (b) Number of candidate *CAV1/CAV2* enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) considered to be active in lymphoblastoid cells, as determined by the presence of TT-seq coverage over enhancer regions.

To further investigate transcriptional differences at candidate *CAV1/CAV2* enhancers in lymphoblastoid cell lines, we quantified TT-seq read coverage across these enhancer regions (Figure 5.24). We found that transcription levels were largely similar between the two ALS patient-derived cell lines, SP3363 and SP3388, while the healthy control cell line, SC3504, exhibited generally lower transcriptional activity across the *CAV1/CAV2* enhancer landscape. Specifically, the average TT-seq coverage over enhancers for SP3363 and SP3388 was 0.161 CPM and 0.155 CPM, respectively, whereas SC3504 displayed a markedly reduced mean coverage of 0.099 CPM, indicating a relative decrease in enhancer transcription in the healthy control.

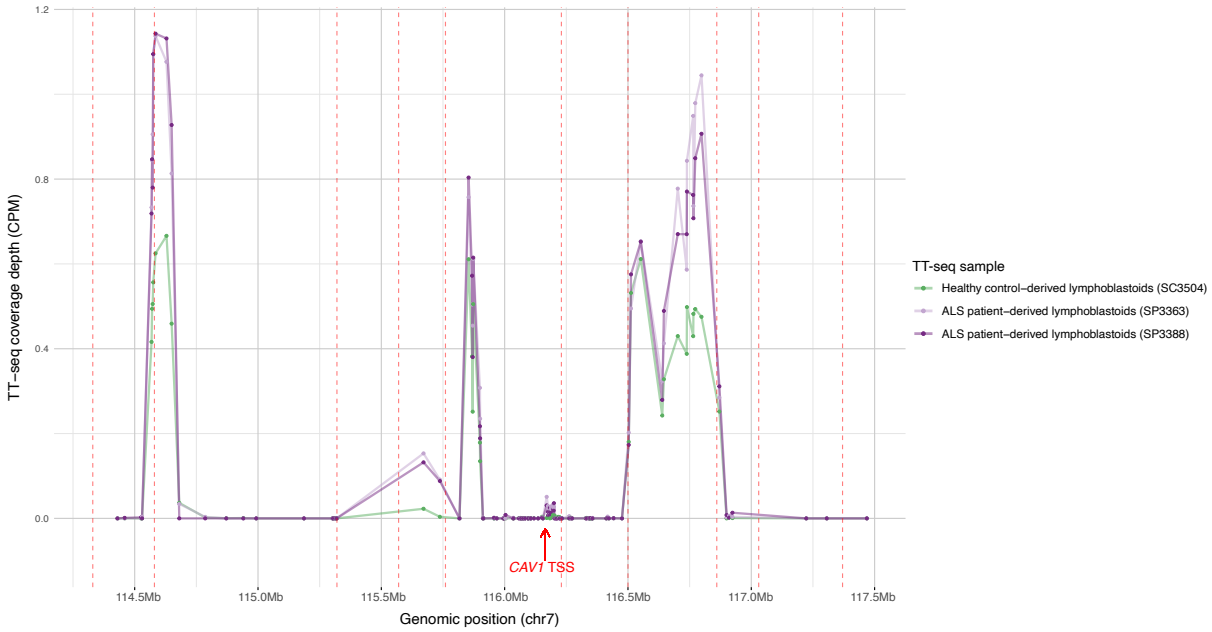



FIGURE 5.24: Levels of nascent transcription over candidate *CAV1/CAV2* enhancers in lymphoblastoid cells
TT-seq read coverage in CPM at the *CAV1/CAV2* candidate enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) in healthy control- and ALS patient-derived lymphoblastoid cells. Red dashed lines denote TAD boundaries

Interestingly, we observed a ∼2.4-fold decrease in the average transcript coverage over the candidate enhancers containing the chr7:116,222,625T>C mutation in the SP3363 lymphoblastoid cells, compared to the healthy control-derived SC3504 lymphoblastoid cells (Figure 5.25a). Conversely, while the candidate enhancers containing the chr7:115,994,269:C>T SNP demonstrated an average transcriptional coverage of ∼6.04e-04 CPM in the SC3504 cell line, we did not observe any transcription over these enhancers in the respective lymphoblastoid cell line containing this mutation, SP3388 (Figure 5.25b).

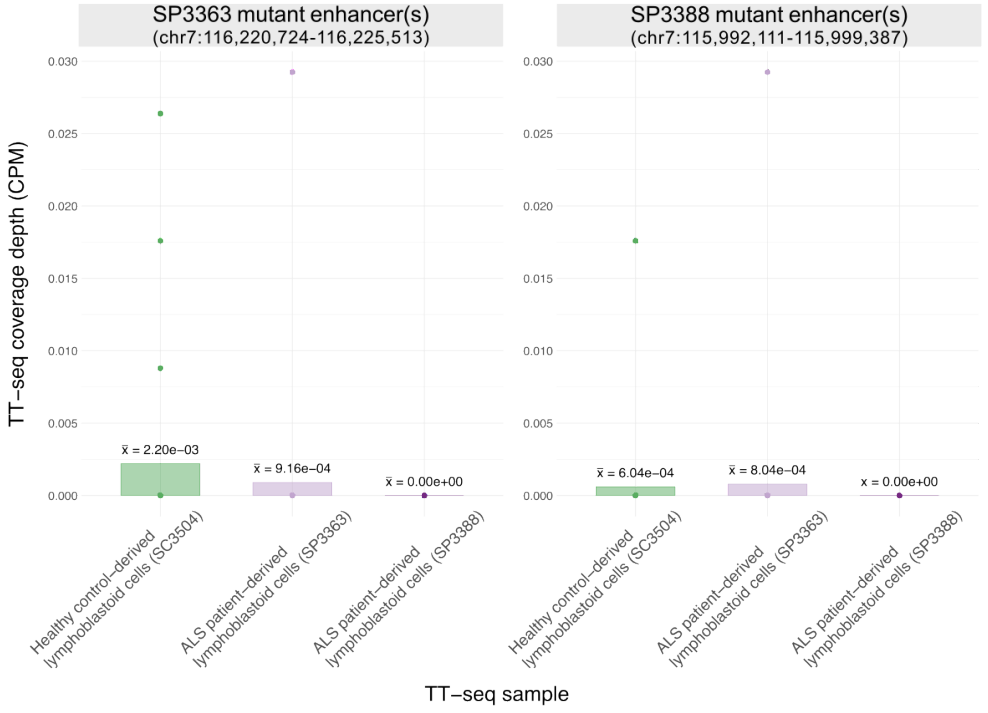



FIGURE 5.25: eRNA transcription at ALS-associated *CAV1/CAV2* enhancers in lymphoblastoid cells

(a) TT-seq read coverage (CPM) in healthy control- and ALS patient-derived lymphoblastoid cells at the candidate *CAV1/CAV2* enhancer(s) containing the chr7:116,222,625T>C mutation (chr7:116,220,724-116,225,513). (b) TT-seq read coverage (CPM) in healthy control- and ALS patient-derived lymphoblastoid cells at the candidate *CAV1/CAV2* enhancer(s) containing the chr7:115,994,269:C>T mutation (chr7:115,992,111-115,999,387). Bars represent mean coverage. Scatter represents TT-seq coverage at single-nucleotide resolution across the enhancers.

## 5.7 Identification of novel enhancer RNAs in human-derived lymphoblastoid cells

As for SH-SY5Y cells, we also assembled *de novo* RNA transcripts from healthy control- and ALS patient-derived lymphoblastoid TT-seq data, filtering as before to identify potential eRNA transcripts. We identified 7 novel potential eRNA transcripts in the SC3504 cell line, 11 in the SP3363 cell line and 10 in the SP3388 cell line (Figure 5.26). Thus, in agreement with the relative reduction in bulk TT-seq read coverage in the healthy control-derived cell line compared to the ALS patient-derived cell line, we also observed a reduction in novel assembled eRNA transcripts in the healthy control-derived cells.

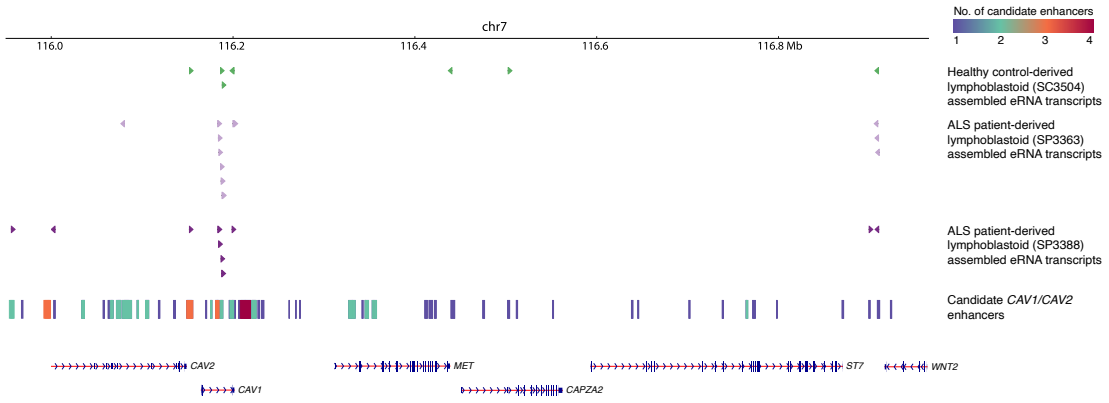



FIGURE 5.26: eRNA transcript assembly in lymphoblastoid cell lines
Trackline view of chr7:115,950,000-116,965,000 (hg19) demonstrating the genomic locations of *de novo* assembled eRNA transcripts from TT-seq in control- and patient-derived lymphoblastoid cells, relative to candidate *CAV1/CAV2* enhancers (Cooper-Knock et al., 2021, Fulco et al., 2019) and genic locations. Candidate enhancers are coloured for clarity due to collapse in the track line.

### 5.7.1 eRNA transcription in healthy control-derived lymphoblastoid cells

Firstly, in order to discern eRNAs that may be associated with healthy control participants, in the absence of disease phenotype, we identified eRNAs that were assembled solely in the healthy control-derived lymphoblastoid cell line, concentrating specifically on enhancers that do not produce transcripts that could be assembled into novel eRNAs in the ALS patient-derived cell lines. Of this nature, we found 2 eRNAs, one antisense eRNA at chr7:116,440,664-116,440,720, transcribed from the *CAV1* enhancer upstream of the *CAPZA2* gene (chr7:116,439,855-116,443,746), hereafter referred to as *eCAPZA2_antisense*, and one sense eRNA at chr7:116,502,612-116,502,934, transcribed from the intronic enhancer within the *CAPZA2* gene (chr7:116,502,110-116,504,190), hereafter referred to as *eCAPZA2_sense* (Figure 5.27). *eCAPZA2_antisense* was particularly short in length and yielded a MFE of -11.35kcal/mol but a MFEden of -1.22kcal/mol (Figure 5.28a). As expected of a short RNA, the mean z-score across the length of *eCAPZA2_antisense* was 0.08, with a minimum z-score of only -1.24 (Figure 5.28b). Despite a positive mean z-score, *eCAPZA2_antisense* demonstrated a more negative MFEden than the expected MFEden of -0.12kcal/mol, based on its GC content. In contrast, *eCAPZA2_sense* yielded an MFE of -142.01kcal/mol and a MFEden of -19.40kcal/mol (Figure 5.28c). *eCAPZA2_sense* also appeared to have increased stability in comparison to the antisense eRNA, with a mean z-score of -0.32 across its length and a minimum z-score of -6.78 (Figure 5.28d). Further, *eCAPZA2_sense* demonstrated a

more pronounced reduction in MFEden value compared to the expected MFEden value of -15.73kcal/mol.

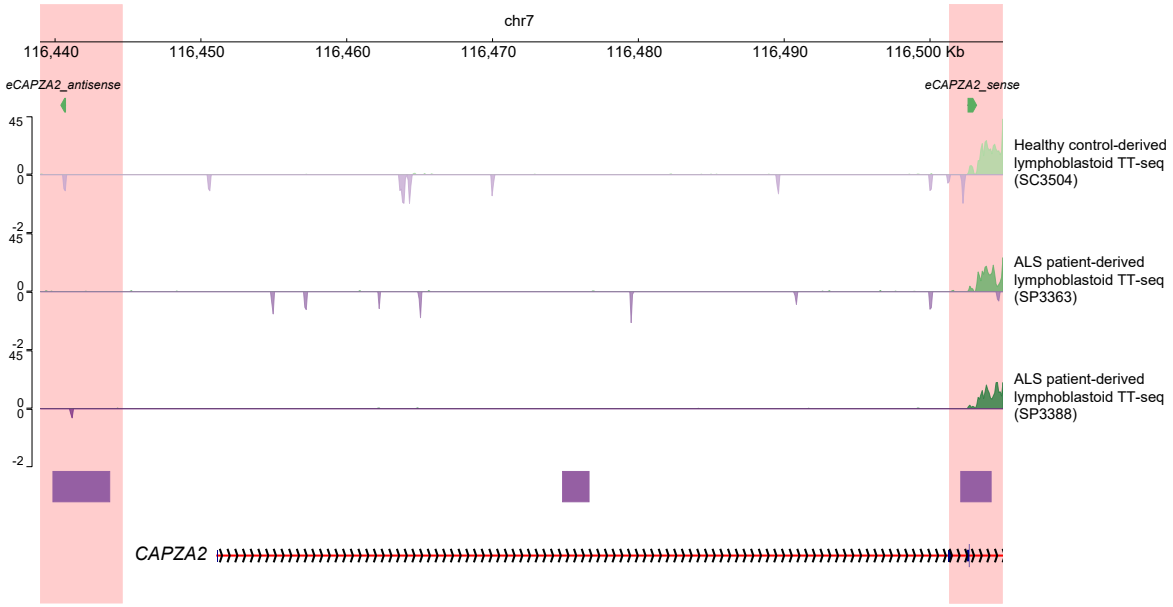



FIGURE 5.27: eRNA transcription at healthy control-associated enhancers Trackline at chr7:116,339,000-116,605,000 (hg19) demonstrating assembled eRNA transcripts from healthy control-derived lymphoblastoid cells, TT-seq coverage in lymphoblastoid cells, including both healthy control- and ALS patient-derived cell lines, candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019). Enhancers of interest highlighted in red.

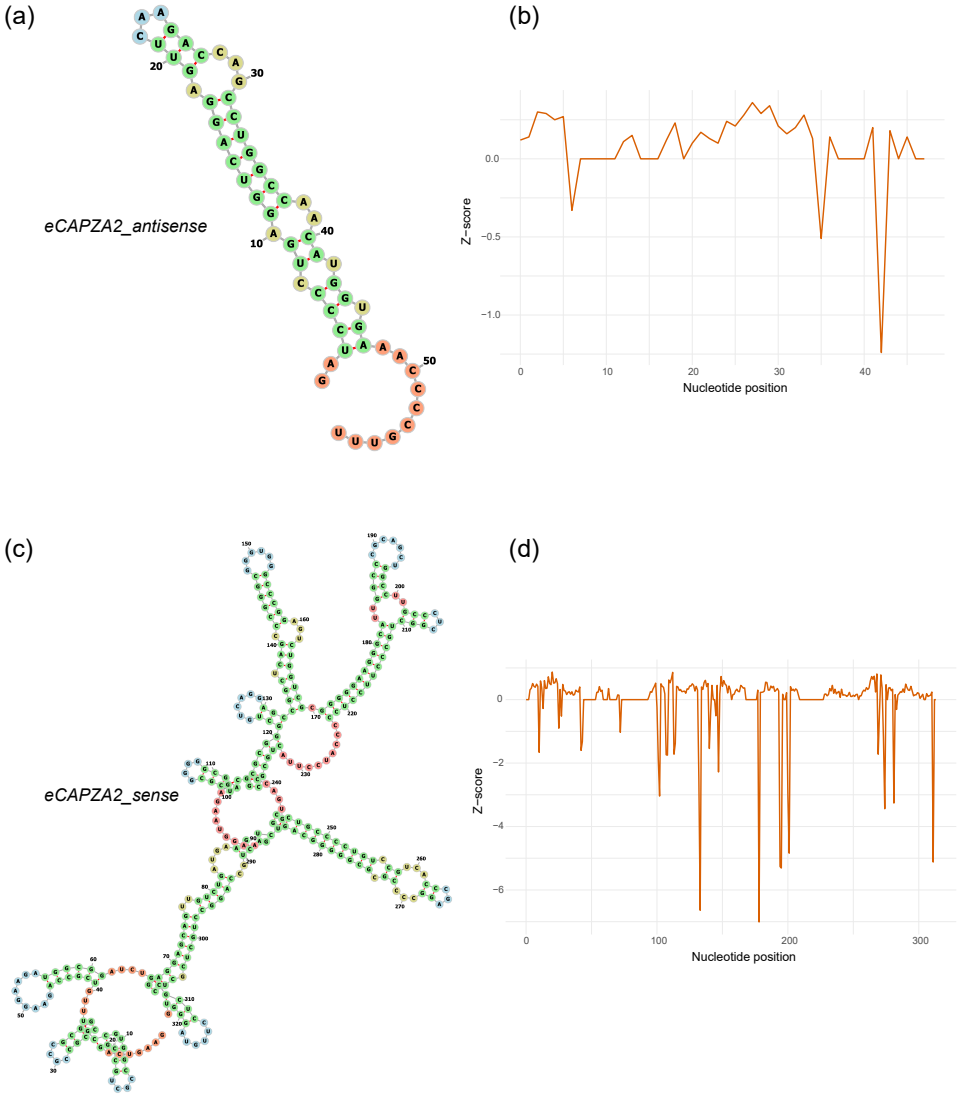



FIGURE 5.28: eRNA structure and stability predictions for eRNAs identified only in healthy control-derived lymphoblastoid cells

(a) MFE structure prediction for the novel antisense eRNA at chr7:116,440,664-116,440,720. (b) Thermodynamic z-score across the length of the antisense eRNA. (c) MFE structure prediction for the novel sense eRNA at chr7:116,502,612-116,502,934. (d) Thermodynamic z-score across the length of the sense eRNA.

TABLE 5.4: Predicted stability of eRNAs in healthy control-derived lymphoblastoid cells

| **RNA** | **MFE (kcal/mol)** | **Observed MFEden (kcal/mol)** | **Expected MFEden (kcal/mol)** | **Mean shuffled MFE (kcal/mol)** | **Mean z-score** | **Min. z-score** |
|---|---|---|---|---|---|---|
| *eCAPZA2_antisense* | -11.35 | -1.22 | -0.12 | -11.75 | 0.08 | -1.24 |
| *eCAPZA2_sense* | -142.01 | -19.40 | -15.73 | -149.98 | -0.32 | -6.78 |

To further assess the stability of these eRNAs assembled from healthy control-derived lymphoblastoid cells, we compared the MFE values to that of 10 random shuffled RNA sequences of the same nucleotide compositions and lengths. Neither of the eRNAs generated an MFE of more negative magnitude than that of the shuffled sequences (Figure 5.29, with the sense eRNA producing a MFE that was more positive on average than that of the shuffled sequences.

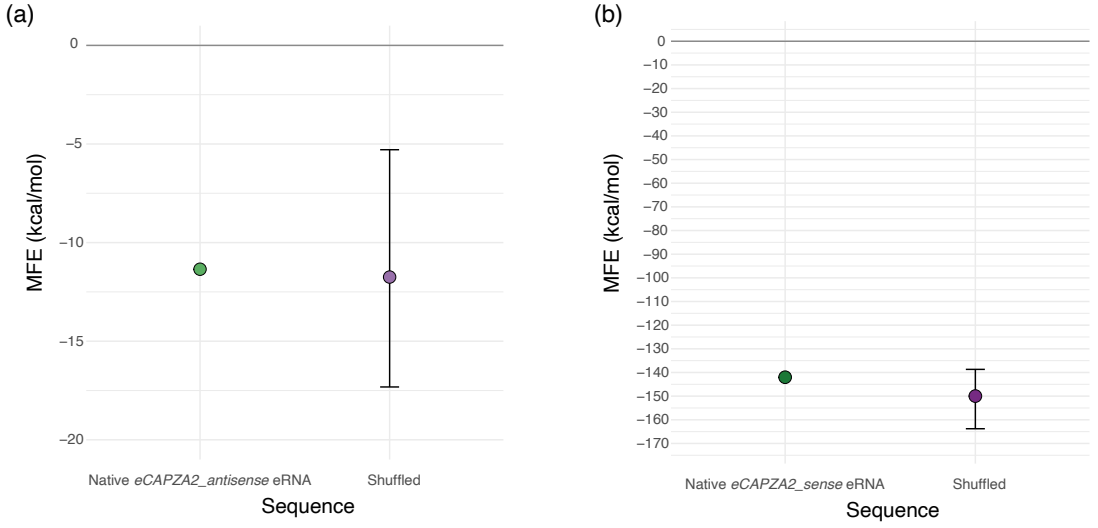



FIGURE 5.29: MFE predictions of native eRNAs assembled from SC3504 TT-seq data compared to shuffled randomised control MFE values
(a) Native antisense eRNA (chr7:116,440,664-116,440,720) MFE compared to MFE values of shuffled RNA sequences (n=10). (b) Native sense eRNA (chr7:116,502,612-116,502,934) MFE compared to MFE values of shuffled RNA sequences (n=10).

Despite identifying novel assembled eRNA transcripts over these enhancers, it is important to note that we still identified some nascent transcription over these enhancers in the SP3363 and SP3388 cell lines. Interestingly, the level of TT-seq read coverage

over these enhancers was somewhat similar when comparing the healthy control-derived lymphoblastoid cells to the ALS patient-derived lymphoblastoid cells (Figure 5.30). At the candidate enhancer upstream of *CAPZA2* that produced *eCAPZA2_antisense*, there was only a 1.2 fold and 1.3 fold more TT-seq read coverage over this enhancer in the SC3504 cells than in the SP3363 and SP388 cells, respectively. Further, at the enhancer within the *CAPZA2* intron that produced the *eCAPZA2_sense*, the level of nascent transcription was approximately equal in all cell lines. Notably, of the 14 enhancers that had TT-seq read coverage exclusively in SC3504 cells, no eRNA transcripts were assembled over these enhancers.

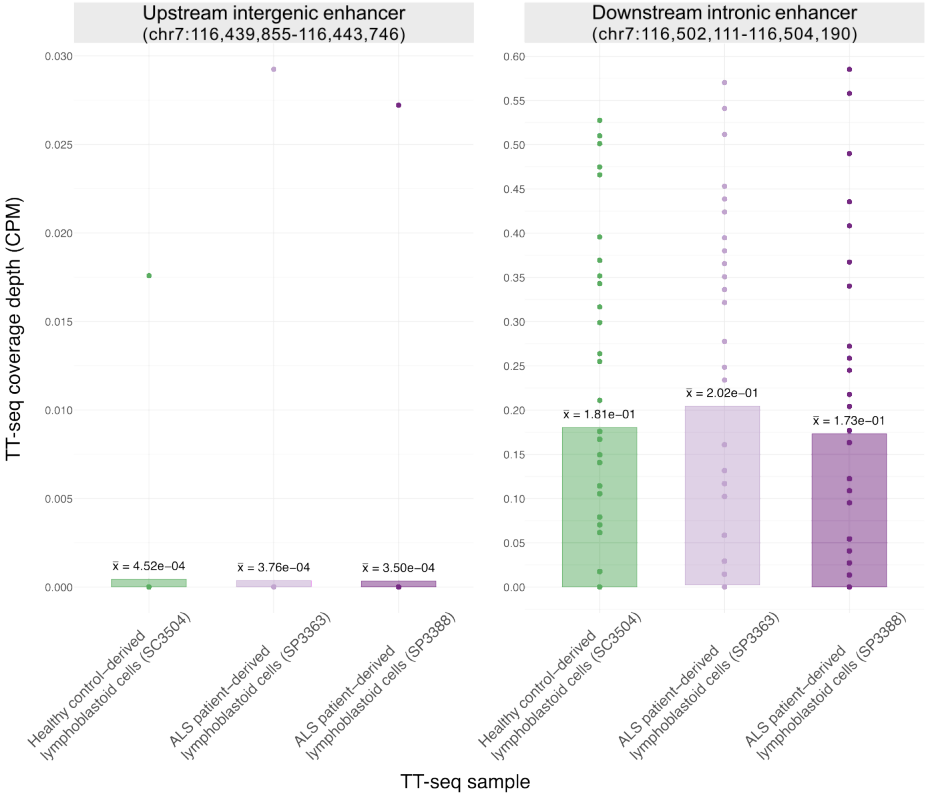



FIGURE 5.30: TT-seq coverage at healthy control-associated enhancers
TT-seq read coverage (CPM) in healthy control- and ALS patient-derived lymphoblastoid cells at the intergenic enhancer upstream of *CAPZA2* (chr7:116,439,855-116,443,746) and at the intronic enhancer within *CAPZA2* (chr7:116,502,110-116,504,190), that generate eRNAs in healthy control-derived lymphoblastoid cells. Bars represent mean coverage. Scatter represents TT-seq coverage at single-nucleotide resolution across the enhancers.

While we have identified novel eRNAs specific to the healthy control-derived lymphoblastoid cell line, potentially indicative that these eRNAs are specific to a healthy state, we observed similar coverage over the enhancers yielding these eRNAs in ALS patient-derived lymphoblastoid TT-seq data. These findings therefore raise questions about the specificity of these eRNAs to a healthy state.

### 5.7.2 eRNA transcription in ALS patient-derived lymphoblastoid cells

Secondly, we explored the structure and stability of novel eRNAs assembled solely in ALS patient-derived lymphoblastoid cells, identifying eRNAs transcribed from enhancers that do not produce any assembled eRNA transcripts in healthy control-derived cells. However, we acknowledge that a key limitation of this comparison is the use of only one healthy control-derived line, compared to two distinct ALS patient-derived lines. As such, we cannot definitively conclude that the absence of certain eRNAs in the control reflects a general feature of the healthy state, rather than an individual-specific genotype effect. Without additional control lines, it is difficult to generalise these findings to the broader healthy population.

At the overlapping intronic candidate enhancers within *CAV1* at chr7:116,180,928-116,184,847 and chr7:116,182,955-116,183,855, predicted to be a tissue-agnostic *CAV2* enhancer and a CNS-specific *CAV1* enhancer, respectively, we identified multiple novel assembled eRNA transcripts, that appear to be specific to the ALS patient-derived cell lines. We identified overlapping sense eRNA transcripts at chr7:116,183,048-116,183,530 in SP3363 cells and at chr7:116,183,164-116,183,736 in SP3388 cells, hereafter referred to as *eCAV1_3363_us* and *eCAV1_3388_us*, respectively, both of which overlap an assembled sense eRNA transcript from WT SH-SY5Y cells. Downstream, we identified a further two overlapping sense eRNA transcripts at chr7:116,183,914-116,184,117 in SP3363 cells and chr7:116,183,989-116,184,385 in SP3388 cells, hereafter referred to as *eCAV1_3363_ds* and *eCAV1_3388_ds*, respectively (Figure 5.31).

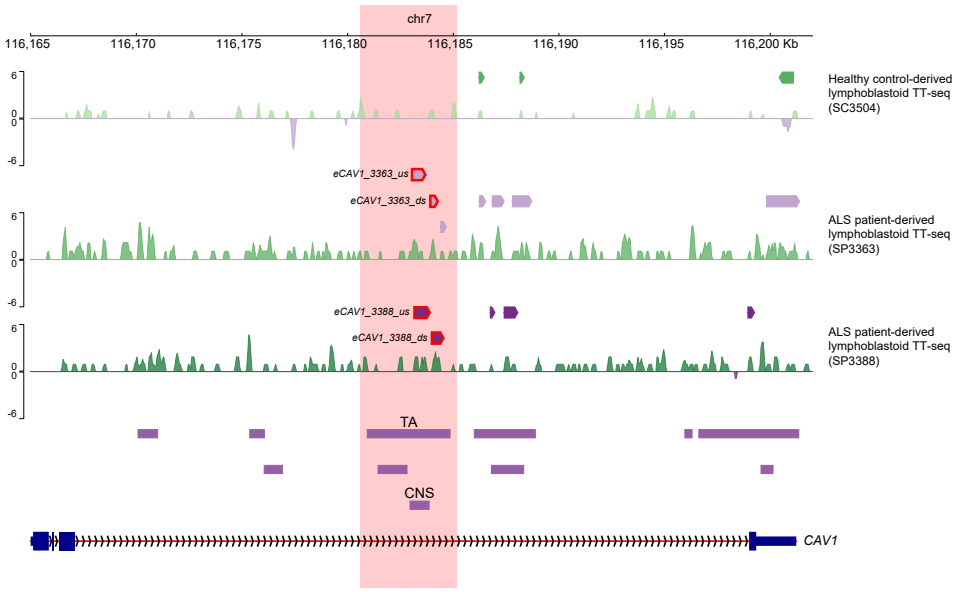



FIGURE 5.31: eRNA transcription at ALS patient-associated enhancers
Trackline at chr7:116,165,000-116,202,000 (hg19) demonstrating assembled eRNA transcripts and TT-seq coverage in both healthy control- and ALS patient-derived lymphoblastoid cells, candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019). Enhancers of interest highlighted in red. Abbreviations: TA - tissue-agnostic; CNS - CNS-specific

*eCAV1_3363_us* yielded a MFE of -90.65kcal/mol, with a MFEden of 5.74kcal/mol, a value more negative than the expected MFEden of 12.13kcal/mol (Figure 5.32a). This eRNA exhibited a mean z-score of -0.002 and a minimum z-score of -7.68 (Figure 5.32b). By comparison, *eCAV1_3388_us* appeared to have increased stability, yielding a MFE of -113.97kcal/mol and a lower MFEden of 5.38kcal/mol, despite a higher expected MFEden of 13.00kcal/mol (Figure 5.32c). In agreement, this eRNA also demonstrated a mean z-score of -0.02 and a minimum z-score of -10.1 (Figure 5.32d).

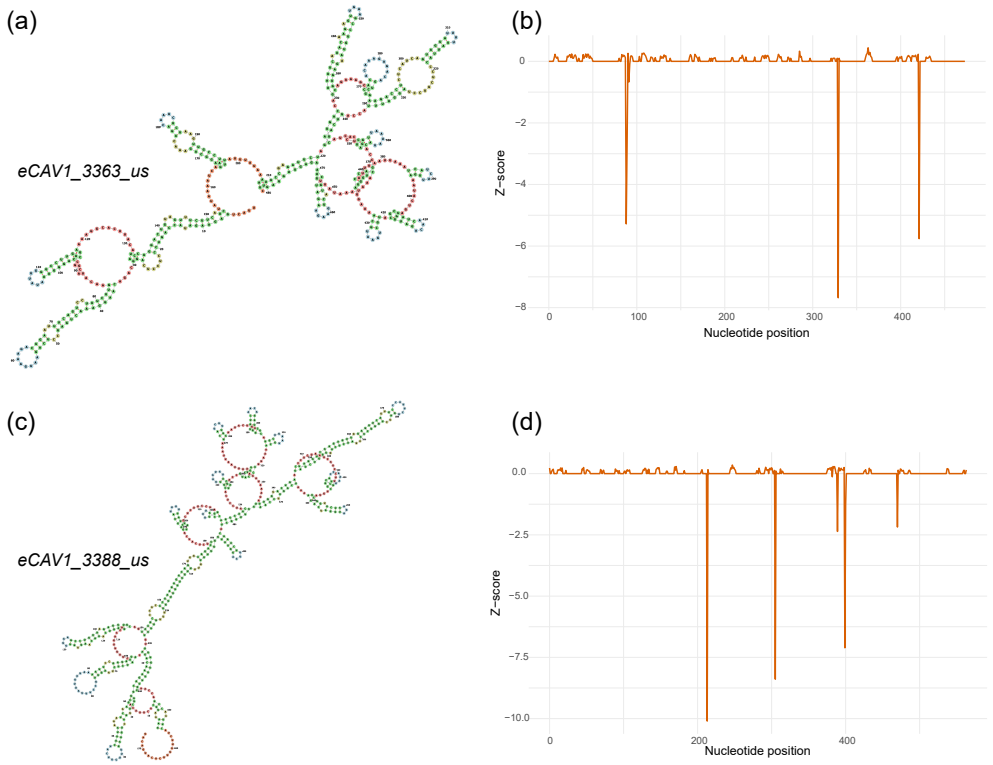



FIGURE 5.32: eRNA structure and stability predictions for eRNAs identified only in ALS patient-derived lymphoblastoid cells
(a) MFE structure prediction for the novel *eCAV1_3363_us* eRNA (chr7:116,183,048-116,183,530), assembled from TT-seq in SP3363 cells. (b) Thermodynamic z-score across the length of *eCAV1_3363_us*. (c) MFE structure prediction for the novel *eCAV1_3388_us* eRNA (chr7:116,183,164-116,183,736), assembled from TT-seq in SP3388 cells. (d) Thermodynamic z-score across the length of *eCAV1_3388_us*.

TABLE 5.5: Predicted stability of upstream eRNAs in ALS patient-derived lymphoblastoid cells

| **RNA** | **MFE (kcal/mol)** | **Observed MFEden (kcal/mol)** | **Expected MFEden (kcal/mol)** | **Mean shuffled MFE (kcal/mol)** | **Mean z-score** | **Min. z-score** |
|---|---|---|---|---|---|---|
| *eCAV1_3363_us* | -90.65 | 5.74 | 12.13 | -78.51 | 0.00 | -7.68 |
| *eCAV1_3388_us* | -113.97 | 5.38 | 13.00 | -98.75 | -0.02 | -10.1 |

While we could not discern whether the *eCAV1_3363_us* and *eCAV1_3388_us* eRNAs were transcribed as a result of activation of the tissue-agnostic *CAV2* enhancer or the CNS-specific *CAV1* enhancer, the *eCAV1_3363_ds* and *eCAV1_3388_ds* eRNAs intersect solely with the tissue-agnostic *CAV2* enhancer. These eRNAs were predicted to be overall less stable than the upstream eRNAs. *eCAV1_3363_ds* yielded a MFE prediction of -27.75kcal/mol and a MFEden of 9.41kcal/mol, compared to an expected

MFEden of 12.35kcal/mol (Figure 5.33a). Regarding predicted stability, this eRNA exhibited a mean z-score of -0.02, with a minimum z-score of only -1.30 (Figure 5.33b). *eCAV1_3388_ds* yielded a MFE of -52.60kcal/mol, with a MFEden of 11.80kcal/mol compared to the expected MFEden of 14.24kcal/mol (Figure 5.33c). Further this eRNA exhibited a mean z-score of 0.01 and a minimum z-score of -3.10 (Figure 5.33d).

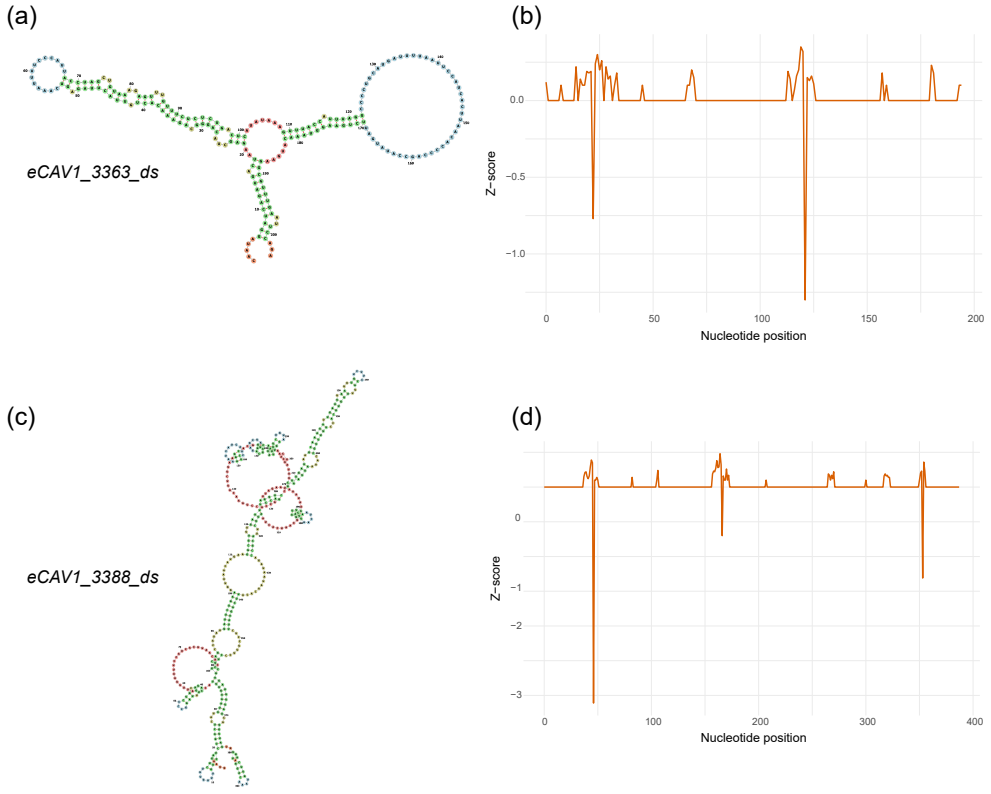



FIGURE 5.33: eRNA structure and stability predictions for eRNAs identified only in ALS patient-derived lymphoblastoid cells
(a) MFE structure prediction for the novel *eCAV1_3363_ds* eRNA (chr7:116,183,914-116,184,117), assembled from TT-seq in SP3363 cells. (b) Thermodynamic z-score across the length of *eCAV1_3363_ds*. (c) MFE structure prediction for the novel *eCAV1_3388_ds* eRNA (chr7:116,183,989-116,184,385), assembled from TT-seq in SP3388 cells. (d) Thermodynamic z-score across the length of *eCAV1_3388_ds*.

TABLE 5.6: Predicted stability of downstream eRNAs in ALS patient-derived lymphoblastoid cells

| **RNA** | **MFE (kcal/mol)** | **Observed MFEden (kcal/mol)** | **Expected MFEden (kcal/mol)** | **Mean shuffled MFE (kcal/mol)** | **Mean z-score** | **Min. z-score** |
|---|---|---|---|---|---|---|
| *eCAV1_3363_ds* | -27.75 | 9.41 | 12.35 | -28.02 | -0.02 | -1.30 |
| *eCAV1_3388_ds* | -52.60 | 11.80 | 14.24 | -52.79 | 0.01 | -3.10 |

To further assess the stability of these eRNAs assembled from ALS patient-derived lymphoblastoid TT-seq data, we compared the MFE values of each eRNA to that of 10 random shuffled RNA sequences of the same nucleotide composition and length. The upstream eRNAs, *eCAV1_3363_us* and *eCAV1_3388_us*, both generated MFE values consistently more negative than that of the shuffled sequences (Figure 5.34a,b). In contrast, the downstream eRNAs, *eCAV1_3363_ds* and *eCAV1_3388_ds*, generated MFE values similar to that of the shuffled sequences (Figure 5.34c,d)

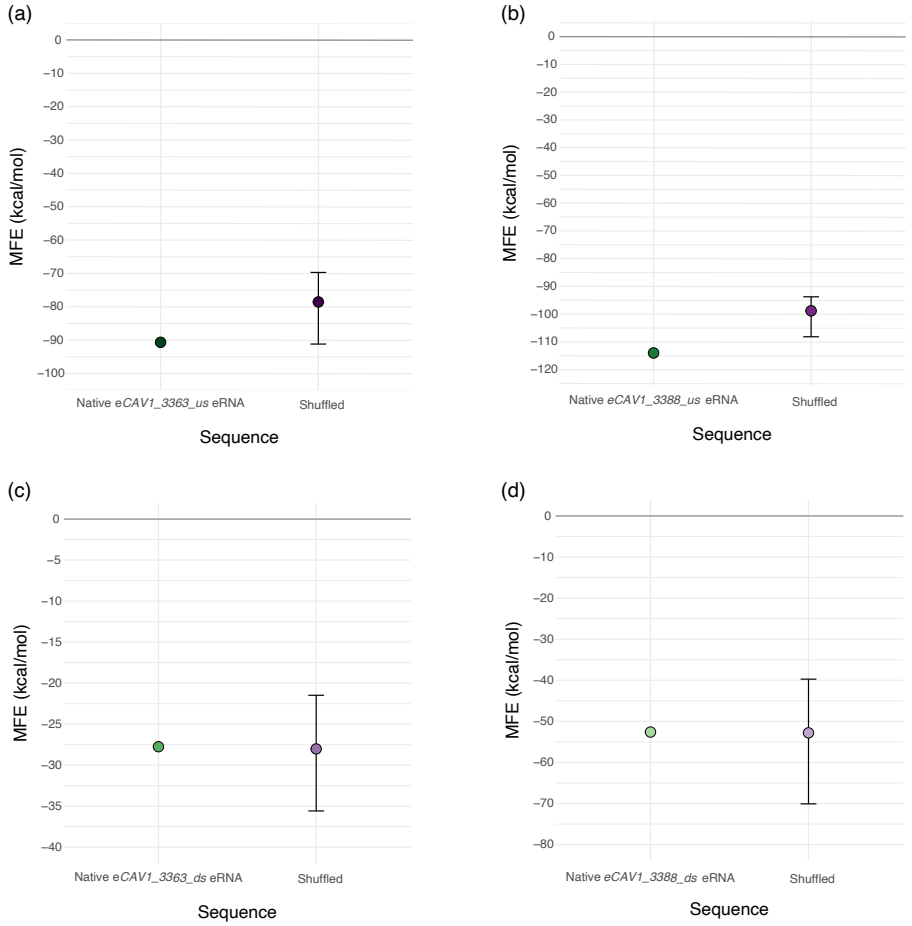



FIGURE 5.34: MFE predictions of native eRNAs assembled from ALS patient-derived lymphoblastoid TT-seq data compared to shuffled randomised control MFE values (a) The MFE of the native upstream eRNA (chr7:116,183,048-116,183,530), assembled from SP3363 TT-seq data compared to MFE values of shuffled RNA sequences (n=10). (b) The MFE of the native upstream eRNA (chr7:116,183,164-116,183,736), assembled from SP3388 TT-seq data compared to MFE values of shuffled RNA sequences (n=10). (c). The MFE of the native downstream eRNA (chr7:116,183,914-116,184,117), assembled from SP3363 TT-seq data compared to MFE values of shuffled RNA sequences (n=10). (d) The MFE of the native downstream eRNA (chr7:116,183,989-116,184,385), assembled from SP3388 TT-seq data compared to MFE values of shuffled RNA sequences (n=10).

Unlike the eRNAs assembled from SC3504 TT-seq data, the ALS patient-derived cell lines exhibited considerably higher relative TT-seq read coverage, compared to healthy control-derived cells, over the enhancers intersecting the eRNAs assembled here. For the tissue-agnostic *CAV2* enhancer, there was on average a 4.7 fold and 4.8 fold increase in read coverage over this region in SP3363 and SP3388 TT-seq data compared to SC3504, respectively. Moreover, for the smaller CNS-specific *CAV1* enhancer region, intersecting only the *eCAV1_3363_us* and *eCAV1_3388_us* eRNAs, there was a 2.8 and 2.4 fold increase in coverage over this region in SP3363 and SP3388 TT-seq data compared to SC3504, respectively (Figure 5.35).

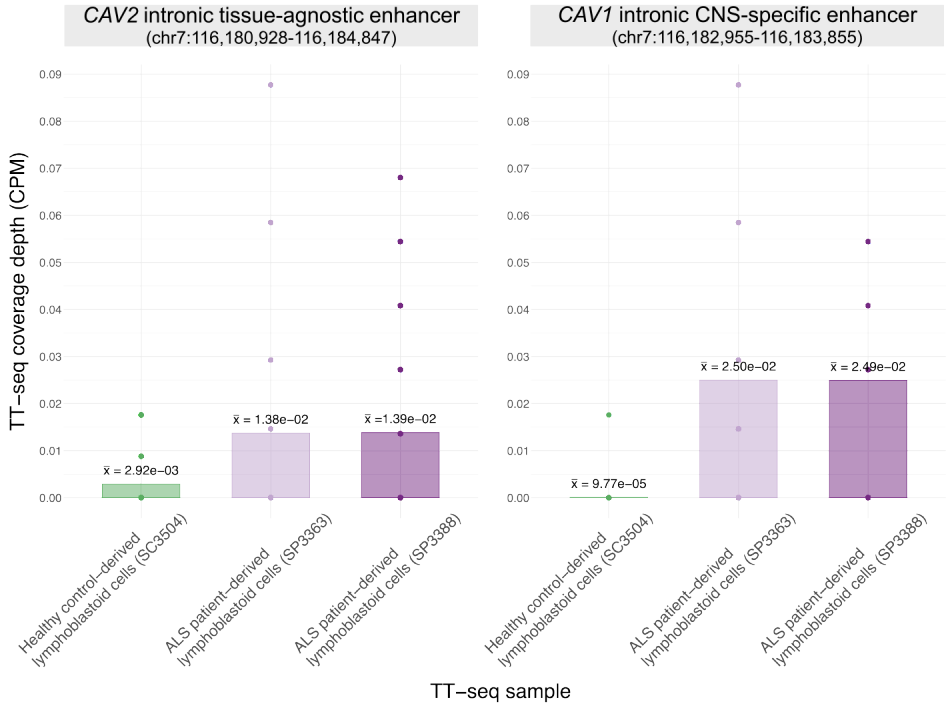



FIGURE 5.35: TT-seq coverage at ALS patient-associated enhancers
TT-seq read coverage (CPM) in healthy control- and ALS patient-derived lymphoblastoid cells at the *CAV2* tissue-agnostic enhancer (chr7:116,180,928-116,184,847) and the *CAV1* CNS-specific enhancer (chr7:116,182,955-116,183,855), both located within the *CAV1* intron, that generate eRNAs in ALS patient-derived lymphoblastoid cells. Bars represent mean coverage. Scatter represents TT-seq coverage at single-nucleotide resolution across the enhancers.

#### 5.7.2.1 ALS patient-specific assembled eRNA transcripts

While we have provided examples of novel *de novo* eRNA transcripts assembled exclusively from TT-seq data in healthy control-derived lymphoblastoid cells or ALS patient-derived lymphoblastoid cells, the examples we have provided thus far are not transcribed from enhancers that demonstrate transcription exclusively in the TT-seq dataset that they arise from. Therefore, we cannot confirm that these eRNAs are not present in all

lymphoblastoid cell lines, as transcription occurs over these enhancer regions across all three datasets. The lack of assembled transcript in these regions may therefore be due to the transcripts not meeting the bioinformatic thresholds required for assembly.

In the search to identify eRNA transcripts in enhancers that were exclusively transcribed in the cell line from which the eRNA originated, we only identified one eRNA transcript, assembled from TT-seq in SP3388 lymphoblastoid cells. This sense eRNA transcript at chr7:115,956,292-115,956,372, hereafter referred to as *eTES_3388*, intersected the candidate enhancers at chr7:115,954,125-115,959,325 and chr7:115,955,578-115,958,367, identified as a tissue-agnostic *CAV2* enhancer and a CNS-specific *CAV1* enhancer, respectively. Following RNA structure prediction, this eRNA yielded a MFE of -7.31 kcal/mol, with a MFEden of 10.50 kcal/mol and a expected MFEden of 12.56kcal/mol (Figure 5.36a). Despite its short length, at only 81 nt, this eRNA had a surprisingly low mean z-score of -0.12, and a minimum z-score of -5.77 (Figure 5.36b), likely attributed to its predicted stem loop structure at the beginning of the sequence.

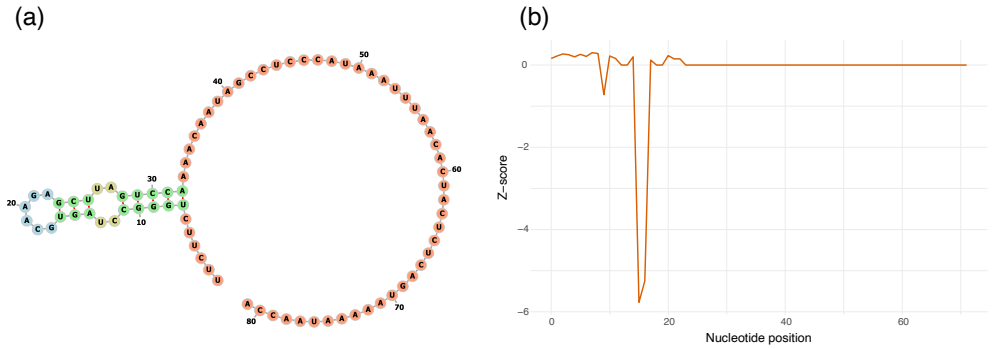



FIGURE 5.36: eRNA structure and stability predictions for the sense eRNA identified only in ALS patient-derived lymphoblastoid cells containing the chr7:115,994,269:C>T mutation
(a) MFE structure prediction for the novel *eTES_3388* eRNA (chr7:115,956,292-115,956,372), assembled from TT-seq in SP3388 cells. (b) Thermodynamic z-score across the length of *eTES_3388*.

TABLE 5.7: Predicted stability of ALS-associated eRNAs in ALS patient-derived lymphoblastoid cells

| **RNA** | **MFE (kcal/mol)** | **Observed MFEden (kcal/mol)** | **Expected MFEden (kcal/mol)** | **Mean shuffled MFE (kcal/mol)** | **Mean z-score** | **Min. z-score** |
|---|---|---|---|---|---|---|
| *eTES_3388* | -7.31 | 10.50 | 12.56 | -4.02 | -0.12 | -5.77 |

Finally, to further assess the stability of this ALS patient-associated eRNA and gain insights into its potential functionality in the ALS disease landscape, we compared the

MFE of the native *eTES_3388* eRNA to that of 10 random shuffled RNA sequences of the same nucleotide composition and length. The MFE of the native eRNA sequence was considerably more negative than the MFE values of all of the shuffled sequences (Figure 5.37).

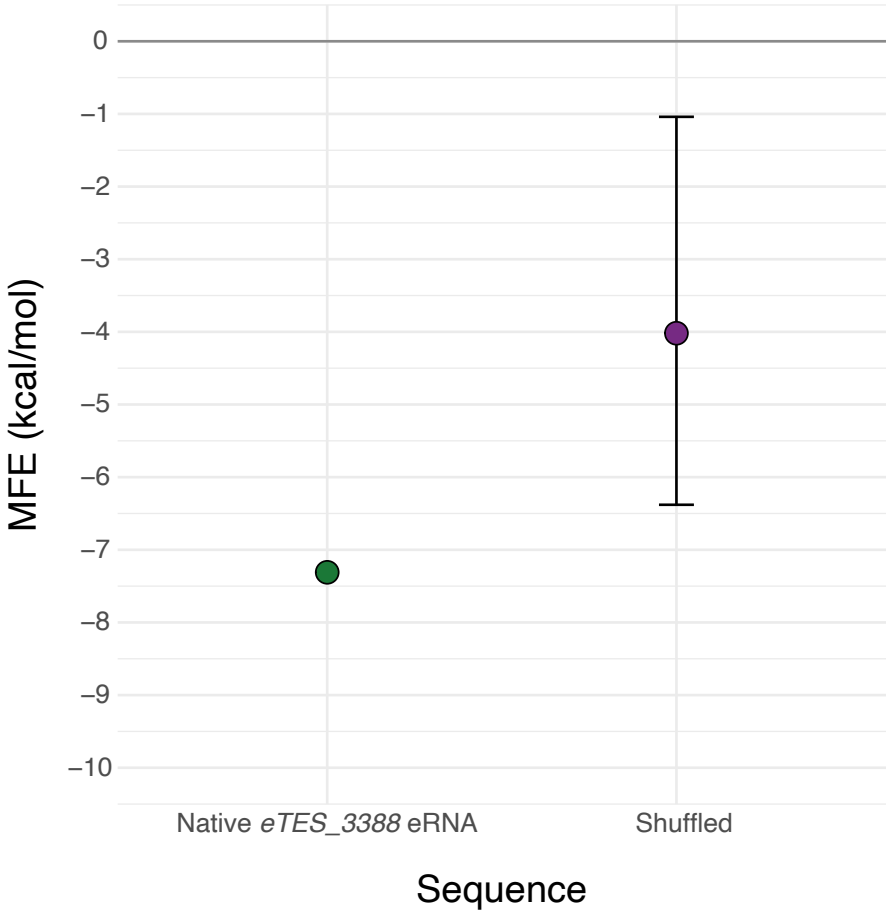



FIGURE 5.37: MFE predictions of native ALS patient-specific eRNA assembled from SP3388 TT-seq data compared to shuffled randomised control MFE values
The MFE of the native sense eRNA (chr7:115,956,292-115,956,372), assembled from SP3388 TT-seq data compared to MFE values of shuffled RNA sequences (n=10).

Taken together, we have provided examples of eRNAs assembled from TT-seq from ALS patient-derived lymphoblastoid cell lines, potentially indicative of the production of ALS-specific eRNAs. In concordance, we observed increased levels of nascent transcription over enhancers producing these ALS-specific eRNAs in ALS patient-derived cell lines compared to healthy control-derived cells. Further, we observed a novel eRNA transcript arising from an enhancer exclusively transcribed in the SP3388 ALS patient-derived cell line, exemplifying both an active enhancer, and a corresponding eRNA specific to the ALS genotype underlying this cell line.

## 5.8 Discussion

### 5.8.1 TT-seq as means to capture eRNA transcription

We have confirmed, as demonstrated in Schwalb et al. (2016), that TT-seq in SH-SY5Y cells effectively captures non-coding transcripts (Figure 5.5). Beyond unstable intronic transcripts, TT-seq in SH-SY5Y cells also detected considerably more transcription across the non-coding genome in comparison to standard total RNA-seq (Figure 5.6). To assess whether eRNAs were captured in our SH-SY5Y TT-seq data, we compared TT-seq read coverage over enhancers identified by the ABC model (Fulco et al., 2019) to non-exonic shuffled control regions (Figure 5.7, Figure 5.8). As expected, we observed more enhancers exhibiting TT-seq coverage than shuffled control regions (Figure 5.9). While transcriptional activity was significantly higher at enhancers compared to shuffled control regions, rank-biserial correlations indicated a modest effect size (Figure 5.10). This may be attributed to a number of factors. Firstly, TT-seq does not exclusively capture eRNAs, and will capture any RNA species transcribed within the five minute metabolic labelling pulse, including pre-mRNAs and other species of ncRNA (Schwalb et al., 2016). In addition, enhancers are known to exhibit cell and tissue type-specificity (reviewed in Heinz et al., 2015), and the candidate enhancers identified by the ABC model were defined in the human cell lines, K562, GM12878, NCCIT, and LNCaP, human liver tissue and mouse ESCs (Fulco et al., 2019). Therefore, it is likely that some of these ABC enhancers would exhibit cell or tissue type-specificity, and thus be inactive in SH-SY5Y cells. Moreover, with none of these cell or tissue types of neuronal origin, it is likely that neuronal-specific enhancers would be absent from the ABC enhancer catalogue, limiting our ability to fully capture the SH-SY5Y enhancer landscape in these analyses.

### 5.8.2 Integration of eRNA transcription data with additional enhancer epigenomic features

Given that eRNA transcription is frequently utilised as a proxy for enhancer activity level (De Santa et al., 2010, Hah et al., 2011, Kaikkonen et al., 2013, Kim et al., 2010, Li et al., 2013), we expected to observe co-occurrence of enhancer transcriptional activity with additional epigenomic features associated with active enhancers, including open chromatin and deposition of H3K27ac. This comparison of nascent RNA transcription over enhancers to additional assays that can capture enhancer epigenomic characteristics, such as ATAC-seq and CUT&RUN, is of importance given that TT-seq non-specifically identifies all nascent RNA transcripts genome-wide. Thus, transcription at an enhancer

region may not always signify eRNA transcription. However, we observed no significant difference in transcription levels between candidate *CAV1/CAV2* enhancers located in open versus closed chromatin, as determined by presence and absence of ATAC-seq signal at enhancers, respectively (Figure 5.13). This suggests that we observed TT-seq coverage at enhancers in a closed chromatin state, as defined by ATAC-seq in SH-SY5Y cells. As heterochromatin impedes TF and RNAPII access to DNA (Georges et al., 2002, Miller and Widom, 2003), it is highly unlikely that these regions were genuinely in a closed chromatin state. There is a possibility that we may have missed less highly accessible regions, potentially due to insufficient sequencing depths or suboptimal sample preparation, thereby reducing the sensitivity of the assay to detect open chromatin regions.

Alternatively, ATAC-seq may have failed to capture regions of transiently accessible chromatin (reviewed in Mansisidor and Risca, 2022). Chromatin dynamics can shift over short time frames, potentially leaving transient windows of accessibility undetected in our ATAC-seq data. Especially when we consider that the tagmentation reaction in ATAC-seq occurs over a 30 minute incubation, and in TT-seq we specifically capture nascent RNA transcription within a 5 minute window, these differences in time-scale may result in discrepancies in our results. ATAC-seq, while widely used to measure chromatin accessibility, primarily reflects nucleosome positioning, as measured by quantifying the protection of DNA from enzymatic digestion by pAG-MNase. However, through MNase titration, regions with high nucleosome occupancy have been shown to maintain high chromatin accessibility, contradicting the assumption that nucleosomes strictly block access to DNA (Mieczkowski et al., 2016). In fact, changes in chromatin accessibility at enhancers do not always correlate with changes in nucleosome occupancy, with high nucleosome occupancy not necessarily precluding high accessibility. Thus, nucleosome positioning alone, as measured by ATAC-seq, cannot fully capture enhancer accessibility, as chromatin states can be dynamic and context-dependent. Enhancers may retain nucleosomes while still remaining functionally accessible for TF and RNAPII binding (Mieczkowski et al., 2016). This may imply, that even enhancers that are identified to be in a closed chromatin state by ATAC-seq, may be transiently accessible, allowing for transcription. This dynamic nature of chromatin, coupled with the specific, transient RNA-specific time-scale of TT-seq, may explain why we observe transcriptional activity in enhancer regions without significant ATAC-seq signal.

On the other hand, when comparing active and inactive enhancers, as determined by presence or absence of H3K27ac CUT&RUN signal, respectively, we did observe a significant increase in transcription levels at active enhancers versus inactive (Figure 5.14). In contrast to chromatin accessibility, which merely indicates whether DNA is in a state permissive for transcriptional activation, there is a tighter link between RNA transcription

and H3K27ac. The deposition of H3K27ac is closely associated with transcriptional activation and is widely recognised as a key indicator of enhancer activation (Allfrey et al., 1964, Creyghton et al., 2010, Rada-Iglesias et al., 2011, Zentner et al., 2011). In support, the majority of candidate enhancers significantly enriched for H3K27ac CUT&RUN signal exhibited TT-seq coverage, and similarly the majority of untranscribed enhancers were not significantly enriched for H3K27ac signal (Figure 5.15). However, this leaves the majority of the candidate enhancers that while exhibiting TT-seq coverage, are not enriched for significant H3K27ac CUT&RUN signal, unaccounted for. Again, this may be a technical issue. During library preparation clean-up, we had to perform an additional round of size exclusion to remove sub-nucleosomal fragments, and therefore viable library fragments may have potentially been lost during these additional purification steps. Additionally, the fold enrichment cut-off we utilised for these analyses required a 2-fold H3K27ac signal enrichment over H3 signal, which may, in reality, be too stringent to capture all regions of H3K27ac deposition. Highlighting this issue, we identified 84 transcriptionally active candidate enhancers, but only 20 enhancers that were enriched for significant H3K27ac signal (Figure 5.15).

Alternatively, this may raise questions about the capacity for enhancers to produce eRNAs without the presence of H3K27ac. While H3K27ac is a common marker of active enhancers, its presence is not strictly required for enhancer activity. Additional histone PTMs are also associated with enhancers, particularly those implicated in nucleosome dynamics and chromatin opening, essential for promoting transcription, including H3K64ac, H3K122ac and H4K16ac (Di Cerbo et al., 2014, Pradeepa et al., 2016, Simon et al., 2011, Taylor et al., 2013, Tropberger et al., 2013). The neutralisation of positive charges on histone tails through lysine acetylation plays a more critical role in establishing a permissive chromatin structure, via weakened interactions between the histone tails and DNA, than the specific patterns of acetylation at individual lysine residues (reviewed in Eberharter and Becker, 2002). Indeed, the global depletion of H3K27ac in mouse ESCs by replacing lysine 27 with a non-modifiable arginine residue has minimal impact on the transcriptome, suggesting that gene transcription is not sensitive to the substantial ablation of H3K27ac at enhancers (Zhang et al., 2020b). Since H3K27ac alone cannot functionally determine enhancer activity, limiting our analyses to only H3K27ac CUT&RUN may only offer a minimal snapshot of enhancers in an active chromatin state, overlooking enhancers where the active chromatin conformation is established by other forms of histone acetylation.

Overall, when comparing active candidate enhancers identified by ATAC-seq and H3K27ac CUT&RUN to those producing eRNAs, as indicated by TT-seq read coverage, H3K27ac CUT&RUN demonstrates a stronger correlation with eRNA production, and, consequently, with enhancer activity, than ATAC-seq. While combining H3K27ac deposition,

chromatin accessibility, and eRNA transcription provides a more comprehensive view of the enhancer landscape, it has been established that eRNA transcription is a more reliable indicator of enhancer activity than H3K27 deposition (Zhu et al., 2013). Thus, while TT-coverage alone cannot fully define enhancer activity, limiting our analyses to enhancers that exhibit ATAC-seq signal, H3K27ac CUT&RUN signal and TT-seq coverage may result in a more restricted representation of the true enhancer landscape.

### 5.8.3 Assembly of *de novo* eRNA transcripts

Given that TT-seq captures all nascent RNA transcripts, we conducted *de novo* transcript assembly to identify a subset of our TT-seq reads that may have represented genuine eRNAs. This not only validated use of TT-seq as a method to identify eRNA transcripts, but also provided means to predict RNA secondary structures and stability. As expected, due to the cell type-specificity of enhancers, and the large proportion of the enhancers in our candidate *CAV1/CAV2* enhancer dataset considered to be CNS-specific, we identified more novel eRNA transcripts from TT-seq data in SH-SY5Y cells than we did from TT-seq data in human-derived lymphoblastoid cells (Figure 5.16, Figure 5.26).

Despite the increased number of candidate enhancers exhibiting TT-seq read coverage in the healthy control-derived lymphoblastoid cells, compared to the ALS patient-derived lymphoblastoid cells (Figure 5.23), we identified not only lower overall transcription in this cell line (Figure 5.24), but also identified fewer *de novo* eRNA transcripts (Figure 5.26). Therefore, while there were fewer active enhancers in ALS patient-derived lymphoblastoid cells, we observed more transcription over those that were active.

Of importance, while we identified some enhancers that only produce novel eRNAs in the healthy control-derived cell line, we identified approximately equal coverage over these enhancers in ALS patient-derived TT-seq data (Figure 5.30), suggesting equal levels of enhancer activity, regardless of disease state. However, considering we only successfully assembled eRNA transcripts from healthy control-derived TT-seq data, this suggests that although the transcription coverage over these enhancers in ALS patient-derived cell lines was roughly equal to that of the healthy control-derived cell line, the actual abundance of these particular eRNAs may have been lower in ALS patients. This may suggest that transcription from these enhancers in ALS patient-derived lymphoblastoid cells was more sporadic, generating a more diverse transcriptome that we were unable to assemble into eRNA transcripts of 50nt or longer.

In contrast, the enhancers that produce novel eRNAs in both ALS patient-derived cell lines, as determined by overlapping assembled transcripts, exhibited considerably more

TT-seq coverage in the ALS patient-derived cell lines than in the healthy control-derived cell line (Figure 5.35). In addition to the reduced coverage over these enhancers in healthy control-derived cells suggesting a lower level of enhancer activity, this may suggest that these eRNAs are not expressed at detectable levels.

For a more robust example of an eRNA specific to ALS patient-derived lymphoblastoid cells, we identified a novel sense eRNA that was solely assembled from TT-seq data in SP3388 cells. Despite being able to assemble an eRNA transcript from this enhancer region in SP3388 cells, we detected no TT-seq coverage over this region in either SC3504 or SP3363 cells, suggesting that both the activity of this enhancer, and the eRNA that was produced as a result were unique to this cell line.

### 5.8.4 eRNA structure and stability prediction

The stability, and hence the MFE, of an RNA molecule is dictated by the combined contributions of its length, nucleotide order and nucleotide composition (reviewed in Trotta, 2014). Longer sequences are on average more stable, due to the increased potential for forming stacking and hydrogen bond interactions. Nucleotide order influences structure stability, governing the number and extension of loop structures and helices. Regarding nucleotide content, guanine-cytosine (GC)-rich RNAs are generally more stable than adenine-uracil (AU)-rich RNAs (reviewed in Trotta, 2014). To assess the predicted stability and potential functionality of *de novo* eRNA transcripts, we have utilised multiple metrics, each offering unique insights into the free energy components that influence RNA structure, and by extension, stability.

Highlighting the contribution of nucleotide order to RNA structure stability, MFE mRNA secondary structures are predicted to be more stable than expected by chance, with MFE values of a greater negative magnitude than those of randomised sequences with the same base composition and length (Seffens and Digby, 1999). This suggests that beyond overall nucleotide composition, the local secondary structures formed by the specific native RNA sequence form local secondary structures that provide stability to the secondary RNA structure. While, in this context, this may reflect the models underlying MFE prediction algorithms - the nearest-neighbour model (Mathews et al., 2004) and dynamic programming algorithms (Zuker, 1989, Zuker and Stiegler, 1981) - the prevalence of local nucleotide interactions in establishing RNA structure does have a basis in experimental data. Short-range base pairing and stacking interactions, forming local structural motifs such as hairpin loops, are crucial for stabilising RNA secondary structures, while long-range interactions often form the basis of tertiary structures (reviewed in Vicens and Kieft, 2022). Thus, MFE prediction algorithms leverage these

structural characteristics. In the nearest-neighbour model, the free energy of an RNA secondary structure is determined by the interactions between adjacent base pairs, with the stability of a given base pair further influenced by the identity and arrangement of neighbouring base pairs (Mathews et al., 2004). Dynamic programming algorithms, such as the Zuker algorithm, systematically explore all possible secondary structures by dividing the RNA sequence into subunits and calculating the optimal structure for each subunit (Zuker, 1989, Zuker and Stiegler, 1981).

### 5.8.5 Predicted structure and stability of benchmark eRNAs

To provide a benchmark for our RNA structure and stability analyses, we predicted the structure and stability of three example eRNA transcripts. We predicted the secondary structure and calculated the predicted stability of the $^{DRR}$eRNA, as this eRNA is known to be spliced and and have in *trans* functionality (Cichewicz et al., 2018, Przanowska et al., 2022, Tsai et al., 2018), both of which are suggestive of a high level of stability. The $^{DRR}$eRNA demonstrated a MFE that was both more negative than the mean shuffled MFE, a MFEden more negative than expected based on GC content and a negative z-score, overall suggestive of a stable eRNA (Figure 5.18, Figure 5.20a). However, it is worth noting that while the MFE was more negative than the mean shuffled MFE, some shuffled sequences did yield a more negative MFE than that of the native sequence, suggesting that overall sequence length and GC content may have a considerable impact on the overall stability of the RNA structure.

In addition, the *TAL1* eRNAs, which have been studied in the context of disease-specific super-enhancer function and transcription (Danko et al., 2015, Lidschreiber et al., 2021), appear to demonstrate questionable stability. While the sense *TAL1* eRNA contained structural motifs predicted to produce stable structures, as demonstrated by a negative z-score, the MFE metrics suggest that these structures may not be stable under physiological conditions (Figure 5.19a,b). Interestingly, of all the eRNAs we analysed this was the only eRNA with an observed MFEden more positive than the expected MFEden. Compared to mRNAs, which typically have MFEden values close to expected, reflecting a greater contribution of nucleotide composition to folding energy, and pre-miRNAs, which often have more negative MFEden values than expected, indicating a stronger role of nucleotide order, an MFEden higher than expected suggests that neither nucleotide order nor composition significantly contribute to RNA stability (Trotta, 2014). This implies that, despite an overall negative z-score, structural elements within the *TAL1* sense eRNA may reduce the formation of stable base pairs, relative to expected from its inherent GC content. In support, shuffled RNA sequences of the same nucleotide composition consistently yielded more negative MFE values that of the native eRNA sequence

(Figure 5.20b). Whereas, the antisense *TAL1* eRNA, while producing a more preferable free energy profile, does not seem to generate the secondary structures necessary for *in vivo* stability, as suggested by a positive z-score (Figure 5.19c,d).

This underscores the importance of utilising both MFE calculations and thermodynamic z-scores to predict overall RNA stability. While the *TAL1* sense eRNA forms stable local structures, as indicated by the negative z-score, in the context of the folding energy of the RNA sequence as a whole, these structures do not appear to stabilise the RNA overall. Conversely, the *TAL1* antisense eRNA exhibited more negative MFE and MFEden values than expected, suggesting enhanced stability based on nucleotide order and content. However, the positive z-score indicates a lack of stable local structures, implying that in action, this predicted stability does not manifest in the formation of stable local structures.

### 5.8.6 The impact of ALS-associated SNPs on predicted eRNA structure and stability

To model the impact of ALS-associated SNPs on eRNA structure and stability, we predicted the eRNA structure and thus, stability, of the antisense eRNA, *eCAV1_WT* (chr7:116,199,437-116,199,680) assembled from SH-SY5Y TT-seq data (Figure 5.16), for both the native WT transcript, and the *eCAV1_ALS* eRNA transcript following computational induction of the ALS-associated SNP chr7:116,199,522T>C (Figure 5.21). We identified not only a considerable change in the predicted MFE structure, but also changes in the predicted stability of the eRNA upon induction of the base change in the input RNA sequence.

While both the WT and the mutated sequence demonstrated MFE values which were more negative that the mean shuffled MFE (Figure 5.22), a MFEden more negative than expected based on GC content and a negative z-score (Figure 5.21), the *eCAV1_ALS* sequence yielded consistently more negative values for every metric. For example, while the MFE of the *eCAV1_WT* sequence fell within the range of the shuffled sequence MFE values, the *eCAV1_ALS* sequence generated a MFE that was consistently lower than that of the randomised shuffled sequences. This may suggest that nucleotide order contributes more greatly to overall folding energy for the mutated sequence than it does for the WT sequence. Likely reflecting this is the observed formation of a stem-loop structure, otherwise absent from the *eCAV1_WT* structure, upon induction of the ALS-associated SNP (Figure 5.21). This highlights the importance of nucleotide order and composition in this eRNA, with a singular base change resulting in a considerable structural difference. Further, formation of this stem-loop structure appears to contribute

to an increased stability, as reflected in the reduction in both the mean and minimum z-scores for the *eCAV1_ALS* sequence, compared to the *eCAV1_WT* native sequence. For *eCAV1_ALS*, we observed minimal changes to the z-scores attributed to the most stable local structures in the *eCAV1_WT* sequence, while also observing the addition of a much more stable local structure, the hairpin stem-loop, contributing a z-score of -10.10. Taken together, this provides an example of an ALS-associated SNP that could considerably alter both the structure and stability of an eRNA.

### 5.8.7 eRNAs in human-derived lymphoblastoid cells

To study the differences in the *CAV1/CAV2* enhancer-associated eRNA transcriptome, between healthy and disease states, we identified eRNA transcripts assembled from TT-seq data in control- and patient-derived lymphoblastoid cells. We identified two *de novo* eRNA transcripts that were only successfully assembled from the healthy control-derived lymphoblastoid TT-seq data, despite identifying similar levels of TT-seq coverage over these enhancers in both the ALS patient-derived lymphoblastoid TT-seq datasets. Further, we identified two sets of overlapping *de novo* eRNA transcripts in ALS-patient derived lymphoblastoid cells, with one eRNA in each set assembled from SP3363 TT-seq data, and the other from SP3388 TT-seq data. Again, while there was TT-seq read coverage over these enhancers in the healthy control-derived SC3504 TT-seq dataset, we did not successfully assemble any eRNA transcripts from the SC3504 TT-seq data at these enhancers. These comparisons allowed us to identify differences in predicted stability of eRNAs present in only healthy control-derived cells versus eRNAs that were present in both ALS patient-derived cell lines, regardless of underlying ALS-associated SNP. We use these comparisons to generally discuss differences in predicted eRNA stability and their contributions to the *CAV1/CAV2* enhancer landscape in both healthy and disease states.

#### 5.8.7.1 eRNAs in healthy control-derived lymphoblastoid cells

For the eRNAs specific to the healthy control participant, the antisense eRNA, *eCAPZA2_antisense* (chr7:116,440,664-116,440,720) does not appear to be stable according to almost any metric assessed (Figure 5.28a,b). Likely due to its short length, it yielded an MFE of only -11.35kcal/mol, a value more positive than the mean shuffled MFE (Figure 5.29a). This may suggest that the nucleotide composition, as opposed to the nucleotide order, contributes more to the resulting free energy profile of the eRNA. In agreement, with a MFEden only slightly more negative than the expected based on

GC content, again, this implies that overall nucleotide content, instead of order, provides more stability to the structure. This eRNA also yielded a positive mean z-score, with visually, the majority of the RNA sequence yielding positive z-scores, suggesting that randomised sequences were largely more stable than the native *eCAPZA2_antisense* eRNA sequence (Figure 5.28b). Taken together, the observed lack of local stable structures in combination with the comparatively more negative MFE and z-score indices of shuffled sequences, suggest that this eRNA is likely unstable.

Meanwhile, the sense eRNA, *eCAPZA2_sense* (chr7:116,502,612-116,502,934) had a particularly negative MFE, MFEden and z-score (Figure 5.28c,d), suggestive of stability, however this likely arose from the high GC content of the transcript, with a 71.2% GC content. Reflecting this, we observed that the MFE of shuffled sequences was almost always more negative than the MFE of the native sequence (Figure 5.29b). While this does not mean that the eRNA itself is not stable, it suggests that the stability can be largely attributed to the inherent strong base pairing that exists due to the high GC content, not due to the stability of specific structural motifs in the RNA sequence. Again, in support, of all the eRNAs we studied here, *eCAPZA2_sense* demonstrated the most positive maximum z-score of 0.87, compared to the maximum z-scores of 0.45, 0.55 and 0.55 of the example eRNAs, *TAL1* sense, *TAL1* antisense and $^{DRR}$eRNA, respectively. While overall the sequence of the native *eCAPZA2_sense* eRNA clearly produces some stable local structures, as demonstrated by the negative mean z-score of -0.32, there are also regions within the sequence that form structures less stable than expected, as evidenced by a positive z-score (Figure 5.28d). Taken together, the *eCAPZA2_sense* eRNA may be somewhat stable, but likely due to its length and high GC content.

The GC content of eRNAs may provides insights into their potential cellular functions. Specifically, the epigenetic modulator complex PRC2 preferentially binds to RNA sequence motifs comprised of consecutive guanine residues (Wang et al., 2017a), suggesting that GC-rich eRNAs may be more likely to sequester PRC2, preventing the deposition of H3K27me3 at enhancers and therefore promoting enhancer activation (Beltran et al., 2019, Mao et al., 2018). Additionally, PRC2 exhibits a structural preference for RNA G-quadruplexes, binding these complex structures with high affinity (Wang et al., 2017a). Therefore, to predict whether the *eCAPZA2_sense* eRNA has the potential to interact with PRC2 and epigenetically influence enhancer activity, we could evaluate whether this eRNA forms G-quadruplex structures. This could be achieved using predictive algorithms or structure-based sequencing approaches, including chemical probing of *in vitro* transcribed transcripts, offering insights into the formation and folding of G-quadruplexes in this candidate eRNA sequence (reviewed in Kwon et al., 2022).

#### 5.8.7.2 eRNAs in ALS patient-derived lymphoblastoid cells

For the eRNAs observed across disease state cells, the set of upstream sense eRNAs *eCAV1_3363_us* (chr7:116,183,048-116,183,530) and *eCAV1_3388_us* (chr7:116,183,164-116,183,736), assembled from SP3363 and SP3388 TT-seq data, respectively, appeared to be stable, with negative MFE values consistently more negative than MFE values of shuffled sequences (Figure 5.34a,b) and MFEden values more negative than expected based on GC content and negative z scores (Figure 5.32). Taking the z-score metric alone, these ALS patient-associated eRNAs generated more predicted local stable secondary structures, compared to the healthy control-associated eRNAs, *eCAPZA2_antisense* and *eCAPZA2_sense* (Figure 5.32b,d). Not only do fewer windows across the sequences produce a positive z-score, but their minimum z-scores are more negative. Therefore, not only do the ALS patient-associated eRNAs produce local structures that are more stable than that of randomised sequences, but they form fewer local structures that would be unfavourable to stability, as observed for the healthy control-associated eRNAs.

Conversely, the set of downstream sense eRNAs *eCAV1_3363_ds* (chr7:116,183,914-116,184,117) and *eCAV1_3388_ds* (chr7:116,183,989-116,184,385), assembled from SP3363 and SP3388 TT-seq data, respectively, appeared much less stable, with much less negative MFE values that lie within the range of shuffled MFE values (Figure 5.33, Figure 5.34c,d). While these eRNAs have MFEden values more negative than the expected MFEden values based on GC content, the differences between the observed and expected MFEden values are less pronounced than observed for the upstream set of ALS patient-derived eRNAs. Taken together, this suggests that GC content may provide more stability to these structures than local stable structures formed due to nucleotide order.

When we compare the apparent stability of the upstream ALS patient-associated eRNAs to the apparent instability of the downstream ALS patient-associated eRNAs, this raises questions about the underlying causes of these differences in stability. The upstream eRNAs arose from a genomic region associated with both a tissue-agnostic *CAV2* candidate enhancer and a CNS-specific *CAV1* candidate enhancer, whereas the downstream eRNAs originated from within a region associated with just the tissue-agnostic *CAV2* enhancer. Interestingly, the upstream eRNAs intersect with a sense eRNA transcript from WT SH-SY5Y TT-seq data at chr7:116,182,983-116,183,329, suggesting that similar eRNAs might be transcribed in neuronally relevant cells, potentially implicating activity of the CNS-specific *CAV1* enhancer, even in non-neuronal lymphoblastoid cells. Given their predicted stability and mirrored production in SH-SY5Y cells, we hypothesise that the upstream eRNAs, *eCAV1_3363_us* and *eCAV1_3388_us*, may have more functional relevance in the context of the neuronal *CAV1/CAV2* enhancer landscape, in comparison to the downstream eRNAs, *eCAV1_3363_ds* and *eCAV1_3388_ds*.

#### 5.8.7.3 Genotype-specific eRNA structure and stability in ALS

To provide insight into the *CAV1/CAV2* enhancer landscape in specific ALS genotypes, we identified *de novo* assembled eRNA transcripts that arise from enhancers solely transcribed in one of the ALS patient-derived lymphoblastoid cell lines. We assembled the novel sense *eTES_3388* eRNA (chr7:115,956,292-115,956,372) from SP3388 TT-seq, which intersects with the candidate *CAV1/CAV2* enhancers at chr7:115,954,125-115,959,325 and chr7:115,955,578-115,958,367, neither of which exhibited coverage in SC3504 or SP3363 TT-seq data. This may suggest that activity at either or both of these enhancers is linked to the ALS genotype of chr7:115,994,269:C>T. Thus, investigating the RNA structure and stability of the resultant eRNA may provide some insight into the functional consequences of enhancer activity at this locus.

Despite its short length, and thus relatively high MFE of -7.31kcal/mol, according to our utilised metrics, this eRNA demonstrated some level of stability (Figure 5.36). While the MFEden is not considerably lower than the expected MFEden based on GC content, potentially suggesting that the nucleotide order has less of an impact on folding stability that overall nucleotide composition, we observed trends in other metrics that counter this notion. The MFE of the native eRNA sequence was consistently, and considerably, more negative than the MFE of shuffled sequences (Figure 5.37), suggesting that the nucleotide order does specifically impart stability to this structure. In agreement, we observed a negative mean z-score of -0.12 and a minimum z-score of -5.77, implying that local structures within the RNA sequence are more stable than expected. Taken together, this provides an example of an ALS genotype-specific eRNA, that despite its short length, appears to be surprisingly stable according to all metrics tested.

### 5.8.8 Conclusion

Overall, our findings underscore the relevance of the nascent transcriptome when evaluating enhancer activity. Combining eRNA transcription with additional methods of probing for enhancer-associated characteristics can provide insights into the dynamic nature of the *CAV1/CAV2* enhancer landscape, with methods such as ATAC-seq and CUT&RUN providing a more static view of enhancer activity than that provided by methods such as TT-seq. We identified that ALS-associated SNPs may be capable of altering eRNA structure and stability, as observed with the alterations in secondary structure and increased apparent stability of the eRNA at chr7:116,199,437-116,199,68 upon induction of chr7:116,199,522T>C. Using multiple metrics to assess the contribution of distinct structural elements to overall eRNA folding stability, we identified key differences in the predicted stability of eRNAs specific to healthy control-derived

cells as compared to ALS patient-derived cells. While both eRNAs identified from SC3504 TT-seq data appeared to be somewhat unstable, the overlapping *eCAV1_3363_us* and *eCAV1_3388_us* eRNAs, from SP3363 and SP3388 TT-seq data, respectively, appeared to be more stable by comparison. Finally, the *eTES_3388* eRNA, specific to the chr7:115,994,269:C>T genotype was predicted to be surprisingly stable according to the metrics utilised, again, supporting the notion that disease-relevant eRNAs have overall increased stability. This raises the question - if the transcription of stable eRNAs is a characteristic of the *CAV1/CAV2* enhancer landscape in the context of ALS, does the transcription of such stable eRNAs have functional consequences regarding disease progression?

# Chapter 6

# How do Amyotrophic Lateral Sclerosis SNPs impact regulatory activity?

## 6.1 Introduction

### 6.1.1 CAV1 neuroprotective function

CAV1 is a key protein component of caveolae, submicroscopic invaginations in the plasma membranes abundant in many mammalian cell types (reviewed in Parton and del Pozo, 2013). Each caveola contains an estimated average of 144 CAV1 molecules (Pelkmans and Zerial, 2005). Different tissues display distinct caveolar density, with Caveolin expression levels dramatically different from tissue to tissue. While neuronal expression of Caveolin is considered low in comparison to other cell types (Head et al., 2011, 2010), especially when compared to to the abundance in endothelial cells and adipocytes (Thorn et al., 2003), neuronal Caveolin expression is still of functional importance (Head et al., 2011, 2010). Converging evidence suggests a neuroprotective role for CAV1 in ALS (Cooper-Knock et al., 2021, Egawa et al., 2018, Head et al., 2011, 2008, Sawada et al., 2019, Takayasu et al., 2010). Indeed, *CAV1* and *CAV2* expression is consistently significantly higher in post-mortem brain tissues of ALS patients versus healthy control individuals (Adey et al., 2023).

A key example of the potential neuroprotective role of CAV1 in neurodegenerative disease is via its roles in endocytosis, of relevance to the trafficking of integral membrane proteins such as A$\beta$ precursor protein (APP) and subsequent A$\beta$ production (reviewed in

Vetrivel and Thinakaran, 2006). APP produces neurotoxic A$\beta$ via amyloidogenic proteolysis, defined by its sequential proteolytic cleavage by $\beta$-secretase and then $\gamma$-secretase. Initial cleavage of APP by $\beta$-secretase produces soluble APP-$\beta$ and membrane-bound APP C-terminal domain. The APP C-terminal domain is subsequently cleaved by $\gamma$-secretase, releasing cytotoxic A$\beta$ peptides extracellularly and generating the APP intracellular domain for nuclear translocation (Figure 6.1a) (reviewed in Wolfe and Guénette, 2007). CAV1 is thought to be a negative regulator of both $\beta$- and $\gamma$-secretase activity (Hattori et al., 2006, Kapoor et al., 2010). CAV1 overexpression results in a significant suppression of $\beta$-secretase function, as demonstrated by a dramatic decrease in soluble APP-$\beta$ levels (Hattori et al., 2006). CAV1 overexpression similarly attenuates $\gamma$-secretase activity, potentially by promoting the dynamic partitioning of $\gamma$-secretase into caveolae (Figure 6.1b) (Kapoor et al., 2010), highlighting the importance of CAV1 expression levels.

In an alternative, non-amyloidogenic pathway, APP undergoes proteolytic cleavage by $\alpha$-secretase, generating the soluble extracellular domain of APP and the APP C-terminal domain. The APP C-terminal domain is further cleaved by $\gamma$-secretase producing the non-toxic products, the APP intracellular domain and p3 peptide, thus precluding the generation of A$\beta$ (Figure 6.1c) (reviewed in Wolfe and Guénette, 2007). CAV1 may play a pivotal role in $\alpha$-secretase-mediated proteolysis, whereby CAV1 physically interacts with the C-terminal cytoplasmic domain of APP to sequester APP in caveolae and compartmentalise the $\alpha$-secretase-mediated cleavage of APP. As such, $\alpha$-secretase processing of APP is significantly promoted by CAV1 overexpression, resulting in increased secretion of the soluble extracellular domain of APP (Figure 6.1d) (Ikezu et al., 1998). This suggests a dual function for CAV1 in APP processing, functioning to both inhibit amyloidogenic processing of APP while promoting the use of the non-amyloidogenic processing pathway via synergistic effects on $\alpha$-, $\beta$- and $\gamma$-secretases, thus suggesting a role for CAV1 in preventing the production of neurotoxic A$\beta$. This likely has *in vivo* relevance, as evidenced by the increased *CAV1* expression in post-mortem brain tissue of both ALS and AD patients, compared to healthy controls (Adey et al., 2023, Gaudreault et al., 2004, Ikezu et al., 1998), again suggesting a neuroprotective role for CAV1.

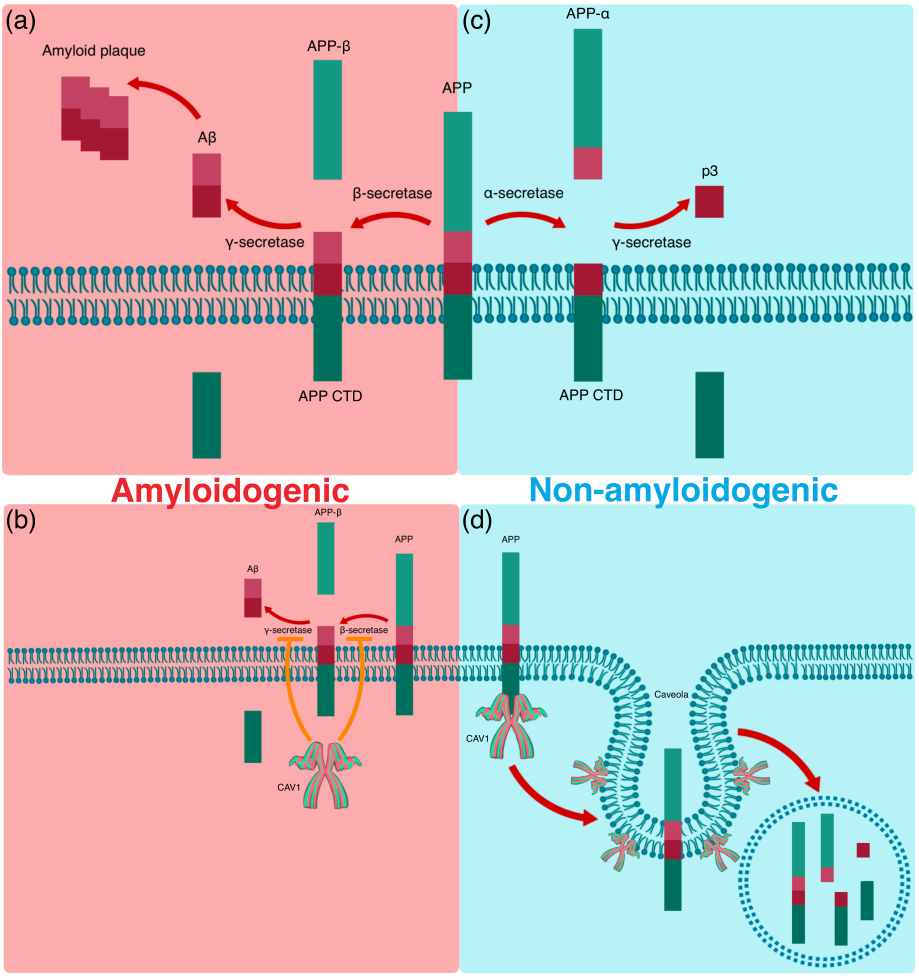



FIGURE 6.1: CAV1 in APP proteolysis
(a) APP undergoes amyoloidogenic proteolysis ultimately generating neurotoxic A$\beta$ that culminates in the production of amyloid plaques. APP is sequentially cleaved by $\beta$-secretase and $\gamma$-secretase. (b) APP undergoes non-amyloidogenic proteolysis, precluding the generation of cytotoxic A$\beta$. APP is sequentially cleaved by $\alpha$-secretase and $\gamma$-secretase to ultimately produce the p3 peptide. (c) CAV1 is may function as a negative regulator of $\beta$-secretase and $\gamma$-secretase, potentially reducing levels of A$\beta$. (d) CAV1 interacts with the C-terminal domain of APP, compartmentalising the non-amyoloidogenic proteolysis of APP within caveolae.

### 6.1.2 Caveolin-1 isoforms

CAV1 is expressed in two isoforms, CAV1$\alpha$ and CAV1$\beta$, that differ only in their extreme N-terminal protein sequence, yielding proteins of 24kDa and 21kDa, respectively (Scherer et al., 1995). The ratio of the $\alpha$ to $\beta$ isoform can be regulated at the level of translation, through alternative translation initiation at methionine positions 1 or 32, respectively (Scherer et al., 1995), or at the level of transcription, with the different isoforms generated by the translation of distinct mRNAs (Kogo and Fujimoto, 2000). The

two isoforms demonstrate distinct but overlapping cellular distributions (Scherer et al., 1995). *De novo* caveolae formation appears to be dependent on expression levels of the two isoforms, with CAV1$\alpha$ more efficient than CAV1$\beta$ at forming caveolae (Fujimoto et al., 2000).

CAV1$\alpha$ and CAV1$\beta$ differ not only in their structure, but also in their functional roles in the cell. Namely, CAV1 undergoes phosphorylation by tyrosine kinases at tyrosine 14 (Lee et al., 2000, Li et al., 1996), in a manner that is thought to be tightly regulated in the cell (Mastick et al., 1995). Due to the N-terminal truncation of the CAV1$\beta$ isoform, relative to CAV1$\alpha$, only the longer CAV1$\alpha$ isoform undergoes phosphorylation at tyrosine 14 (Lee et al., 2000, Li et al., 1996). Phosphorylation of CAV1 regulates spatial organisation within Caveolin oligomers, with the negatively charged N-terminal phosphotyrosine residues promoting the separation of neighbouring CAV1 molecules within the oligomer. The resultant swelling of caveolae is linked to the destabilisation of CAV1 oligomers, increasing the velocity of cargo trafficking and ultimately, the release of caveolae from the plasma membrane (Zimnicka et al., 2016). This may reflect earlier reports suggesting that CAV1$\alpha$ is more prominent in deep caveolae, while CAV1$\beta$ is more prominent in shallow caveolae (Fujimoto et al., 2000), which could potentially be the result of the phosphorylation-regulated separation of the CAV1 oligomer.

With the balance between the CAV1 isoforms, and the dynamics of CAV1 phosphorylation, affecting both caveolae structure and function, altering the ratio between isoforms and abundance of N-terminal tyrosine phosphorylation can have influence over disease pathology. As evidenced largely in cancer, through interactions with growth factor signalling cascades and focal adhesion dynamics (Hau et al., 2019, Lee et al., 2000, Meng et al., 2017), phosphorylated CAV1 is critically implicated in tumour cell migration, invasion and metastasis (Hau et al., 2019, Meng et al., 2017). Understanding the specific contributions of the distinct CAV1 isoforms and thus the transcriptional regulatory mechanisms that determine *CAV1* mRNA isoform usage, is crucial for elucidating the complex biology of caveolae and their involvement in human health and disease.

### 6.1.3 The pathological impacts of *CAV1* enhancer mutations

Following evidence of potential *in vivo* neuroprotective functions for CAV1, with the suggested downregulation of *CAV1* expression caused by *CAV1* enhancer mutations (Cooper-Knock et al., 2021), it would be expected that enhancer mutation carriers would lack such neuroprotective compensatory mechanisms, leading to impaired neurotrophic signalling, and ultimately more aggressive disease progression. On the contrary, ALS patients with *CAV1* enhancer mutations show an increased survival rate compared to

non-carried within Project MinE, with a median increase in survival of 345 days (Adey et al., 2023). Therefore, the proposed compensatory reaction to neurotoxicity does little to explain the observed improved survival in patients carrying *CAV1* mutations.

Given the complexity of enhancer and eRNA functionality, we should not expect disease-associated SNPs in enhancers to simply cause a direct pathogenic up- or downregulation of causative gene expression. To understand how mutations in enhancers contribute to the aetiology of disease, we must consider various factors that control enhancer identity and activity. At the chromatin level, mutations in enhancers can affect 3D interactions between enhancers and promoters, chromatin compaction, and PTMs. At the DNA sequence level, enhancer mutations can influence the binding of TFs, cofactors, and DNA methylation levels. Additionally, the transcription of eRNAs from disease-associated enhancers may not only provide insight into how mutations can impact enhancer activity, but also yield aberrant eRNAs with altered structures and functions.

In support of a function for non-coding RNAs (ncRNAs) in ALS pathogensis, there is evidence of the involvement of lncRNAs in ALS. Several lncRNAs demonstrate interactions with the established ALS-related RBPs, TDP-43 and FUS (Lagier-Tourenne et al., 2012, Nishimoto et al., 2013, Tollervey et al., 2011, Wang et al., 2008b). With further evidence suggesting that FUS, upon allosteric modulation by promoter-associated ncRNAs, can inhibit the histone acetyltransferase activity of the key enhancer-associated cofactors, CBP and p300 (Wang et al., 2008b). Therefore, with a mechanistic role arising for ALS-associated ncRNAs, and with evidence of involvement of eRNAs in disease in general, we can imagine that eRNAs could potentially play a role in ALS pathogenesis.

### 6.1.4 CRISPR interference

CRISPR interference (CRISPRi) enables repression of mammalian gene expression. By utilising a catalytically inactive Cas9 (dCas9) tethered to functional inhibitory domains to interfere with transcription, CRISPRi can be utilised to occlude regulatory elements by sgRNA-programmed targeting of dCas9 (reviewed in Doench, 2018). Unlike the use of catalytically active Cas9 for the study of DNA function, CRISPRi does not induce DNA double-strand breaks so is therefore unimpaired by the effects of genomic rearrangements and DNA damage-associated toxicity (Kosicki et al., 2018, Meyers et al., 2017). Further, CRISPRi tends to yield a more homogeneous repression of gene function, while knockout studies utilising Cas9 often generate subpopulations of cells displaying distinct differences in residual protein abundance, due to the heterogeneity of nonsense-mediated decay efficiency (Smits et al., 2019).

Next generation CRISPRi effectors, including Zim3-KRAB, enables strong and consistent knockdown of gene expression, through the recruitment of endogenous epigenetic modulators to the target region. Zim3-KRAB-dCas9 provides both strong on-target knockdown effects and minimal off-target effects on the transcriptome (Alerasool et al., 2020, Replogle et al., 2022). In addition, the small size of Zim3-KRAB effector constructs relative to other CRISPRi effectors is advantageous for viral delivery methods, promoting highly efficient repression of target DNA, of particular importance for the targeting of non-coding regulatory elements (Figure 6.2)(Alerasool et al., 2020).

CRISPRi is an established approach for identifying the gene targets of *cis*-regulatory elements, with large-scale systematic genetic screens pairing CRISPRi with sgRNA libraries underpinning much of our current knowledge on gene-enhancer interactions (Fulco et al., 2016, 2019, Gasperini et al., 2019).

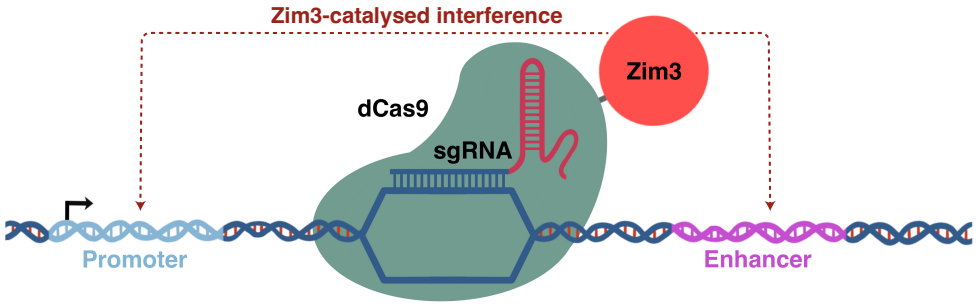



FIGURE 6.2: CRISPRi enables repression of gene expression via targeting of *cis*-regulatory elements
CRISPRi can characterise the regulatory effects of *cis*-regulatory elements, including promoters and enhancers, by targeted recruitment of an inhibitory effector, such as Zim3, by dCas9, to a region of interest.

### 6.1.5 Chapter aims

With both *CAV1* upregulation (Adey et al., 2023) and downregulation (Cooper-Knock et al., 2021) implicated in ALS, the overarching aim of this chapter is to understand the differential regulatory impacts of both individual enhancers, and ALS-associated SNPs on *CAV1* expression. We aimed to identify the impact of ALS-associated SNPs on enhancer function by characterising the regulatory function of individual ALS-associated enhancers, independently. We particularly focus on two enhancers, one distal located ~20kb downstream of the *CAV1* promoter, and one proximal, located within the *CAV1* 3' UTR. Firstly, we characterised how epigenetic manipulation of these enhancers using CRISPRi impacts the expression of *CAV1*. We establish how both the knockout of the

distal enhancer and ALS-associated SNPs influence *CAV1* mRNA transcription characteristics and eRNA expression from the proximal enhancer. Further, ALS-associated SNPs located within the 3'UTR-associated enhancer are tested in reporter assays to provide insight into their effect on regulatory potential. Additionally, this chapter includes computational prediction work, whereby motif discovery is performed on both enhancer DNA sequences, all in the context of ALS-associated SNPs.

## 6.2 Selection of candidate *CAV1* enhancers

To obtain a deeper understanding of the *CAV1* regulatory landscape in the context of ALS aetiology, we aimed to functionally characterise two model enhancers with the ultimate goal to understand the impact that ALS-associated SNPs may have on their enhancer function.

### 6.2.1 Proximal enhancer

Based on epigenomic characterisation and eRNA transcription data, we have identified a proximal enhancer of interest located across the 3' UTR of the *CAV1* gene. This candidate enhancer region, spanning chr7:116,196,622-116,201,351, is considered a tissue-agnostic *CAV1/CAV2* enhancer that also encompasses a CNS-specific *CAV1* enhancer at chr7:116,199,565-116,200,132. Both enhancers were in open chromatin in all iPSC-derived *in vitro* differentiated motor neuron samples, regardless of disease state, as indicated by ATAC-seq signal enrichment. Furthermore, these enhancers were co-enriched for H3K4me1 and H3K27ac ChIP-seq signal in all ALS patient-derived *in vitro* differentiated motor neuron lines, and the eldest healthy control-derived sample (Figure 6.3).

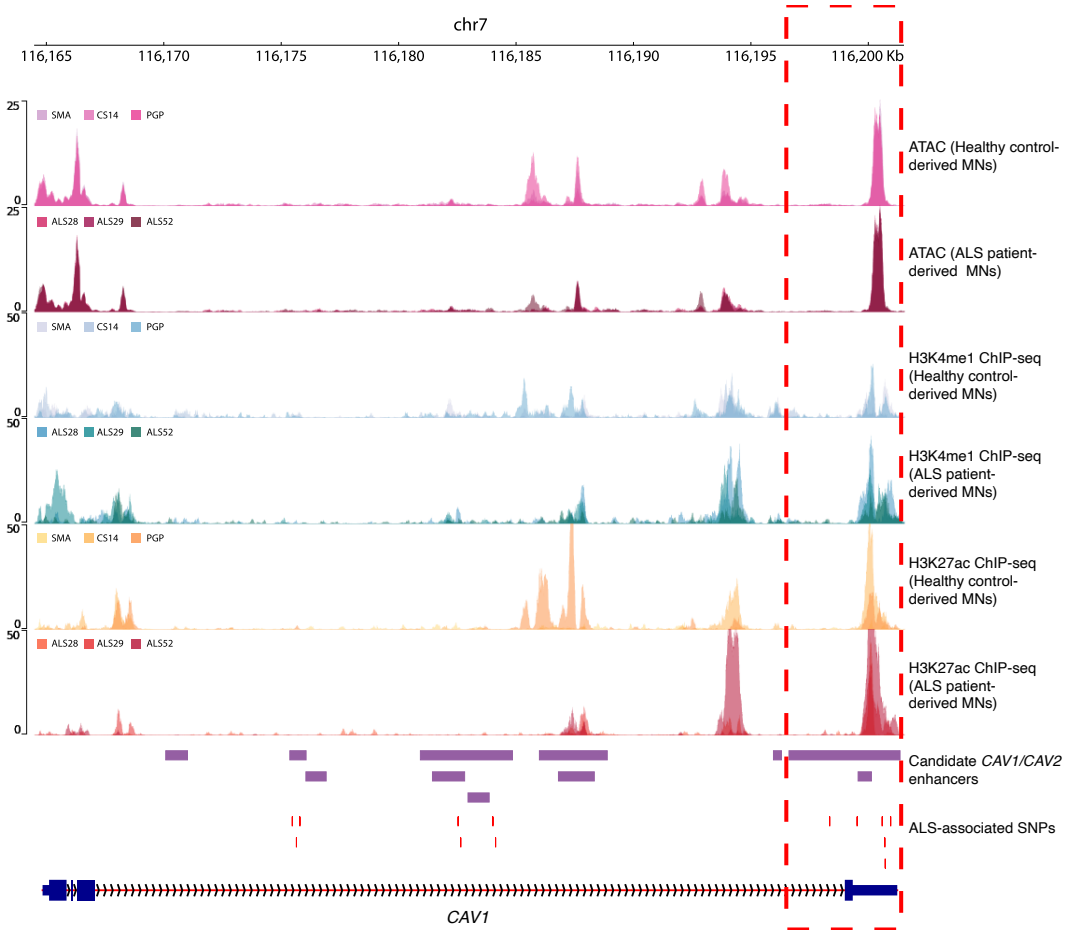



FIGURE 6.3: Enhancer characteristics at the *CAV1/CAV2* proximal enhancer in iPSC-derived motor neurons

Trackline at chr7:116,164,500-116,201,500 (hg19) demonstrating ATAC-seq in iPSC-derived motor neurons, H3K4me1 ChIP-seq in iPSC-derived motor neurons, H3K27ac ChIP-seq in iPSC-derived motor neurons (Zhang et al., 2020a), candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019) and ALS-associated SNPs (Cooper-Knock et al., 2021). Position of the *CAV1/CAV2* proximal enhancer (chr7:116,196,622-116,201,351) is highlighted by red dashed box.

While these enhancers did not exhibit significant ATAC-seq signal in SH-SY5Y cells, they demonstrated the highest H3K27ac CUT&RUN signal enrichment of all candidate *CAV1/CAV2* enhancers, with 13.4-fold and 14.4-fold enrichment in H3K27ac CUT&RUN signal over H3 signal, for the tissue-agnostic and CNS-specific enhancers, respectively. This suggests that the proximal enhancer region is active in SH-SY5Y cells. In support, TT-seq in WT SH-SY5Y cells revealed transcriptional activity at this region, enabling the assembly of the novel antisense eRNA, *eCAV1_WT* (chr7:116,199,437-116,199,680), which demonstrated increased stability following the induction of the ALS-associated SNP chr7:116,199,522T>C (Figure 6.4).

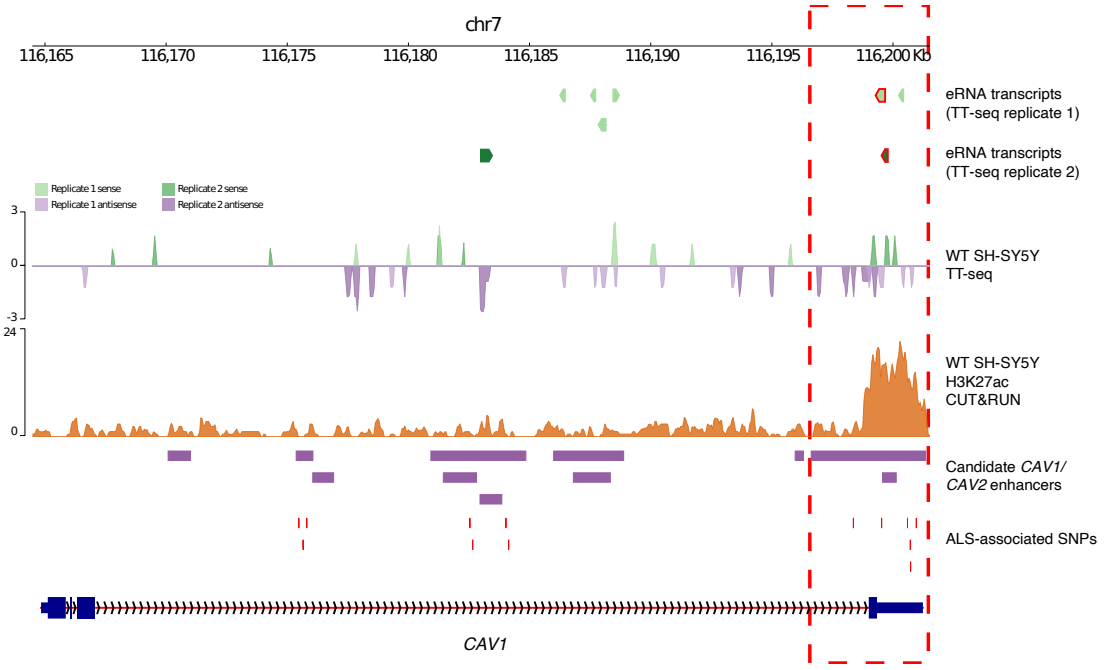



FIGURE 6.4: Enhancer characteristics at the *CAV1/CAV2* proximal enhancer in SH-SY5Y cells

Trackline at chr7:116,164,500-116,201,500 (hg19) demonstrating TT-seq in WT SH-SY5Y cells, including both assembled eRNA transcripts and coverage, H3K27ac CUT&RUN in WT SH-SY5Y cells, candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019) and ALS-associated SNPs (Cooper-Knock et al., 2021). Position of the *CAV1/CAV2* proximal enhancer (chr7:116,196,622-116,201,351) is highlighted by red dashed box. *eCAV1_WT* transcripts highlighted in red.

Interestingly, in human-derived lymphoblastoid cells, we observed significant transcriptional differences in this region. Average TT-seq coverage over the proximal enhancer was 7.9-fold and 6.3-fold higher in the ALS patient-derived lymphoblastoid cell lines, SP3363 and SP3388, respectively, compared to the healthy control-derived cell line, highlighting the potential functional relevance of this enhancer in ALS pathology (Figure 6.5).

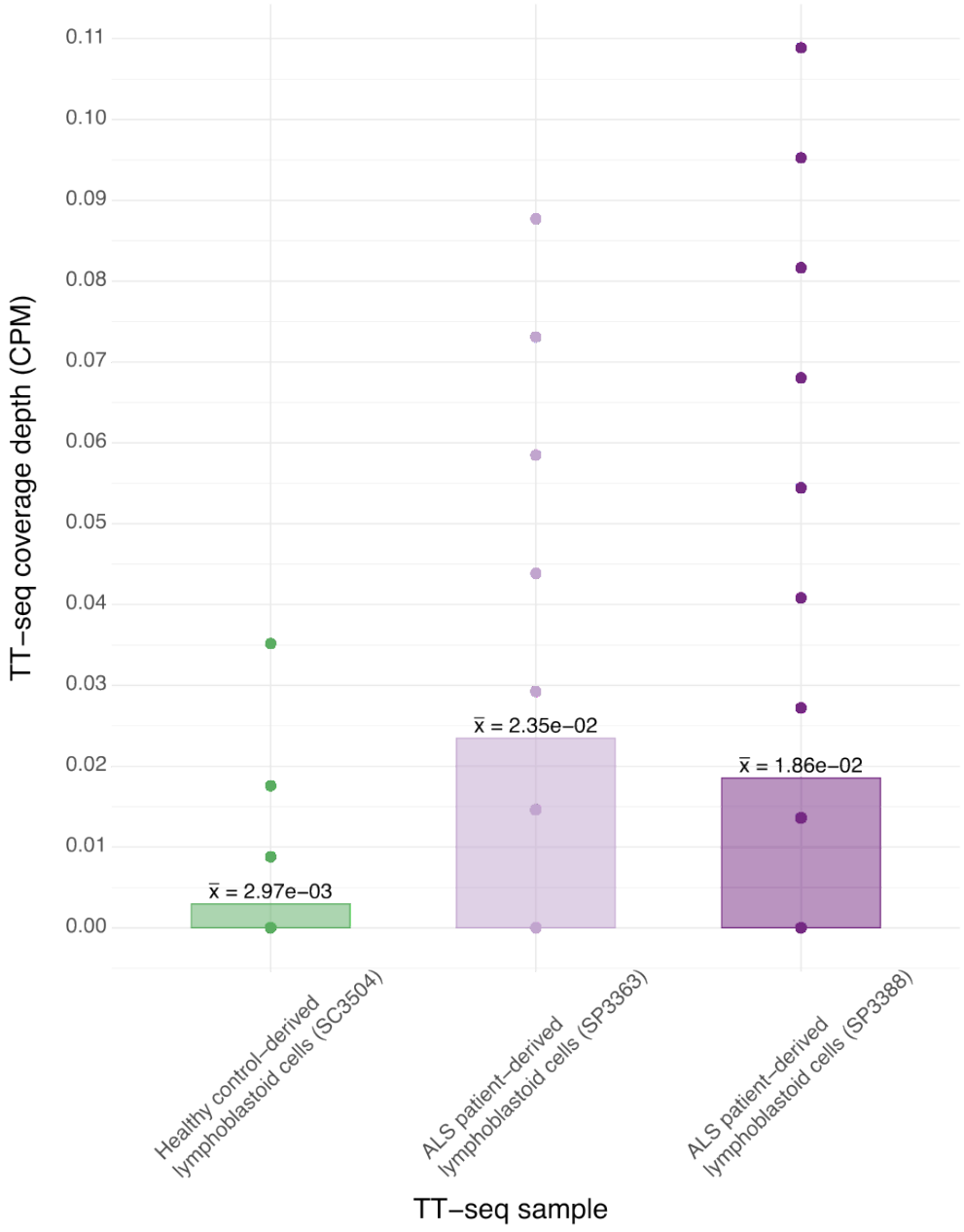



FIGURE 6.5: eRNA transcription at the *CAV1/CAV2* proximal enhancer in lymphoblastoid cells
TT-seq read coverage (CPM) in control- and patient-derived lymphoblastoid cells at the proximal *CAV1/CAV2* enhancer (chr7:116,196,622-116,201,351).

### 6.2.2 Distal model enhancer

To model the activity of distal *CAV1* enhancers, we utilised a candidate enhancer region ~20kb downstream of the *CAV1* promoter. Within this distal region there are two overlapping candidate enhancers, one tissue-agnostic *CAV1/CAV2* enhancer at chr7:116,220,523-116,224,531 and one CNS-specific *CAV1* enhancer at chr7:116,221,274-116,225,513. There is an established regulatory link between this region and *CAV1* and

*CAV2* transcription, with CRISPR/Cas9-mediated disruption of this region resulting in a significant decrease in both *CAV1* and *CAV2* mRNA expression (Cooper-Knock et al., 2021). Of relevance, the patient-derived lymphoblastoid cell line, SP3363, has the ALS-associated SNP, chr7:116,222,625T>C that lies within this distal model enhancer region. Interestingly, we observed a 2.4-fold decrease in TT-seq read coverage in the SP3363 cell line, compared to the healthy control line, for the distal enhancer region (Figure 5.25a). This contrasts the trend observed for the proximal model enhancer region, and may allow further insight into the functional relevance of distal enhancers in ALS aetiology.

### 6.2.3 Genomic location of the model enhancers

To explore how the epigenomic landscape may influence the activity of our model enhancers, we analysed chromosome topology, CTCF binding sites, and genic locations around these enhancer regions. Additionally, we examined the presence of ALS-associated SNPs within these enhancers to gain further insights into their potential role in ALS progression (Figure 6.6). Both of these enhancers are situated within the *CAV1/CAV2* TAD. Notably, the distal enhancer contains a CTCF binding site, likely due to its proximity to the TAD boundary, as supported by CTCF ChIP-seq. Of particular significance to the functional characterisation of enhancers implicated in ALS pathology, six ALS-associated SNPs are located within the proximal enhancer, while five are found within the distal enhancer region.

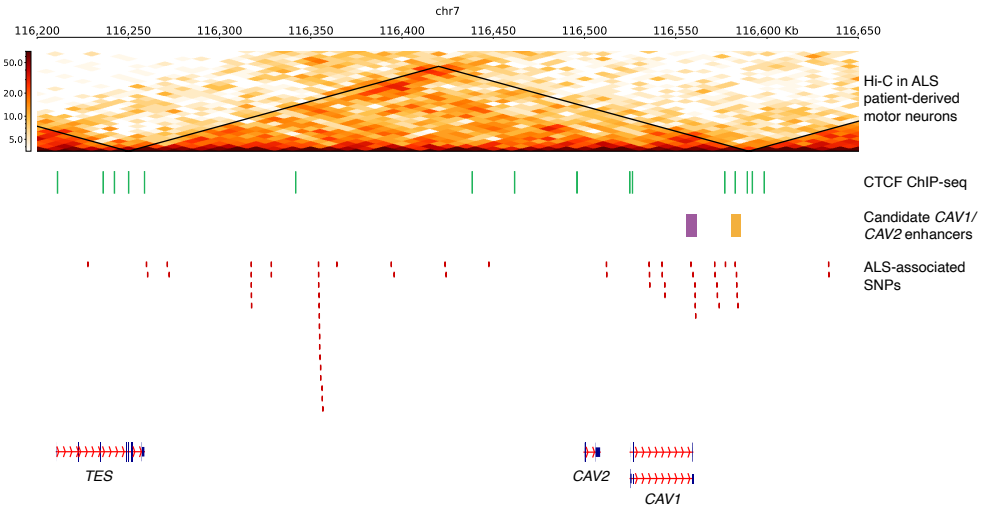



FIGURE 6.6: Epigenomic landscape surrounding proximal and distal model *CAV1/CAV2* enhancers

Chromosome topology, as determined by Hi-C contact frequencies at chr7:116,200,000-116,650,000 (hg38) in patient fibroblast-derived iPSCs differentiated *in vitro* into motor neurons (Zhang et al., 2022), CTCF binding as determined by CTCF ChIP-seq in SK-N-SH neuroblastoma cells (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020), the model candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019) and ALS-associated SNP within candidate *CAV1/CAV2* enhancers identified by GWAS (Cooper-Knock et al., 2021). Proximal enhancer is shown in purple and distal enhancer is shown in yellow.

## 6.3 ALS-associated enhancers and their impact on *CAV1* transcription

### 6.3.1 How does repression of ALS-associated enhancers affect the expression of *CAV1*?

Despite multiple lines of evidence supporting enhancer activity in our model enhancers - such as the enrichment of epigenomic features characteristic of enhancers and nascent transcription - this alone does not establish a direct gene-enhancer relationship with *CAV1*. To validate that our model enhancers do indeed regulate *CAV1* expression, we targeted the CRISPRi effector, dCas9-Zim3 to our model enhancers to assess the impact on *CAV1* mRNA expression.

#### 6.3.1.1 Validation of dCas9 expression in dCas9-Zim3 SH-SY5Y cells

We generated a SH-SY5Y cell line stably expressing the dCas9-Zim3 CRISPRi effector (via transduction of the pHR-UCOE-SFFV-Zim3-dCas9-P2A-Hygro construct) for

further transduction with lentiGuide-Puro constructs containing sgRNA sequences targeting regions of interest. To demonstrate successful stable expression of dCas9 in the SH-SY5Y cell line expressing dCas-Zim3, we performed a western blot for Cas9, comparing expression relative to the WT SH-SY5Y cell line (Figure 6.7). Clear bands were observed for the dCas9-Zim3 SH-SY5Y cell line, while there was no band present for the WT SH-SY5Y cell line. This suggested that dCas9-Zim3 was successfully expressed in our dCas9-Zim3 SH-SY5Y cell line.

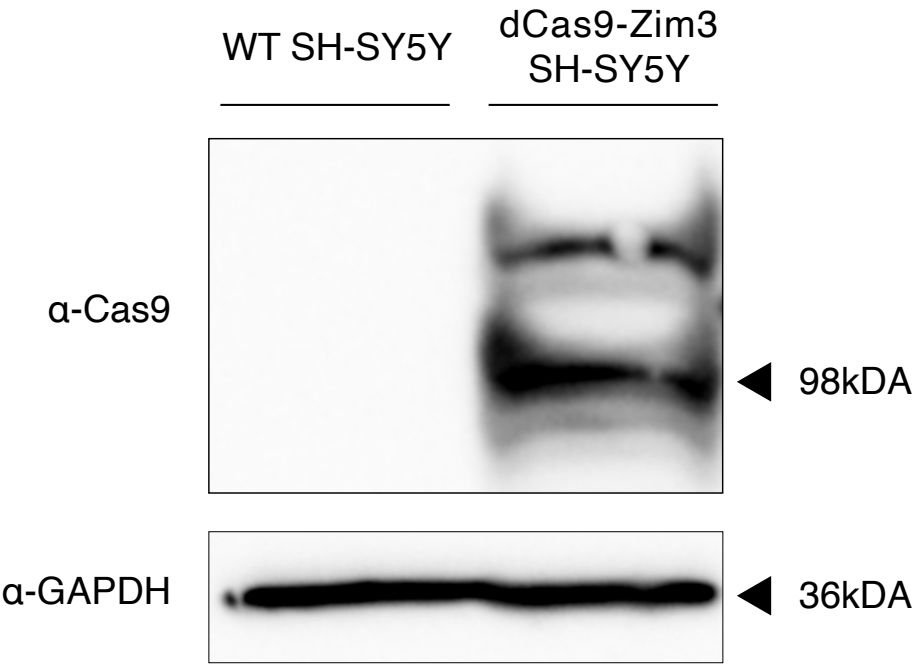



FIGURE 6.7: dCas9-Zim3 SH-SY5Y cells express Cas9
Western blot demonstrating stable expression of dCas9 in dCas9-Zim3 SH-SY5Y cells, confirmed by probing with $\alpha$-Cas9 (Abcam, ab191468) and $\alpha$-GAPDH (Proteintech, 60004-1-Ig). n=1.

#### 6.3.1.2 CRISPRi of model *CAV1/CAV2* significantly reduces *CAV1* expression

We targeted the CRISPRi effector domain to both the proximal and distal *CAV1* enhancers, the *CAV1* promoter (positive control) and with a scrambled sgRNA (negative control). We subsequently measured the expression of *CAV1* mRNA via RT-qPCR. As expected of regulatory elements implicated in *CAV1* expression, *CAV1* mRNA expression was significantly reduced relative to the scrambled control cell line for both the promoter-targeting CRISPRi line and all enhancer-targeting CRISPRi lines (Figure 6.8).

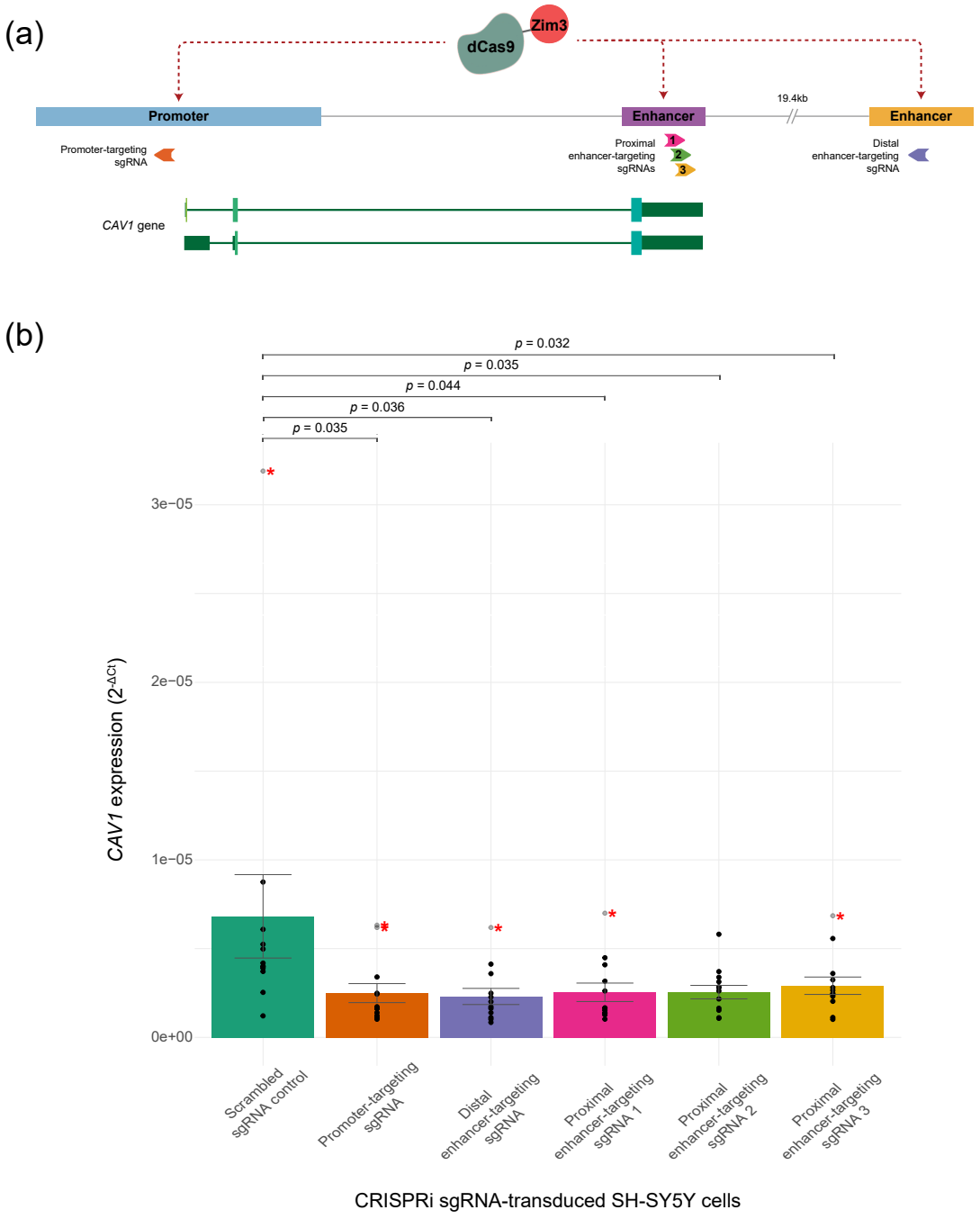



FIGURE 6.8: CRISPRi of candidate *CAV1* enhancers resulted in reduced *CAV1* expression

(a) sgRNA target locations utilised for targeting of dCas9-Zim3 to candidate *CAV1* promoter and enhancers. (b) Expression of *CAV1* mRNA in CRISPRi SH-SY5Y cell lines, as determined by RT-qPCR. Outliers identified based on the interquartile range upper bound and denoted by red asterisk. Error bars determined by standard error, *p*-values determined by Student's t-test. n=6

### 6.3.2 How does knockout of the distal *CAV1* enhancer impact *CAV1* transcriptional characteristics?

To model the impact that the chr7:116,222,625T>C ALS-associated SNP may have on *CAV1* expression in neuronally-relevant cells, Cooper-Knock et al. (2021) employed CRISPR/Cas9 to introduce indel mutations proximal to the site of mutation in SH-SY5Y cells. The gRNAs used for editing introduced a cut 13bp downstream of the chr7:116,222,625T>C mutation site, yielding predominantly a single-nucleotide insertion of chr7:116,222,638T>TT, with an overall editing efficiency of 72%. Additionally, they generated a control SH-SY5Y cell line using CRISPR/Cas9 to introduce a nonsense mutation into *CAV1* exon 2, allowing for comparison of the impact of enhancer disruption and exonic disruption on *CAV1* expression. In both enhancer knockout and exon knockout cells, a significant reduction in *CAV1* expression, compared to control cells, was observed (Cooper-Knock et al., 2021). As in Cooper-Knock et al. (2021), we performed RT-qPCR to measure *CAV1* mRNA expression in both the *CAV1* enhancer and exon knockout SH-SY5Y cell lines relative to WT expression. Due to the implications of CAV1 phosphorylation in both protein function and disease, we tested whether the enhancer knockout had an effect on transcription of the *CAV1α* mRNA isoform. We designed intron spanning qPCR primers specific to *CAV1α*, with the forward spanning across the junction of exon 1 and exon 2 of the *CAV1α* isoform. We also included primers targeting the CDS in the final exon shared by both mRNA isoforms to assess overall *CAV1* pre-mRNA production (Figure 6.9a,b).

As expected, we observed significantly less *CAV1* pre-mRNA expression in enhancer and exon knockout cells than in WT SH-SY5Y cells, with more expression observed in the enhancer knockout than exon knockout cells (Figure 6.9c). Conversely, when measuring expression of the *CAV1α* mature mRNA isoform, we actually observed significantly higher expression in the enhancer knockout cells, compared to WT SH-SY5Y cells (Figure 6.9c). This may suggest that while enhancer knockout results in a reduction in overall *CAV1* expression, it favours the production of the *CAV1α* isoform.

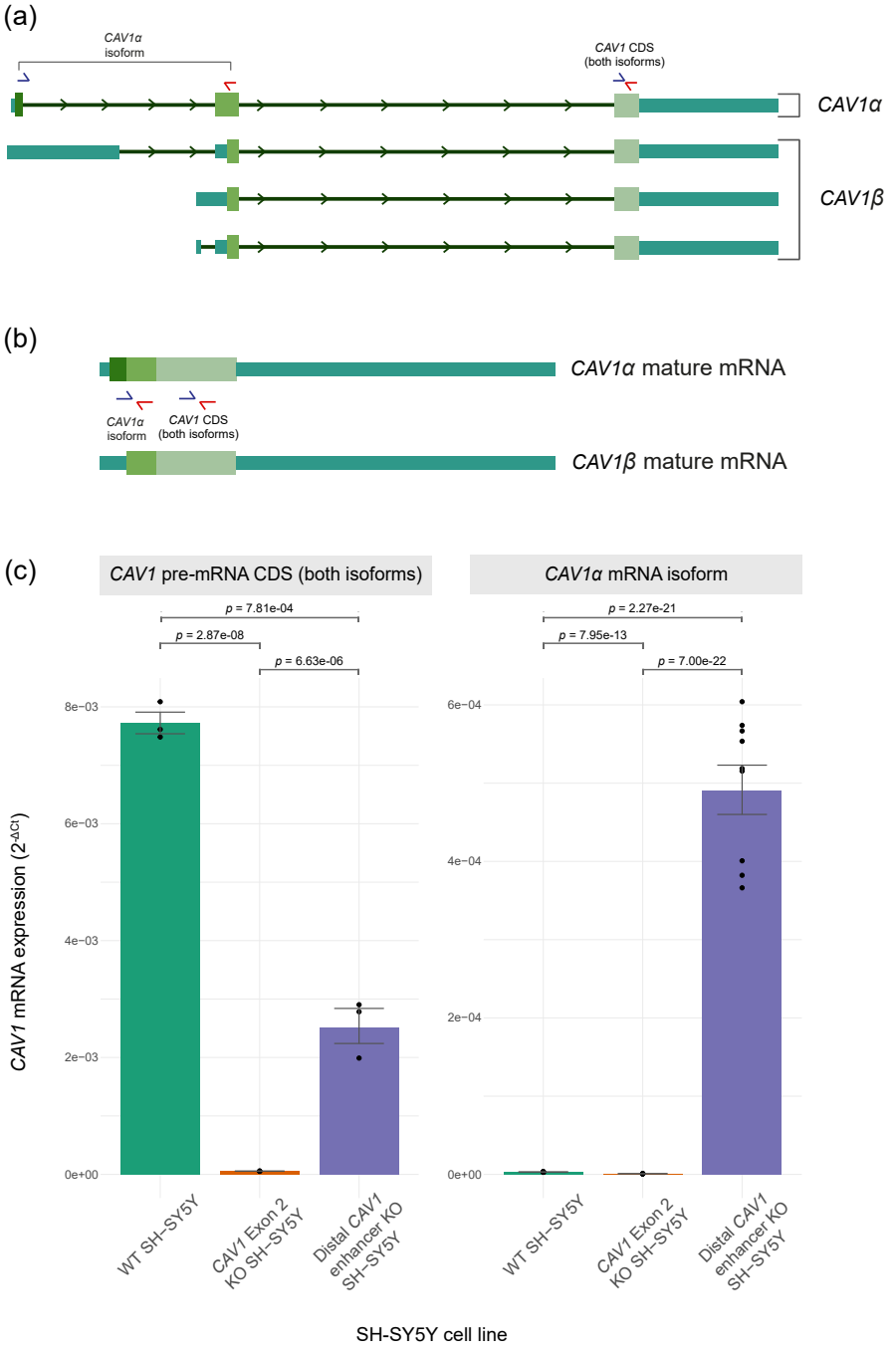



FIGURE 6.9: The impact of *CAV1* enhancer knockout on *CAV1* expression
(a) Schematic demonstrating the RT-qPCR primer locations relative to the genic location, demonstrating the genic structure of the *CAV1α* and *CAV1β* isoforms. (b) Schematic demonstrating the RT-qPCR primer locations relative to the structure of the *CAV1α* and *CAV1β* mature mRNA isoforms. (c) RT-qPCR for overall *CAV1* expression and *CAV1α* mature mRNA isoform expression in WT SH-SY5Y, exon 2 knockout and distal enhancer knockout cells. Error bars determined by standard error, $p$-values determined by unpaired t-test. n=3

### 6.3.3 *CAV1* transcriptional characteristics in patient-derived lymphoblastoid cells

The distal *CAV1* enhancer knockout in SH-SY5Y cells was generated to model the chr7:116,222,625T>C ALS-associated SNP, but did not contain the base change present in the patient sample. We utilised patient-derived lymphoblastoid cells to measure the impact that the bona fide SNPs had on *CAV1* expression. Using this, we can model the impact that the SNPs at chr7:116,222,625T>C (in SP3363 cells) and chr7:115,994,269C>T (in SP3388 cells) have on enhancer activity, and thus *CAV1* expression. This also allows us to compare the impact of the chr7:116,222,625T>C ALS-associated SNP to that of the CRISPR/Cas9-induced indel in SH-SY5Y cells, that not only appears to knockout the enhancer, significantly reducing overall *CAV1* mRNA expression, but also results in alterations in mRNA isoform bias.

We observed notable differences in overall *CAV1* mRNA expression across both ALS patient-derived lymphoblastoid cell lines compared to the healthy control-derived SC3504 cell line. The SP3363 ALS patient-derived lymphoblastoid cell line demonstrated a significant increase in both overall *CAV1* pre-mRNA expression and *CAV1α* mature mRNA expression, with a particularly substantial increase in total expression compared to the healthy control. While the SP3388 ALS patient-derived cell line showed no significant difference in *CAV1α* expression relative to the control, it did exhibit a significant increase in overall *CAV1* mRNA expression (Figure 6.10). Therefore, it appears that the presence of ALS-associated SNPs in *CAV1* enhancers actually results in an increase in *CAV1* expression. Contrary to the reduced overall *CAV1* expression coupled with increased *CAV1α* expression observed upon knockout of the model distal enhancer, presence of chr7:116,222,625T>C resulted in both an increase in overall *CAV1* expression and *CAV1α* mature mRNA.

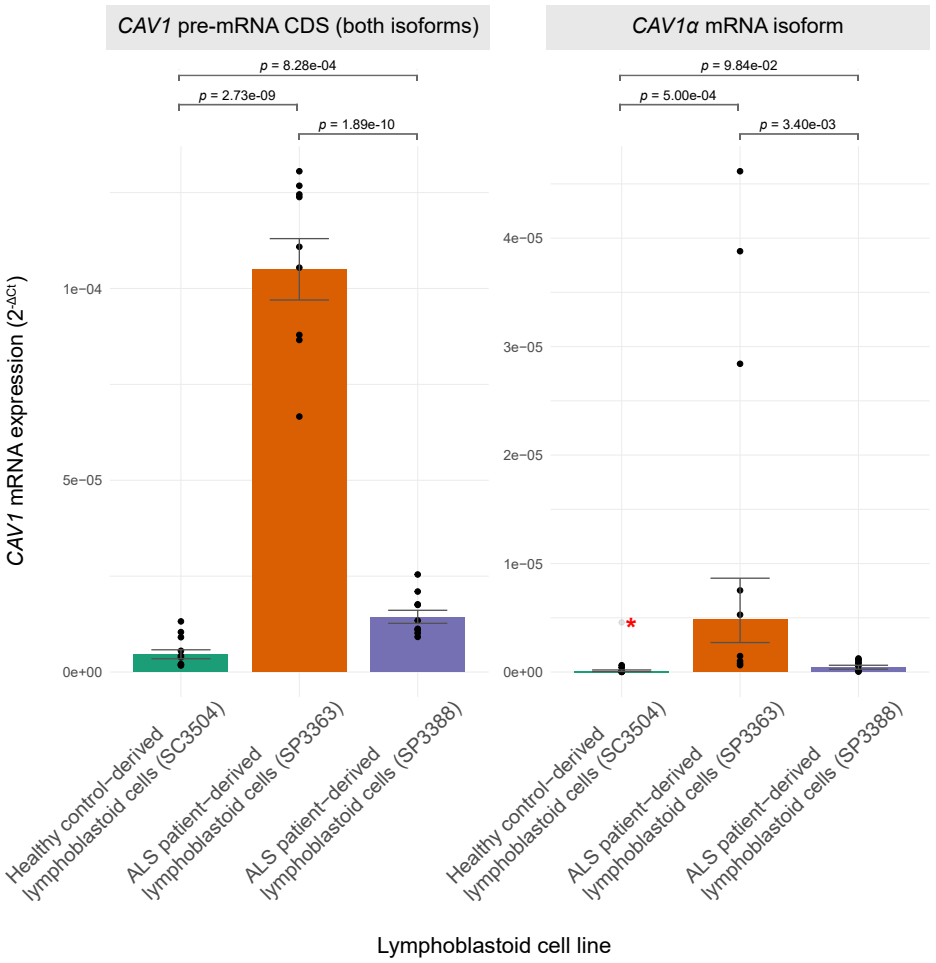



FIGURE 6.10: *CAV1* transcription characteristics in lymphoblastoid cells RT-qPCR for overall *CAV1* expression and *CAV1α* mature mRNA isoform expression in lymphoblastoid cells of healthy control and ALS patient origin. Outliers identified based on the interquartile range upper bound and denoted by red asterisk. Error bars determined by standard error, *p*-values determined by unpaired t-test. n=3

## 6.4 eRNA transcription from the proximal enhancer

### 6.4.1 Identification of *de novo* transcription start sites

In order to investigate eRNA transcription from the model *CAV1* proximal enhancer, we firstly examined the *CAV1* gene transcription termination site (TTS) in SH-SY5Y cells. Taking transcripts assembled from WT SH-SY5Y TT-seq data, and comparing to the hg19 reference genome (Pertea and Pertea, 2020), we identified that genic transcription terminated at chr7:116,199,152 and chr7:116,199,297, for replicates 1 and 2, respectively. This indicates that any transcription observed downstream could be attributed to enhancer activity from the model proximal enhancer.

To corroborate this, we identified TSSs throughout the length of the *CAV1* genic sequence, utilising published CAGE-seq data from WT SK-N-SH cells, downloaded from the ENCODE portal with the identifier ENCSR000CLC (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020). CAGE-seq specifically captures the 5' $m^7G$-cap of RNAs (Shiraki et al., 2003), making it a suitable method for detecting capped RNAs such as eRNAs, which like mRNAs, are often 5'-capped (Kristjánsdóttir et al., 2020). This approach allowed us to map not only the TSSs attributed to the *CAV1* mRNA, but also sites of *de novo* transcription within the model proximal *CAV1* enhancer. We identified 5 CAGE-seq peaks across the *CAV1* gene: chr7:116,165,035-116,165,125; chr7:116,166,411-116,166,457; chr7:116,199,286-116,199,998; chr7:116,200,558-116,200,616; chr7:116,200,716-116,200,776 (Figure 6.11a).

The peak at chr7:116,165,035-116,165,125 likely corresponded to the first exon of the *CAV1α* isoform, the peak at chr7:116,166,411-116,166,457 likely represented either the second exon of *CAV1α* or the first exon of *CAV1β*, and the peak at chr7:116,199,286-116,199,998 likely marked the final *CAV1* exon. The large peak at chr7:116,199,286-116,199,998 appeared to be attributed to multiple potential TSSs, as visually we observed peaks within both the coding exon and the 3' UTR. Additionally, we identified two *de novo* TSSs located further downstream within the *CAV1* 3' UTR, represented by the CAGE-seq peaks at chr7:116,200,558-116,200,616 and chr7:116,200,716-116,200,776, likely reflecting the presence of proximal enhancer-derived transcripts (Figure 6.11b).

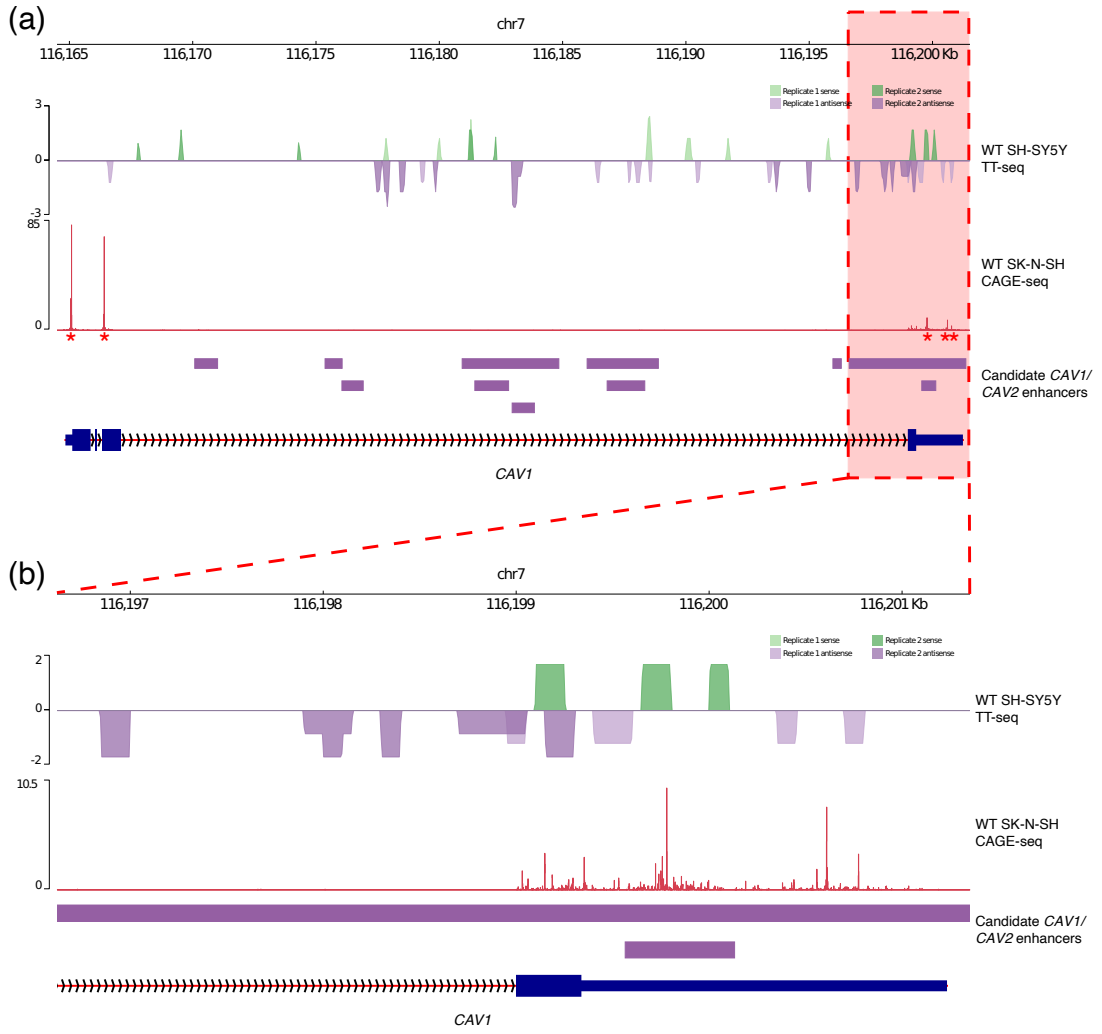



FIGURE 6.11: *De novo* transcription in SH-SY5Y cells at the *CAV1* gene locus (a) Trackline at chr7:116,164,500-116,201,500 (hg19) demonstrating TT-seq in WT SH-SY5Y cells, CAGE-seq in SK-N-SH neuroblastoma cells (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020) and the model candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019). Transcription start sites, as determined by CAGE-seq peaks, denoted by red asterisks. (b) Magnified view of trackline at chr7:116,196,622-116,201,351 (hg19) demonstrating TT-seq in WT SH-SY5Y cells, CAGE-seq in SK-N-SH neuroblastoma cells (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020) and the model candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019).

To provide confidence that the TT-seq coverage, and the corresponding CAGE-seq peaks that we observed across the 3' UTR are evidence of eRNA transcription, we compared TT-seq read coverage and TSSs, as determined by CAGE-seq, to the known canonical polyA sites for the *CAV1* gene (Figure 6.12). Using polyA_DB (Wang et al., 2018), we identified 6 polyA sites in the final exon of *CAV1*: chr7:116,199,356, chr7:116,199,431, chr7:116,199,495, chr7:116,199,682, chr7:116,201,233 and chr7:116,201,261. We observed a clear peak in transcription across the coding sequence of the final *CAV1* exon, with

transcription dropping to background levels around the most proximal polyA site. This, alongside the end coordinates of the *CAV1* mRNA transcripts assembled from TT-seq data, may suggest that *CAV1* genic transcription ends prior to the 3' UTR. In agreement with proximal polyA usage for *CAV1* in neuronally-relevant cells, publicly available polyA-seq in the human brain identified the *CAV1* polyA site at chr7:116,199,156-116,199,556 (Derti et al., 2012)(data accessible at NCBI GEO database (Edgar et al., 2002), accession GSM747470).

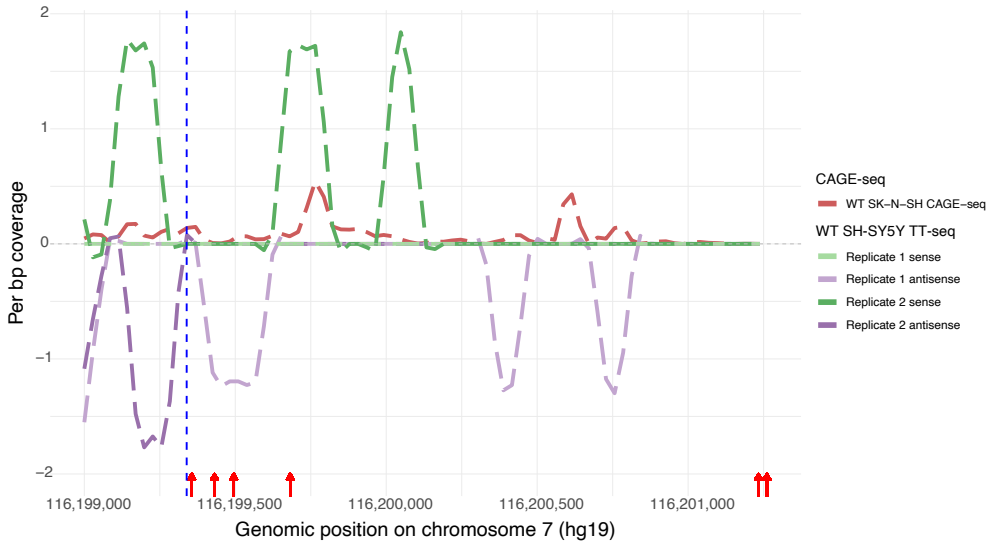



FIGURE 6.12: SH-SY5Y cells show preference for the proximal polyA site
Per bp coverage of WT SH-SY5Y TT-seq and WT SK-N-SH CAGE-seq (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020) across the final exon of *CAV1*(chr7:116,199,000:116,201,239). polyA sites are denoted by red arrows (Wang et al., 2018). End of the coding exon denoted by blue dashed line

### 6.4.2 Quantification of eRNA expression from the model proximal *CAV1/CAV2* enhancer

Based on the *de novo* TSSs observed in both our TT-seq data and the published CAGE-seq data, we investigated whether we could quantify these *de novo* eRNA transcripts, using RT-qPCR. We designed RT-qPCR primers to capture sense eRNAs at 116,200,698-116,200,855 and antisense eRNAs at 116,200,472-116,200,555 (Figure 6.13a).

#### 6.4.2.1 Quantification of eRNA expression in SH-SY5Y cells

We observed significantly higher sense eRNA transcription in WT SH-SY5Y cells than in SH-SY5Y cells containing the distal enhancer or the *CAV1* exon 2 KO, with transcription also significantly higher in the enhancer KO cells than in the exon KO cells. Conversely, while we observed significantly higher antisense eRNA transcription in WT SH-SY5Y

cells than in the exon 2 KO SH-SY5Y cells, there was no significant transcriptional difference between WT SH-SY5Y cells and distal enhancer KO SH-SY5Y cells. Further, there was no significant difference between transcription of the antisense eRNA between the distal enhancer KO cells and the exon 2 KO cells (Figure 6.13b).

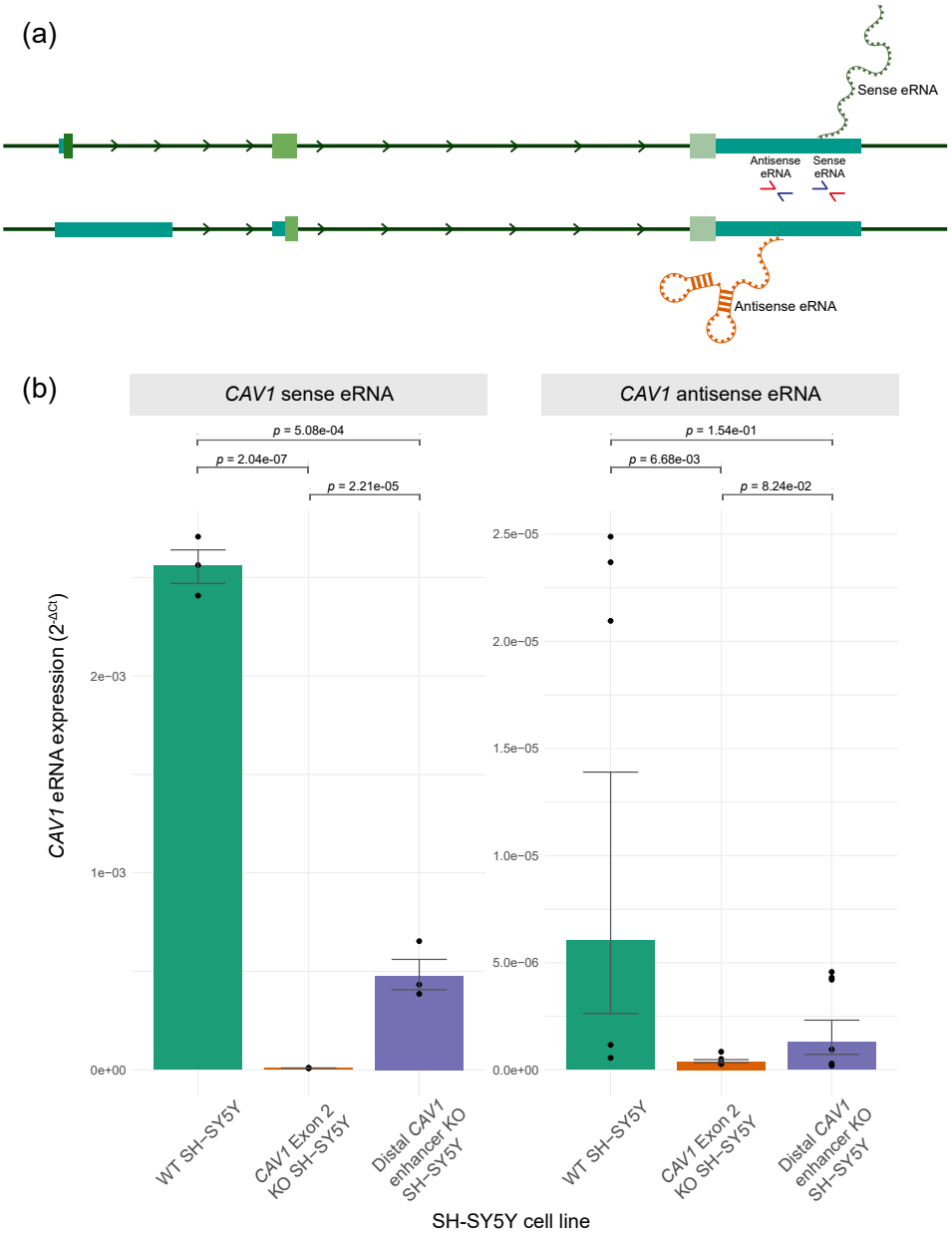



FIGURE 6.13: The impact of *CAV1* enhancer knockout on *CAV1* eRNA expression RT-qPCR for *CAV1* sense eRNA and antisense eRNA expression in WT SH-SY5Y, exon 2 knockout and distal enhancer knockout cells. Error bars determined by standard error, $p$-values determined by unpaired t-test. n=3

#### 6.4.2.2 Quantification of eRNA expression in lymphoblastoid cells

To explore the disease relevance of eRNA transcription at the proximal enhancer, we investigated the differences in sense and antisense eRNA expression in control- and patient-derived lymphoblastoid cells. As for SH-SY5Y cells, we investigated nascent transcription across the *CAV1* gene in lymphoblastoid cells compared to CAGE-seq peaks (Figure 6.14). Conversely to SH-SY5Y cells, we observed moderately consistent sense transcription across the 3' UTR. While this highlights differential PAS usage in distinct cell types, it also may suggest transcriptional read-through of *CAV1* gene transcription into the 3' UTR.

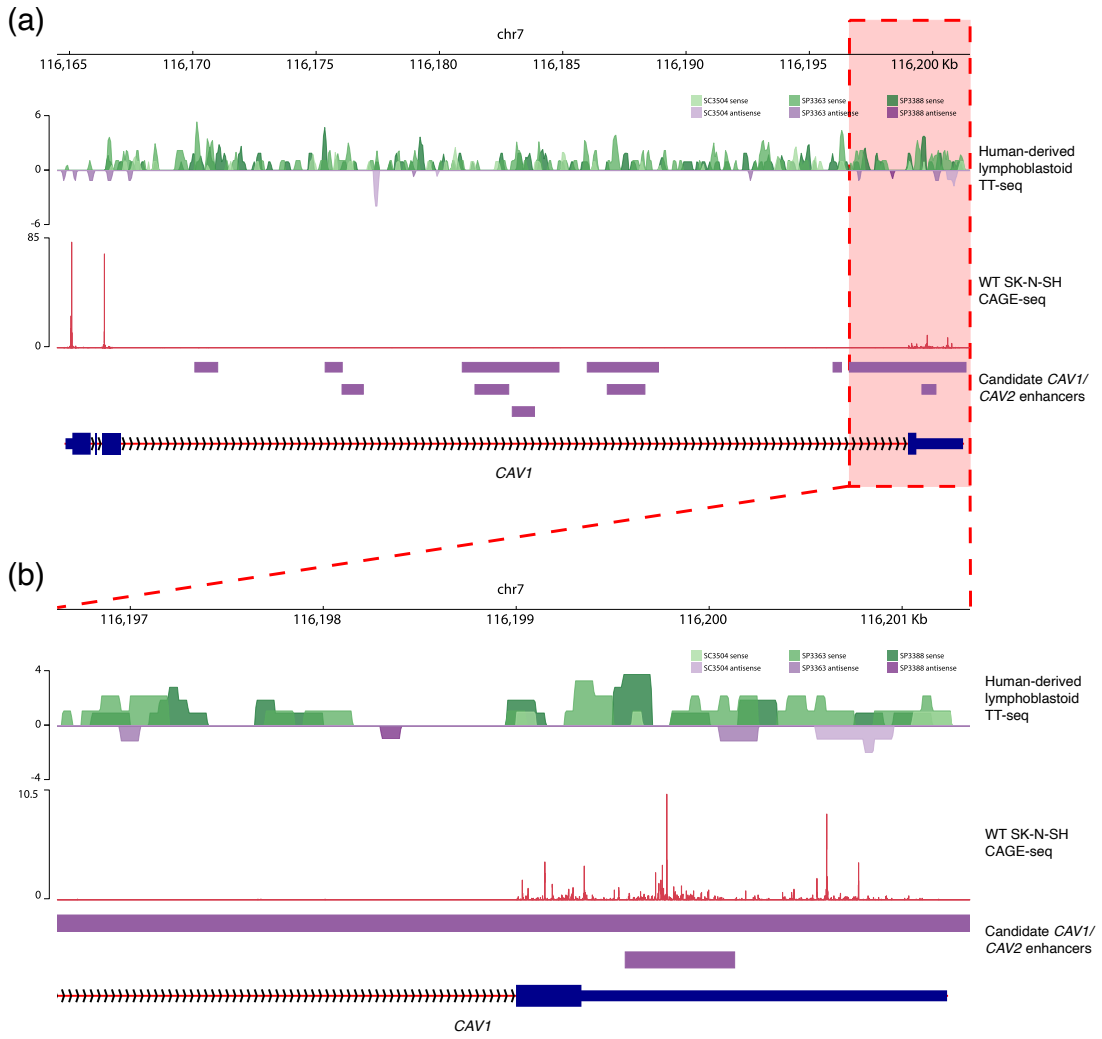



FIGURE 6.14: *De novo* transcription in lymphoblastoid cells at the *CAV1* gene locus (a) Trackline at chr7:116,164,500-116,201,500 (hg19) demonstrating TT-seq in lymphoblastoid cells, CAGE-seq in SK-N-SH neuroblastoma cells (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020) and the model candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019). (b) Magnified view of trackline at chr7:116,196,622-116,201,351 (hg19) demonstrating TT-seq in lymphoblastoid cells, CAGE-seq in SK-N-SH neuroblastoma cells (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020) and the model candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019).

We next examined eRNA expression levels by RT-qPCR (Figure 6.15). For both the sense and the antisense eRNA, we observed significantly more expression in the ALS patient-derived SP3363 cell line than in the healthy control-derived SC3504 cell line. When comparing both ALS patient-derived lymphoblastoid cell lines, for both eRNAs, we observed significantly more expression in the SP3363 cell line relative to the SP3388 cell line. Further, while there was no significant difference in the level of sense eRNA expression between the SC3504 cell line and the ALS patient-derived SP3388 cell line, for

the antisense eRNA, there was significantly more expression in the ALS patient-derived cell line compared to WT expression.

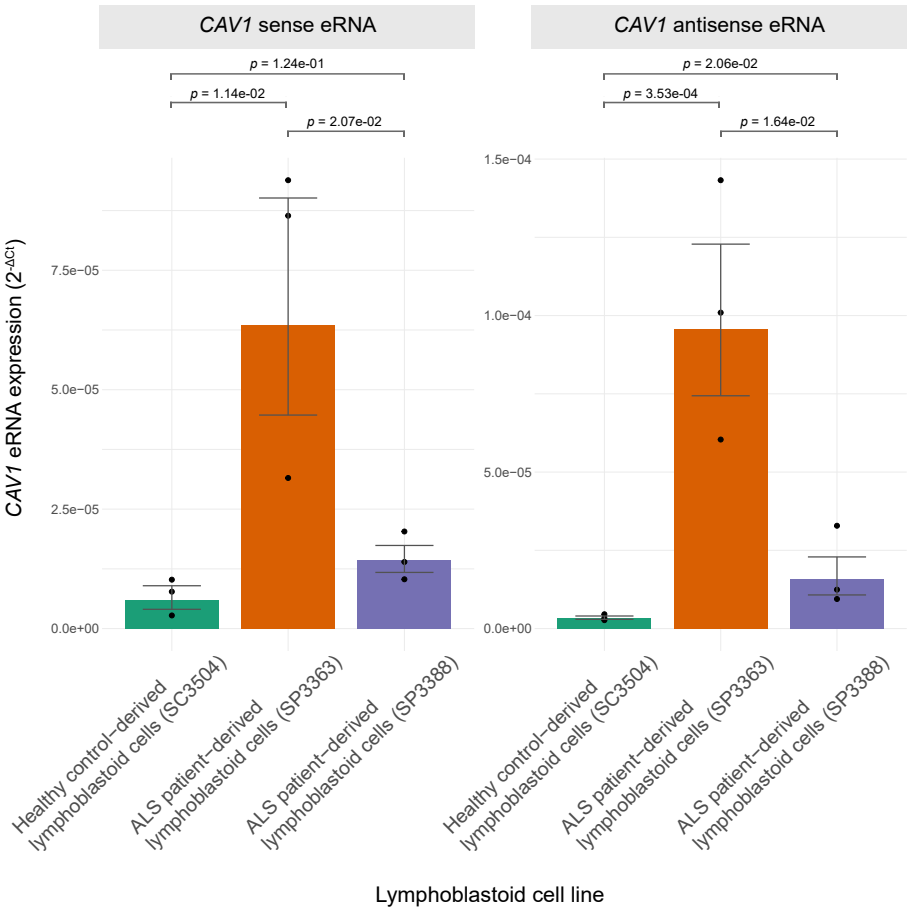



FIGURE 6.15: *CAV1* eRNA expression in lymphoblastoid cells
RT-qPCR for *CAV1* sense eRNA and antisense eRNA expression in control- and patient-derived lymphoblastoid cells of healthy control and ALS patient origin. Error bars determined by standard error, *p*-values determined by unpaired t-test. n=3

Taken together, we have established a direct functional link between the proximal and distal *CAV1/CAV2* enhancers and *CAV1* expression, highlighting a potential role for the distal enhancer in *CAV1* mRNA isoform choice. Further, by identifying sites of *de novo* eRNA transcription within the *CAV1/CAV2* proximal enhancer, we have assessed the potential cooperativity between the two enhancers, revealing alterations in eRNA transcription from the proximal enhancer following knockout or mutation of the distal enhancer.

## 6.5 The impact of CBP inhibition on *CAV1* mRNA and eRNA transcription

As previously established, the model proximal enhancer exhibits strong enrichment of H3K27ac CUT&RUN signal, which closely overlaps with sites of *de novo* eRNA transcription (Figure 6.16). With CBP a key histone acetyltransferase responsible for depositing acetyl groups on histones flanking enhancers, underpinning their dynamic activation (Hilton et al., 2015, Miao et al., 2022, Narita et al., 2021), we were interested to investigate the impact that CBP inhibition may have on *CAV1* mRNA and eRNA expression from the model proximal enhancer. We treated SH-SY5Y cells with A-485, a potent catalytic inhibitor of CBP/p300 (Lasko et al., 2017), and subsequently performed RT-qPCR to quantify *CAV1* mRNA and eRNA expression relative to untreated cells (Figure 6.17).

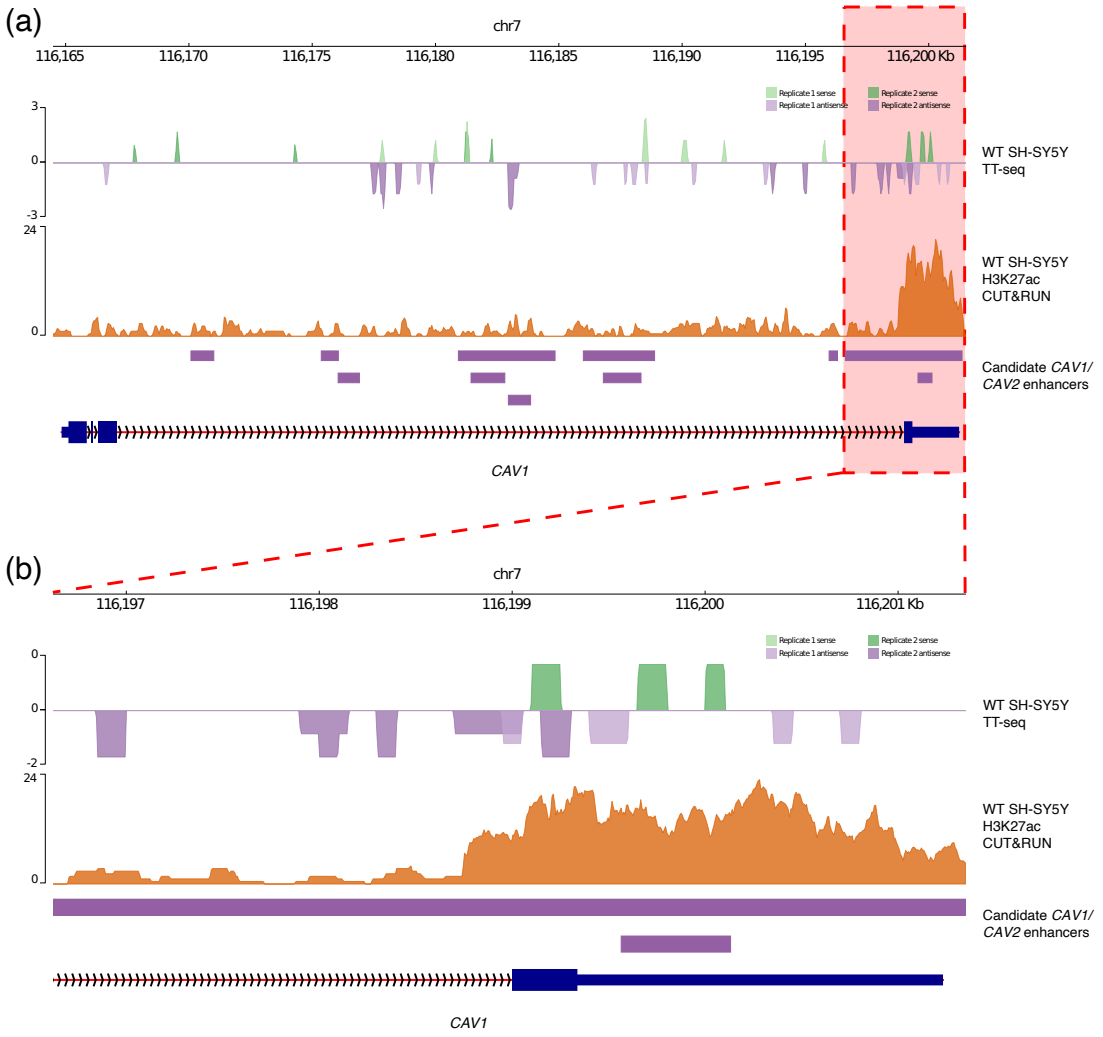



FIGURE 6.16: Enrichment of H3K27ac CUT&RUN signal across the model proximal *CAV1* enhancer

(a) Trackline at chr7:116,164,500-116,201,500 (hg19) demonstrating TT-seq in WT SH-SY5Y cells, H3K27ac CUT&RUN in WT SH-SY5Y cells and the model candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019). (b) Trackline at chr7:116,196,622-116,201,351 (hg19) demonstrating TT-seq in WT SH-SY5Y cells, H3K27ac CUT&RUN in WT SH-SY5Y cells and the model candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019)

CBP inhibition using A-485 resulted in a significant reduction in overall *CAV1* mRNA expression, relative to untreated SH-SY5Y cells. However, no significant change in the production of mature *CAV1α* mRNA was observed (Figure 6.17a). Indicative of CBP-dependent activity at the proximal model *CAV1* enhancer, we also detected a significant decrease in the expression of both sense and antisense eRNAs from the proximal enhancer upon, compared to WT SH-SY5Y cells (Figure 6.17b).

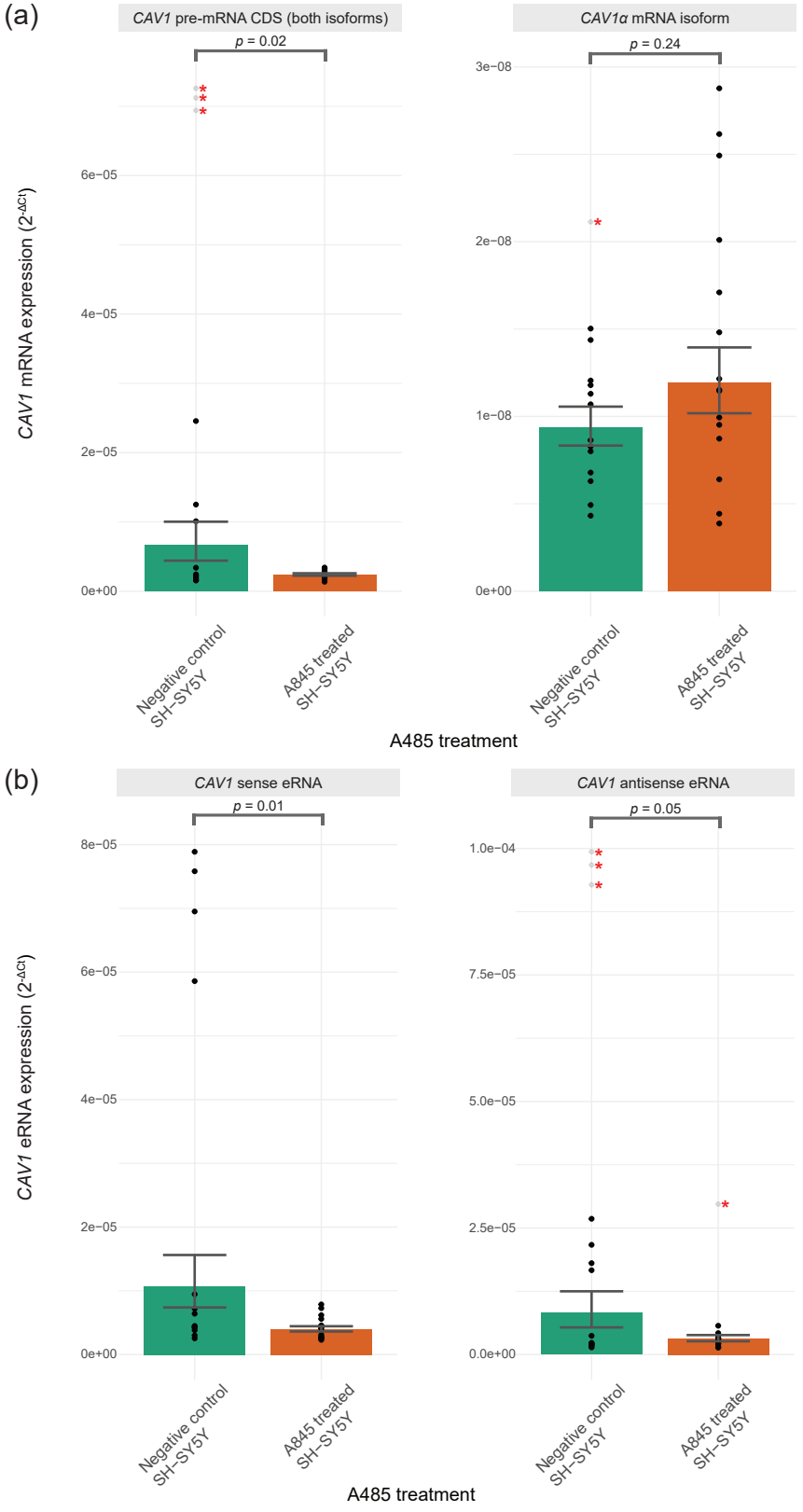



FIGURE 6.17: The impact of CBP inhibition on *CAV1* mRNA and eRNA expression (a) RT-qPCR for overall *CAV1* expression and *CAV1α* mature mRNA isoform expression in SH-SY5Y cells treated with A485 compared to WT SH-SY5Y cells. Error bars determined by standard error, $p$-values determined by unpaired t-test. n=3. (b) RT-qPCR for *CAV1* sense eRNA and antisense eRNA expression in SH-SY5Y cells treated with A485 compared to WT SH-SY5Y cells. Outliers identified based on the interquartile range upper bound and denoted by red asterisk. Error bars determined by standard error, $p$-values determined by unpaired t-test. n=3.

## 6.6 The impact of ALS-associated SNPs of *CAV1* enhancer regulatory potential

Reporter assays are a powerful tool for experimentally characterising enhancer function by directly measuring the impact of enhancer sequences on *in vitro* gene expression (reviewed in Alam and Cook, 1990). To quantitatively evaluate the regulatory potential of the proximal *CAV1/CAV2* enhancer, we performed a luciferase reporter assay in SH-SY5Y cells. In reporter assays, enhancer sequences are linked to reporter genes that produce a quantitative output, such as luminescence, providing insight into the functional activity of the enhancer (reviewed in Alam and Cook, 1990). In this work, we cloned the proximal model *CAV1/CAV2* enhancer sequence upstream of a simian virus 40 (SV40) promoter and a modified firefly (*Photinus pyralis*) luciferase gene, generating a luminescent signal proportional to the transcriptional activity driven by the enhancer. This allows for quantitative evaluation of enhancer activity, with the output of the reporter gene reflecting the regulatory capacity of the enhancer in driving gene expression in mammalian cells.

Beyond testing the regulatory function of the WT proximal *CAV1/CAV2* enhancer, we can also use the luciferase reporter system to test the basal regulatory impact of ALS-associated SNPs on enhancer function. Using site-directed mutagenesis, we introduced 5 ALS-associated SNPs into the enhancer sequence of the reporter construct: chr7:116,199,522T>C, chr7:116,200,589T>A, chr7:116,200,705C>T, chr7:116,200,719C>A and chr7:116,200,953A>T.

While the sensitivity of firefly luciferase reporter assays allows for effective monitoring of transcriptional upregulation by genetic regulatory elements, they can present challenges when assessing the impact of SNPs that may reduce regulatory efficiency. A reduced reporter signal can be ambiguous, potentially resulting either specific effects on the reporter or from global effects, such as cell death or variable transfection efficiency. To address this, normalisation of experimental results with a control reporter, such as *Renilla* luciferase, can aid in differentiation between specific regulatory effects and global effects (reviewed in Alam and Cook, 1990). In this work, we employed a dual-luciferase approach, using firefly and *Renilla* luciferase co-reporter constructs. Therefore, with the addition of equal amounts of the *Renilla* control reporter to each reaction, the expression of the firefly luciferase can be normalised to *Renilla* expression, accounting for potential variability. As neither luciferase enzyme is endogenously expressed in mammalian cells, any background luminescence detected is likely a characteristic of the luminometer or the luminescent substrate, ensuring accurate interpretation of enhancer activity.

We performed luciferase reporter assays to compare the regulatory effect of the WT proximal *CAV1/CAV2* enhancer to that of the enhancer containing each of the 5 ALS-associated SNPs (Figure 6.18a). As controls, we included a negative control reporter construct lacking both the SV40 promoter and enhancer (Figure 6.18b) and a positive control construct containing both the SV40 promoter and enhancer (Figure 6.18c). While the proximal *CAV1/CAV2* enhancer appeared to be a comparatively weak enhancer, relative to the strong SV40 enhancer, it drove significantly higher firefly luciferase expression than the negative control vector. Further, two ALS-associated mutations, chr7:116,200,589T>A and chr7:116,200,705C>T, significantly impaired enhancer-driven luciferase expression *in vitro*. Notably, the enhancer containing the chr7:116,200,589T>A SNP failed to drive luciferase expression beyond the levels observed in the negative control, which lacked an enhancer altogether (Figure 6.18d).

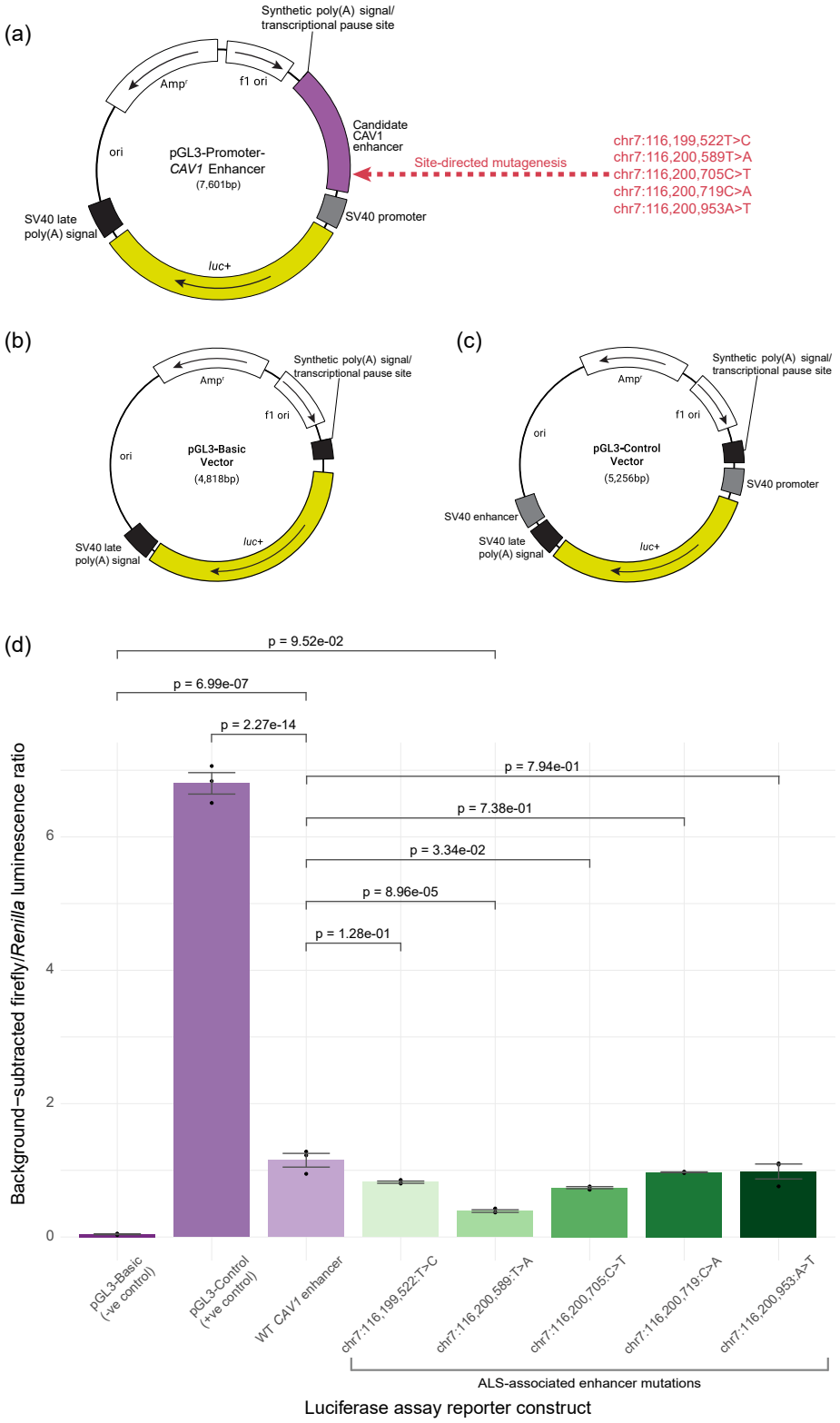



FIGURE 6.18: ALS-associated SNPs alter the regulatory function of the model proximal *CAV1/CAV2* enhancer

(a) Plasmid map of the luciferase reporter vector containing the model proximal *CAV1/CAV2* enhancer to test the potential of both the WT enhancer sequence and the enhancer sequence containing ALS-associated SNPs to drive mammalian gene expression. (b) Plasmid map of the pGL3-Basic negative control reporter vector. (c) Plasmid map of the pGL3-Control positive control reporter vector. (d) Background-subtracted firefly/*Renilla* luminescence ratio for the firefly reporter construct containing WT and mutated enhancer sequences in SH-SY5Y cells. Error bars determined by standard error, *p*-values determined by one-way ANOVA followed by Tukey's Honestly Significant Difference (HSD) test . n=3

## 6.7 Prediction of the impact ALS-associated SNPs have on protein binding motifs at enhancers

Utilising the JASPAR database for TF binding profiles (Rauluseviciute et al., 2024), we investigated whether any of the ALS-associated SNPs within the model proximal and distal *CAV1* enhancers interrupt putative TF binding sites. We filtered JASPAR TF binding sites to identify binding sites with a score of more than 400 ($p = 10^{-4}$), where the score indicates the statistical significance of the match between the TF binding profile and the sequence at the genomic position of interest. Following filtering, we observed TF binding site motifs for ZNF354A482, MEF2A, MEF2B, MEF2C and MEFD that overlap with the site of the ALS-associated SNP chr7:116,199,522T>C. Of particular interest were the MEF2 family of DNA-binding proteins, as MEF2 is implicated in neuronal survival (Gong et al., 2003, Mao et al., 1999, Yang et al., 2009) and is a key regulator of activity-dependent synapse development (Flavell et al., 2006, Shalizi et al., 2006). Homo- and heterodimers of MEF2 proteins bind directly to DNA regions possessing the MEF2 response element consensus sequence YTA(A/T)4TAR (Gossett et al., 1989). Almost exact matches to this consensus sequence were found on the reverse strand at chr7:116,199,520-116,199,531. Hence, the mutation site lies directly within the consensus binding motif (Figure 6.19).

MEF2 proteins serve dual functions, acting as transcriptional activators or repressors depending on their interaction with co-factor complexes, including epigenetic co-regulators that modify chromatin structure. Calcium-dependent mechanisms dictate the role of MEF2 in gene regulation, modulating the competitive binding of histone acetyltransferases and histone deacetylases. In the resting state, MEF2 associates with co-repressors and histone deacetylases to suppress the expression of target genes (Lu et al., 2000, Miska et al., 1999, Sparrow et al., 1999). Upon activation, calcium-dependent mechanisms trigger the dissociation of co-repressors, allowing MEF2 to recruit co-activators such as CBP/p300 to initiate transcription (He et al., 2011, Youn et al., 2000).

The position weight matrix (PWM) analysis for each of the MEF2 TFs provides insight into the impact that the chr7:116,199,522T>C SNP may have on these binding motifs. PWMs quantify the DNA-binding specificity of TFs, with each value in the matrix indicating the frequency or probability of a nucleotide occurring at a given position within the motif, based on experimental data (Rauluseviciute et al., 2024). At chr7:116,199,522, both MEF2B and MEF2D show a maximum binding preference for A, with a frequency of 1.00, suggesting that the chr7:116,199,522T>C SNP could disrupt their binding. In contrast, MEF2A and MEF2C exhibit frequencies of 0.96 and 0.99 for A at the same position, respectively. With the chr7:116,199,522T>C SNP resulting in a base change to

a G, there is potential that the binding potential for MEF2A and MEF2C may remain, though diminished, with frequencies of 0.03 and 0.01 for G, respectively. This indicates that MEF2A and MEF2C may still be able to bind at this position, albeit with reduced affinity, in the presence of the SNP.

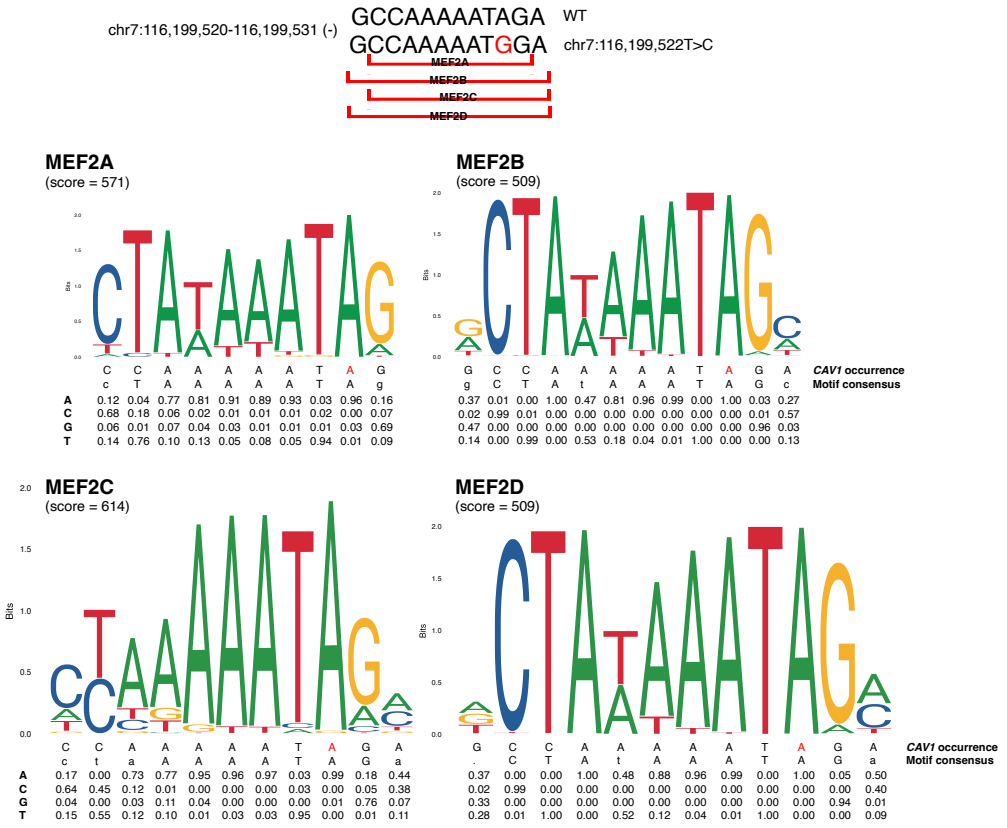



FIGURE 6.19: MEF2 binding motifs within the model proximal *CAV1* enhancer
MEF2 binding site motifs and position weight matrices from the JASPAR database (Rauluseviciute et al., 2024) at chr7:116,199,520-116,199,531 (-), comparing the *CAV1* sequence occurrence and the motif consensus, with sites of the chr7:116,199,522T>C shown in red.

### 6.7.1 Transcription factor motifs in the model distal *CAV1/CAV2* enhancer

When considering TF binding to the model distal *CAV1/CAV2* enhancer, we observed CTCF ChIP-seq peaks at chr7:116,222,506-116,222,756, suggesting that the chr7:116,222,625T>C SNP lies in the centre of a CTCF binding site (Figure 6.20). However, upon investigating the genomic positions of the predicted CTCF binding motifs, we identified the primary sequence motif at chr7:116,222,682-116,222,714 and the shorter secondary motif at chr7:116,222,682-116,222,696 (Rauluseviciute et al., 2024, Schmidt et al., 2012), suggesting that the chr7:116,222,625T>C SNP does not disrupt the CTCF binding site.

Interestingly, we did observe predicted binding motifs across chr7:116,222,625, namely ETV5::FOXO1, FOXO1::ELK3 and FOXO1::FLI1, suggesting that FOXO1, while interacting with ETV5, ELK3 or FLI1, may be able to bind at this region (Figure 6.21).

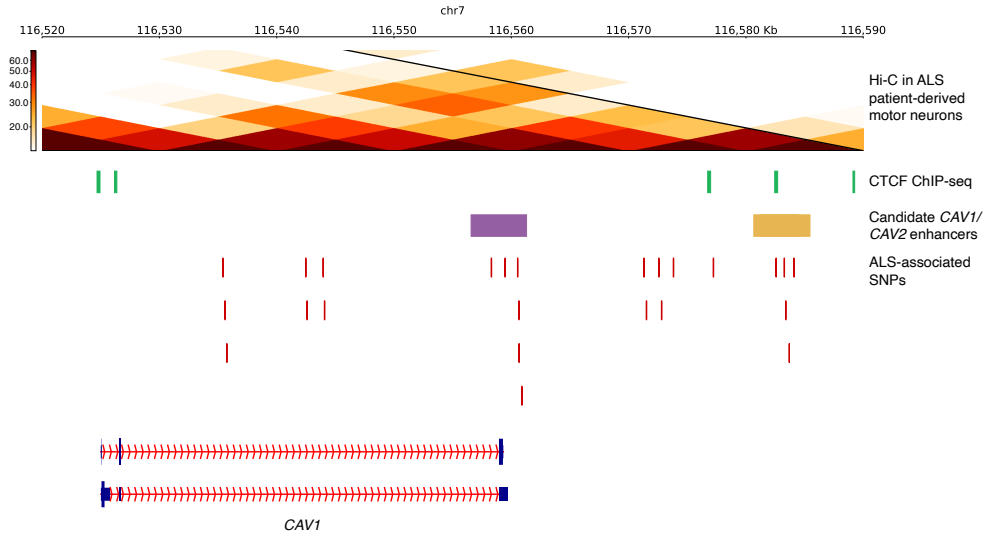



FIGURE 6.20: CTCF binding at the model distal *CAV1* enhancer
Chromosome topology, as determined by Hi-C contact frequencies at chr7:116,520,000-116,590,000 (hg38) (Zhang et al., 2022); CTCF binding as determined by CTCF ChIP-seq in SK-N-SH neuroblastoma cells (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020); the model candidate *CAV1/CAV2* enhancers identified by the ABC model (Cooper-Knock et al., 2021, Fulco et al., 2019); ALS-associated SNPs within candidate *CAV1/CAV2* enhancers identified by GWAS (Cooper-Knock et al., 2021). Proximal enhancer is shown in purple and distal enhancer is shown in yellow.

The PWM analysis for each motif demonstrated that at chr7:116,222,625, the WT T nucleotide is the preferred base at this position, with a frequency of 0.55, 0.49 and 0.55 for ETV5::FOXO1, FOXO1::ELK3 and FOXO1::FLI1, respectively. The ETV5::FOXO1 motif is the most likely motif disrupted by chr7:116,222,625T>C, with the mutant C nucleotide being the nucleotide with the lowest frequency at only 0.02, suggesting that the SNP may disrupt binding here. For both FOXO1::ELK3 and FOXO1::FLI1 the frequency of C at this position is 0.15, therefore the SNP is less likely to have an impact on these binding motifs, and thus may have a lesser effect.

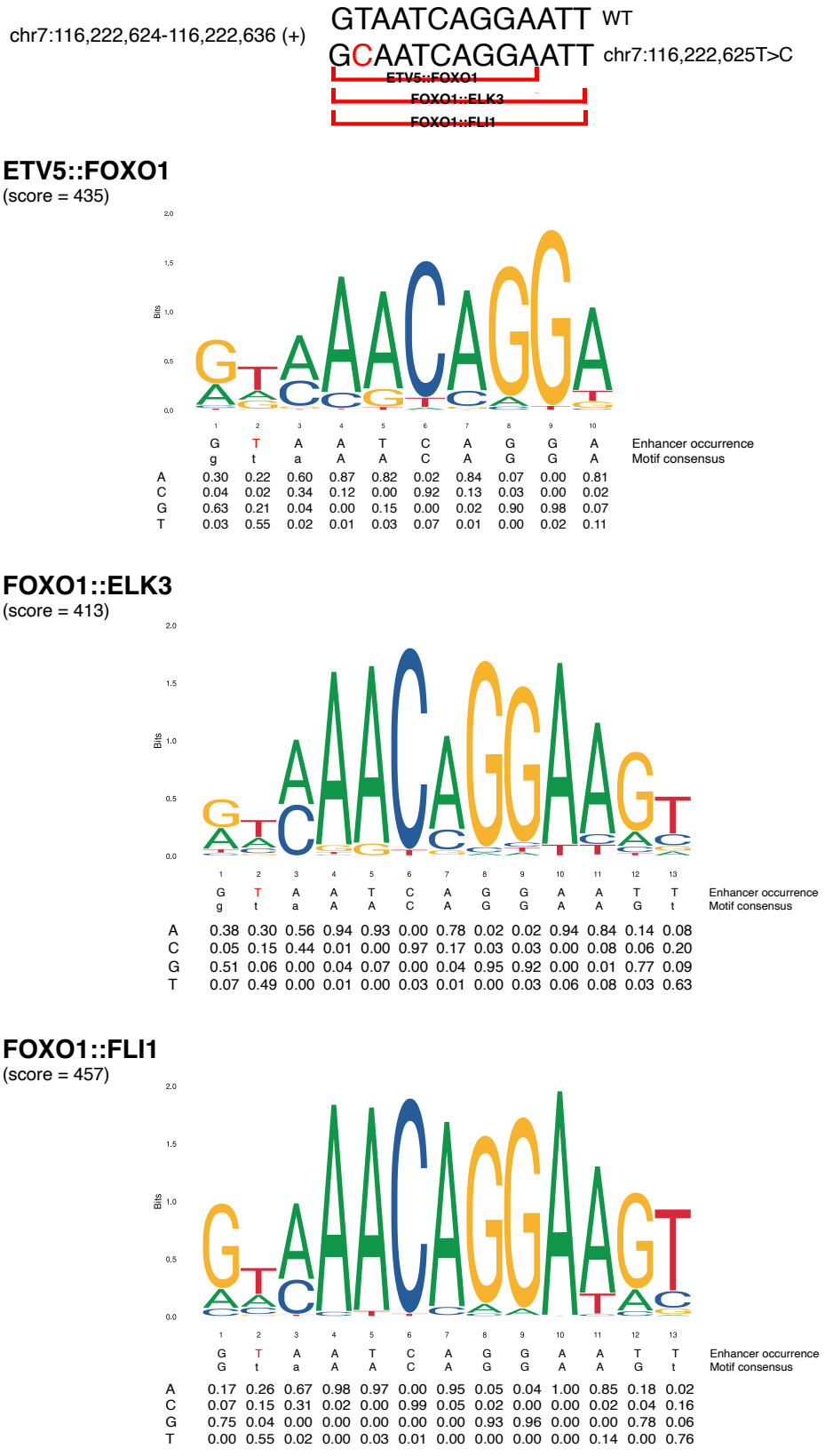



FIGURE 6.21: FOXO1::ETS binding motifs within the model distal *CAV1* enhancer ETV5::FOXO1, FOXO1::ELK3, FOXO1:FLI1 binding site motifs and position weight matrices from the JASPAR database (Rauluseviciute et al., 2024) at chr7:116,22,624-116,222,636 (+), comparing the *CAV1* enhancer sequence occurrence and the motif consensus, with sites of the chr7:116,222,625T>C shown in red.

## 6.8 Discussion

### 6.8.1 Establishing a functional link between the model *CAV1* enhancers and *CAV1* gene expression

Despite strong evidence supporting the enhancer activity of our model enhancers, the presence of characteristic epigenomic marks and eRNA transcription alone does not establish a direct functional link between these enhancers and *CAV1* expression. To address this, we confirmed the identity of the proximal enhancer at chr7:116,196,622-116,201,351 and the distal enhancer at chr7:116,220,523-116,225,513 as *CAV1* enhancers using CRISPRi (Figure 6.8b). Targeting the Zim3-KRAB transcriptional effector to our model enhancers reduced *CAV1* expression, relative to the negative control, to the same extent as targeting the *CAV1* promoter. This demonstrates that these enhancers significantly impact *CAV1* expression to a comparable magnitude as the *CAV1* promoter.

### 6.8.2 Do *CAV1* enhancers play a role in *CAV1* mRNA isoform choice?

#### 6.8.2.1 Differential *CAV1* mRNA isoform expression

The two isoforms of CAV1, CAV1$\alpha$ and CAV1$\beta$, have distinct functional cellular roles (Fujimoto et al., 2000), including in human disease (Hau et al., 2019, Meng et al., 2017). The ratio of the two isoforms can be controlled at the transcriptional level, through the translation of distinct mRNAs (Kogo and Fujimoto, 2000), formed through the use of alternative TSSs.

Following model distal enhancer knockout, despite an overall decrease in total *CAV1* expression, we observed an increase in *CAV1$\alpha$* mature mRNA expression, relative to expression in WT SH-SY5Y cells (Figure 6.9c). The increased production of *CAV1$\alpha$* following enhancer knockout suggests that the model distal enhancer may regulate differential mRNA isoform usage. With the *CAV1* mRNA isoforms formed by alternative initiation of gene transcription (Kogo and Fujimoto, 2000), there is a possibility that the distal enhancer may promote gene transcription from distinct TSSs.

Many mammalian promoters are composite elements composed of several TSSs, whereby regulatory inputs determine the selection of specific TSSs to generate cell type-specific expression profiles in a dynamic and temporal manner (Chia et al., 2021, FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014). When multiple promoters and enhancers work in concert to control gene expression, their interactions often give rise to complex and varied regulatory outcomes. These regulatory hubs enable interactions that may underpin multiple modes of enhancer-mediated transcription

regulation. Enhancers can function promiscuously, regulating multiple genes (Symmons et al., 2014), or conversely may have an inherent preference for a particular promoter (Juven-Gershon et al., 2008). Furthermore, enhancers can function in synergy (Ghavi-Helm et al., 2014, Oudelaar et al., 2018), whereby their regulatory effects can combine in an additive, superadditive, subadditive or even repressive fashion (reviewed in Kvon et al., 2021). Alternatively, enhancers may function in a mutually exclusive manner, competing for promoter activation (Bothma et al., 2015, Scholes et al., 2019).

Knockout of the model distal enhancer does not entirely abrogate *CAV1* expression, in fact, we observed significantly more *CAV1* expression in the enhancer knockout SH-SY5Y cell line than in the *CAV1* exon 2 knockout cell line (Figure 6.9c). This suggests that the distal enhancer alone does not govern *CAV1* mRNA expression, and multiple enhancers are likely working cooperatively to control *CAV1* expression. In fact, we have established the combinatorial nature of enhancer-mediated *CAV1* expression regulation, whereby targeting the inhibitory epigenetic modulator, Zim3-KRAB, to both the proximal and distal enhancers produced uniform reductions in *CAV1* gene expression, compared to the negative control (Figure 6.8b). With CAV1$\alpha$ and CAV1$\beta$ having functionally distinct roles in caveolae (Fujimoto et al., 2000), tight control of differential *CAV1* mRNA isoform expression may be required to maintain the precise balance of the different isoforms. Therefore, synergism or competition between the distal enhancer with additional enhancer elements may control the balance of *CAV1$\alpha$* versus *CAV1$\beta$* expression.

In the loop extrusion model, architectural proteins such as CTCF play a crucial role. CTCF binding sites within *cis*-regulatory elements can anchor long-range chromatin looping interactions (Rao et al., 2014), thus the presence of a CTCF ChIP-seq peak within the distal enhancer may provide insight into how this enhancer contacts the cognate promoter. Additionally, we identified CTCF ChIP-seq peaks at the first two exons of *CAV1* — the first exon of *CAV1$\alpha$* and the second exon of *CAV1$\alpha$* or first exon of *CAV1$\beta$* (Figure 6.20). CTCF binding sites can constrain enhancer-promoter interactions within regulatory hubs (Oudelaar et al., 2019), thus loop extrusion between the distal enhancer and these exonic CTCF binding sites may regulate differential isoform expression by modulating chromatin architecture. CTCF has also been implicated in differential mRNA isoform expression through its roles in alternative splicing and exon inclusion. For example, exonic CTCF binding within the *CD45* gene impedes RNAPII elongation, allowing assembly of the spliceosome at a weak upstream splice site. This facilitates the inclusion of the CTCF-bound exon, thus contributing to the generation of different mRNA isoforms (Shukla et al., 2011).

#### 6.8.2.2 CAV1$\alpha$ phosphorylation

The distinct functional roles of the two isoforms of CAV1, CAV1$\alpha$ and CAV1$\beta$, are largely governed by the distinct capacities of the two isoforms to undergo phosphorylation at the N-terminal end. Only the longer isoform, CAV1$\alpha$ undergoes phosphorylation at tyrosine 14 (Lee et al., 2000, Li et al., 1996).

Phosphorylation of CAV1 regulates the spatial organisation within the Caveolin oligomer, with negatively charged phosphotyrosine residues promoting the separation of neighbouring CAV1 monomers (Zimnicka et al., 2016). Thus, with the apparent bias towards expression of the *CAV1α* mRNA isoform following enhancer knockout, more CAV1 monomers within the Caveolin complex may be able to undergo phosphorylation. This may result in spatial instability within the Caveolin complex, as a result of increased number of phosphotyrosine residues, potentially contributing to the disrupted membrane lipid raft integrity observed in ALS patient-derived lymphoblastoid cells carrying the chr7:116,222,625T>C SNP reported in Cooper-Knock et al. (2021). In support, ALS patient-derived lymphoblastoid cells carrying the chr7:116,222,635T>C mutation within the distal model enhancer also exhibited increased *CAV1α* mRNA expression compared to healthy controls (Figure 6.10), suggesting the ALS-associated SNP in the enhancer may influence differential *CAV1* isoform expression. This may provide a mechanism for the disrupted membrane lipid raft integrity observed in this cell line (Cooper-Knock et al., 2021).

#### 6.8.2.3 Increased *CAV1* expression in ALS patient-derived lymphoblastoid cells

Conversely to what was observed in enhancer knockout SH-SY5Y cells, cells carrying chr7:116,222,635T>C demonstrated increased overall *CAV1* expression, compared to healthy control-derived cells (Figure 6.10), suggesting the ALS-associated SNP in this enhancer may drive elevated *CAV1* expression. Using these analyses alone, we cannot determine whether the observed increase in total *CAV1* expression is solely due to increased *CAV1α* expression. Moreover, RT-qPCR cannot distinguish between overall *CAV1* pre-mRNA and the mature *CAV1β* isoform expression, as at the mRNA level, the differences between the two isoforms is based on transcription from the first exon encoding *CAV1α*.

Despite a comparatively less pronounced change in expression, ALS patient-derived lymphoblastoid cells carrying the chr7:115,994,269C>T SNP also exhibited an increase in total *CAV1* expression, compared to healthy control-derived cells (Figure 6.10). This

may suggest that the upregulation of *CAV1* is a pathological mechanism underlying ALS progression for patients carrying *CAV1* enhancer mutations. Considering the proposed neuroprotective properties of *CAV1* (Cooper-Knock et al., 2021, Egawa et al., 2018, Head et al., 2011, 2008, Sawada et al., 2019, Takayasu et al., 2010), taken with the increased median survival of ALS patients carrying *CAV1* enhancer mutations (Adey et al., 2023), these findings posit a mechanism where enhancer mutations drive *CAV1* overexpression, potentially providing enhanced neuroprotection and improved survival outcomes.

When comparing *CAV1* expression in SH-SY5Y cells and lymphoblastoid cells, it is important to note that these distinct cell types likely exhibit different levels of basal *CAV1* transcription. To better understand the neuronal impact of ALS-associated SNPs, generating SH-SY5Y cell lines carrying these mutations would provide more relevant insights into their impact.

### 6.8.3 eRNA transcription from the proximal enhancer provides insight into enhancer cooperativity

Using TT-seq and published CAGE-seq data (ENCODE Project Consortium, 2012, Hitz et al., 2023, Luo et al., 2020), we identified sites of *de novo* eRNA transcription within the model proximal enhancer (Figure 6.11, Figure 6.12). We quantified expression over these site of *de novo* transcription by RT-qPCR. Upon knockout of the distal enhancer, we observed a decrease in sense eRNA expression, compared to expression in WT SH-SY5Y cells (Figure 6.13b). This suggests potential synergism between the proximal and distal enhancers, whereby knocking out the distal enhancer disrupts this cooperativity, reducing the transcriptional output, and hence the eRNA production, from the proximal enhancer. However, we did not observe any significant change in the level of antisense eRNA expression upon distal enhancer knockout, relative to WT SH-SY expression. This could suggest that the transcription we observed when targeting sense eRNA expression may not be eRNA transcription, but in fact transcriptional read-through from the *CAV1* gene. Therefore, with the reduction in *CAV1* gene expression following distal enhancer knockout, we would observe less transcriptional read-through. Alternatively, the variability in eRNA detection, especially for the antisense eRNAs, could be due to their generalised transience and instability, whereby consistently detecting their expression from total RNA may be challenging, as indicated by the high standard error for antisense eRNA expression in WT SH-SY5Y cells.

Interestingly, for expression of both the sense and antisense eRNA, we observed a significant decrease in expression following *CAV1* exon 2 knockout, compared to expression

in WT SH-SY5Y cells (Figure 6.13b). Of importance, the candidate promoter spans chr7:116,160,447-116,171,095 and thus overlaps exon 2, including the locus targeted by the gRNA at chr7:116,166,668-116,166,687 (Fishilevich et al., 2017). This raises the possibility that knockout out of the exon may have also disrupted the activity of the promoter. Transcription at enhancers and promoters is often concurrent (Andersson et al., 2014, Hah et al., 2013, Kim et al., 2010, Michel et al., 2017), with enhancer-promoter contacts largely correlating with transcriptional activation (Carullo et al., 2020, Sanyal et al., 2012), suggesting that the act of transcription at enhancers and promoters is highly coordinated and mutually reinforcing (Du et al., 2024, Panigrahi et al., 2023). In this model of transcriptional interdependence, the activity at both the enhancer and its target promoter relies on their mutual physical availability; direct enhancer-promoter contact; and shared transcriptional output (Panigrahi et al., 2023). Therefore, with the disruption of *CAV1* promoter activity, this may mutually impact activity at the proximal enhancer, resulting in a sharp reduction in eRNA transcription.

To determine whether ALS-associated SNPs influence eRNA transcription from the proximal enhancer, we quantified eRNA expression in control- and patient-derived lymphoblastoid cells (Figure 6.15). We observed a significant increase in the expression of both the sense and antisense eRNA in ALS patient-derived lymphoblastoid cells carrying the distal enhancer mutation chr7:116,222,625T>C, compared to the healthy control. This increase in expression parallels the elevated *CAV1* expression observed in this cell line, suggesting that activation of the proximal enhancer may contribute to the observed increase in *CAV1* mRNA expression. Although we cannot entirely eliminate the possibility that the sense eRNA RT-qPCR primers may be capturing transcriptional read-through of the gene - especially given the increased transcription over the 3' UTR in lymphoblastoid cells compared to SH-SY5Y cells (Figure 6.14) - we also observed a comparable increase in antisense eRNA expression. Since the antisense eRNA RT-qPCR primers target the reverse strand, they should not detect mRNA read-through, confirming an increase in eRNA transcription, at least on the antisense.

### 6.8.4 CBP is a key regulator of enhancer activity at the proximal *CAV1* enhancer

The CBP-catalysed deposition of H3K27ac at enhancers is a well-established hallmark of enhancer activity (Creyghton et al., 2010, Heintzman et al., 2007, Rada-Iglesias et al., 2011). With the significant enrichment of H3K27ac CUT&RUN signal over the model proximal enhancer (Figure 6.16), we hypothesised that activity of this enhancer may be dependent on the histone acetyltransferase activity of CBP. CBP catalytic inhibition in

SH-SY5Y cells via A-485 treatment resulted in a significant decrease in *CAV1* mRNA expression, relative to untreated cells (Figure 6.17a). This may suggest that the regulatory impact of the proximal enhancer on *CAV1* expression is, at least partially, dependent on the histone acetyltransferase activity of CBP. Beyond the acetylation-mediated opening of chromatin, CBP-catalysed acetylation promotes BRD4 recruitment, driving the activation of P-TEFb to release RNAPII from promoter-proximal pausing, thus facilitating mRNA transcription (reviewed in Core and Adelman, 2019).

CBP activity is a crucial regulator of RNAPII recruitment, initiation and pause release at enhancers, bestowing CBP with utmost control over dynamic enhancer activation (Narita et al., 2021). With the interconnectivity between CBP-regulated histone acetylation and enhancer transcriptional activity, it is unsurprising that upon CBP inhibition we observed a reduction in both sense and antisense eRNA expression at the proximal enhancer, compared to expression in untreated cells (Figure 6.17b). The observed reduction in eRNA expression further supports the idea that the decrease in *CAV1* mRNA expression following CBP inhibition is as a result of inhibited enhancer activity of the proximal enhancer.

### 6.8.5 ALS-associated SNPs have differential impacts on proximal enhancer activity

There is a multifarious complexity to understanding the mechanism underlying how *CAV1* enhancer mutations may cause increased survival in ALS patients. Essentially, while these enhancer mutations purportedly result in *CAV1* downregulation (Cooper-Knock et al., 2021), this may not be the case for all mutations in *CAV1* enhancers, as indeed, there have been reports of SNPs in enhancers increasing cognate gene expression (Gao et al., 2018, Korneev et al., 2020, Padhy et al., 2017, Soldner et al., 2016, Xia and Wei, 2018). Hence, mutations in *CAV1* enhancers may therefore promote increased *CAV1* expression, akin to the neuroprotective phenotype observed in ALS patients carrying such mutations (Adey et al., 2023). Additionally, studying the impacts of enhancer mutations on *CAV1* and *CAV2* expression in patient-derived neuronal-like cells (Cooper-Knock et al., 2021) fails to capture the potentially complex global impact of such mutations across the heterogeneous cell and tissue types in the brain and CNS. If *CAV1* enhancer mutations do indeed result in downregulation of *CAV1* expression, we cannot make assumptions on the global impact of *CAV1* downregulation. While increased *CAV1* expression in the neurons is considered to be neuroprotective (Cooper-Knock et al., 2021, Egawa et al., 2018, Head et al., 2011, 2008, Sawada et al., 2019, Takayasu et al., 2010), this does not take into account the impact on other Caveolin species, such as endothelial Caveolins, which may also impact brain health, with *CAV1*

upregulation linked to loss of BBB integrity and neuroinflammation (Errede et al., 2012, Knowland et al., 2014, Lutz et al., 2017, Trevino et al., 2023, Wu et al., 2016, Yang et al., 2020). Thus, reduced *CAV1* expression in carriers of *CAV1* enhancer mutations could potentially mitigate ALS-induced neuroinflammation, a known key contributor to ALS disease progression (reviewed in Liu and Wang, 2017).

To characterise the potential regulatory impacts of ALS-associated SNPs on *CAV1/CAV2* enhancer function, we performed luciferase reporter assays to quantitatively measure the regulatory capacity of the proximal *CAV1/CAV2* and further survey how the presence of ALS-associated SNPs within this enhancer may alter its regulatory activity. The reporter construct containing the WT *CAV1/CAV2* enhancer drove significantly higher luminescence compared to the negative control, which lacked both the SV40 promoter and enhancer (Figure 6.18d), demonstrating the capacity for the enhancer to promote mammalian gene expression in this system. Moreover, comparatively to the reporter containing the WT enhancer sequence, reporters containing enhancer sequences with the ALS-associated chr7:116,200,589T>A and chr7:116,200,705C>T mutations produced significantly less luminescence (Figure 6.18d). Notably, the chr7:116,200,589T>A appeared to abolish enhancer activity in this reporter system, with the level of luminescence produced from this reporter construct not significantly different from that of the negative control. This may suggest that the presence of the chr7:116,200,589T>A mutation in the proximal *CAV1/CAV2* enhancer eliminates the regulatory potential of the enhancer in this system, reducing its activity to a level indistinguishable from a reporter absent of regulatory elements.

The significantly higher luciferase activity observed in cells transfected with the positive control reporter construct (Figure 6.18d), containing the SV40 promoter and enhancer, can be attributed to the inherent strength of these viral regulatory elements (Banerji et al., 1981). The SV40 promoter is a strong constitutive promoter commonly used in mammalian expression systems due to its capacity to drive high levels of gene expression (reviewed in Makrides, 2003). Additionally, the SV40 enhancer is known for its robust transcriptional activation (Banerji et al., 1981), further amplifying gene expression when paired with the promoter. Together, these viral regulatory elements create a highly efficient expression system, resulting in elevated levels of luciferase activity compared to the more moderate effect of the endogenous proximal *CAV1/CAV2* enhancer. As an endogenous enhancer, the *CAV1/CAV2* enhancer likely functions under more precise and context-specific conditions, fine-tuning gene expression in response to specific cellular or environmental signals. Moreover, in their *in vivo* genomic context, enhancers largely function via cooperative interactions with additional regulatory elements (Gschwind et al., 2023). When studied in isolation, such as in reporter assays, enhancers may only exhibit moderate activity, but in their physiological environment,

their full regulatory potential may be altered through synergistic activity with other enhancers. This contextual specificity contrasts with the constitutive activation observed of SV40 elements.

With the SV40 promoter exogenous to mammalian cells, it may have less potential to function in synergy with the *CAV1/CAV2* enhancer than the endogenous *CAV1* promoter (Bergman et al., 2022). The endogenous *CAV1* promoter would reflect the *in vivo* transcriptional context of the enhancer, allowing for a more accurate assessment of enhancer-promoter cooperativity. Critically, the use of the SV40 promoter in our reporter assay, in addition to the transient transfection approach used, deprive the system of the chromatin context and epigentic regulatory components of the transcriptional machinery that would be present *in vivo* (Jeong and Stein, 1994, Rojas-Fernandez et al., 2015). Further, recent reports have also raised questions about the correlation between active enhancers identified through chromatin-based assays and those characterised using reporter assays (Lindhorst and Halfon, 2023). To best model the regulatory impact of the model proximal *CAV1/CAV2* enhancer on *CAV1* expression, the reporter system would need to engage all components that regulate endogenous gene expression, including the endogenous regulatory elements and the native chromatin architecture (Rojas-Fernandez et al., 2015).

### 6.8.6 Predicted impacts of ALS-associated SNPs on TF binding at model enhancers

To investigate whether any of the ALS-associated SNPs in the model proximal and distal *CAV1* enhancers have potential to disrupt putative TF binding sites, and thereby impact enhancer activity through diminished TF binding, we identified high-confidence TF binding motifs within these enhancers using the JASPAR database (Rauluseviciute et al., 2024).

#### 6.8.6.1 Predicted impacts of ALS-associated SNPs on the proximal enhancer

We identified that the chr7:116,199,522T>C ALS-associated SNP within the proximal enhancer has the potential to disrupt the binding motif of the MEF2 family of TFs (Figure 6.19). Diminished binding of MEF2 may have downstream impacts on the chromatin state at this enhancer. Based on the enrichment of H3K27ac CUT&RUN signal over the proximal enhancer and the impact on eRNA production following CBP inhibition(Figure 6.16,Figure 6.17), we hypothesised that CBP is recruited to this enhancer.

MEF2 can recruit epigenetic co-regulators, including CBP, to modulate chromatin structure (He et al., 2011, Youn et al., 2000). Therefore, disrupted MEF2 binding at this enhancer may result in less efficient recruitment of CBP, having downstream impacts on chromatin accessibility, RNAPII recruitment and thus may impair enhancer activity. To test MEF2-mediated effects on proximal enhancer function, we could perform an siRNA-mediated knockdown of the *MEF2* gene family (Back and Manfredi, 2021), followed by immunoprecipitation-based probing of the proximal enhancer to interrogate the downstream impacts on CBP binding and H3K27ac deposition. Furthermore, to investigate the impacts of the chr7:116,199,522T>C ALS-associated SNP on MEF2 binding, we could perform an electrophoretic mobility shift assay (EMSA), visualising capacity for both the WT and mutated enhancer DNA sequence to bind to MEF2 proteins.

#### 6.8.6.2 Predicted impacts of ALS-associated SNPs on the distal enhancer

Further, we identified putative binding sites for composite binding of FOXO1 with members of the ETS family of TFs at the distal enhancer, with potential to be disrupted by the ALS-associated SNP at chr7:116,222,625T>C (Figure 6.21). FOXO1, a member of the forkhead family of TFs, has established pioneer factor function, capable of binding nucleosomal DNA and driving chromatin decompaction (Hatta and Cirillo, 2007). The function of FOXO1 on chromatin can be further fine-tuned by its association with co-factors, epigenetic regulators and other TFs. FOXO1 has dual functions in transcriptional control, implicated in both transcriptional activation in concert with CBP/p300 recruitment, and in transcriptional repression. Further, the function of FOXO1 can be modulated by PTM, with acetylation modifying the DNA binding and transcriptional regulatory potential of FOXO1 (reviewed in Lam et al., 2013). ETV5, ELK3 and FLI1 are all members of the E26 transformation specific (ETS) family of TFs. Cooperative binding of FOXO1 with ETV5, ELK3 and FLI1 is not surprising, given the established TF cooperativity between Forkhead proteins and ETS proteins (Ibarra et al., 2020). This interplay likely shapes the functional outcome of FOXO1-ETS cooperative binding, fine-tuning transcriptional programs and enabling dynamic response to cellular signals.

Considering the pioneer factor activity of FOXO1 in isolation (Hatta and Cirillo, 2007), we would expect that binding of FOXO1 to the distal model enhancer would likely result in activation of the enhancer, and thus an increase in *CAV1* expression. Therefore, disruption of the FOXO1 binding site, and the resultant diminished FOXO1 binding affinity would therefore prevent the opening of chromatin at this enhancer, resulting in reduced *CAV1* expression. However, we observed an increase in *CAV1* mRNA expression in the patient-derived lymphoblastoid cell line carrying the chr7:116,222,625T>C SNP

(Figure 6.10). This highlights the potential modulating role of ETS proteins in FOXO1-ETS cooperativity. Whereby, if the ETS TF partner had a transcriptionally repressive role, upon mutation, we would observe the increased *CAV1* expression, as seen in our RT-qPCR data. With FOXO1 interacting with co-activators, co-repressors and other protein partners to modulate its activity(reviewed in Lam et al., 2013), there are multifarious mechanisms by which the disrupted binding here may influence enhancer activity, and thus *CAV1* gene expression.

Furthermore, the pioneer factor activity of FOXO1, by influencing chromatin accessibility at the enhancer, may affect downstream CTCF-dependent interactions between the distal enhancer and the *CAV1* promoter. The composite FOXO1-ETS binding motifs are located within a SK-N-SH CTCF ChIP-seq peak, less than 100 bp upstream of the CTCF binding motifs. CTCF has been shown to co-bind with pioneer factors at enhancers, whereby lineage-specific pioneer factors open chromatin to facilitate CTCF occupancy, driving the transcriptional regulation that governs cell identity during development and differentiation (Liu et al., 2023).

DNA sequence alone does not fully determine TF binding specificity; DNA shape also plays a crucial role. TFs may exhibit a preference for specific DNA conformations, which are influenced by the DNA backbone and minor groove geometry, further modulating TF-DNA interactions (reviewed in Rohs et al., 2010). We must take DNA shape into consideration when predicting cooperative TF binding affinity, whereby DNA shape has been shown to be of particular importance for cooperative TF pairs between Forkhead and ETS TFs (Ibarra et al., 2020). Thus, if a base change alone has little impact on the TF binding specificity at the sequence level, it may impact the binding specificity due to alterations in the DNA shape specific to the cooperative TF pair. If in reality the impact of the SNP on the binding motif is minimal, with preference for the T nucleotide not very strong at chr7:116,222,625, with a maximal frequency of 0.55, then DNA shape changes may have an impact here (Figure 6.21).

Although the cooperative roles of FOXO1 and ETS family TFs are not extensively described in the literature, the ETS family member, ETS1, has been shown to modulate the regulatory activity of FOXO1 through interactions with FOXO1 and CBP. FOXO1 function is largely influenced by PTMs, with CBP, responsible for depositing acetyl residues on FOXO1 to regulate its activity. ETS1 enhances this acetylation by forming a complex with CBP, thereby inhibiting the binding of FOXO1 to regulatory DNA and modulating its transcriptional potential (Li et al., 2019). If ETV5, ELK3 and FLI1 have a similar capacity to modulate acetylation of FOXO1, we can hypothesise a mechanism by which reduced binding of FOXO1-ETS pairs results in increased enhancer activation. In WT cells, FOXO1-ETS TF pairs would bind at the predicted binding

motif, recruiting CBP and inducing ETS-mediated FOXO1 acetylation, thus preventing transcriptional activation of the enhancer. However, in the presence of the SNP, there may be diminished binding affinity of the FOXO1-ETS pair, reducing the ETS-mediated acetylation of FOXO1 and allowing transcriptional activation at the enhancer.

### 6.8.7 Conclusion

We have functionally characterised two *CAV1* enhancers: one located proximally, within the 3' UTR of the *CAV1* gene, and one located distally to the *CAV1* gene, ∼20kb downstream from the *CAV1* promoter. Using CRISPRi to investigate their regulatory potential, we have confirmed the *CAV1* enhancer identity of these model enhancers, whereby both the proximal and distal enhancers influence *CAV1* expression to an extent comparable to the *CAV1* promoter.

Moreover, the differential expression of *CAV1* mRNA isoforms in distal enhancer knockout SH-SY5Y cells, as well as in ALS patient-derived lymphoblastoid cells harbouring a mutation in the distal enhancer, suggests a potential role for the distal enhancer in TSS choice and differential mRNA isoform expression. This supports a complex regulatory framework whereby enhancers cooperate or compete to control transcriptional dynamics and isoform choice. Notably, knockout of the distal enhancer also resulted in reduced eRNA expression from the proximal enhancer, providing further evidence of enhancer cooperativity. Additionally, the potential impact of enhancer-mediated *CAV1* mRNA choice on CAV1 phosphorylation, which has functional consequences for caveolae stability and membrane lipid raft integrity, further underscores the significance of enhancer-mediated regulation in cellular function.

In the context of ALS-associated SNPs within *CAV1* enhancers, we identified how genetic variants may disrupt TF binding, impair enhancer function, and consequently affect *CAV1* expression. In support, ALS-associated SNPs within the model proximal *CAV1/CAV2* enhancer significantly reduce its regulatory activity in reporter assays, positing a mechanism by which ALS-associated mutations may disrupt enhancer-mediated regulation of *CAV1*, potentially contributing to disease mechanisms. In conjunction, we observe some evidence of enhancer cooperativity, whereby mutations in the distal model enhancer may result in compensatory upregulation of proximal enhancer activity. Quantification of eRNA expression in patient-derived lymphoblastoid cells demonstrated increased eRNA production from the proximal *CAV1* enhancer in ALS patient-derived cells, suggesting an enhancer-mediated mechanism contributing to the observed upregulation of *CAV1*. Notably, the increased expression of *CAV1* in ALS patient-derived cells aligns with the proposed neuroprotective properties of CAV1.

Further, the increased expression of the *CAV1α* isoform in ALS patient-derived cells, paired with the known functional impact of phosphorylation on CAV1 oligomerisation, mirrors the observed impact of ALS-associated SNPs on membrane lipid raft integrity (Cooper-Knock et al., 2021). These findings raise numerous possibilities about how enhancer mutations may contribute to the pathophysiology of ALS through altered gene regulation.

Additionally, the observed alterations in enhancer activity following CBP inhibition, and the predicted impact of ALS-associated SNPs on TF binding further emphasise the complex regulatory mechanisms controlling *CAV1* expression and their potential relevance to ALS pathology. The dependence of proximal enhancer activity on CBP further implicates histone acetylation as a key regulatory mechanism underlying *CAV1* expression, highlighting the importance of considering chromatin modifications when investigating the impact of disease-associated non-coding mutations.

While multiple lines of evidence support our identification of *de novo* eRNA transcription from the model proximal *CAV1* enhancer, we cannot ignore the implications of its genic location when studying this enhancer. We must address the complexity of disentangling the function of the *CAV1* 3' UTR from the model proximal enhancer to better understand their respective roles in *CAV1* gene regulation.

# Chapter 7

# Disentangling *CAV1* enhancer and 3' UTR function

## 7.1 Introduction

### 7.1.1 Alternative polyadenylation yields differential 3' UTR isoforms

More than half of all human genes utilise APA to yield mRNA transcripts that differ in 3' UTR length, modulating the post-transcriptional fate the resultant mRNA in a manner dependent on 3' UTR isoform expression. While tissue-specific genes are largely expressed as a single 3' UTR isoform, ubiquitously expressed genes express multiple alternative 3' UTR isoforms in a cell type-specific manner, altering the ratios of distinct 3' UTR isoforms to achieve tissue specificity, in a manner largely independent from changes in mRNA abundance (Lianoglou et al., 2013). Protein expression is typically regulated through transcriptional changes that affect mRNA abundance in genes with a single 3' UTR isoform, and through APA site selection, followed by post-transcriptional regulation, in constitutively transcribed genes with multiple 3' UTR isoforms. Thus, the presence of multiple permissive PASs within the 3' UTR can increase the breadth of mRNA expression, differentially including or excluding the host of regulatory elements within the 3' UTR to either promote or escape post-transcriptional control (Lianoglou et al., 2013).

Considering the cell type-specificity of 3' UTR isoform usage, it is unsurprising that enhancers can regulate differential 3' UTR isoform expression. The endogenous enhancer associated with the *PTEN* gene, which expresses multiple 3' UTR isoforms, exerts little control over total mRNA production but plays a key role in 3' UTR isoform switching by increasing the 3' end processing activity at the proximal, intrinsically weak PAS. Thus,

cell type-specific enhancers can integrate cellular signals to modulate 3' UTR isoform expression through differential PAS usage (Kwon et al., 2022).

Mature mRNA production can therefore be determined by the regulatory activity of enhancers at two levels; firstly by the extent of transcript production, and via transcript processing. For genes expressing a single 3' UTR isoform, enhancers function solely to increase transcript production, whereas for genes expressing multiple 3' UTR isoforms, enhancers have the potential to impact both transcript production and 3' end processing (Kwon et al., 2022).

### 7.1.2 The impact of genetic variants in 3' UTRs

We cannot ignore the potential pathological relevance of ALS-associated SNPs in the *CAV1* 3' UTR, independent of the potential impact they may have on the enhancer located within the 3' UTR. Genetic variants have the capacity to modulate APA regulation, with SNPs identified to alter the relative expression of alternative 3' UTR isoforms for thousands of human genes (Mariella et al., 2019). Downstream, APA can cause changes in gene expression due to the isoform-dependent presence of regulatory elements in the 3' UTR that have consequences for mRNA stability (Tian and Manley, 2017, reviewed in). In these cases, alternative polyadenylation trait loci (apaQTL), may also be considered eQTLs (Mariella et al., 2019). However, APA may also have effects that have little to no impact on gene expression levels, for example through changes in mRNA translation efficiency and localisation (An et al., 2008, Floor and Doudna, 2016, Spies et al., 2009). Thus, apaQTLs that impact these downstream translation and localisation pathways, while having negligible impacts on gene expression itself, may have substantial phenotypic effects.

A large proportion of the identified apaQTLs are located within 3' UTRs, as to be expected, given 3' UTRs contain several elements implicated in APA and mRNA stability. SNPs located within the 3' UTR can cause the creation or destruction of functional PAS motifs and binding motifs for miRNAs and RBPs. Further, SNPs that bias the production of a particular 3' UTR isoform may result in the inclusion or exclusion of functional miRNA binding sites, impacting mRNA stability (Mariella et al., 2019). Simplistically, if an mRNA has two 3' UTR isoforms, one short and one long, miRNA binding sites in the short isoform would be common to both isoforms, whereas miRNA binding sites 3' of the proximal PAS would only be included in the long isoform.

Interestingly, apaQTLs are preferentially enriched within regions of active chromatin, including promoters and enhancers, even when discounting genetic variants in potential linkage disequilibrium with variants located in 3' UTRs (Mariella et al., 2019). While

historically it has been believed that APA regulation is governed by *cis*-elements located within mRNA transcripts, up- and downstream of PASs, there have been some examples of DNA regulatory elements contributing to APA regulation (Kwon et al., 2022, Oktaba et al., 2015). Hence, the identification of apaQTLs within promoters and enhancers further corroborates the involvement of DNA regulatory elements in APA regulation (Mariella et al., 2019), coupling transcription and polyadenylation.

Given the prevalence of apaQTLs in both 3' UTRs and enhancers (Mariella et al., 2019), it would not be surprising if ALS-associated SNPs within the *CAV1* 3' UTR resulted in alterations in *CAV1* 3' UTR isoform usage.

### 7.1.3 Chapter aims

In this chapter we consider the implications of ALS-associated SNPs within the *CAV1* 3' UTR, independent of enhancer function. We further discuss how eRNA transcription at the proximal *CAV1/CAV2* enhancer may be interpreted as isolated 3' UTR expression, with potential influences over protein-coding expression, in a cell-type specific manner. In order to disentangle the 3' UTR expression from the eRNA expression, we provide a CRISPR/Cas9-based genetic editing approach to transcriptionally isolate *CAV1* gene expression from expression of the 3' UTR-associated enhancer.

## 7.2 Characteristics of the CAV1 3' UTR

### 7.2.1 Characterising *CAV1* 3' UTR isoform expression

Diversity in 3' UTR length, arising from APA, is largely cell type-specific, specifically with cells of the CNS predominantly expressing long 3' UTR isoforms as a result of preferential distal PAS usage (Guvenek and Tian, 2018, Miura et al., 2013, Zhang et al., 2005). We therefore investigated *CAV1* transcript lengths in iPSC-derived motor neurons (Zhang et al., 2022). We identified variation in *CAV1* transcript length across samples in a manner that positively correlated with patient age (Pearson correlation coefficient, $r = 0.84$), regardless of disease state. (Figure 7.1).

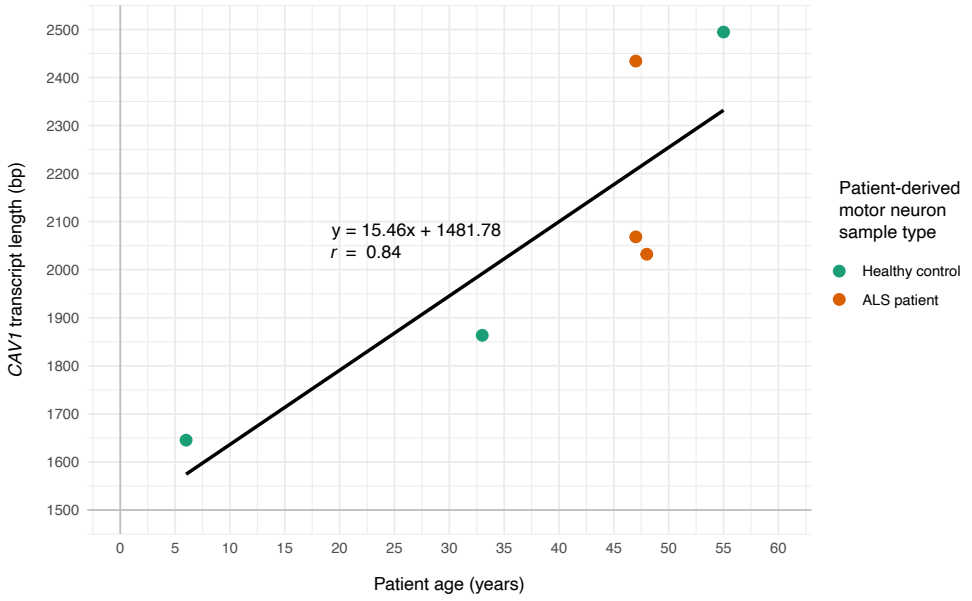



FIGURE 7.1: *CAV1* transcript length increases with age
*CAV1* mRNA transcript length, as determined by RSEM transcription quantification (Zhang et al., 2022) versus patient age, regardless of disease state. Correlation determined by Pearson correlation coefficient.

There were minimal trends between levels of *CAV1* expression and 3' UTR length when considering all patient samples, largely due to similarities in expression levels between younger healthy control individuals, namely samples SMA and CS14, and expression in ALS patients. However, when considering ALS patients independently from controls, *CAV1* transcript length was negatively correlated with *CAV1* expression ($r$ = -0.90) (Figure 7.2). This greatly contrasts what was observed for the oldest control sample (PGP), which expressed the longest observed *CAV1* transcript (2495bp) with *CAV1* expression levels significantly higher than all other other samples, control and ALS patient alike. Interestingly, the ALS patient expressing the longest *CAV1* transcripts, ALS29, with a length comparable to the transcript length for PGP, with lengths of 2434bp and 2495bp, respectively, demonstrated the lowest expression of *CAV1*. Despite the similarities in *CAV1* transcript length, likely produced through use of the same PAS, PGP demonstrated a more than five-fold increase in *CAV1* expression compared to ALS29.

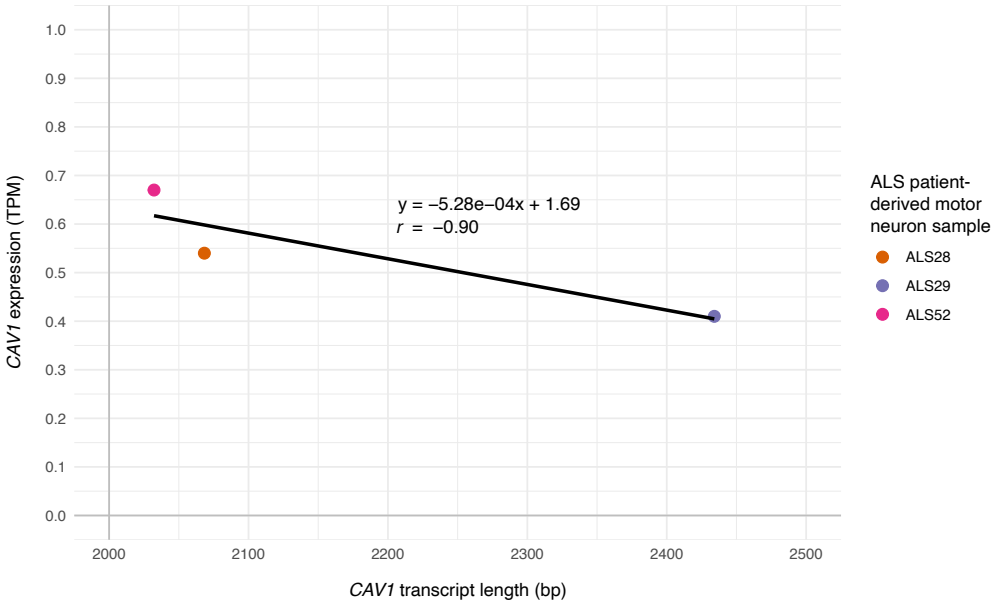



FIGURE 7.2: *CAV1* expression decreases with *CAV1* transcript length in ALS patients *CAV1* mRNA transcript length versus *CAV1* expression, both determined by RSEM transcript quantification (Zhang et al., 2022), in ALS patient-derived motor neurons. Correlation determined by Pearson correlation coefficient.

Importantly, distinguishing between the *CAV1α* and *CAV1β* mRNA isoforms based solely on transcript length is challenging. *CAV1β* transcripts often surpass *CAV1α* transcripts in length, primarily due to the extended 5' UTRs present in *CAV1β* isoforms (Harrison et al., 2024).

### 7.2.2 The impact of ALS-associated SNPs on post-transcriptional regulatory elements in the *CAV1* 3' UTR

To investigate whether ALS-associated SNPs may impact 3' UTR function, we identified AU-rich elements (AREs) that overlap SNPs within the *CAV1* 3' UTR, using AREsite2 (Fallmann et al., 2016). AREs are fundamental to the regulatory function of 3' UTRs. When bound by *trans*-acting factors, they can influence mRNA stability — either negatively or, in some instances, positively — thereby affecting translation efficiency (reviewed in Mayr, 2017). Of the 109 AREs identified in the *CAV1* 3' UTR, only one AU-rich motif overlaps with an ALS-associated SNP. We identified a WWTTTWW motif at chr7:116,199,522-116,199,528, of the sequence TATTTTT. Upon introduction of chr7:116,199,522T>C, this sequence motif would be altered to CATTTTT, effectively abolishing the motif (Figure 7.3). Therefore, the presence of ALS-associated SNPs within the *CAV1* 3' UTR may provide a mechanism by which disease-associated mutations can impact post-transcriptional regulation of *CAV1*.

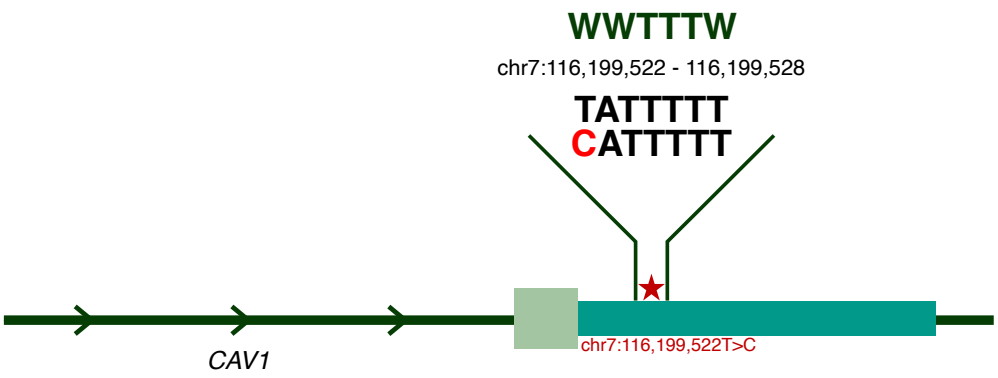



FIGURE 7.3: AU-rich element in the *CAV1* 3' UTR is disrupted by ALS-associated SNP
Schematic representing the ARE site in the *CAV1* 3' UTR. The WWTTTWW motif at chr7:116,199,522-116,199,528 is disrupted by the chr7:116,199,522T>C.

## 7.3 Isolated 3' UTR expression

Due to the widespread imbalance in the expression of 3' UTRs and their corresponding protein-coding sequences in neurons (Kocabas et al., 2015), along with the high 3' UTR expression and relatively low *CAV1* mRNA expression in undifferentiated SH-SY5Y cells, we investigated the 3' UTR to CDS expression ratio of *CAV1*.

For simplicity, we defined the *CAV1* CDS as the full genomic coding region—that is, the region encompassing both coding exons and introns—spanning chr7:116,165,091–116,199,607. The *CAV1* 3' UTR was defined as the region spanning chr7:116,199,521–116,201,233. We acknowledge that defining such a broad region as the CDS could capture intronic transcripts, particularly in nascent RNA-seq data such as TT-seq. However, we adopted this approach in alignment with the GNF Gene Expression Atlas 2 (Su et al., 2004), which utilises distinct microarray probes to separately measure *CAV1* expression from the CDS and 3' UTR region.

### 7.3.1 Isolated *CAV1* 3' UTR expression in model cell lines

Firstly, we determined the 3' UTR to CDS expression ratio across our SH-SY5Y and lymphoblastoid cell lines using our TT-seq data (Figure 7.4). Using nascent RNA-seq data provides an insight into the dynamic nature of isolated 3' UTR expression (Ji et al., 2021, Kocabas et al., 2015, Mercer et al., 2011). Across all cell lines, we observed elevated mean 3' UTR expression compared to cognate CDS expression. While the presence of ALS-associated SNPs in the ALS patient-derived lymphoblastoid cell lines results in

increased *CAV1* gene expression, they appeared to have negligible impacts on *CAV1* 3' UTR to CDS expression ratios. In the context of these SNPs, the mechanism driving increased *CAV1* expression does not appear to be linked to isolated 3' UTR expression.

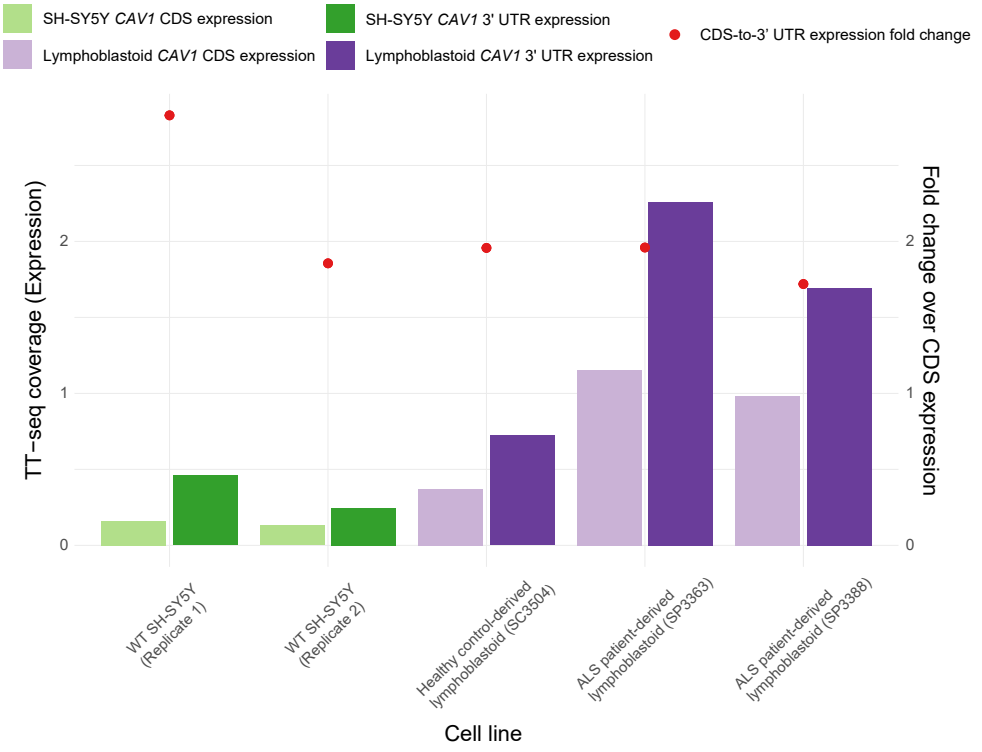



FIGURE 7.4: Isolated *CAV1* 3' UTR expression in SH-SY5Y and lymphoblastoid cell lines

Mean expression over the *CAV1* CDS and 3' UTR in SH-SY5Y and lymphoblastoid cell lines, as determined by TT-seq read coverage. Fold change over CDS expression is denoted by red points.

### 7.3.2 Isolated *CAV1* 3' UTR expression in patient derived motor neurons

To further investigate whether isolated 3' UTR expression has influence over *CAV1* expression, particularly in the context of disease states, we determined the 3' UTR to CDS expression ratio in iPSC-derived motor neurons of human origin (Figure 7.5). We observed considerably higher 3' UTR to CDS expression for the healthy control sample, PGP, compared to the other samples. As earlier established, this is the sample that also exhibited considerably more overall *CAV1* gene expression, compared to the other samples. We cannot make assumptions on the resultant protein expression levels based on mRNA expression alone, so we cannot exclude the possibility that high isolated 3' UTR expression in the PGP sample functions to regulate CAV1 protein expression in *cis* (Kocabas et al., 2015, Mercer et al., 2011).

It is important to note that 3'UTR to CDS expression ratio here is determined by total RNA-seq, therefore these expression ratios represent total steady-state RNA levels, as opposed to TT-seq that would capture nascent RNA species, including transient RNAs, such as eRNAs. Moreover, even at steady-state levels, we observed more expression over the *CAV1* 3' UTR than the CDS. This may suggest that somewhat stable RNA species are conferring high expression of the *CAV1* 3' UTR in these samples.

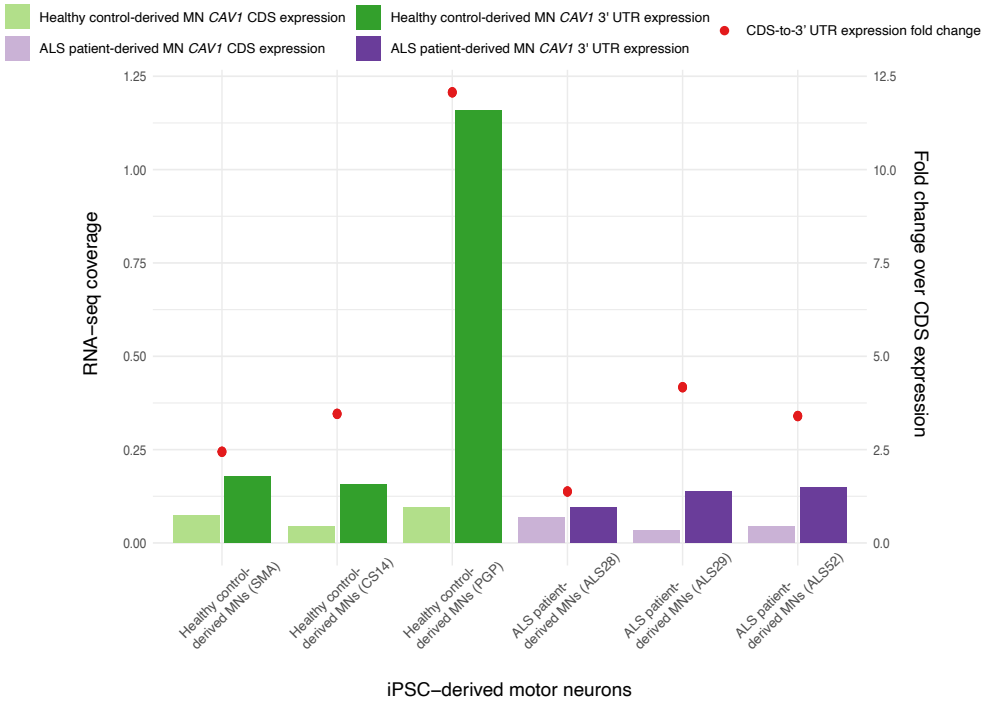



FIGURE 7.5: Isolated *CAV1* 3' UTR expression in iPSC-derived motor neurons
Mean expression over the *CAV1* CDS and 3' UTR in iPSC-derived *in vitro* differentiated motor neurons, as determined by total RNA-seq read coverage (Zhang et al., 2022). Fold change over CDS expression is denoted by red points.

### 7.3.3 Tissue-specific isolated *CAV1* 3' UTR expression

As we have established a high 3' UTR to CDS expression ratio in both model cell lines and iPSC-derived motor neurons, we next investigated isolated 3' UTR in whole tissues. With the cell and tissue-type specificity of isolated 3' UTR expression (Ji et al., 2021, Kocabas et al., 2015, Mercer et al., 2011) and the common high 3' UTR to CDS ratio in neurons (Kocabas et al., 2015), we aimed to identify if such expression ratios exist for *CAV1* in ALS-associated tissues. Given the low expression of CAV1 protein in the brain (Head et al., 2011, 2010), and the potential influence of the isolated 3' UTR to control protein expression (Kocabas et al., 2015, Mercer et al., 2011), we further compared expression ratios in the brain to those of lung tissue and cells, that demonstrate high CAV1 protein expression (reviewed in Maniatis et al., 2012).

Utilising microarray data from the GNF Gene Expression Atlas 2 (Su et al., 2004), we investigated the differential expression of the *CAV1* 3' UTR and CDS in ALS-associated CNS tissues (Figure 7.6a), and in lung-associated tissues and cells (Figure 7.6b). For all CNS-associated tissues, with the exception of the cerebellum, we observed more expression over the *CAV1* 3' UTR than the cognate CDS. The spinal cord exhibited the highest 3' UTR to CDS expression ratio, and of note, demonstrated relative underexpression across the CDS and relative overexpression of the 3' UTR, compared to the 79 human tissues surveyed in the GNF Gene Expression Atlas 2. Conversely, in lung-associated tissues and cells, with the exception of foetal lung, we observed roughly equal *CAV1* 3' UTR to CDS expression. The distinct trend observed for foetal lung tissue is not necessarily surprising, given the established dynamic differential 3' UTR to CDS expression during development (Ji et al., 2021, Kocabas et al., 2015, Mercer et al., 2011). With the inverse correlation of high 3' UTR to CDS expression ratio and protein expression (Kocabas et al., 2015), the trends observed here may corroborate similar mechanisms for CAV1, with a high 3' UTR to CDS expression ratio in the majority of CNS-related tissues compared to the equal 3' UTR and CDS expression in the majority of lung-related tissues and cells.

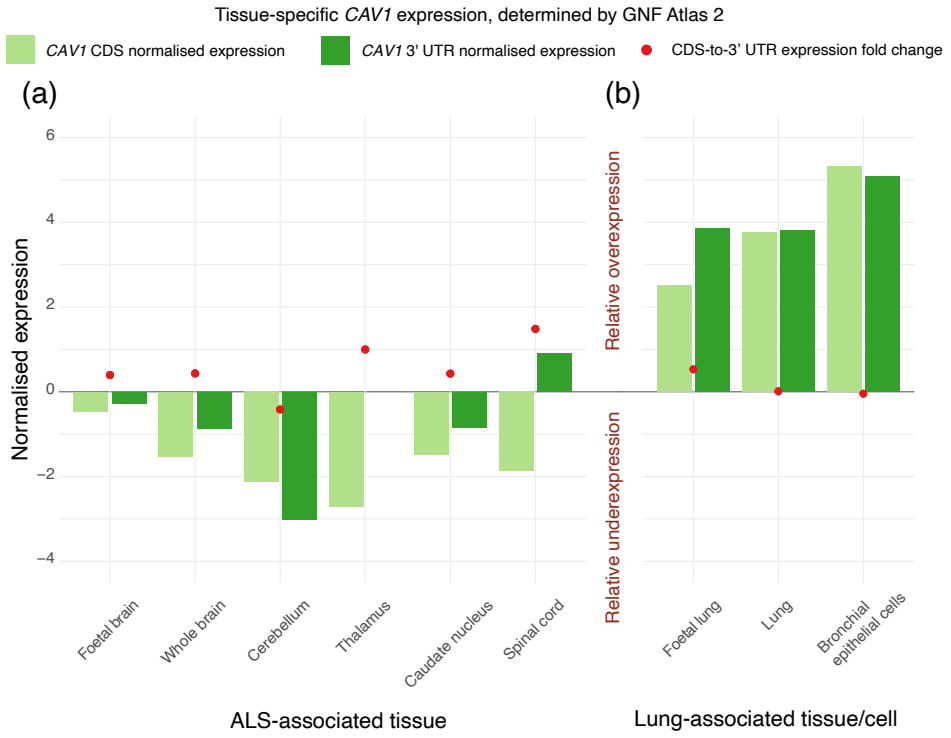



FIGURE 7.6: Isolated *CAV1* 3' UTR expression in tissue
Mean expression over the *CAV1* ALS-associated brain tissues and lung tissues and cells, as determined microarray expression as part of the GNF Gene Expression Atlas 2 (Su et al., 2004). Expression is normalised based on the average level of CAV1 expression across 79 human tissues. Fold change over CDS expression is denoted by red points.

However, we cannot ignore the potential influence that the CNS-specific *CAV1/CAV2* enhancer within the *CAV1* 3' UTR may have over isolated 3' UTR expression in CNS-relevant cells and tissues. Cell and tissue type-specific eRNA expression across the *CAV1* 3' UTR may indeed contribute to the observed high 3' UTR to CDS expression ratio in both neuronally-relevant model cell lines and in CNS tissues.

## 7.4 UTR-associated enhancers

Enhancers within exonic regions, including UTRs, are largely understudied, likely due to the difficulty in experimentally dissecting their function without also interrupting genic function. Due to the presence of the ALS-associated enhancer within the *CAV1*, we were interested to identify the prevalence of UTR-localised enhancers throughout the genome. Using the ABC enhancer catalogue (Fulco et al., 2019), we investigated the location of enhancers relative to genic features, including coding exons, introns, intergenic regions and UTRs (Figure 7.7). As expected, the majority of ABC candidate enhancers localised to intronic regions, followed by intergenic regions. Interestingly, more enhancers were located within coding exons than in UTRs, with UTR-localised enhancers making up the smallest proportion of the ABC enhancer catalogue. We further segregated enhancers within UTRs into those that localise to 5' UTRs and 3' UTRs, yielding approximately equal proportions (6763 enhancer within 5' UTRs and 6650 enhancers within 3' UTRs). Therefore, while 3' UTRs are the genomic feature least enriched for candidate enhancers, they still constitute a notable number of predicted enhancers genome-wide. This raises questions about the involvement of enhancers, and thus eRNA transcription, in differential 3' UTR and CDS expression.

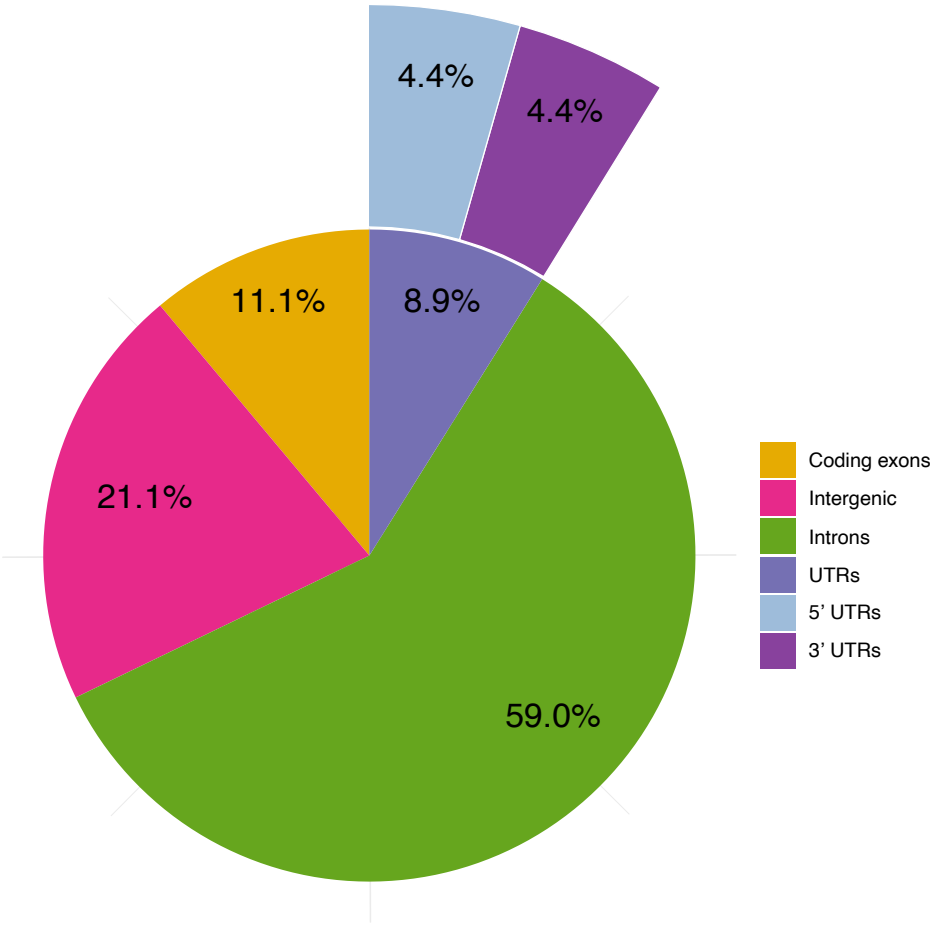



FIGURE 7.7: Genomic location of ABC candidate enhancers
The genomic location of ABC candidate enhancers (Fulco et al., 2019), relative to genic features, including coding exons, introns, intergenic regions and UTRs.

## 7.5 Disentangling the regulatory functions of the candidate *CAV1* enhancer from the *CAV1* 3' UTR

The *CAV1* 3' UTR demonstrates enhancer characteristics, as per the enrichment of enhancer-associated histone modifications (Chapter 4) and *de novo* transcription initiation within this region (Chapter 5), in addition to the capacity of its sequence to drive gene expression in reporter assays (Chapter 6). This enhancer sequence appears to have regulatory control over *CAV1* expression, as observed in CRISPRi assays (Figure 6.8), but we cannot be sure that the regulatory activity we observe is due to the enhancer activity and not the 3' UTR independently without isolating the enhancer sequence from the functions of the 3' UTR.

In order to remove the influence of the *CAV1* 3' UTR when investigating candidate *CAV1* enhancer activity, we have designed an experimental approach for the generation of SH-SY5Y cell lines that isolate the *CAV1* enhancer within the 3' UTR from the genic sequence. We aimed to generate such cell lines by inserting the SV40 PAS sequence at the end of the *CAV1* coding sequence (CDS), 5' to the 3' UTR sequence, using CRISPR/Cas9. This would ensure that transcription of *CAV1* mRNA terminates at the end of the CDS, preventing readthrough transcription from the *CAV1* CDS into the 3' UTR. By removal of the 3' UTR sequence from *CAV1* pre-mRNAs, we would be able to identify which of the transcriptional events seen in the 3' UTR are eRNAs and thus, would be able to identify the nature of the regulatory activity of the enhancer within the *CAV1* 3' UTR without the influence of 3' UTR regulatory activity.

### 7.5.1 sgRNA validation

We designed sgRNAs targeting proximal to the end of the *CAV1* CDS, with the target sites of all designed gRNAs located <200bp away from the site to be targeted for SV40 PAS sequence insertion. Each of the gRNAs, plus a scrambled negative control gRNA were cloned into pSpCas9(BB)-2A-Puro (PX459) V2.0, and validated in HEK293FT cells.

We utilised the T7 endonuclease I (T7E1) assay to evaluate the efficiency of Cas9 cleavage using different sgRNAs. The T7E1 assay allows for the detection of indels that arise from DNA double strand break repair via non-homologous end joining (NHEJ), that often occur following Cas9-mediated DNA cleavage. Following transfection with sgRNA-targeted Cas9, the targeted locus is amplified by PCR. The resultant amplicons are then reannealed and if an aberrant NHEJ event has occurred following Cas9 cleavage, the mutations will yield mismatched DNA heteroduplexes (Guschin et al., 2010). T7E1 is a structure-selective nuclease, and thus selectively cleaves upon detection of structural deformities in the heteroduplexed DNA (Mashal et al., 1995). Visualisation of the resulting product therefore allows for detection of efficient Cas9 cleavage (Guschin et al., 2010).

The T7E1 assay was initially performed in HEK293FT cells due to ease of liposome-based transfection (Figure 7.8). This assay suggested successful Cas9-mediated cleavage in HEK293FT cells transfected with pX459_gRNA4, yielding bands of approximately 160bp and 404bp, as expected, following T7E1 cleavage. However, we cannot determine with confidence whether the Cas9 cleaved the DNA efficiently in HEK293FT cells transfected with pX459_gRNA1 and pX459_gRNA3. Firstly, the banding pattern between between the untransfected gDNA versus the transfected gDNA shows many similarities

and secondly, the suboptimal amplicon lengths (284bp and 230bp for pX459_gRNA1; 254bp and 325bp for pX459_gRNA3) caused difficulties in distinguishing a banding pattern indicative of T7E1 cleavage.

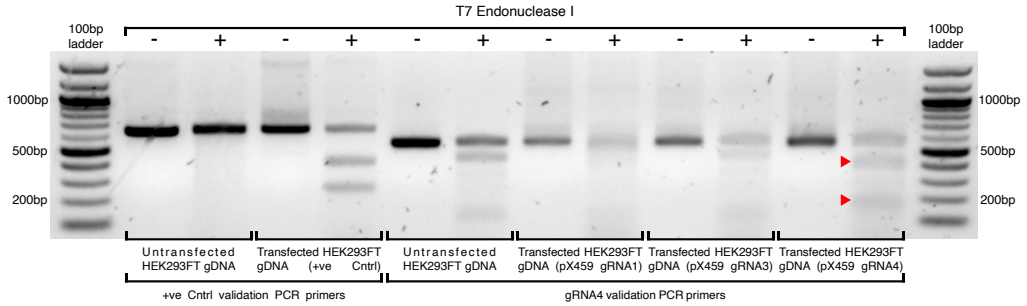



FIGURE 7.8: T7E1 assay for gRNA validation in HEK293FT cells
Validation of sgRNAs in HEK293FT cells for incorporation of SV40 PAS sequence downstream of *CAV1* 3' UTR. Red arrows indicate cleavage of DNA heteroduplexes by T7E1 following cleavage by Cas9 in cells transfected with pX459_gRNA4, yielding bands of ∼160bp and ∼404bp. n=2

#### 7.5.1.1 sgRNA validation in SH-SY5Y cells

As we aimed to generate CRISPR/Cas9-edited cells derived from the SH-SY5Y cell line, we repeated the T7E1 assay in SH-SY5Y cells, to ensure efficient Cas9-mediated cleavage in our model cell line (Figure 7.9a), again comparing to T7E1 assay using control sgRNA constructs (Figure 7.9b). We repeated the experiment as for HEK293FT cells but the sgRNA-containing pX459 plasmids were transfected using electroporation and the validation primers were optimised to ensure distinct bands following T7E1 cleavage (190bp and 325bp for pX459_gRNA1; 182bp and 398bp for pX459_gRNA3; 146bp and 399bp for pX459_gRNA4) (Figure 7.9c). Distinctly from the T7E1 assay in HEK293FT cells, this assay suggested successful Cas9-mediated cleavage in SH-SY5Y cells transfected with pX459_gRNA1, yielding bands of approximately 190bp and 325bp following T7E1 cleavage (Figure 7.9a,b).

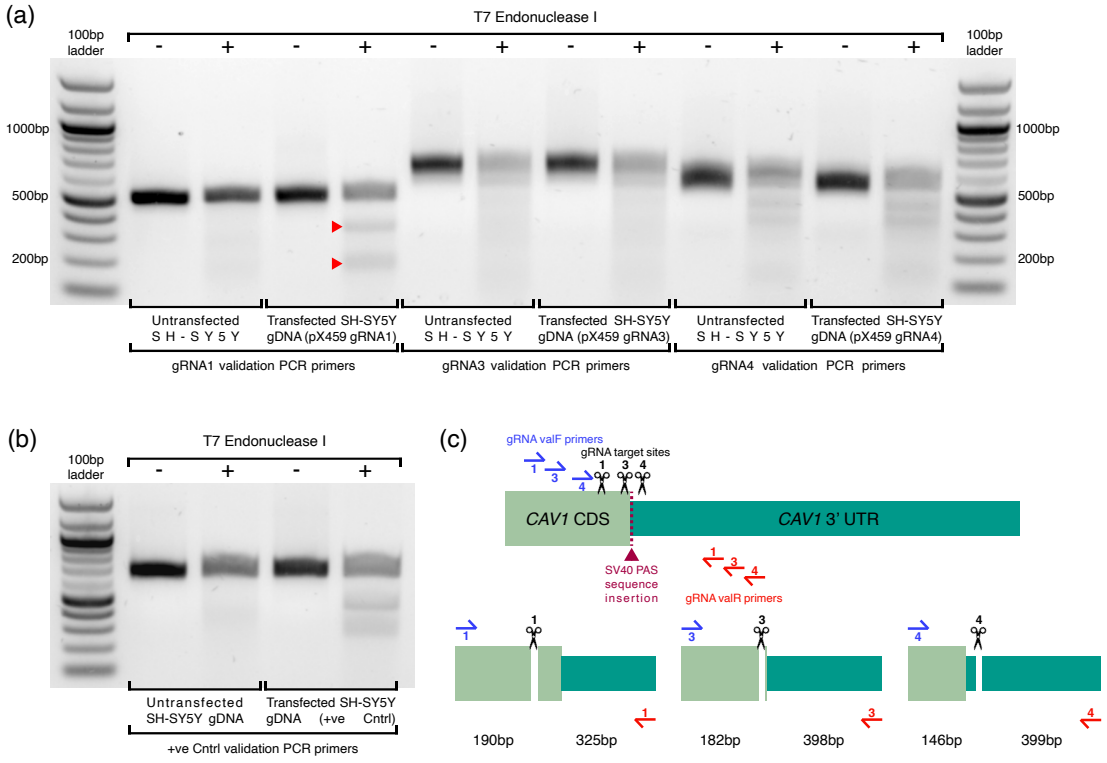



FIGURE 7.9: T7E1 assay for gRNA validation in SH-SY5Y cells
(a) Validation of sgRNAs in SH-SY5Y cells for incorporation of SV40 PAS sequence downstream of *CAV1* 3' UTR. Red arrows indicate cleavage of DNA heteroduplexes by T7E1 following cleavage by Cas9 in cells transfected with pX459_gRNA1, yielding bands of ~190bp and ~325bp. n=1. (b) Validation of positive control sgRNAs in SH-SY5Y cells. n=1. (c) Schematic demonstrating the expected band sizes following cleavage of DNA heteroduplexes by T7E1 for each sgRNA.

### 7.5.2 Generation of homology-directed repair template for insertion of SV40 PAS

In order to insert the SV40 PAS sequence downstream of the *CAV1* CDS, we generated a donor vector containing the SV40 PAS sequence flanked by homology arms homologous to *CAV1* up- and downstream of the insertion site. To do so, we modified the CAV1_enhancer_WT G-block vector (CAV1_enhancer_WT, Appendix A.1). The backbone of this vector is appropriate for this application as it already contained the *CAV1* sequence around the insertion site. We utilised a G-block of the SV40 PAS sequence flanked by homology to the *CAV1* sequence containing unique restriction sites, *Eco*RI and *Mlu*I (SV40_gBlock_RD, Appendix A.2). The SV40 PAS sequence was ligated into the CAV1_enhancer_WT plasmid, between the *CAV1* final exon CDS and the 3' UTR, following sequential single digest of both the CAV1_enhancer_WT and SV40_gBlock_RD G-blocks using *Eco*RI and *Mlu*I, generating the homology-directed repair (HDR) construct (CAV1_SV40_gRNA4, Appendix A.3).

To prevent Cas9 from targeting repaired DNA, additional silent blocking mutations were incorporated into the resultant HDR template via mutation of the PAM sequence. SV40_gBlock_RD was designed to include a PAM mutation for pX459_gRNA4 (containing an ATG, as opposed to the WT AGG). However, considering the successful T7E1 assay for pX459_gRNA1 in SH-SY5Y cells, we also wanted to generate an appropriate HDR template for this sgRNA. We performed sequential site-directed mutagenesis, firstly to convert the gRNA4 PAM sequence back to WT (converting ATG to AGG) and then mutated the gRNA1 PAM sequence (converting GGG to GGA). Following site-directed mutagenesis, we performed whole plasmid sequencing of the resultant plasmids to confirm correct PAM site mutations and identify any abnormalities (Figure 7.10). As expected, CAV1_SV40PAS_WT contained no PAM site mutations, while CAV1_SV40PAS_gRNA1 and CAV1_SV40PAS_gRNA4 contained the appropriate PAM site mutations. The sequence of CAV1_SV40PAS_gRNA4 was free of abnormalities. However, an intermediate fragment of the homology arm downstream of the SV40 PAS insertion site was missing from the plasmids CAV1_SV40PAS_WT and CAV1_SV40PAS_gRNA1, resulting in discontinuity in homology to the *CAV1* 3' UTR in these plasmids. Given the observed Cas9-mediated cleavage in SH-SY5Y cells transfected with pX459_gRNA1, we must repeat the sequential site-directed mutagenesis experiments to generate the appropriate HDR template for this sgRNA, enabling the insertion of the SV40 PAS into the *CAV1* locus in SH-SY5Y cells.

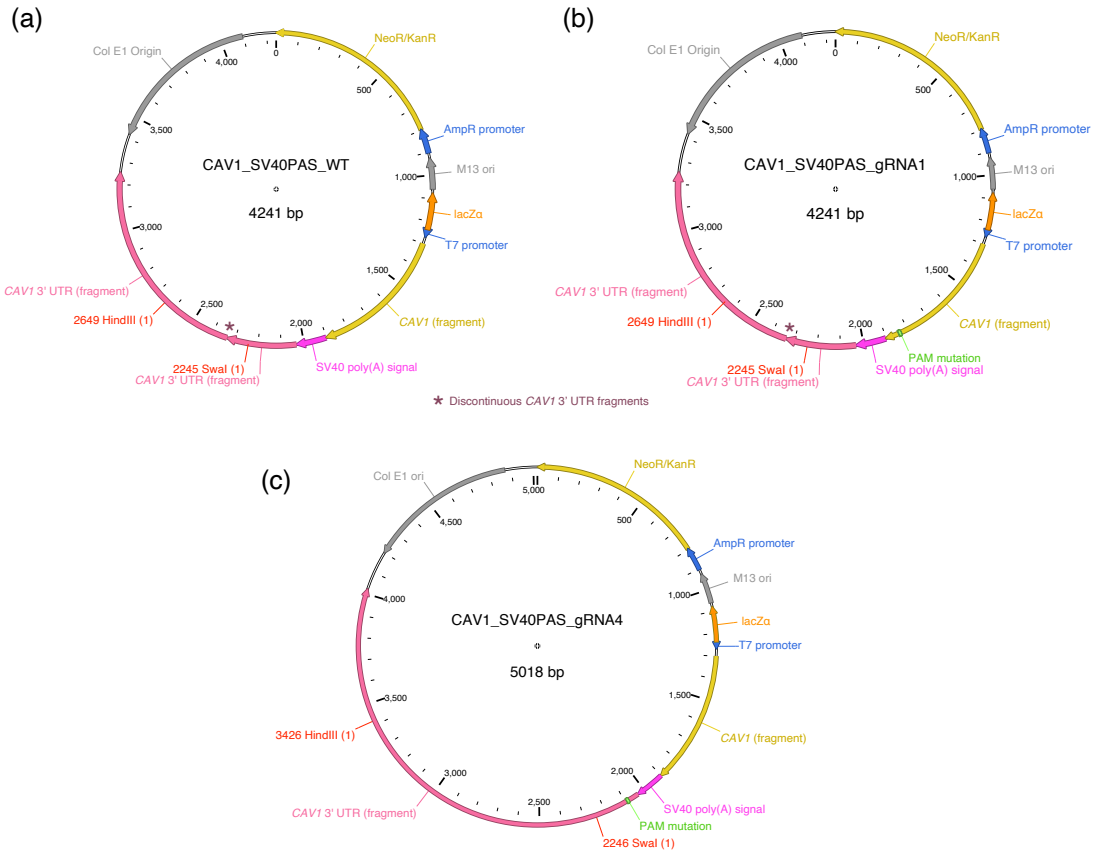



FIGURE 7.10: Plasmid maps of HDR templates following whole plasmid sequencing Plasmid maps generated following whole plasmid sequencing of HDR templates containing WT *CAV1* sequence (a), PAM mutation for pX459_gRNA1 (b) and pX459_gRNA4 (c). Regions of discontinuous homology to *CAV1* are denoted by purple asterisk

### 7.5.3 Test transfection of SV40 polyadenylation signal into *CAV1* in HEK293FT cells

To demonstrate that the SV40 PAS HDR template can successfully be incorporated into the genome, we performed a liposome-based test transfection in HEK293FT cells. Based on the T7E1 assay in HEK293FT cells, suggesting successful Cas9-mediated cleavage using pX459_gRNA4, we transfected cells with this plasmid and the corresponding HDR template containing the gRNA4 PAM site mutation (CAV1_SV40PAS_gRNA4). To optimise ratios of Cas9/sgRNA to the HDR template, we tested two different ratios of pX459_gRNA4 to SV40 PAS HDR template, namely 1:2 and 1:5 pX459:HDR ratios. 48 hours following transfection, we performed a gDNA extraction and a subsequent PCR across the insertion site in *CAV1*, using the gRNA 4 validation primers, as used for the T7E1 assay.

Due to the 132 bp size of the SV40 PAS insert, we expected to be able to resolve the difference between PCR products with and without the insert on an agarose gel. The expected sizes of PCR products would be 565 bp and 697 bp, dependent on the absence or presence of the SV40 PAS insert, respectively. To demonstrate that the the PCR products run at different sizes dependent on the inclusion of the SV40 PAS into the *CAV1* genic sequence, in addition to PCR of transfected HEK293FT gDNA, we also performed a PCR on the the WT enhancer vector (CAV1_enhancer_WT) and the HDR template (CAV1_SV40PAS_gRNA4) (Figure 7.11).

As expected, for both the WT enhancer vector and untransfected HEK293FT gDNA, we observed PCR products of ~565 bp. Suggestive of successful incorporation of the SV40 PAS insert into *CAV1* DNA in HEK293FT cells, we observed PCR products of ~697 bp for both transfected HEK293FT gDNA samples, yielding PCR products of the same size as the HDR template. We detected no visible difference in SV40 PAS incorporation dependent on pX459:HDR ratio (Figure 7.11).

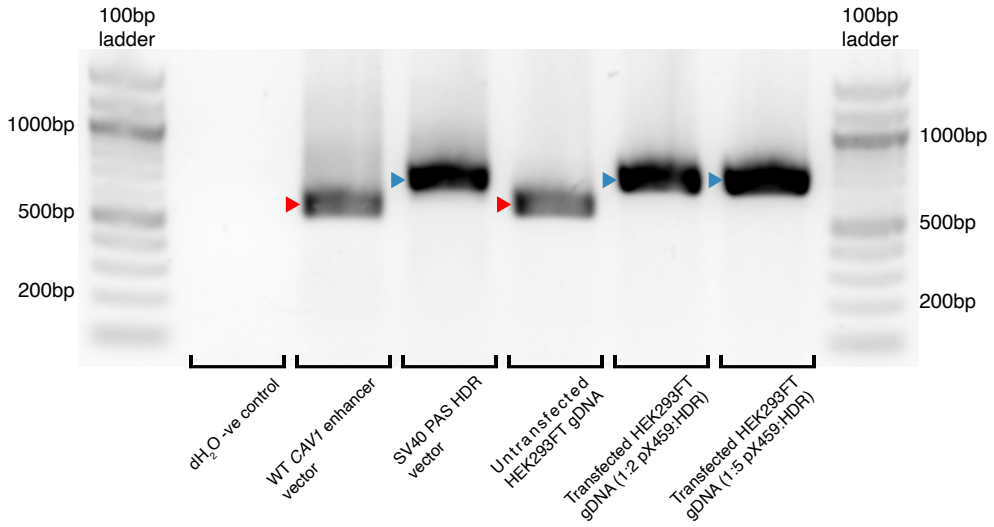



FIGURE 7.11: Successful incorporation of the SV40 PAS into *CAV1* sequence in HEK293FT cells

Validation of SV40 PAS insertion into *CAV1* DNA sequence following transfection in HEK293FT cells. PCR over the insertion site using gRNA 4 validation primers yields PCR products of distinct sizes, dependent on the insertion of the SV40 PAS. PCR product of WT *CAV1* sequence yields a PCR product of 565 bp (red arrows), while PCR product of *CAV1* sequence with SV40 PAS insertion yields a PCR product of 697 bp (blue arrows). For comparison, PCR was performed on both the WT enhancer vector (CAV1_enhancer_WT) and the HDR template (CAV1_SV40PAS_gRNA4), to demonstrate the expected PCR product size change, and resultant band shift.

## 7.6 Discussion

In this chapter, we examined the differential expression of *CAV1* 3' UTRs and devised a study to specifically test the role of transcripts arising from the 3' UTR-localised enhancer.

### 7.6.1 Differential *CAV1* 3' UTR expression

While in SH-SY5Y cell presumably produce short *CAV1* 3' UTR isoforms, based on preferential usage of the proximal polyA site, this is contrary to what we would expect from a neuronal-like cell, with neurons typically biased towards long 3' UTR isoform expression (Guvenek and Tian, 2018, Miura et al., 2013, Zhang et al., 2005). With the assumption that SH-SY5Y cells utilise the most proximal polyA site at chr7:116,199,356 (Wang et al., 2018), *CAV1* transcripts in SH-SY5Y cells are likely between 550 and 600 bp in length, based on the *CAV1* transcript isoforms identified by Ensembl (Harrison et al., 2024). This greatly contrasts what we observed in iPSC-derived motor neurons, which displayed *CAV1* transcript lengths ranging from 1645 to 2434 bp (Figure 7.1), suggesting preferential usage of more distal polyA sites, typical of neurons (Guvenek and Tian, 2018, Miura et al., 2013, Zhang et al., 2005). This likely reflects the stronger neuronal identity of iPSC-derived motor neurons compared to undifferentiated SH-SY5Y cells.

In iPSC-derived motor neurons, we observed an increase in *CAV1* transcript length, likely due to 3' UTR lengthening, with age (Figure 7.1), mirroring the global shift towards long 3' UTR isoform expression observed in ageing mice (Shafik et al., 2021). However, given the established inverse relationship between 3' UTR length and protein expression level (Mayr and Bartel, 2009, Sandberg et al., 2008), this directly opposes the observed increase in CAV1 protein expression observed with age (Ha et al., 2021). If CAV1 expression was solely dictated by 3' UTR-mediated post-transcriptional regulation of *CAV1* mRNA, we would expect to observe a reduction in 3' UTR length with age. Interestingly, we do however observe a negative correlation between *CAV1* expression and transcript length in ALS patient-derived motor neurons (Figure 7.2). While mRNA abundance alone is insufficient to predict protein abundance (reviewed in Liu et al., 2016), this provides insight into the relationship between 3' UTR length and mRNA expression, specifically in disease states.

By utilising total RNA-seq to predict *CAV1* transcript abundance, we captured steady-state mRNA levels, influenced by post-transcriptional regulation by regulatory sequences in the 3' UTR. Comparing the iPSC-derived motor neuron samples expressing the

longest *CAV1* 3' UTR isoforms, PGP and ALS29, provides insight into potential post-transcriptional regulation of *CAV1* in healthy and disease states. With the healthy control-derived motor neurons exhibiting more than a five-fold increase in *CAV1* expression compared to the ALS patient-derived motor neurons, this may suggest reduced *CAV1* mRNA stability in ALS patients, potentially as a result of increased 3' UTR-mediated post-transcriptional regulation.

### 7.6.2 The impact of ALS-associated SNPs on AU-rich elements

To investigate how ALS-associated SNPs within the *CAV1* 3' UTR may impact *CAV1* expression, independent on their impact on the proximal enhancer, we identified AREs within the *CAV1* 3' UTR. Despite the presence of 5 SNPs across the 3' UTR only one SNP, chr7:116,199,522T>C, overlapped with ARE motifs (Figure 7.3). While AREs are, in majority, repressive to mRNA expression, we cannot predict the impact of abolishing the ARE site in the 3' UTR, as proteins both with stabilising and destabilising functions have been shown to bind at these motifs (Brennan and Steitz, 2001, Chen et al., 2001, Kruys et al., 1989, Lai et al., 2000, Lykke-Andersen and Wagner, 2005b, Meijlink et al., 1985, Peng et al., 1998, Sandler et al., 2011, Shaw and Kamen, 1986). Therefore, this ALS-associated SNP has potential to exert effect on *CAV1* mRNA expression, either by disrupting destabilising effects and resulting in *CAV1* upregulation, or through disrupting stabilising effects and resulting in *CAV1* downregulation.

Intriguingly, while we identified one ALS-associated SNP overlapping an ARE, the remaining 4 ALS-associated SNPs within the 3' UTR (chr7:116,200,589T>A; chr7:116,200,705C>T; chr7:116,200,719C>A; chr7:116,200,953A>T) do not overlap with AREs. This suggests that these SNPs likely impact *CAV1* expression through disruption of the enhancer sequence, and likely have no functional impact on the *CAV1* 3' UTR in isolation.

### 7.6.3 3' UTR expression isolated from cognate CDS expression

Across human tissues, and particularly in neurons, there is widespread non-random differential expression of mRNA 3' UTRs and their cognate protein-coding regions. Many genes exhibit high 3' UTR expression, with patterns of expression often reciprocal to expression of the cognate CDS (Ji et al., 2021, Kocabas et al., 2015, Mercer et al., 2011). Such observations resulted in the idea that the 3' UTRs of human genes may be expressed in isolation, separately from their associated protein coding sequence, as a result of *de novo* transcription within the 3' UTR (Ji et al., 2021, Mercer et al., 2011). RNA transcripts originating from isolated 3' UTRs may function in *cis* to regulate linked

protein expression (Mercer et al., 2011), with high 3' UTR to CDS expression ratio in neurons inversely correlated with associated protein expression, positing a mechanism where increased 3' UTR expression relative to the CDS may titrate protein expression (Kocabas et al., 2015).

Isolated 3' UTR expression typically occurs in a spatiotemporal, cell type-specific manner (Ji et al., 2021, Mercer et al., 2011). Furthermore, many of the characteristics of isolated 3' UTRs - namely *de novo* transcription start sites, H3K27ac enrichment, DNaseI hypersensitivity and dynamic cell type-specific functionality (Ji et al., 2021, Kocabas et al., 2015, Mercer et al., 2011) - mirror many of the established characteristics of enhancers. For this reason, we propose that in some cases, the phenomena of isolated 3' UTR expression may be arise due to the presence of enhancer elements within the 3' UTR sequence.

### 7.6.4 Isolated *CAV1* 3' UTR expression

Isolated 3' UTR expression, as a result of unbalanced expression from the 3' UTR and its cognate CDS, is implicated in genetic regulation (Ji et al., 2021, Kocabas et al., 2015, Mercer et al., 2011). Due to the observed enrichment of nascent RNA transcription across the *CAV1* 3' UTR in our TT-seq datasets, we investigated whether we observe isolated 3' UTR expression not only in our model cell lines, but also in iPSC-derived motor neurons (Zhang et al., 2020a). Across both SH-SY5Y cells (Figure 7.4) and iPSC-derived motor neurons (Figure 7.5), we observe higher 3' UTR expression relative to CDS expression. In neurons, 3' UTR to CDS expression ratios are somewhat predictive of protein expression levels (Kocabas et al., 2015). Therefore, the expression ratios we observe in SH-SY5Y cells and iPSC-derived motor neurons may reflect the established low expression of CAV1 protein in the brain (Head et al., 2011, 2010).

Paired with our analysis of 3' UTR length, investigating isolated 3' UTR expression may provide some insight into *de novo* transcription within the *CAV1* 3' UTR. Despite the predicted preferential expression of short *CAV1* 3' UTR isoforms in SH-SY5Y cells, we observed higher levels of 3' UTR expression relative to the CDS. This is predictive of *de novo* transcription initiation within the *CAV1* 3' UTR. On the other hand, some of the increased 3' UTR expression observed in the PGP healthy control sample may be attributed to the expression of a longer *CAV1* 3' UTR isoform, potentially skewing mean expression across the 3' UTR. To assess the influence of 3' UTR length on expression, we compared the 3' UTR to CDS expression ratios between the PGP sample and the ALS29 patient-derived sample, which produces *CAV1* transcripts of a similar length. Although ALS29 has the highest 3' UTR to CDS ratio among ALS-derived samples,

PGP still exhibits a considerably higher fold change relative to CDS expression. This suggests that the high 3' UTR to CDS ratio in PGP is not solely due to the generation of longer 3' UTR isoforms, but possibly reflects elevated enhancer activity in this sample, leading to increased eRNA transcription at the *CAV1* 3' UTR.

Isolated 3' UTR expression has been proposed to have power over protein expression via titration of cognate mRNA expression (Kocabas et al., 2015, Mercer et al., 2011). To explore whether high relative *CAV1* 3' UTR expression may have influence over CAV1 protein in whole human tissues, particularly in disease-relevant tissues, we compared the 3' UTR to CDS expression ratios between CNS-relevant tissues, known to exhibit low CAV1 expression, to expression ratios in lung-relevant cells and tissues, known to exhibit high CAV1 expression (Figure 7.6). As expected, the majority of CNS-associated tissues exhibited higher *CAV1* 3' UTR expression relative to CDS expression, while lung-associated cells and tissues demonstrated more balanced expression between the 3' UTR and CDS. This may support a mechanism where differential 3' UTR to CDS expression ratios may regulate CAV1 protein levels, with equal CDS and 3' UTR expression driving higher protein levels, and elevated 3' UTR to CDS expression ratios correlating with lower protein levels.

Of particular interest, we observed notable differential 3' UTR to CDS expression in spinal cord tissue, with CDS relatively underexpressed compared to all human tissues studied, versus the relative overexpression observed for the 3' UTR (Figure 7.6). Spinal cord atrophy is a key hallmark of ALS (Nigri et al., 2023), with spinal cord degeneration potentially serving as a predictor of survival due to its strong correlation with disease progression (reviewed in El Mendili et al., 2019, Turner and Benatar, 2015). The degree of spinal cord atrophy reflects of the loss of both upper and lower motor neurons, providing a valuable metric for predicting disease severity and progression (El Mendili et al., 2014). Overall, these findings highlight a clear bias in the expression across the *CAV1* gene in key tissues relevant to ALS pathology.

Taken together, we cannot ignore the potential influence of the enhancer within the *CAV1* 3' UTR, particularly given the presence of the CNS-specific enhancer at chr7:116,199,565-116,200,132. The high 3' UTR expression observed across neuronal-like cell types and CNS-relevant human tissues may be driven by eRNA transcription from the CNS-specific enhancer. In contrast, the equal 3' UTR to CDS expression ratio in lung tissues may indicate transcriptional readthrough from the *CAV1* gene into the 3' UTR rather than *de novo* transcription events.

Using *CAV1* as an example of a gene that demonstrates differential 3' UTR to CDS expression across distinct tissues, we consider the potential implications of 3' UTR-localised enhancers on isolated 3' UTR expression. Whilst candidate enhancers in the ABC

enhancer catalogue largely reside within intronic and intergenic regions (Figure 7.7), >6,000 candidate enhancers were identified within 3' UTR regions genome-wide. Although this constitutes a small proportion of the total candidate enhancers, it highlights the possibility that the 3' UTRs of many genes may possess enhancer functionality.

The characteristic non-continuous RNA expression across isolated 3' UTRs directly aligns with the enrichment of H3K27ac deposition, DNAseI hypersensitive sites and CpG islands (Ji et al., 2021). For this reason, we propose that in some instances, the phenomena of isolated 3' UTRs may be a result of enhancer functionality within the 3' UTR sequence. The *Sox* gene family has readily been used to model unequal 3' UTR and CDS expression, with *Sox11* and *Sox12* specifically cited for exhibiting isolated 3' UTR expression (Ji et al., 2021, Kocabas et al., 2015). For both of these genes, we identify putative enhancer elements from the ABC enhancer catalogue (Fulco et al., 2019) within their respective 3' UTRs. This supports the potential for underlying enhancer functionality within the 3' UTR to contribute to observed isolated 3' UTR expression. expression.

### 7.6.5 Enhancers in 3' UTRs and disease relevance

There is an abundance of examples of SNPs within 3' UTRs having pathological consequence, however there are comparatively limited examples of disease-associated SNPs within 3' UTR-located enhancers with phenotypic outcome. Along with the rs7873784 SNP in the *TLR4* enhancer within the *TLR4* 3' UTR, which creates a *de novo* binding site for PU.1 that drives pathological inflammation (Korneev et al., 2020), there is also evidence that polymorphisms in 3' UTR-located enhancers are linked to neurodegenerative diseases.

*SNCA* is positioned as a major risk gene in PD pathology, with *SNCA* polymorphisms commonly associated with the development of sporadic PD (Nalls et al., 2014). GWAS in PD patients have shown strong signals in the *SCNA* 3' UTR (Krohn et al., 2020, Nalls et al., 2014, Rudakou et al., 2021, Simón-Sánchez et al., 2009, Soldner et al., 2016), implying that variants in the 3' UTR of *SNCA* may contribute to PD risk. SNPs within the *SNCA* 3' UTR result in an increase in gene expression in luciferase reporter assays (Sotiriou et al., 2009). In line with the increase in *SCNA* expression in autosomal dominant PD (Chartier-Harlin et al., 2004, Ibáñez et al., 2009), this may suggest a comparable pathological increase in *SCNA* expression may result from SNPs within the 3' UTR. Rare variant burden testing identified two PD-associated variants within the 3' UTR of *SCNA*, rs1045722 and rs3857053 (Rudakou et al., 2021). Interestingly, these 3' UTR variants are located within a strong enhancer region and are further associated

with altered splicing of *SCNA* in the brain cortex (Blažeković et al., 2021, Rudakou et al., 2021).

In both of the above examples, it is not possible to distinguish wholly whether the resulting pathological phenotype is a result of mutation of the enhancer sequence or the 3' UTR sequence. While PU.1 binding to enhancers is well-establised (Heinz et al., 2010), there is some evidence of PU.1 binding at 3' UTRs (Hosokawa et al., 2018). Regarding the altered splicing of *SCNA*, enhancers, in cooperation with their cognate promoter, have been shown to regulate alternative splicing (Bhat-Nakshatri et al., 2013, Kadener et al., 2002), largely through impacts on transcriptional initiation and elongation. 3' UTRs may also have some control over alternative splicing through the preferential binding of splicing factors and RBPs to regulatory elements within the 3' UTR. For example, RBPs such as TIA1/TIAL1 and HuR that are historically implicated in mRNA stability and post-transcriptional control through binding to the 3' UTR (Díaz-Muñoz et al., 2017, Gray et al., 2010, Liu et al., 2014, Wang et al., 2010a, Yu et al., 2011), also play roles in alternative splicing (Del Gatto-Konczak et al., 2000, Díaz-Muñoz et al., 2017, Förch et al., 2000, Lee et al., 2021, Osma-Garcia et al., 2022, Wang et al., 2010b, Zhang et al., 2020c).

Due to the difficulties in identifying how disease-associated SNPs within 3' UTR-localised enhancer elements may differentially impact both enhancer- and 3' UTR-mediated regulatory mechanisms, we have devised a CRISPR/Cas9-based strategy to generate SH-SY5Y cells lines that functionally isolate the *CAV1* CDS from the 3' UTR.

### 7.6.6 Generation of cell lines to study the *CAV1* proximal enhancer independent of *CAV1* mRNA

Due to the unexpected regulatory activity of the 3'UTR-associated enhancer, we aimed to disentangle the enhancer function from the 3' UTR function by inducing transcriptional termination immediately downstream of the *CAV1* CDS. Not only would this allow us to identify if the repressive functions of the 3' UTR are impacting enhancer activity, but it would allow us to study the impact that ALS-associated SNPs within this enhancer have on *CAV1* expression without impacting the genic sequence of *CAV1*.

Although we successfully generated the HDR template for SV40 PAS insertion using pX459_gRNA4 (Figure 7.10c), in order to generate SH-SY5Y cells with the SV40 insert we must repeat site-directed mutagenesis of the CAV1_SV40PAS_gRNA4 HDR template to generate the appropriate PAM mutations for Cas9-mediated SV40 PAS insertion using the pX459_gRNA1 vector. The discontinuity of the CAV1_SV40PAS_WT and CAV1_SV40PAS_gRNA1 HDR constructs following site-directed mutagenesis likely arise

due to the repetitive nature of the *CAV1* 3' UTR. Due to dual functionality of the *CAV1* 3' UTR, it contains both AT-rich motifs characteristic of 3' UTRs (Chen et al., 2001, Kruys et al., 1989, Meijlink et al., 1985, Shaw and Kamen, 1986), and GC-rich motifs characteristic of enhancers (Xiong et al., 2019). PCR amplification of repetitive sequences is error-prone, whereby repetitive DNA sequences can form complex structures which promote dissociation of the polymerase enzyme from the template DNA, yielding incomplete amplification (Riet et al., 2017). To optimise this reaction, we would repeat the site-directed mutagenesis reactions, testing different denaturation and annealing temperatures, as suggested in Riet et al. (2017).

If we wanted to further validate the incorporation of the SV40 PAS insert into *CAV1* DNA beyond the visual interpretation of the PCR product band shift (Figure 7.11), we could performed a test digest of PCR products from transfected gDNA. The SV40 insert includes a unique *Sal*I restriction site. Downstream, upon transfection with the sgRNA and the donor template, we could use this to screen for SV40 insertion into cells. However, restriction digest of linear DNA can prove difficult, especially when restriction digest yields products of distinct lengths. Type II restriction enzymes can demonstrate kinetic preference for recognition sites central to linear DNA substrates (Nardone et al., 1986). Restriction digest of a PCR product containing the SV40 PAS insert would yield products of 111bp and 586bp, therefore the relative terminal location of the *Sal*I restriction site may hinder this as a screening technique. To rectify this, designing PCR primers that would centralise the *Sal*I restriction site in the PCR product may aid in restriction digest of linear PCR products (Nardone et al., 1986), however we would again be reliant on observation of a band shift on the agarose gel to confirm SV40 PAS insertion.

### 7.6.7 Conclusion

Taken together, we have discussed how differential *CAV1* 3' UTR isoform expression and *de novo* 3' UTR transcription events may contribute to overall *CAV1* gene expression, considering the implications of ALS-associated SNPs in the 3' UTR, independent from the impacts of the proximal enhancer located within the 3' UTR. Across all model cell lines, iPSC-derived motor neuron samples, and in ALS-associated whole tissue, we observe high *CAV1* 3' UTR expression relative to CDS expression, directly reflecting the commonly observed isolated 3' UTR expression in neurons (Kocabas et al., 2015). We considered how differential 3' UTR isoform transcription may impact the observed isolated 3' UTR expression, especially in the context of how eRNA transcription from the *CAV1/CAV2* proximal enhancer may shape the transcriptional profile of the 3' UTR.

Whereby, despite the preferential expression of the short *CAV1* 3' UTR isoform in SH-SY5Y cells, we established high relative levels of 3' UTR expression, a characteristic that may be directly linked to the presence of an actively transcribing enhancer within the 3' UTR.

Furthermore, we have demonstrated the effective insertion of the SV40 PAS directly downstream of the *CAV1* CDS in HEK293FT cells, providing an experimental framework for the future generation of genetically modified SH-SY5Y cells. Generation of the SH-SY5Y cell line would provide an effective screening approach for the impact of the ALS-associated SNPs within the 3' UTR, enabling investigation into the impact of on the proximal model *CAV1/CAV2* enhancer, without influencing the *CAV1* genic sequence.

# Chapter 8

# Discussion

## 8.1 Enhancer selectivity and specificity

Complicating the fundamental role for enhancer elements in underpinning regulatory networks, enhancers are largely cell and tissue type-specific (Heintzman et al., 2009, Visel et al., 2009). Understanding the cell and tissue type specificity of the *CAV1* enhancer landscape, as well as the functionality of *CAV1*-associated enhancers in distinct neuronal and CNS-relevant cells and tissues, is crucial for comprehending both the impact and *in vivo* relevance of ALS-associated SNPs. In fact, enhancers containing disease-associated SNPs demonstrate activity in pathologically relevant cell types (Andersson et al., 2014, ENCODE Project Consortium, 2012), and genetic variants linked to neurological diseases exhibit cell type-specific patterns of enhancer enrichment (Nott et al., 2019).

The location, orientation, strength, and TF motif grammar at a given enhancer constitute the so-called "enhancer code", which determines the cell- and tissue-specificity of enhancer activity (Taskiran et al., 2024). Enhancers with tissue-specific activity are predominantly enriched in intronic regions, whereas ubiquitously expressed housekeeping genes are primarily regulated by intergenic enhancers (Borsari et al., 2021). The cooperative binding of TFs at enhancers specifies their activity and accessibility in distinct cell types, facilitating the existence of "tissue-agnostic" enhancers that are active across multiple cell types (Taskiran et al., 2024).

Recent computational approaches have systematically analysed the enhancer-promoter interactome across different cell types based on enhancer-promoter interactions predicted by the ABC model, identifying distinct patterns of enhancer-promoter regulation across diverse cell types (Wu and Huang, 2024). As expected, specific enhancer-promoter interactions limited to individual cell types and conserved enhancer-promoter interactions

shared among the majority of cell types were observed. However, an additional enhancer-promoter regulation pattern has emerged where each gene is associated with multiple enhancers, but these enhancers rarely coexist in the same cell type. This regulatory pattern is established in one of two ways: either through a single cell type-invariant enhancer that functions across the majority of cell types, with additional enhancers forming cell type-specific interactions, or through maintained chromatin accessibility, which enables interactions between distinct enhancers in different cell types (Figure 8.1) (Wu and Huang, 2024). Therefore, understanding which enhancers are active in pathogenically-relevant cell types and tissues, and particularly those analogously active in experimental model systems is essential for the study of enhancer architecture in disease.

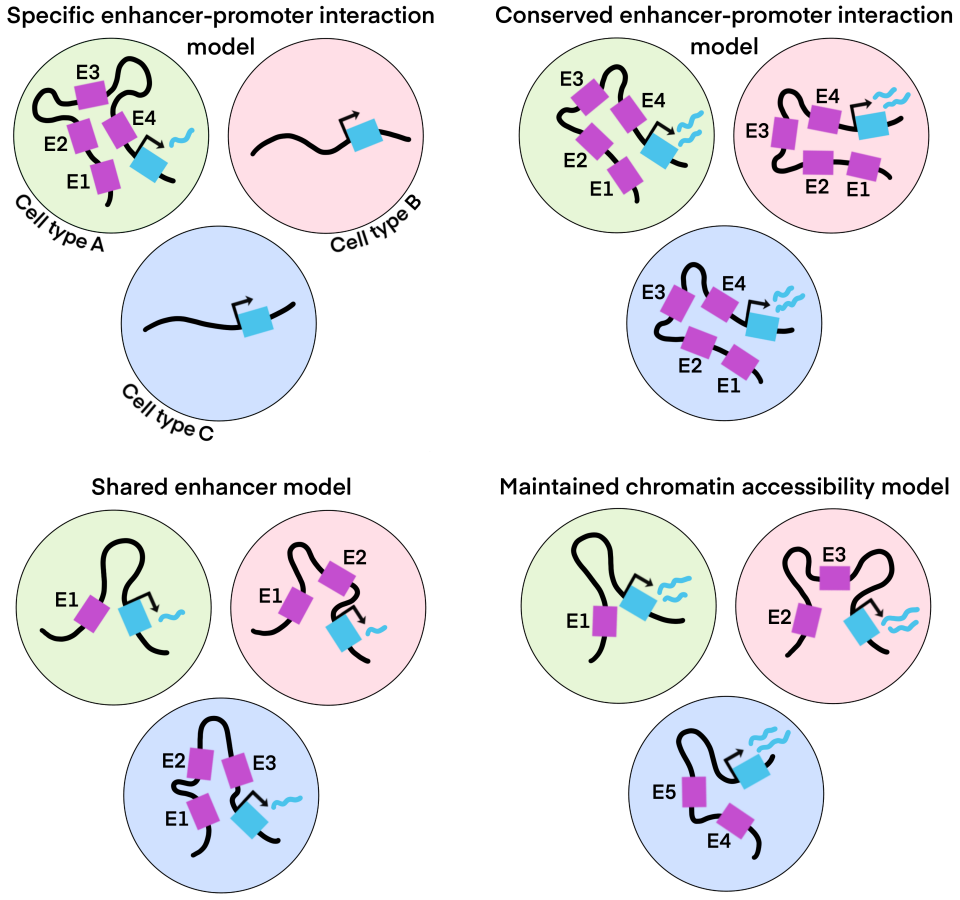



FIGURE 8.1: Models of enhancer-promoter interaction
Distinct patterns of enhancer-promoter regulation are predicted to exist *in vivo*. The specific enhancer-promoter interaction model limits regulatory interactions to a given cell type, while the conserved model features interactions consistent across diverse cell types. The shared enhancer and maintained chromatin accessibility models involve regulatory patterns where the same gene is regulated by distinct enhancers in different cell types. Shared enhancers can establish stable interactions with target genes in most cells, while cell-specific enhancers modulate expression. Through maintained chromatin accessibility at both enhancers and promoters, gene expression can be maintained across different cell types.

Comparison of chromatin accessibility and H3K27ac deposition between neuronal-like SH-SY5Y cells and iPSC-derived motor neurons revealed examples of enhancers in an active state that are either unique to each cell type or common to both cell types, highlighting enhancer specificity. The majority of the candidate *CAV1/CAV2* enhancers examined in this work are considered to be CNS-specific, with 88 out of the 114 total enhancers exhibiting this specificity (Table B.1). Beyond the generalised CNS-specificity of the studied *CAV1/CAV2* enhancers, there may be further specificity that distinguishes the neuronal-like nature of SH-SY5Y cells from iPSC-derived motor neurons, as evidenced by minimal overlaps in H3K27ac enrichment between SH-SY5Y cells and iPSC-derived motor neurons (Figure 4.18). Comparatively, we observed more overlaps in chromatin accessibility between cell types, with 12 enhancers in an active chromatin state across all iPSC-derived motor neuron samples, regardless of disease state, and SH-SY5Y cells (Figure 4.7). This may reflect a model where enhancers that demonstrate invariant chromatin accessibility across cell types may contact the promoter and recruit additional cell type-specific enhancers (Wu and Huang, 2024), which could be better identified through H3K27ac deposition in their respective cell types.

Further highlighting conserved versus cell type-specific enhancer activity, visually, we observed both conserved and distinct patterns of nascent transcription across *CAV1/CAV2* enhancers between SH-SY5Y cells and human-derived lymphoblastoid cells (Figure 5.11, Figure 5.23). Due to this, we question the identity of many of these enhancers as CNS-specific. Surprisingly, we observed a greater number of transcribed enhancers in healthy control-derived lymphoblastoid cells compared to SH-SY5Y cells. This suggests that some of the candidate enhancers previously thought to be CNS-specific may have broader regulatory roles beyond the CNS, potentially reflecting their involvement in more general transcriptional regulation across cell types. Given the inherent complexity of regulatory interactions in 3D space, with potential interactions between tissue-agnostic and cell type-specific enhancers driving gene expression from a shared promoter, understanding the long-range chromatin interactions of the *CAV1/CAV2* enhancer landscape, through methods such as proximity-based fluorescence in situ hybridization (FISH) (Finn and Misteli, 2021) or 3C (Deng and Blobel, 2017), may provide further clarity on the true cell type-specificity and regulatory mechanisms of these enhancers.

## 8.2 Model cell lines for studying enhancer activity

### 8.2.1 SH-SY5Y cells as a neuronal model

The applications of mammalian primary neurons derived from embryonic tissue are limited because, once terminally differentiated into mature neurons, they can no longer be propagated in culture (Kovalevich and Langford, 2013). To overcome this limitation, transformed neuronal-like cell lines, including the neuroblastoma-derived SH-SY5Y cell line, are commonly utilised in neurobiology research (Biedler et al., 1978, Kovalevich and Langford, 2013). As SH-SY5Y cells are of human origin, a number of human-specific genes are expressed that would not be inherently abundant in rodent primary neuronal cultures (Kovalevich and Langford, 2013). This is of significant relevance to the study of human transcriptional regulation, especially when considering the cell type-specificity of enhancer function (Heintzman et al., 2009, Visel et al., 2009).

#### 8.2.1.1 Differentiation of SH-SY5Y cells

Both the parental SK-N-SH line, and the SH-SY5Y subcloned line comprise of two morphologically distinct phenotypes: neuroblast-like cells and epithelial-like cells (Encinas et al., 2000, Ross et al., 1983). In the undifferentiated cell state, these morphologically neuroblastic cells are characterised by non-polarised cell bodies with few, underdeveloped projections (Kovalevich and Langford, 2013, Ross et al., 1983). Upon sequential exposure of SH-SY5Y cells to retinoic acid and brain-derived neurotrophic factor (BDNF), neuroblastic cells differentiate into a neuronal-like phenotype, expressing neuronal markers and exhibiting morphological characteristics typical of primary neuronal cultures, including neuritic outgrowth and growth cones (Encinas et al., 2000, Fagerström et al., 1996, Kovalevich and Langford, 2013). Physiologically, both undifferentiated and differentiated SH-SY5Y cells exhibit markers indicative of neuronal phenotype, albeit undifferentiated cells display characteristics of more immature neuronal cells. Undifferentiated SH-SY5Y cells express Nestin (Lopes et al., 2010), a key marker developmental marker characteristic of neural progenitor cells (reviewed in Palmer et al., 2000). Interestingly, although to a lesser extent than differentiated cells, undifferentiated SH-SY5Y cells also express neuronal nuclear antigen (NeuN) (Lopes et al., 2010), a protein abundant in the nucleus of neurons, commonly utilised as a biomarker of neuronal differentiation (reviewed in Gusel'nikova and Korzhevskiy, 2015). Upon differentiation, SH-SY5Y cells also express the neuronal markers tyrosine hydroxylase (Lopes et al., 2010, Ross et al., 1983), a key enzyme in catecholamine biosynthesis (reviewed in Molinoff and Axelrod, 1971),

and neuron-specific enolase (Lopes et al., 2010), the predominant enolase-isoenzyme in neuronal tissues and a highly specific neuronal biomarker (reviewed in Isgrò et al., 2015).

In line with the functionally more mature neuronal phenotype of differentiated SH-SY5Y cells, following differentiation, SH-SY5Y cells also lose their proliferative potential as cells are withdrawn from the cell cycle and arrested in the G1 stage of mitosis (Encinas et al., 2000). Distinctly, mature neurons are paused in the G0 of the cell cycle (reviewed in Frade and Ovejero-Benito, 2015), only reactivating cell cycle progression and proceeding to G1 in response to DNA damage (Tomashevski et al., 2010). This distinction between G1 and G0 arrest may reflect the distinction between more immature and terminally differentiated neurons, where arrest in quiescent G0 phase in mature neurons marks exit from the cell cycle, conferring stability and preventing proliferation (reviewed in Hardwick and Philpott, 2014). The increased proliferation rate of undifferentiated SH-SY5Y cells, relative to differentiated SH-SY5Y cells, makes them a more convenient model system for the study of nascent RNA transcription, given the reduction in transcriptional elongation in quiescent cells compared to those that are actively proliferating (reviewed in Young et al., 2017). Given the large cellular input for TT-seq (Schwalb et al., 2016), required to successfully capture appropriate amounts nascent RNA for next-generation sequencing, using differentiated SH-SY5Y cells could be time-consuming and costly. Especially when we consider the probable reduction in transcription in differentiated cells, using differentiated SH-SY5Y cells as input for TT-seq may require significant experimental scale-up compared to undifferentiated cells in order to garner enough nascent RNA for readout.

#### 8.2.1.2 Transcriptional similarities between SH-SY5Y cells and neurons

Clear similarities between neuronal cells and SH-SY5Y cells have been identified regarding mRNA stability, abundance and isoform choice. mRNA abundance and stability in SH-SY5Y cells appears to be linked to genes implicated in neurogenesis, whereby mRNAs transcribed from genes commonly upregulated during neurogenesis were of relative high abundance and stability in SH-SY5Y cells compared to other non-neuronal cell lines (Guvenek et al., 2022). Compared to other cell types, neuronal cells preferentially select the PAS distal to the gene promoter (Guvenek and Tian, 2018, Miura et al., 2013, Zhang et al., 2005). SH-SY5Y cells express longer 3' UTR mRNA isoforms compared to the non-neuronal HEK293T and HepG2 cell lines (Guvenek et al., 2022). This mirrors the preferential production of long 3' UTR isoforms in neuronal cells (Guvenek and Tian, 2018, Miura et al., 2013, Zhang et al., 2005). These studies establish a somewhat neuronal identity for SH-SY5Y cells, specifically regarding post-transcriptional regulation.

Despite the suggestion that SH-SY5Y cells generally produce longer 3' UTR isoforms (Guvenek and Tian, 2018), this opposes the observed short *CAV1* 3' UTR isoform production identified in our SH-SY5Y TT-seq data (Figure 6.12). While Guvenek and Tian (2018) studied PAS usage in undifferentiated SH-SY5Y cells, it is interesting to consider how *CAV1* 3' UTR expression may alter upon SH-SY5Y differentiation. With SH-SY5Y cells exhibiting a phenotype more characteristic of primary neurons upon differentiation (Encinas et al., 2000, Fagerström et al., 1996, Kovalevich and Langford, 2013, Lopes et al., 2010, Ross et al., 1983), we may expect to observe the production of a longer *CAV1* 3' UTR isoform following differentiation, more similar to what we observed for iPSC-derived motor neurons (Figure 6.12).

The use of TT-seq to measure 3' UTR isoform expression has inherent limitations, wherein this method specifically captures nascent RNA species, and therefore may not accurately represent steady-state mRNA levels. To gain a more accurate understanding of 3' UTR isoform expression, methods such as total RNA-seq, which assess steady-state mRNA levels, may provide a more accurate view of 3' UTR isoform expression. Alternatively, to avoid potential biases introduced by transcript assembly, long-read sequencing, which does not fragment the RNA during library preparation, may provide a more robust approach to identify *CAV1* transcripts as a whole.

### 8.2.2 Model systems to capture the cell type-specific impact of *CAV1* enhancer mutations

Regarding our transcriptomic and epigenetic analyses of the *CAV1/CAV2* enhancer landscape, none of our model cell lines fully recapitulate both the specific effects of *CAV1* enhancer mutations and the cell type-specificity of enhancers. While the iPSC-derived *in vitro* differentiated motor neurons from ALS patients offer a valuable model to directly investigate the *CAV1/CAV2* enhancer landscape in the context of ALS, they may not be optimal for studying the specific effects of *CAV1* enhancer mutations. The mutation underlying ALS pathogenesis in these patients was the *C9orf72* hexanucleotide repeat expansion, therefore the relevance of these cell lines in the study of *CAV1* enhancer-specific regulatory mechanisms may be limited. However, with the proposed neuroprotective properties of CAV1 (Cooper-Knock et al., 2021, Egawa et al., 2018, Head et al., 2011, 2008, Sawada et al., 2019, Takayasu et al., 2010), these patient-derived motor neurons may still offer insights into how the *CAV1* regulatory landscape is altered in ALS in the absence of causative enhancer mutations.

On the other hand, while ALS patient-derived lymphoblastoid cells may capture the effects of specific *CAV1* enhancer mutations, they lack CNS-specific enhancer activity,

limiting their relevance in modeling the neuron-specific regulatory landscape. Due to the distinct differences between our model cell lines, a major limitation of our study is the necessity to extrapolate what the results from each cell line might mean in the context of the native cellular environment in ALS disease states. To best recapitulate the combined impact of *CAV1* enhancer mutations and neuronal enhancer activity, the ideal cellular model would be genetically modified SH-SY5Y cells or iPSCs, further differentiated into motor neurons. Generating a model system that combines both the neuronal context and models the effect of specific *CAV1* enhancer mutations would enable a more precise investigation into how enhancer mutations contribute to disease mechanisms, providing a clearer understanding of their role in neurodegeneration.

### 8.2.3 Enhancer cooperativity in model cell lines

Clustering of *CAV1/CAV2*-associated enhancer elements proximal to the *CAV1* gene may be a consequence of either distance from the target gene, or of distance from the TAD boundary (Figure 3.5, Figure 3.9). Enhancers, while capable of contacting their target gene at a distance, typically localise proximal to their target gene (Zuin et al., 2022), with many gene-enhancer regulatory predictions often based on gene-enhancer distance (Fishilevich et al., 2017, Fulco et al., 2019). In addition, enhancers may cluster proximal to TAD boundaries to enable inter-TAD interactions (Hung et al., 2024). TAD boundaries form hierarchical interactions, whereby high-level TAD boundaries mediate interactions with a number of additional TAD boundaries. A high-level TAD boundary can gather numerous enhancers by interacting with other TAD boundaries (Xu et al., 2024). Thus, the clustering of enhancers at the downstream *CAV1* TAD boundary may enable complex cooperative enhancer-enhancer interactions that span across TADs (Figure 3.5). Such clustering of regulatory elements may be considered analogous to the formation of super-enhancers (Xu et al., 2024). This counters the idea that super-enhancers typically cluster central to TADs (Schmitt et al., 2016). However, with the recent identification of super-enhancers that can overcome TAD insulation and form inter-TAD regulatory interactions to regulate gene expression across TAD boundaries (Kessler et al., 2023), coupled with the phenomenon of TAD boundary "stacking" (Hung et al., 2024), localisation of super-enhancers proximal to TAD boundaries may be more prevalent than previously appreciated.

For the enhancer cluster proximal to *CAV1*, we observed considerable enrichment of H3K27ac in a number of these enhancers in both ALS patient-derived motor neurons and in SH-SY5Y cells (Figure 4.18). H3K27ac deposition is a key determinant of super-enhancer activity (Zhang et al., 2016), potentially suggestive of super-enhancer function at this locus. Interestingly, when comparing healthy to disease states in lymphoblastoid

cells, we observed an increase in nascent transcription at this enhancer cluster, largely exclusive to disease states (Figure 5.23). This may demonstrate the activation of the super-enhancer at this locus in disease states. This mirrors what has been observed in tumour cells, where cancer cells can acquire super-enhancers at oncogenes to drive tumoriogenesis (Hnisz et al., 2013).

## 8.3 ALS-associated SNPs produce distinct effects on *CAV1* expression

### 8.3.1 Mutations in the model proximal enhancer

We have provided evidence that, even within the same enhancer, ALS-associated SNPs may have distinct impacts on *CAV1* expression. Through computational prediction, the chr7:116,199,522T>C SNP was implicated in inducing eRNA structural changes in SH-SY5Y cells (Figure 5.21) and altering TF binding sequence motifs (Figure 6.19). Conversely, in *in vitro* luciferase reporter assays, this SNP did not significantly alter the ability of the enhancer to drive reporter gene expression compared to the WT enhancer sequence, whereas the downstream chr7:116,200,589T>A and chr7:116,200,705C>T SNPs significantly reduced reporter activity (Figure 6.18). Taken together, these finding suggest that different SNPs may affect *CAV1* expression via distinct mechanisms. The chr7:116,199,522T>C SNP may influence both enhancer DNA and its transcribed eRNA, illustrating how SNPs can differential impact DNA and RNA components of an enhancer.

#### 8.3.1.1 Impacts of ALS-associated SNPs on eRNA structure and function

Despite the predicted increased stability conferred by the chr7:116,199,522T>C mutation to the antisense eRNA at chr7:116,199,437-116,199,680, with all tested parameters indicating greater stability (Figure 5.21), understanding the precise functional impact of this structural change remains complex. One key consideration is the increase in the stem loop size following mutation. Stem loop structures in RNA can provide a kinetic barrier to unfolding (Rissone et al., 2022), thus larger stem loops may enhance the stability of the overall secondary RNA structure. However, while larger stem loops may contribute to greater structural integrity for the eRNA as a whole, the local structure of the enlarged stem loop itself may be less stable due to reduced intraloop interactions (Groebe and Uhlenbeck, 1988, Kuznetsov et al., 2008). As stem loop size increases, unpaired nucleotides within the loop structure have fewer opportunities to form stabilising base stacking and hydrogen bonding interactions, resulting in a more flexible, less

rigid structure (Fulle and Gohlke, 2008). This greater conformational freedom allows larger stem loops to explore more structural states within the free energy landscape, which can increase folding times relative to smaller structures (Kuznetsov et al., 2008). Thus, while larger stem loops may stabilise the overall RNA structure, locally they are thermodynamically less stable due to their inherent flexibility and ability to fluctuate between multiple structural conformations. In order to test the potential impacts of ALS-associated SNPs on eRNA stability, techniques such as road-block qPCR (Watson et al., 2020) or thiol(SH)-linked alkylation for the metabolic sequencing of RNA (SLAM-seq) (Herzog et al., 2017) could be employed to provide an experimental measure of how ALS-associated SNPs affect eRNA half-life.

Due to the predicted dichotomous effects of the chr7:116,199,522T>C ALS-associated SNP on eRNA structure — differentially affecting overall and local RNA stability — we must consider whether the overall half-life of the eRNA or the stability of local structures is more critical to eRNA functionality. For instance, if the stem loop structure is necessary for structure-specific protein interactions (reviewed in Svoboda and Di Cara, 2006), the mutation could hinder these interactions. Conversely, for sequence-independent functions of eRNAs, such as CBP binding (Bose et al., 2017), the increased overall stability could enhance eRNA functionality by improving resistance to degradation. While we have not explored the sequence-specific impacts of SNPs on eRNA function, eRNAs largely bind to epigenetic modulators in a locus-specific manner, as opposed to a sequence-specific manner, as observed for the key epigenetic modulator enyzmes CBP and BRD4 (Bose et al., 2017, Carullo et al., 2020, Rahnamoun et al., 2017).

### 8.3.2 DNA sequence-dependent impacts of ALS-associated SNPs on enhancer activity

The predicted disruption of MEF2 TF family binding site motifs by the chr7:116,199,522T>C SNP (Figure 6.19), alongside the significant impact of the chr7:116,200,589T>A and chr7:116,200,705C>T SNPs on the capacity of the enhancer to drive reporter gene expression (Figure 6.18), provide compelling evidence for DNA sequence-dependent mechanisms that underlie the regulatory potential of the proximal model enhancer. The proximal model *CAV1/CAV2* enhancer may be more prone to DNA sequence-dependent changes in activity, due to the inherent evolutionary conservation of 3' UTR sequences (Siepel et al., 2005, Xie et al., 2005). Disease-associated variants are enriched in evolutionarily conserved putative enhancer elements (Hujoel et al., 2019). Sequences under strong negative selection are more likely to harbour causal genetic variants, potentially due to the critical component of sequence in determining their function. Consequently,

variants within conserved regions are more likely to exert a phenotypic effect, paralleling the impact observed for mutations in conserved protein-coding regions (Reva et al., 2011).

Beyond generalised conservation of 3' UTRs, AREs also demonstrate evolutionary conservation (reviewed in Chen and Shyu, 1995). We observed that the ALS-associated SNP chr7:116,199,522T>C overlaps with, and potentially disrupts a WWTTTWW motif (Figure 7.3), further suggesting functionality for the DNA sequence at this nucleotide position and linking the disruption of evolutionarily conserved sequences to ALS pathogenesis.

#### 8.3.2.1 Assessing DNA-dependent enhancer functionality by luciferase reporter assay

While the transient transfection of luciferase reporter constructs serves as an effective approach to directly link DNA sequence to function, we cannot overlook the potential experimental artefacts intrinsic to this experimental system. Transient transfection of reporter constructs results in departure from the native genomic context of the enhancer DNA sequences, potentially eliminating key cellular features, including epigenomic characteristics, such as histone modifications, and features of the native chromatin architecture, such as long-range *cis*-regulatory looping interactions (Cooper et al., 2006, Jeong and Stein, 1994, Rojas-Fernandez et al., 2015). Plasmids undergo differential chromatinisation upon transient transfection compared to endogenous chromatin, consequently, they do not replicate the full complexity of native chromatin structure (Jeong and Stein, 1994). Nonetheless, reporter assays utilising transient transfection often yield cell type-specific results (Cooper et al., 2006, Landolin et al., 2010).

With luminescence intensity the only output of luciferase reporter assays, beyond the knowledge that output is linked to DNA sequence-dependent functionality, it is not possible to identify the mechanisms underlying observed changes in luminescence. In addition, due to the departure from the native genomic context of the enhancer, alterations in reporter activity may not reflect the true impact of SNPs in the native cellular environment. For example, ALS-associated SNPs may alter enhancer activity through impeded enhancer-promoter compatibility and chromatin organisation. Given the reporter system did not employ the endogenous promoter, observed reduced enhancer activity due to loss of enhancer-promoter compatibility between the *CAV1/CAV2* enhancer and the SV40 promoter may not reflect regulatory interactions in the *in vitro* cellular environment. To better recapitulate the native genomic environment of the

*CAV1/CAV2* enhancer, we could utilise a reporter system that includes the endogenous promoter, with the *CAV1/CAV2* enhancer downstream of the reporter gene. This would better model the enhancer-promoter interactions as they exist in the endogenous *CAV1* locus. However, with activity no longer driven by the highly active SV40 promoter (reviewed in Makrides, 2003), capable of measuring the subtle regulatory effects of mutations, this may limit the identification of regulatory alterations due to SNPs.

Reporter assays may most strongly reflect TF binding and RNAPII recruitment to the enhancer, in fact, luciferase reporter assays have been utilised to assess the impact of TF binding motif disruption on promoter activity (Whitfield et al., 2012). However, chromatin environment strongly influences both TF binding and RNAPII recruitment (reviewed in Srivastava and Mahony, 2020), so we are hesitant to draw conclusions about the effect of SNPs on TF binding and RNAPII recruitment at the enhancer based on reporter assay results alone. Despite the inherent limitations of reporter assays for evaluating the regulatory interactions at occur *in vivo*, they remain a valuable and efficient screening method for understanding the regulatory capacity of DNA sequences and thus, the impact of SNPs on intrinsic regulatory function. As such, although computational predictions did not reveal any effects of the chr7:116,200,589T>A and chr7:116,200,705C>T SNPs on TF binding or eRNA structure, reporter assays have provided rationale to further investigate the potential impact of these SNPs on sequence-dependent enhancer function.

### 8.3.3 Mutations in the model distal enhancer

#### 8.3.3.1 ALS-associated mutations may result in a compensatory upregulation in *CAV1* expression

Despite the predicted inhibitory impact of ALS-associated SNPs on *CAV1* expression (Cooper-Knock et al., 2021), in ALS patient-derived lymphoblastoid cell lines, the ALS-associated SNPs chr7:116,222,625T>C and chr7:115,994,269C>T appeared to upregulate *CAV1* expression, as measured by RT-qPCR (Figure 6.10). When investigating nascent transcription from the mutated enhancers in patient-derived lymphoblastoid cells, we observed a ∼2-fold decrease in the average transcript coverage over the candidate enhancers containing the chr7:116,222,625T>C in the SP3353 cell line, compared to healthy control-derived cells (Figure 5.25). Similarly, presence of the chr7:115,994,269C>T in the SP3388 cell line appeared to completely eliminate the nascent transcription observed over its corresponding enhancer in the healthy control-derived lymphoblastoid cell line (Figure 5.25). When considering the candidate *CAV1/*

*CAV2* enhancer landscape as a whole, we observed a uniform increase in nascent transcription from *CAV1/CAV2* candidate enhancers in ALS patient-derived lymphoblastoid cell lines, compared to healthy control-derived cells (Figure 5.23).

The transcriptional upregulation at enhancers, indicative of increased enhancer activity, across the the remainder of the enhancer landscape, despite the relative transcriptional downregulation at SNP-containing enhancers, may suggest a compensatory mechanism at play. Enhancer redundancy may indeed be coupled with compensatory mechanisms that counteract the loss of function of individual elements (Zhang et al., 2023). Further, multipartite enhancer ensembles controlling *CAV1* expression may provide means to fine-tune gene dosage. Beyond additive enhancer activity, which would likely lead to gene downregulation upon enhancer loss, enhancers can exhibit sub-additive activity, serving to buffer against gene overexpression. In such cases, the loss of an individual enhancer would likely result in gene upregulation (Bothma et al., 2015).

In support of the upregulation of enhancer activity across the enhancer landscape following ALS-associated *CAV1* enhancer mutation, a mechanism that is potentially compensatory in nature, we observed a trend of predicted increased eRNA structural stability for eRNA transcripts assembled from TT-seq data in ALS patient-derived cells, compared to healthy control cells (Figure 5.28, Figure 5.32, Figure 5.33, Figure 5.36). This enhanced stability may offer a mechanism for the observed increase in enhancer activity, as more stable eRNAs may be able to drive eRNA-mediated enhancer activation prior to degradation by the RNA exosome. eRNAs can modulate enhancer activity indirectly - through creation of a chromatin environment conducive to transcription activation - or directly - by promoting RNAPII pause release and TF binding. Generally, eRNA-mediated control of enhancer activity operates through a positive-feedback loop, in which permissive chromatin state and TF binding at the enhancer drive enhancer activity, and thus eRNA production, and subsequently the eRNA transcripts themselves reinforce activating chromatin modifications and TF occupancy at the enhancer (reviewed in Harrison and Bose, 2022).

We must acknowledge that the proposed compensatory mechanism observed in lymphoblastoid cells may be cell type-specific. However, we also observe evidence of enhancer cooperativity in SH-SY5Y cells. Upon knockout of the model distal *CAV1/CAV2* enhancer element, we observed a significant decrease in expression of the sense eRNA transcribed from the proximal enhancer, compared to WT SH-SY5Y cells (Figure 6.13). In contrast, the opposite trend was observed in lymphoblastoid cells, where cells harbouring the chr7:116,222,625T>C ALS-associated SNP in the distal enhancer exhibited increase sense and antisense eRNA transcription from the proximal enhancer, compared

to healthy control cells (Figure 6.15). While this may be an artefact of differential enhancer activity between distinct cell types, it could also underscore the dependence of enhancer cooperativity on 3D chromatin interactions between enhancer elements.

Physical contact between synergistic enhancers contributes to enhancer-enhancer cooperative regulation. When enhancers loop to the same protein-coding gene simultaneously, they can come into contact, allowing their synergistic activation by TFs, so in turn, the enhancers can cooperatively activate target gene expression (Chen et al., 2019). The single base change conferred by the ALS-associated SNP in lymphoblastoid cells, while diminishing the activity of the mutated enhancer, may still allow enhancer-enhancer contacts, enabling the compensatory actions of other enhancers to drive, and even upregulate, gene expression. Conversely, the mutation that knocks out the distal enhancer in SH-SY5Y cells may disrupt enhancer-enhancer interactions, similar to the observed interruption of enhancer-promoter 3D contacts by disease-associated SNPs (Sanchez-Mut et al., 2018, Visser et al., 2012). As a result, the loss of synergistic enhancer activity may prevent compensatory upregulation of *CAV1* gene expression, as posited for lymphoblastoid cells. Introduction of the SNPs present in ALS patient-derived lymphoblastoid cells into SH-SY5Y cells would clarify whether the differences in proximal enhancer activity following either distal enhancer mutation or knockout are due to distinct cell type-specific enhancer-enhancer interactions or different experimental systems.

#### 8.3.3.2 SNPs can impact *CAV1* isoform usage

While we observe alterations in *CAV1* mRNA isoform usage following both distal *CAV1/CAV2* enhancer knockdown and mutation (Figure 6.9, Figure 6.10), RNA expression alone cannot confirm whether these changes have physiological consequences downstream. While RT-qPCR allows us to measure *CAV1α* isoform expression, we can only infer the *CAV1α* to *CAV1β* isoform ratio based on total *CAV1* expression and *CAV1α* levels. In order to accurately assess differential mRNA isoform expression, performing total RNA-seq followed by transcript assembly or utilising long-read sequencing would provide a more robust approach to definitively determine whether the distal enhancer influences *CAV1* isoform expression.

Additionally, CAV1 protein isoform expression may be regulated not only at the transcriptional level but also through differential translation initiation. Thus, even if we observe increased *CAV1α* mRNA expression, this does not guarantee exclusive translation into CAV1$\alpha$, as CAV1$\beta$ could still be produced via downstream translational initiation. To address this, a western blot probing for phosphorylated CAV1 could determine whether increased *CAV1α* mRNA expression correlates with changes in CAV1

phosphorylation levels, which could disrupt the spatial organisation of CAV1 monomers within the caveolin complex. This would help clarify whether the ALS-associated mutation of the distal enhancer has downstream functional impacts on CAV1.

Furthermore, we have proposed a potential CTCF-dependent mechanism by which the distal *CAV1/CAV2* enhancer may influence differential *CAV1* mRNA isoform expression (Figure 6.20). To test this, we could employ immunoprecipitation-based techniques to examine the effects of enhancer knockout or mutation on CTCF binding at the enhancer, providing further insight into the regulatory role of CTCF in *CAV1* isoform expression.

### 8.3.4 The impact of intra-enhancer location of SNP impact

The relative position of genetic variants within enhancers may have influence over their impact on gene expression, with mutations located centrally within enhancers more likely to disrupt enhancer activity, often through alterations in TF binding (Yang et al., 2022). This is of relevance to our work, with the ALS-associated SNP chr7:116,199,522T>C located the most centrally in both the proximal tissue-agnostic *CAV1/CAV2* (chr7:116,196,622-116,201,351) and the proximal CNS-specific *CAV1* enhancer (chr7:116,199,565-116,200,132). The central positioning of this SNP could have power over its impacts on TF binding, specifically disrupting a predicted MEF2 TF family binding site (Figure 6.19). Similarly, the chr7:116,222,625T>C SNP lies within 2 bp of the centre of the distal tissue-agnostic *CAV1/CAV2* enhancer (chr7:116,220,523-116,224,531), which may likewise affect its influence on FOXO1-ETS binding sites (Figure 6.21).

Taken together, the parallel investigation of both the proximal and distal *CAV1/CAV2* enhancers provide evidence of ALS-associated SNPs that both up- and downregulate *CAV1* expression. This demonstrates that the fine balance of *CAV1* expression within the CNS may be implicit in ALS progression.

## 8.4 Future perspectives

### 8.4.1 Established techniques for further enhancer interrogation

While we have characterised the entire *CAV1/CAV2* landscape, at the level of epigenetic modification and nascent transcription, in this work we have only functionally characterised two model enhancers. Production of the Zim3-dCas9 SH-SY5Y cell line has provided us with a platform to systematically test any number of candidate enhancers and their impact of *CAV1* gene expression. Beyond linking enhancers to *CAV1*

expression, through establishment of the luciferase reporter assay, we now have means to test the impact of ALS-associated SNPs on enhancer regulatory activity at the level of the DNA sequence.

Regarding the genetic uncoupling of the proximal *CAV1/CAV2* enhancer from the *CAV1* 3' UTR, we have demonstrated successful insertion of the SV40 PAS between the *CAV1* CDS and 3' UTR in HEK293FT cells, and thus providing an experimental method for generating the equivalent genetic modification in SH-SY5Y cells. Further introduction of ALS-associated SNPs into this SH-SY5Y cell line would allow for the assessment of the impact of ALS-associated SNPs on *CAV1* expression, independent of the potential impact of SNPs on 3' UTR regulatory activity. Moreover, performing TT-seq on these modified cell lines would allow for validation of the nascent transcripts arising from the 3' UTR as bona fide eRNAs, enabling accurate profiling of enhancer activity for the proximal *CAV1/CAV2* enhancer.

### 8.4.2 Complexity of *CAV1/CAV2* enhancer regulatory mechanisms

To further elucidate the regulatory mechanisms controlling *CAV1* expression, we further aim to identify the TFs and co-activators binding at *CAV1/CAV2*. Beyond implied CBP activity at the proximal *CAV1/CAV2* model enhancer, based on H3K27ac enrichment and reduced eRNA expression after CBP inhibition (Figure 6.16, Figure 6.17b), CBP CUT&RUN or ChIP-seq would allow us to confirm binding of CBP at this region. Given the observed effects of CBP inhibition of eRNA transcription from this enhancer, combining CBP inhibition using A-485 with ChIP-qPCR for H3K27ac at this locus would allow for validation that CBP is indeed depositing activating histone modifications at this region. Further, pairing CBP inhibition with TT-seq could enable the capture of transient eRNAs, which may be missed by RT-qPCR, further confirming the role of CBP in eRNA transcription at this locus. This would allow us to test our hypothesis that CBP inhibition disrupts a positive feedback loop at the proximal *CAV1/CAV2* enhancer, wherein loss of CBP enzymatic activity prevents the deposition of activating chromatin modifications at this enhancer, resulting in a reduction in both eRNA and *CAV1* mRNA transcription.

In the context of disease, with ALS-associated SNPs predicted to alter TF binding site motifs, genetically modified cell lines containing these SNPs would enable the experimental assessment of these predicted TF binding disruptions. Immunoprecipitation-based techniques, such as ChIP-qPCR, could then be employed to determine whether the SNPs impair TF binding *in vitro*, providing insight into how these genetic variants might influence enhancer function in ALS.

## 8.5 Concluding remarks

With *CAV1* established as a risk gene underpinning ALS, we have highlighted the complexities governing ALS aetiology in patients harbouring *CAV1* enhancer mutations, with ALS-associated SNPs in *CAV1* enhancers differentially impacting *CAV1* expression. We have explored the *CAV1/CAV2* enhancer landscape in neuronally-relevant cell lines, both at the level of chromatin architecture and the level of enhancer activity, as determined by transcriptomic and epigenomic analyses. Our parallel investigation of proximal and distal enhancer elements emphasises the intricate nature of *CAV1* enhancer-mediated regulatory processes, revealing distinct differences in their activity and the regulation of *CAV1*, both basally, and in the context of ALS-associated mutation. Finally, with limited published research on the prevalence of enhancer elements within 3' UTRs, we provide insight into a key mechanism that may contribute to the unique transcriptional profiles of 3' UTRs.

# Appendix A

# Molecular Cloning

## A.1 Proximal *CAV1/CAV2* enhancer G-block

For the generation of luciferase reporter constructs and genetically modified cell lines, a GeneArt G-block vector (ThermoFisher) containing the sequence of the proximal *CAV1/CAV2* was utilised. The sequence of CAV1_enhancer_WT is as follows:

CTAAATTGTAAGCGTTAATATTTTGTTAAAATTCGCGTTAAATTTTTGTTA
AATCAGCTCATTTTTTAACCAATAGGCCGAAATCGGCAAAATCCCTTATAA
ATCAAAAGAATAGACCGAGATAGGGTTGAGTGGCCGCTACAGGGCGCTCC
CATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGTTTCGGTGCGGG
CCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGA
TTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACG
GCCAGTGAGCGCGACGTAATACGACTCACTATAGGGCGAATTGAAGGAAG
GCCGTCAAGGCCGCATCATCATCATGACGCGTAGTGAGCCATGATCATGCC
ACTGCACTCCAGCCTGGGTGACAGAGCAAGACCCTGTCTCAAAAAAATAAA
AATGAAAAAAATCTGTGTTCCCAAGTTCCAAGTGATGCTGATGCTGCTGGT
TGCCTTTAAGCATCTCACAAAGAACGAACTCATAAATGCTAATACAGTATA
TGTCTATGGATACTGAATAGTGGGTTTTTTTTCTCTTTTCTTCTATTCTGT
GCTCATGTTGTGTCACTTCTTCCTTTTAGATTGACTTTGAAGATGTGATTG
CAGAACCAGAAGGGACACACAGTTTTGACGGCATTTGGAAGGCCAGCTTC
ACCACCTTCACTGTGACGAAATACTGGTTTTACCGCTTGCTGTCTGCCCTC
TTTGGCATCCCGATGGCACTCATCTGGGGCATTTACTTCGCCATTCTCTCT
TTCCTGCACATCTGGGCAGTTGTACCATGCATTAAGAGCTTCCTGATTGAG
ATTCAGTGCATCAGCCGTGTCTATTCCATCTACGTCCACACCGTCTGTGAC
CCACTCTTTGAAGCTGTTGGGAAAATATTCAGCAATGTCCGCATCAACTTG

CAGAAAGAAATATAAATGACATTTCAAGGATAGAAGTATACCTGATTTTTT
TTCCTTTTAATTTTCCTGGTGCCAATTTCAAGTTCCAAGTTGCTAATACAG
CAACAATTTATGAATTGAATTATCTTGGTTGAAAATAAAAAGATCACTTTC
TCAGTTTTCATAAGTATTATGTCTCTTCTGAGCTATTTCATCTATTTTTGG
CAGTCTGAATTTTTAAAACCCATTTAAATTTTTTTCCTTACCTTTTTATTTG
CATGTGGATCAACCATCGCTTTATTGGCTGAGATATGAACATATTGTTGAA
AGGTAATTTGAGAGAAATATGAAGAACTGAGGAGGAAAAAAAAAAAAAAG
AAAAGAACCAACAACCTCAACTGCCTACTCCAAAATGTTGGTCATTTTATG
TTAAGGGAAGAATTCCAGGGTATGGCCATGGAGTGTACAAGTATGTGGGC
AGATTTTCAGCAAACTCTTTTCCCACTGTTTAAGGAGTTAGTGGATTACTG
CCATTCACTTCATAATCCAGTAGGATCCAGTGATCCTTACAAGTTAGAAAA
CATAATCTTCTGCCTTCTCATGATCCAACTAATGCCTTACTCTTCTTGAAA
TTTTAACCTATGATATTTTCTGTGCCTGAATATTTGTTATGTAGATAACAA
GACCTCAGTGCCTTCCTGTTTTTCACATTTTCCTTTTCAAATAGGGTCTAA
CTCAGCAACTCGCTTTAGGTCAGCAGCCTCCCTGAAGACCAAAATTAGAAT
ATCCATGACCTAGTTTTCCATGCGTGTTTCTGACTCTGAGCTACAGAGTCT
GGTGAAGCTCACTTCTGGGCTTCATCTGGCAACATCTTTATCCGTAGTGGG
TATGGTTGACACTAGCCCAATGAAATGAATTAAAGTGGACCAATAGGGCT
GAGCTCTCTGTGGGCTGGCAGTCCTGGAAGCCAGCTTTCCCTGCCTCTCAT
CAACTGAATGAGGTCAGCATGTCTATTCAGCTTCGTTTATTTTCAAGAATA
ATCACGCTTTCCTGAATCCAAACTAATCCATCACCGGGGTGGTTTAGTGGC
TCAACATTGTGTTCCCATTTCAGCTGATCAGTGGGCCTCCAAGGAGGGGCT
GTAAAATGGAGGCCATTGTGTGAGCCTATCAGAGTTGCTGCAAACCTGACC
CCTGCTCAGTAAAGCACTTGCAACCGTCTGTTATGCTGTGACACATGGCCC
CTCCCCCTGCCAGGAGCTTTGGACCTAATCCAAGCATCCCTTTGCCCAGAA
AGAAGATGGGGGAGGAGGCAGTAATAAAAAGATTGAAGTATTTTGCTGGA
ATAAGTTCAAATTCTTCTGAACTCAAACTGAGGAATTTCACCTGTAAACCT
GAGTCGTACAGAAAGCTGCCTGGTATATCCAAAAGCTTTTTATTCCTCCTG
CTCATATTGTGATTCTGCCTTTGGGGACTTTTCTTAAACCTTCAGTTATGA
TTTTTTTTTCATACACTTATTGGAACTCTGCTTGATTTTTGCCTCTTCCAG
TCTTCCTGACACTTTAATTACCAACCTGTTACCTACTTTGACTTTTTGCAT
TTAAAACAGACACTGGCATGGATATAGTTTTACTTTTAAACTGTGTACATA
ACTGAAAATGTGCTATACTGCATACTTTTTAAATGTAAAGATATTTTTATC
TTTATATGAAGAAAATCACTTAGGAAATGGCTTTGTGATTCAATCTGTAAA
CTGTGTATTCCAAGACATGTCTGTTCTACATAGATGCTTAGTCCCTCATGC
AAATCAATTACTGGTCCAAAAGATTGCTGAAATTTTATATGCTTACTGATA
TATTTTACAATTTTTTATCATGCATGTCCTGTAAAGGTTACAAGCCTGCAC
AATAAAAATGTTTAACGGTTAAACAGTCAGCTTTATTATTTTTTCCCAAAA
CAGGTGTTTATGTGTCAGAGTCTGTGTATGTCTATGTATTTGTATGTAATG

AGCATGTGCATCTCGAGCCATCATCATCTGGGCCTCATGGGCCTTCCTTTC
ACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAACATGG
TCATAGCTGTTTCCTTGCGTATTGGGCGCTCTCCGCTTCCTCGCTCACTGA
CTCGCTGCGCTCGGTCGTTCGGGTAAAGCCTGGGGTGCCTAATGAGCAAA
AGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTT
TCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTC
AGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTG
GAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACC
TGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCT
GTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGC
ACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTC
TTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTG
GTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTG
AAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGC
GCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCC
GGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAG
ATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACG
GGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCAT
GAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAG
TTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTATTA
GAAAAATTCATCCAGCAGACGATAAAACGCAATACGCTGGCTATCCGGTGC
CGCAATGCCATACAGCACCAGAAAACGATCCGCCCATTCGCCGCCCAGTTC
TTCCGCAATATCACGGGTGGCCAGCGCAATATCCTGATAACGATCCGCCAC
GCCCAGACGGCCGCAATCAATAAAGCCGCTAAAACGGCCATTTTCCACCAT
AATGTTCGGCAGGCACGCATCACCATGGGTCACCACCAGATCTTCGCCATC
CGGCATGCTCGCTTTCAGACGCGCAAACAGCTCTGCCGGTGCCAGGCCCTG
ATGTTCTTCATCCAGATCATCCTGATCCACCAGGCCCGCTTCCATACGGGT
ACGCGCACGTTCAATACGATGTTTCGCCTGATGATCAAACGGACAGGTCGC
CGGGTCCAGGGTATGCAGACGACGCATGGCATCCGCCATAATGCTCACTTT
TTCTGCCGGCGCCAGATGGCTAGACAGCAGATCCTGACCCGGCACTTCGCC
CAGCAGCAGCCAATCACGGCCCGCTTCGGTCACCACATCCAGCACCGCCGC
ACACGGAACACCGGTGGTGGCCAGCCAGCTCAGACGCGCCGCTTCATCCTG
CAGCTCGTTCAGCGCACCGCTCAGATCGGTTTTCACAAACAGCACCGGACG
ACCCTGCGCGCTCAGACGAAACACCGCCGCATCAGAGCAGCCAATGGTCTG
CTGCGCCCAATCATAGCCAAACAGACGTTCCACCCACGCTGCCGGGCTACC
CGCATGCAGGCCATCCTGTTCAATCATACTCTTCCTTTTTCAATATTATTG
AAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTAT
TTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCC
AC

## A.2 SV40 polyadenylation signal G-block

For insertion of a premature PAS into *CAV1* in SH-SY5Y cells, a gene fragment (IDT gBlock) containing the SV40 PAS was designed flanked by complementarity to *CAV1* was designed. The sequence of SV40_gBlock_RD is as follows:

CATCATACGCGTAGTGAGCCATGATCATGCCACTGCACTCCAGCCTGGGTG
ACAGAGCAAGACCCTGTCTCAAAAAAATAAAAATGAAAAAAATCTGTGTTC
CCAAGTTCCAAGTGATGCTGATGCTGCTGGTTGCCTTTAAGCATCTCACAA
AGAACGAACTCATAAATGCTAATACAGTATATGTCTATGGATACTGAATAG
TGGGTTTTTTTTCTCTTTTCTTCTATTCTGTGCTCATGTTGTGTCACTTCT
TCCTTTTAGATTGACTTTGAAGATGTGATTGCAGAACCAGAAGGGACACAC
AGTTTTGACGGCATTTGGAAGGCCAGCTTCACCACCTTCACTGTGACGAAA
TACTGGTTTTACCGCTTGCTGTCTGCCCTCTTTGGCATCCCGATGGCACTC
ATCTGGGGCATTTACTTCGCCATTCTCTCTTTCCTGCACATCTGGGCAGTT
GTACCATGCATTAAGAGCTTCCTGATTGAGATTCAGTGCATCAGCCGTGTC
TATTCCATCTACGTCCACACCGTCTGTGACCCACTCTTTGAAGCTGTTGGG
AAAATATTCAGCAATGTCCGCATCAACTTGCAGAAAGAAATATAAGTCGAC
TTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCATTC
ATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAG
TAAAACCCTCGAGTTCTATAGTGTCATGACATTTCAAGGATAGAAGTATAC
CTGATTTTTTTTCCTTTTAATTTTCATGGTGCCAATTTCAAGTTCCAAGTT
GCTAATACAGCAACAATTTATGAATTGAATTATCTTGGTTGAAAATAAAAA
GATCACTTTCTCAGTTTTCATAAGTATTATGTCTCTTCTGAGCTATTTCAT
CTATTTTTGGCAGTCTGAATTTTTAAAACCCATTTAAATTTTTTTCCTTAC
CTTTTTATTTGCATGTGGATCAACCATCGCTTTATTGGCTGAGATATGAAC
ATATTGTTGAAAGGTAATTTGAGAGAAATATGAAGAACTGAGGAGGAAAA
AAAAAAAAAAGAAAAGAACCAACAACCTCAACTGCCTACTCCAAAATGTTG
GTCATTTTATGTTAAGGGAAGAATTCCATCAT

## A.3 Donor template for generation of genetically modified cell lines

The sequence of the CAV1_SV40_gRNA4 HDR template, generated by insertion of SV40_gBlock_RD into CAV1_enhancer_WT is as follows:

TTAGAAAAATTCATCCAGCAGACGATAAAACGCAATACGCTGGCTATCCGG
TGCCGCAATGCCATACAGCACCAGAAAACGATCCGCCCATTCGCCGCCCAG
TTCTTCCGCAATATCACGGGTGGCCAGCGCAATATCCTGATAACGATCCGC
CACGCCCAGACGGCCGCAATCAATAAAGCCGCTAAAACGGCCATTTTCCAC
CATAATGTTCGGCAGGCACGCATCACCATGGGTCACCACCAGATCTTCGCC
ATCCGGCATGCTCGCTTTCAGACGCGCAAACAGCTCTGCCGGTGCCAGGCC
CTGATGTTCTTCATCCAGATCATCCTGATCCACCAGGCCCGCTTCCATACG
GGTACGCGCACGTTCAATACGATGTTTCGCCTGATGATCAAACGGACAGGT
CGCCGGGTCCAGGGTATGCAGACGACGCATGGCATCCGCCATAATGCTCAC
TTTTTCTGCCGGCGCCAGATGGCTAGACAGCAGATCCTGACCCGGCACTTC
GCCCAGCAGCAGCCAATCACGGCCCGCTTCGGTCACCACATCCAGCACCGC
CGCACACGGAACACCGGTGGTGGCCAGCCAGCTCAGACGCGCCGCTTCATC
CTGCAGCTCGTTCAGCGCACCGCTCAGATCGGTTTTCACAAACAGCACCGG
ACGACCCTGCGCGCTCAGACGAAACACCGCCGCATCAGAGCAGCCAATGGT
CTGCTGCGCCCAATCATAGCCAAACAGACGTTCCACCCACGCTGCCGGGCT
ACCCGCATGCAGGCCATCCTGTTCAATCATACTCTTCCTTTTTCAATATTA
TTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATG
TATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGT
GCCACCTAAATTGTAAGCGTTAATATTTTGTTAAAATTCGCGTTAAATTTT
TGTTAAATCAGCTCATTTTTTAACCAATAGGCCGAAATCGGCAAAATCCCT
TATAAATCAAAAGAATAGACCGAGATAGGGTTGAGTGGCCGCTACAGGGC
GCTCCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGTTTCGGTG
CGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAG
GCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAAC
GACGGCCAGTGAGCGCGACGTAATACGACTCACTATAGGGCGAATTGAAG
GAAGGCCGTCAAGGCCGCATCATCATCATGACGCGTAGTGAGCCATGATCA
TGCCACTGCACTCCAGCCTGGGTGACAGAGCAAGACCCTGTCTCAAAAAAA
TAAAAATGAAAAAAATCTGTGTTCCCAAGTTCCAAGTGATGCTGATGCTGC
TGGTTGCCTTTAAGCATCTCACAAAGAACGAACTCATAAATGCTAATACAG
TATATGTCTATGGATACTGAATAGTGGGTTTTTTTTCTCTTTTCTTCTATT
CTGTGCTCATGTTGTGTCACTTCTTCCTTTTAGATTGACTTTGAAGATGTG
ATTGCAGAACCAGAAGGGACACACAGTTTTGACGGCATTTGGAAGGCCAG
CTTCACCACCTTCACTGTGACGAAATACTGGTTTTACCGCTTGCTGTCTGC
CCTCTTTGGCATCCCGATGGCACTCATCTGGGGCATTTACTTCGCCATTCT
CTCTTTCCTGCACATCTGGGCAGTTGTACCATGCATTAAGAGCTTCCTGAT
TGAGATTCAGTGCATCAGCCGTGTCTATTCCATCTACGTCCACACCGTCTG
TGACCCACTCTTTGAAGCTGTTGGGAAAATATTCAGCAATGTCCGCATCAA
CTTGCAGAAAGAAATATAAGTCGACTTGTAACCATTATAAGCTGCAATAAA
CAAGTTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAG

GTGTGGGAGGTTTTTTAAAGCAAGTAAAACCCTCGAGTTCTATAGTGTCAT
GACATTTCAAGGATAGAAGTATACCTGATTTTTTTTCCTTTTAATTTTCAT
GGTGCCAATTTCAAGTTCCAAGTTGCTAATACAGCAACAATTTATGAATTG
AATTATCTTGGTTGAAAATAAAAAGAATCACTTTCTCAGTTTTCATAAGTA
TTATGTCTCTTCTGAGCTATTTCATCTATTTTTGGCAGTCTGAATTTTTAA
AACCCATTTAAATTTTTTTCCTTACCTTTTTATTTGCATGTGGATCAACCA
TCGCTTTATTGGCTGAGATATGAACATATTGTTGAAAGGTAATTTGAGAGA
AATATGAAGAACTGAGGAGGAAAAAAAAAAAAAGAAAAGAACCAACAACC
TCAACTGCCTACTCCAAAATGTTGGTCATTTTATGTTAAGGGAAGAATTCC
AGGGTATGGCCATGGAGTGTACAAGTATGTGGGCAGATTTTCAGCAAACT
CTTTTCCCACTGTTTAAGGAGTTAGTGGATTACTGCCATTCACTTCATAAT
CCAGTAGGATCCAGTGATCCTTACAAGTTAGAAAACATAATCTTCTGCCTT
CTCATGATCCAACTAATGCCTTACTCTTCTTGAAATTTTAACCTATGATAT
TTTCTGTGCCTGAATATTTGTTATGTAGATAACAAGACCTCAGTGCCTTCC
TGTTTTTCACATTTTCCTTTTCAAATAGGGTCTAACTCAGCAACTCGCTTT
AGGTCAGCAGCCTCCCTGAAGACCAAAATTAGAATATCCATGACCTAGTTT
TCCATGCGTGTTTCTGACTCTGAGCTACAGAGTCTGGTGAAGCTCACTTCT
GGGCTTCATCTGGCAACATCTTTATCCGTAGTGGGTATGGTTGACACTAGC
CCAATGAAATGAATTAAAGTGGACCAATAGGGCTGAGCTCTCTGTGGGCT
GGCAGTCCTGGAAGCCAGCTTTCCCTGCCTCTCATCAACTGAATGAGGTCA
GCATGTCTATTCAGCTTCGTTTATTTTCAAGAATAATCACGCTTTCCTGAA
TCCAAACTAATCCATCACCGGGGTGGTTTAGTGGCTCAACATTGTGTTCCC
ATTTCAGCTGATCAGTGGGCCTCCAAGGAGGGGCTGTAAAATGGAGGCCA
TTGTGTGAGCCTATCAGAGTTGCTGCAAACCTGACCCCTGCTCAGTAAAGC
ACTTGCAACCGTCTGTTATGCTGTGACACATGGCCCCTCCCCCTGCCAGGA
GCTTTGGACCTAATCCAAGCATCCCTTTGCCCAGAAAGAAGATGGGGGAG
GAGGCAGTAATAAAAAGATTGAAGTATTTTGCTGGAATAAGTTCAAATTCT
TCTGAACTCAAACTGAGGAATTTCACCTGTAAACCTGAGTCGTACAGAAAG
CTGCCTGGTATATCCAAAAGCTTTTTATTCCTCCTGCTCATATTGTGATTC
TGCCTTTGGGGACTTTTCTTAAACCTTCAGTTATGATTTTTTTTTCATACA
CTTATTGGAACTCTGCTTGATTTTTGCCTCTTCCAGTCTTCCTGACACTTT
AATTACCAACCTGTTACCTACTTTGACTTTTTGCATTTAAAACAGACACTG
GCATGGATATAGTTTTACTTTTAAACTGTGTACATAACTGAAAATGTGCTA
TACTGCATACTTTTTAAATGTAAAGATATTTTTATCTTTATATGAAGAAAA
TCACTTAGGAAATGGCTTTGTGATTCAATCTGTAAACTGTGTATTCCAAGA
CATGTCTGTTCTACATAGATGCTTAGTCCCTCATGCAAATCAATTACTGGT
CCAAAAGATTGCTGAAATTTTATATGCTTACTGATATATTTTACAATTTTT
TATCATGCATGTCCTGTAAAGGTTACAAGCCTGCACAATAAAAATGTTTAA
CGGTTAAACAGTCAGCTTTATTATTTTTTCCCAAAACAGGTGTTTATGTGT

CAGAGTCTGTGTATGTCTATGTATTTGTATGTAATGAGCATGTGCATCTCG
AGCCATCATCATCTGGGCCTCATGGGCCTTCCTTTCACTGCCCGCTTTCCA
GTCGGGAAACCTGTCGTGCCAGCTGCATTAACATGGTCATAGCTGTTTCCT
TGCGTATTGGGCGCTCTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGT
CGTTCGGGTAAAGCCTGGGGTGCCTAATGAGCAAAAGGCCAGCAAAAGGC
CAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCC
CCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCC
GACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCG
CTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCC
TTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTC
GGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCA
GCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGT
AAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAG
AGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACT
ACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAG
TTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCG
CTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAA
AAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAG
TGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAG
GATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTA
AAGTATATATGAGTAAACTTGGTCTGACAGTTA

# Appendix B

# Bioinformatic analysis

## B.1 Candidate *CAV1* enhancers

Candidate *CAV1/CAV2* enhancers as identified in Cooper-Knock et al. (2021), demonstrating their cell-type specificity (Table B.1), and the number of ALS-associated SNPs within these enhancers (Table B.2).

Table B.1: Enhancer regions associated with *CAV1* and *CAV2* expression, and tissue associations

| Gene Target | Chromosome | Start | Finish | Tissue |
|---|---|---|---|---|
| CAV1 | chr7 | 114429291 | 114431269 | CNS |
| CAV1 | chr7 | 114458233 | 114460475 | CNS |
| CAV1 | chr7 | 114525449 | 114526553 | CNS |
| CAV1 | chr7 | 114529275 | 114530372 | CNS |
| CAV1 | chr7 | 114567780 | 114569045 | CNS |
| CAV1 | chr7 | 114569345 | 114571951 | CNS |
| CAV1 | chr7 | 114572709 | 114573331 | CNS |
| CAV1 | chr7 | 114573849 | 114576684 | CNS |
| CAV1 | chr7 | 114583410 | 114585966 | CNS |
| CAV1 | chr7 | 114627975 | 114630075 | CNS |
| CAV1 | chr7 | 114648866 | 114650794 | CNS |
| CAV1 | chr7 | 114680511 | 114680894 | CNS |
| CAV1 | chr7 | 114785263 | 114786500 | CNS |
| CAV1 | chr7 | 114870292 | 114872289 | CNS |
| CAV1 | chr7 | 114939946 | 114941012 | CNS |

Continued on next page

Table B.1 – continued from previous page

| Gene Target | Chromosome | Start | Finish | Tissue |
|---|---|---|---|---|
| CAV1 | chr7 | 114992058 | 114993494 | CNS |
| CAV1 | chr7 | 115186013 | 115186673 | CNS |
| CAV1 | chr7 | 115301123 | 115304100 | CNS |
| CAV1 | chr7 | 115304366 | 115304820 | CNS |
| CAV1 | chr7 | 115310813 | 115313335 | CNS |
| CAV1 | chr7 | 115317812 | 115318384 | CNS |
| CAV1 | chr7 | 115318462 | 115319029 | CNS |
| CAV1 | chr7 | 115670883 | 115671400 | CNS |
| CAV1 | chr7 | 115737105 | 115737568 | CNS |
| CAV1 | chr7 | 115816693 | 115817206 | CNS |
| CAV1 | chr7 | 115853719 | 115854235 | CNS |
| CAV1 | chr7 | 115867068 | 115868903 | CNS |
| CAV1 | chr7 | 115870180 | 115870885 | CNS |
| CAV1 | chr7 | 115871349 | 115873794 | CNS |
| CAV1:CAV2 | chr7 | 115898310 | 115902716 | Tissue-agnostic |
| CAV1 | chr7 | 115898864 | 115900695 | CNS |
| CAV1 | chr7 | 115910552 | 115913260 | CNS |
| CAV1 | chr7 | 115913518 | 115914249 | CNS |
| CAV2 | chr7 | 115954125 | 115959325 | Tissue-agnostic |
| CAV1 | chr7 | 115955578 | 115958367 | CNS |
| CAV1 | chr7 | 115967434 | 115969066 | CNS |
| CAV1:CAV2 | chr7 | 115992110 | 115999387 | Tissue-agnostic |
| CAV1 | chr7 | 115993415 | 115996411 | CNS |
| CAV1 | chr7 | 115996955 | 115998734 | CNS |
| CAV2 | chr7 | 116003024 | 116004444 | Tissue-agnostic |
| CAV1:CAV2 | chr7 | 116033656 | 116036708 | Tissue-agnostic |
| CAV1 | chr7 | 116034665 | 116036373 | CNS |
| CAV1 | chr7 | 116057349 | 116058637 | Tissue-agnostic |

Continued on next page

Table B.1 – continued from previous page

| Gene Target | Chromosome | Start | Finish | Tissue |
|---|---|---|---|---|
| CAV2 | chr7 | 116062851 | 116064462 | Tissue-agnostic |
| CAV2 | chr7 | 116065132 | 116068292 | Tissue-agnostic |
| CAV1 | chr7 | 116066587 | 116067242 | CNS |
| CAV2 | chr7 | 116071887 | 116076515 | Tissue-agnostic |
| CAV1 | chr7 | 116073579 | 116076380 | CNS |
| CAV2 | chr7 | 116078147 | 116088487 | Tissue-agnostic |
| CAV1 | chr7 | 116080916 | 116084930 | CNS |
| CAV1 | chr7 | 116094659 | 116096095 | CNS |
| CAV2 | chr7 | 116094877 | 116096147 | Tissue-agnostic |
| CAV1 | chr7 | 116103858 | 116107331 | CNS |
| CAV2 | chr7 | 116104521 | 116107038 | Tissue-agnostic |
| CAV1 | chr7 | 116118084 | 116119473 | CNS |
| CAV2 | chr7 | 116134797 | 116136789 | Tissue-agnostic |
| CAV1:CAV2 | chr7 | 116149154 | 116154310 | Tissue-agnostic |
| CAV1 | chr7 | 116149961 | 116153401 | CNS |
| CAV2 | chr7 | 116154567 | 116155823 | Tissue-agnostic |
| CAV1 | chr7 | 116170083 | 116171009 | CNS |
| CAV1 | chr7 | 116175367 | 116176059 | CNS |
| CAV2 | chr7 | 116176051 | 116176914 | Tissue-agnostic |
| CAV2 | chr7 | 116180927 | 116184847 | Tissue-agnostic |
| CAV1 | chr7 | 116181441 | 116182810 | CNS |
| CAV1 | chr7 | 116182954 | 116183855 | CNS |
| CAV1:CAV2 | chr7 | 116185994 | 116188883 | Tissue-agnostic |

Continued on next page

Table B.1 – continued from previous page

| Gene Target | Chromosome | Start | Finish | Tissue |
|---|---|---|---|---|
| CAV1 | chr7 | 116186806 | 116188332 | CNS |
| CAV1 | chr7 | 116195961 | 116196301 | CNS |
| CAV1:CAV2 | chr7 | 116196622 | 116201351 | Tissue-agnostic |
| CAV1 | chr7 | 116199565 | 116200132 | CNS |
| CAV1:CAV2 | chr7 | 116202174 | 116202561 | Tissue-agnostic |
| CAV1 | chr7 | 116206178 | 116207412 | CNS |
| CAV1 | chr7 | 116207998 | 116213360 | CNS |
| CAV1:CAV2 | chr7 | 116209195 | 116214933 | Tissue-agnostic |
| CAV1 | chr7 | 116213507 | 116219235 | CNS |
| CAV1:CAV2 | chr7 | 116215990 | 116219157 | Tissue-agnostic |
| CAV1:CAV2 | chr7 | 116220723 | 116224531 | Tissue-agnostic |
| CAV1 | chr7 | 116221274 | 116225513 | CNS |
| CAV1 | chr7 | 116227556 | 116229293 | CNS |
| CAV2 | chr7 | 116231455 | 116233866 | Tissue-agnostic |
| CAV2 | chr7 | 116261779 | 116262057 | Tissue-agnostic |
| CAV1 | chr7 | 116268838 | 116269440 | CNS |
| CAV1 | chr7 | 116273324 | 116274010 | CNS |
| CAV1 | chr7 | 116327276 | 116331188 | CNS |
| CAV1 | chr7 | 116331287 | 116334384 | CNS |
| CAV1 | chr7 | 116341997 | 116343027 | CNS |
| CAV1 | chr7 | 116345275 | 116345781 | CNS |
| CAV1 | chr7 | 116345958 | 116349052 | CNS |
| CAV1 | chr7 | 116353124 | 116354955 | CNS |
| CAV1 | chr7 | 116355221 | 116357861 | CNS |
| CAV1 | chr7 | 116411044 | 116413659 | CNS |
| CAV1 | chr7 | 116415730 | 116419494 | CNS |
| CAV1 | chr7 | 116421735 | 116423410 | CNS |
| CAV1 | chr7 | 116439854 | 116443746 | CNS |

Continued on next page

Table B.1 – continued from previous page

| Gene Target | Chromosome | Start | Finish | Tissue |
|---|---|---|---|---|
| CAV1 | chr7 | 116474804 | 116476617 | Tissue-agnostic |
| CAV1 | chr7 | 116502110 | 116504190 | CNS |
| CAV1 | chr7 | 116511658 | 116512938 | CNS |
| CAV1 | chr7 | 116551519 | 116552444 | CNS |
| CAV1 | chr7 | 116638294 | 116639676 | CNS |
| CAV1 | chr7 | 116645209 | 116646095 | CNS |
| CAV1 | chr7 | 116701473 | 116702466 | CNS |
| CAV1 | chr7 | 116738137 | 116738529 | CNS |
| CAV1 | chr7 | 116738883 | 116739229 | CNS |
| CAV1 | chr7 | 116764255 | 116764795 | CNS |
| CAV1 | chr7 | 116764919 | 116766070 | CNS |
| CAV1 | chr7 | 116771574 | 116774341 | CNS |
| CAV1 | chr7 | 116797868 | 116798377 | CNS |
| CAV1 | chr7 | 116869939 | 116871410 | CNS |
| CAV1 | chr7 | 116899318 | 116900524 | CNS |
| CAV1 | chr7 | 116908705 | 116911053 | CNS |
| CAV1 | chr7 | 116923096 | 116924277 | CNS |
| CAV1 | chr7 | 117222645 | 117223108 | CNS |
| CAV1 | chr7 | 117305812 | 117306347 | CNS |
| CAV1 | chr7 | 117468242 | 117468792 | CNS |

Table B.2: ALS-associated SNPs within candidate *CAV1/CAV2* enhancer regions

| **Chromosome** | **Start** | **End** | **Closest gene** | **Number of ALS-associated SNPs** |
|---|---|---|---|---|
| chr7 | 114429291 | 114431269 | *FOXP2* | 2 |
| chr7 | 114458233 | 114460475 | *MDFIC* | 0 |
| chr7 | 114525449 | 114526553 | *MDFIC* | 0 |
| chr7 | 114529275 | 114530372 | *MDFIC* | 1 |
| chr7 | 114567780 | 114569045 | *MDFIC* | 0 |
| chr7 | 114569345 | 114571951 | *MDFIC* | 1 |
| chr7 | 114572709 | 114573331 | *MDFIC* | 0 |
| chr7 | 114573849 | 114576684 | *MDFIC* | 3 |
| chr7 | 114583410 | 114585966 | *MDFIC* | 1 |

Table B.2 – continued from previous page

| **Chromosome** | **Start** | **End** | **Closest gene** | **Number of ALS-associated SNPs** |
|---|---|---|---|---|
| chr7 | 114627975 | 114630075 | *MDFIC* | 0 |
| chr7 | 114648866 | 114650794 | *MDFIC* | 1 |
| chr7 | 114680511 | 114680894 | *MDFIC* | 0 |
| chr7 | 114785263 | 114786500 | *MDFIC* | 0 |
| chr7 | 114870292 | 114872289 | *MDFIC* | 9 |
| chr7 | 114939946 | 114941012 | *MDFIC* | 4 |
| chr7 | 114992058 | 114993494 | *MDFIC* | 0 |
| chr7 | 115186013 | 115186673 | *TFEC* | 0 |
| chr7 | 115301123 | 115304100 | *TFEC* | 6 |
| chr7 | 115304366 | 115304820 | *TFEC* | 0 |
| chr7 | 115310813 | 115313335 | *TFEC* | 2 |
| chr7 | 115317812 | 115318384 | *TFEC* | 9 |
| chr7 | 115318462 | 115319029 | *TFEC* | 1 |
| chr7 | 115670883 | 115671400 | *TFEC* | 1 |
| chr7 | 115737105 | 115737568 | *TFEC* | 0 |
| chr7 | 115816693 | 115817206 | *TFEC* | 0 |
| chr7 | 115853719 | 115854235 | *TES* | 0 |
| chr7 | 115867068 | 115868903 | *TES* | 1 |
| chr7 | 115870180 | 115870885 | *TES* | 0 |
| chr7 | 115871349 | 115873794 | *TES* | 0 |
| chr7 | 115898310 | 115902716 | *TES* | 4 |
| chr7 | 115898864 | 115900695 | *TES* | 4 |
| chr7 | 115910552 | 115913260 | *TES* | 2 |
| chr7 | 115913518 | 115914249 | *TES* | 0 |
| chr7 | 115954125 | 115959325 | *TES* | 10 |
| chr7 | 115955578 | 115958367 | *TES* | 10 |
| chr7 | 115967434 | 115969066 | *TES* | 3 |
| chr7 | 115992110 | 115999387 | *TES* | 22 |
| chr7 | 115993415 | 115996411 | *TES* | 21 |
| chr7 | 115996955 | 115998734 | *TES* | 0 |
| chr7 | 116003024 | 116004444 | *TES* | 1 |
| chr7 | 116033656 | 116036708 | *CAV2* | 3 |
| chr7 | 116034665 | 116036373 | *CAV2* | 2 |

Table B.2 – continued from previous page

| **Chromosome** | **Start** | **End** | **Closest gene** | **Number of ALS-associated SNPs** |
|---|---|---|---|---|
| chr7 | 116057349 | 116058637 | *CAV2* | 0 |
| chr7 | 116062851 | 116064462 | *CAV2* | 2 |
| chr7 | 116065132 | 116068292 | *CAV2* | 0 |
| chr7 | 116066587 | 116067242 | *CAV2* | 0 |
| chr7 | 116071887 | 116076515 | *CAV2* | 0 |
| chr7 | 116073579 | 116076380 | *CAV2* | 0 |
| chr7 | 116078147 | 116088487 | *CAV2* | 1 |
| chr7 | 116080916 | 116084930 | *CAV2* | 0 |
| chr7 | 116094659 | 116096095 | *CAV2* | 0 |
| chr7 | 116094877 | 116096147 | *CAV2* | 0 |
| chr7 | 116103858 | 116107331 | *CAV2* | 0 |
| chr7 | 116104521 | 116107038 | *CAV2* | 0 |
| chr7 | 116118084 | 116119473 | *CAV2* | 0 |
| chr7 | 116134797 | 116136789 | *CAV2* | 0 |
| chr7 | 116149154 | 116154310 | *CAV2* | 3 |
| chr7 | 116149961 | 116153401 | *CAV2* | 3 |
| chr7 | 116154567 | 116155823 | *CAV2* | 0 |
| chr7 | 116170083 | 116171009 | *CAV1* | 0 |
| chr7 | 116175367 | 116176059 | *CAV1* | 4 |
| chr7 | 116176051 | 116176914 | *CAV1* | 0 |
| chr7 | 116180927 | 116184847 | *CAV1* | 6 |
| chr7 | 116181441 | 116182810 | *CAV1* | 4 |
| chr7 | 116182954 | 116183855 | *CAV1* | 0 |
| chr7 | 116185994 | 116188883 | *CAV1* | 0 |
| chr7 | 116186806 | 116188332 | *CAV1* | 0 |
| chr7 | 116195961 | 116196301 | *CAV1* | 0 |
| chr7 | 116196622 | 116201351 | *CAV1* | 6 |
| chr7 | 116199565 | 116200132 | *CAV1* | 0 |
| chr7 | 116202174 | 116202561 | *CAV1* | 0 |
| chr7 | 116206178 | 116207412 | *CAV1* | 0 |
| chr7 | 116207998 | 116213360 | *CAV1* | 8 |
| chr7 | 116209195 | 116214933 | *CAV1* | 10 |
| chr7 | 116213507 | 116219235 | *CAV1* | 4 |

Table B.2 – continued from previous page

| **Chromosome** | **Start** | **End** | **Closest gene** | **Number of ALS-associated SNPs** |
|---|---|---|---|---|
| chr7 | 116215990 | 116219157 | *CAV1* | 2 |
| chr7 | 116220723 | 116224531 | *CAV1* | 9 |
| chr7 | 116221274 | 116225513 | *CAV1* | 9 |
| chr7 | 116227556 | 116229293 | *CAV1* | 0 |
| chr7 | 116231455 | 116233866 | *CAV1* | 0 |
| chr7 | 116261779 | 116262057 | *MET* | 0 |
| chr7 | 116268838 | 116269440 | *MET* | 0 |
| chr7 | 116273324 | 116274010 | *MET* | 2 |
| chr7 | 116327276 | 116331188 | *MET* | 4 |
| chr7 | 116331287 | 116334384 | *MET* | 1 |
| chr7 | 116341997 | 116343027 | *MET* | 2 |
| chr7 | 116345275 | 116345781 | *MET* | 1 |
| chr7 | 116345958 | 116349052 | *MET* | 3 |
| chr7 | 116353124 | 116354955 | *MET* | 0 |
| chr7 | 116355221 | 116357861 | *MET* | 4 |
| chr7 | 116411044 | 116413659 | *MET* | 1 |
| chr7 | 116415730 | 116419494 | *MET* | 9 |
| chr7 | 116421735 | 116423410 | *MET* | 6 |
| chr7 | 116439854 | 116443746 | *MET* | 5 |
| chr7 | 116474804 | 116476617 | *CAPZA2* | 1 |
| chr7 | 116502110 | 116504190 | *CAPZA2* | 2 |
| chr7 | 116511658 | 116512938 | *CAPZA2* | 0 |
| chr7 | 116551519 | 116552444 | *CAPZA2* | 0 |
| chr7 | 116638294 | 116639676 | *ST7* | 0 |
| chr7 | 116645209 | 116646095 | *ST7* | 1 |
| chr7 | 116701473 | 116702466 | *ST7* | 0 |
| chr7 | 116738137 | 116738529 | *ST7* | 0 |
| chr7 | 116738883 | 116739229 | *ST7* | 0 |
| chr7 | 116764255 | 116764795 | *ST7* | 0 |
| chr7 | 116764919 | 116766070 | *ST7* | 0 |
| chr7 | 116771574 | 116774341 | *ST7* | 4 |
| chr7 | 116797868 | 116798377 | *ST7* | 1 |
| chr7 | 116869939 | 116871410 | *ST7* | 0 |

Table B.2 – continued from previous page

| **Chromosome** | **Start** | **End** | **Closest gene** | **Number of ALS-associated SNPs** |
|---|---|---|---|---|
| chr7 | 116899318 | 116900524 | *WNT2* | 0 |
| chr7 | 116908705 | 116911053 | *WNT2* | 4 |
| chr7 | 116923096 | 116924277 | *WNT2* | 0 |
| chr7 | 117222645 | 117223108 | *CFTR* | 0 |
| chr7 | 117305812 | 117306347 | *CFTR* | 1 |
| chr7 | 117468242 | 117468792 | *CTTNBP2* | 3 |

## B.2 eRNA seqeunces

For RNA secondary structure prediction the following input nucleotide sequences were utilised.

### B.2.1 $^{DRR}$eRNA

AUUUAUAGCACCUUGGAAGACUAGCCAAGGGAGCUGAAAUGCAAGGCCUG
GAAAGGACAGGGGGAAAUCAAAGGGGCCACCAUGGCGGCAGGAGAACUGA
GCCUCAGGAUGAGCUGUGUGCUUCUCCAGAUUUAAGUCAAAUGAAAGAGC
ACUUAUGAUGAACCCCACAUCCAUCGGCAGCAUACUGUUGGAAUGUUGCA
ACCGACCAAUGGGAGAGAGCACGUCCCAGGCAAACCAGCCUCUGCUUUGC
CUGGGCAGAGGCAGCGUGAGAGCUUGUAUGAGUAAGUACCUACAUAGAGC
CCAGGUUUGCUGGAAUAGAAUGACUUGUAGCACAUUUUGCUAAAUUCAAG
UAUAAGGAUAGAAAUCAAAAGAGCCCAAGACUGUUCAUUCACUCGCUUGA
CACUUAAACCACUGUGCCCGCCGUGGGAUCACUGCUGCAGUGGCUUCCGG
AGACGCCAUGGUGAGCAAAGUACUCCUAUCCAUGGUAUGCUGGUCUUCGU
GUCCCUCGGUGAAUAAC

### B.2.2 *TAL1* sense eRNA

UAUUUUAAGCGCAUGUGCAUUCUCUCUGUCUCUAAGUCUGCUCCUCUUUCC
UACCCCGGCCCUGUCUCUCCAUCUCUGUCUUUAAUCUGCCUCUCAUCAUUG
CCUCCUUCCUCUUUUUUGGUCUCUGUUCAGGCUGUCAAUAAGAGCUCCAGC
UGUGCACAGGUACCCCAUAGAAGCAAGGCAUGCCGAUUCUCUUGCCACAU
UCCCGACAGGCUGGACCUGUUAGAAAGGCAUAAGCUGGUCU

### B.2.3 *TAL1* antisense eRNA

CCUUGUCUACAGAUCCUUUUUUGUGCCCUGUGUGCCUGGCCCUGAGCAGC
AUGAGUUAGACUGUAACGGAGGGCUCCAGGGUAUGCUAAUCAGGAGAGCA
AGUGGUGAGAGGAGGAGAUAUAAAUGCAGCUGAGGACAGUAUUGAUCAGA
AAAGGUUUCCUGGACAACAUAAGAUUGAUGUGGGGCCACAAAGAACGUUU
AUAGUCAAUUUUUUGAGGCAUAAUUUACAUACUAUAAGAUGUACUCAUUU
UAGUGUACAGUUUGUUGAGUUUUGACAAAUGCAUACACUCAUAUAACCAC
CACCACAAUCAAGAUAUAGAAUAUUUCCUUUAUGCAGAAAGUUCCCUGUG
CCUCUUUGCAGUCAUAUGUCCUCUUCCAUCACCCCUGAUCUGAGCUUUGC

### B.2.4 *eCAV1_WT*

UUUUUUUUUUUUCCUCCUCAGUUCUUCAUAUUUCUCUCAAAUUACCUUUCA
ACAAUAUGUUCAUAUCUCAGCCAAUAAAGCGAUGGUUGAUCCACAUGCAA
AUAAAAAGGUAAGGAAAAAAAUUUAAAUGGGUUUUAAAAAUUCAGACUGC
CAAAAAUAGAUGAAAUAGCUCAGAAGAGACAUAAUACUUAUGAAAACUGA
GAAAGUGAUCUUUUUAUUUUCAACCAAGAUAAUUCAAUUCAUA

### B.2.5 *eCAV1_ALS*

UUUUUUUUUUUUCCUCCUCAGUUCUUCAUAUUUCUCUCAAAUUACCUUUCA
ACAAUAUGUUCAUAUCUCAGCCAAUAAAGCGAUGGUUGAUCCACAUGCAA
AUAAAAAGGUAAGGAAAAAAAUUUAAAUGGGUUUUAAAAAUUCAGACUGC
CAAAAAUGGAUGAAAUAGCUCAGAAGAGACAUAAUACUUAUGAAAACUGA
GAAAGUGAUCUUUUUAUUUUCAACCAAGAUAAUUCAAUUCAUA

### B.2.6 *eCAPZA2_antisense*

GAUCCCCUGAGGUCAGGAGUUCAAGACCAGCCUGGCCAACAUGGUGAAAC
CCCGUUU

### B.2.7 *eCAPZA2_sense*

GAAGUCAGGUGGCCGCUGCCGCCGCCGCCGCCGCGGUUUGUCGCCAGAAG
GAAGAUGGCGGAUCUGGAGGAGCAGUUGUCUGAUGAAGAGAAGGUAAGAG
UCGCGGGGGCGACGGCGCGGGCUGUCAGGAGCCGGCUCAGCCCGGGCGGG

UGGGCCCGGAGUCUGGUCGCGGGGGAAGGGCAUUGGCCCGCAGCUGGCCU
UGCCCUCGGCUGCCCUUCCUCCCCCAUCCUUACUGCGCCCAGUGCCUGCCC
CUGUCCGUCACCCGAGGCCCCGCGCGGGGGCAGGUCUCUGCCAGGCCUGC
UCGCUCCGCUCCUUGUAGGGUG

### B.2.8 *eCAV1_3363_us*

UUACUGAGAAGAUACCAGUCCCUCCUCAGAGCAGACAAGGAAACUGAGCC
UCAGAAUGAAAGACUGAAUUUCAGUCCUUUCUUGAACAUGGACCUCCAGG
GUUAUAUUGGGCCUUGGAAAAGGCACUUACACUCUGGACUGUAGUUUCUU
CAUCUAUAAAAUCAAGAGGCAGAAACAGACAAUCUCUAAGUUGCCUUUAU
UUAUAAAAUUCCGAGAUUCUAGUUGACCAGUAUUCAUACAAGAGUUGAAG
CCUGUAAGAGUGCAGAAAGCCCACACAAAGAGACAGUGGAAGACCUCUCA
UCAGUAGUAUUUUUAUUACCCUCUUCCUAGGUUUUACCAGUCAACAUCCUC
ACUGUUAAUAUACAGACCGUGGUAUUUAAUUAAAUCAUCUUUGAAAUACU
GAGCUAUCAACAGAUGGCAUGCUGAAUGCAAAAGGACCACAAAUAAAUAU
UUGGUACUGAAGAAGAUCAAGAGUUGGAGUUC

### B.2.9 *eCAV1_3388_us*

GAAAAGGCACUUACACUCUGGACUGUAGUUUCUUCAUCUAUAAAAUCAAG
AGGCAGAAACAGACAAUCUCUAAGUUGCCUUUAUUUAUAAAAUUCCGAGA
UUCUAGUUGACCAGUAUUCAUACAAGAGUUGAAGCCUGUAAGAGUGCAGA
AAGCCCACACAAAGAGACAGUGGAAGACCUCUCAUCAGUAGUAUUUUUAU
UACCCUCUUCCUAGGUUUUACCAGUCAACAUCCUCACUGUUAAUAUACAGA
CCGUGGUAUUUAAUUAAAUCAUCUUUGAAAUACUGAGCUAUCAACAGAUG
GCAUGCUGAAUGCAAAAGGACCACAAAUAAAUAUUUGGUACUGAAGAAGA
UCAAGAGUUGGAGUUCAUUUCCCAUUCUGAUCUGGGCUCAGAACUCUGUG
GUCUUCCCUCUAAUCAUCCUUGCCACCAAAUUGGCUGUAUCUGUUCUAAGA
UGGAUCAGAAAAUCAGUUCCAAAGUUGGCUACAAACUUUCAGGUUUGGGU
UUUGUUUUGUUUUUUUGUUUUGUUUUGUUUUGUUUUUGCAACCAGCCAAU
UCAUCUUAGUUCACAUGACAG

### B.2.10 *eCAV1_3363_ds*

CAAUAGGAUCAGAUGACUAAGAGCUACAGGGCAGAAACACUGUUACUUAG
AGUCAAAUUUUCCCAUUACCUAGCUGUAAAGAGUUUGUUUCUCUCUGACU
CAUAUAAAGUUUACCAUUUAGGCCCCUGCAUGAUUUUAAUUCCAUCACUUA

ACACCCCAGCCAUAUGAUUCUGAAGGUAAACAUGAAGGCGUUUGAAUUCC
AGA

### B.2.11 *eCAV1_3388_ds*

GUAAAGAGUUUGUUUCUCUCUGACUCAUAUAAAGUUUACCAUUUAGGCCC
CUGCAUGAUUUUAAUUCCAUCACUUAACACCCCAGCCAUAUGAUUCUGAAG
GUAAACAUGAAGGCGUUUGAAUUCCAGACCACCUAAACAUUCUUAAGGAA
AUCAUCAUCUCCACGGGCAGAGCUAUGCCAAAAUCUGUAGGUUUUAACUC
AAAUUUCAUGAUAAGCAAAAAUUGAAUUAAUUUGUCUUCCAUUUUGUUCA
CCUUUUUGCCAAAAUUAUGCCUGGAUUAGAAUAAAUAAAUUCAAUCAAUG
AAUGCAAUCACUAAUUCUUACGCCAGAUAAUAACACAUUCAGAAUUCUCCU
UUCCCUGGGAGAUUUUAUCAGGUUAGUGUUCUUGUAAACAGGAGA

### B.2.12 *eTES_3388*

UUCUUCUGGGCCUAGUGCAAGAGCUUAGUCCAAAACAAUAGCCUCCCAUA
AAUUUAACACUACUCUCAGUAAAAAUAACCA

## B.3 Generation of shuffled RNA sequences

### B.3.1 Generation of shuffled RNA sequences of a given length and GC%

To calculate the MFEden of shuffled sequences for sequence lengths ranging from 50nt to 600nt, in 50nt intervals, for a given GC%, we utilised the following shell script, generating MFE values for the given GC% input.

```
#!/bin/bash

# Check for input parameters
if [ "$#" -ne 1 ]; then
  echo "Usage: $0 <GC_percentage>"
  exit 1
fi

# Input parameter for GC percentage
```

```
gc_percentage=$1

# Function to generate an RNA sequence and calculate the MFE
generate_sequence_and_calculate_mfe() {
  length=$1
  gc_percentage=$2

  # Calculate the number of G and C bases
  gc_count=$((length * gc_percentage / 100))

  # Calculate the number of A and U bases
  au_count=$((length - gc_count))

  # Determine how many Gs and Cs
  G=$((gc_count / 2))
  C=$((gc_count - G))

  # Determine how many As and Us
  A=$((au_count / 2))
  U=$((au_count - A))

  # Generate the sequence
  sequence=""
  sequence+=$(printf "G%.0s" $(seq 1 $G))
  sequence+=$(printf "C%.0s" $(seq 1 $C))
  sequence+=$(printf "A%.0s" $(seq 1 $A))
  sequence+=$(printf "U%.0s" $(seq 1 $U))

  # Shuffle the sequence
  shuffled_sequence=$(echo "$sequence" | fold -w1 | shuf | tr -d '\n')

  # Ensure the sequence length is correct
  shuffled_sequence=${shuffled_sequence:0:$length}

  # Create a temporary file for the sequence
  temp_file=$(mktemp)
  echo $shuffled_sequence > $temp_file

  # Run RNAfold with the provided options and extract the MFE
```

```
  mfe=$(RNAfold -p --noLP --salt 0.1 < $temp_file |
  grep '(' |
  awk '{print $NF}' |
  tr -d '()')

  # Clean up the temporary file
  rm $temp_file

  # Return the MFE
  echo $mfe
}

# Start with lengths from 50 to 600, incrementing by 50
lengths=(50 100 150 200 250 300 350 400 450 500 550 600)

# Print the header
echo -e "Length\t1\t2\t3\t4\t5\t6\t7\t8\t9\t10"

# Loop over each sequence length
for length in "${lengths[@]}"; do
  echo -n "$length"

  # Run 10 iterations for each length
  for i in {1..10}; do
    mfe=$(generate_sequence_and_calculate_mfe $length $gc_percentage)
    echo -n -e "\t$mfe"
  done

  # Print a new line for the next length
  echo ""
done
```

### B.3.2 Generation of shuffled RNA sequences to calculate MFE density

To generate shuffled RNA sequences with equimolar ratios of each nucleotide for a given RNA sequence of length $x$, the following shell script was utilised.

```
#!/bin/bash
```

```
# Desired sequence length
length=x

# Calculate the base count per nucleotide
base_count=$((length / 4))
remainder=$((length % 4))

# Initialize counts for each nucleotide
A=$base_count
U=$base_count
C=$base_count
G=$base_count

# Distribute the remainder randomly amoung the nucleotides
for i in $(seq 1 $remainder); do
  case $((RANDOM % 4)) in
    0) A=$((A + 1)) ;;
    1) U=$((U + 1)) ;;
    2) C=$((C + 1)) ;;
    3) G=$((G + 1)) ;;
  esac
done

# Generate the sequence
sequence=""
sequence+=$(printf "A%.0s" $(seq 1 $A))
sequence+=$(printf "U%.0s" $(seq 1 $U))
sequence+=$(printf "C%.0s" $(seq 1 $C))
sequence+=$(printf "G%.0s" $(seq 1 $G))

# Shuffle the sequence
shuffled_sequence=$(echo "$sequence" | fold -w1 | shuf | tr -d '\n')

# Ensure the sequence length is correct
shuffled_sequence=${shuffled_sequence:0:$length}

# Output the final sequence
echo "Sequence: $shuffled_sequence"
```

## B.4 Generation of quantile bins within TAD domains

To generate quantile bins within TADs across the *CAV1/CAV2* enhancer landscape, the following custom `R` function was utilised, with the TAD domain BED file as input, with each domain given an arbitrary number (domain_number).

```
# Number of bins
num_bins <- 10

# Function to create bins
create_bins <- function(chr, start, end, domain_number, num_bins) {
  bin_size <- (end - start) / num_bins
  bin_starts <- seq(start, end - bin_size, length.out = num_bins)
  bin_ends <- bin_starts + bin_size
  df <- data.frame(chr = rep(chr, num_bins),
                   start = bin_starts,
                   end = bin_ends,
                   identity = rep("domain", num_bins),
                   domain_number = rep(domain_number, num_bins))
  return(df)
}

# List to store the new data frames
new_domains <- list()

# Create bins for each domain
for (i in 1:nrow(domains)) {
  new_domains[[i]] <- create_bins(domains$chr[i],
  domains$start[i],
  domains$end[i],
  domains$domain_number[i],
  num_bins)
}

# Combine all data frames into one
domains_binned <- do.call(rbind, new_domains)
```

# Bibliography


Abhinav, K., Stanton, B., Johnston, C., Hardstaff, J., Orrell, R. W., Howard, R., Clarke, J., Sakel, M., Ampong, M.-A., Shaw, C. E., Leigh, P. N., and Al-Chalabi, A. (2007). Amyotrophic lateral sclerosis in South-East england: a population-based study. the South-East england register for amyotrophic lateral sclerosis (SEALS registry). *Neuroepidemiology*, 29(1-2):44–48.

Adelman, K. and Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.*, 13(10):720–731.

Adey, B. N., Cooper-Knock, J., Al Khleifat, A., Fogh, I., van Damme, P., Corcia, P., Couratier, P., Hardiman, O., McLaughlin, R., Gotkine, M., Drory, V., Silani, V., Ticozzi, N., Veldink, J. H., van den Berg, L. H., de Carvalho, M., Pinto, S., Mora Pardina, J. S., Povedano Panades, M., Andersen, P. M., Weber, M., Başak, N. A., Shaw, C. E., Shaw, P. J., Morrison, K. E., Landers, J. E., Glass, J. D., Vourc'h, P., Dobson, R. J. B., Breen, G., Al-Chalabi, A., Jones, A. R., and Iacoangeli, A. (2023). Large-scale analyses of CAV1 and CAV2 suggest their expression is higher in post-mortem ALS brain tissue and affects survival. *Front. Cell. Neurosci.*, 17:1112405.

Alam, J. and Cook, J. L. (1990). Reporter genes: application to the study of mammalian gene transcription. *Anal. Biochem.*, 188(2):245–254.

Alerasool, N., Segal, D., Lee, H., and Taipale, M. (2020). An efficient KRAB domain for CRISPRi applications in human cells. *Nat. Methods*, 17(11):1093–1096.

Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (1964). Acetylation and methylation of histones and their possible role in the regulation of rna synthesis. *Proc. Natl. Acad. Sci. U. S. A.*, 51(5):786.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal dynamics at the shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell*, 16(1):47–57.

Amin, A., Perera, N. D., Beart, P. M., Turner, B. J., and Shabanpoor, F. (2020). Amyotrophic lateral sclerosis and autophagy: Dysfunction and therapeutic targeting. *Cells*, 9(11).

Amin Al Olama, A., Kote-Jarai, Z., Schumacher, F. R., Wiklund, F., Berndt, S. I., Benlloch, S., Giles, G. G., Severi, G., Neal, D. E., Hamdy, F. C., Donovan, J. L., Hunter, D. J., Henderson, B. E., Thun, M. J., Gaziano, M., Giovannucci, E. L., Siddiq, A., Travis, R. C., Cox, D. G., Canzian, F., Riboli, E., Key, T. J., Andriole, G., Albanes, D., Hayes, R. B., Schleutker, J., Auvinen, A., Tammela, T. L. J., Weischer, M., Stanford, J. L., Ostrander, E. A., Cybulski, C., Lubinski, J., Thibodeau, S. N., Schaid, D. J., Sorensen, K. D., Batra, J., Clements, J. A., Chambers, S., Aitken, J., Gardiner, R. A., Maier, C., Vogel, W., Dörk, T., Brenner, H., Habuchi, T., Ingles, S., John, E. M., Dickinson, J. L., Cannon-Albright, L., Teixeira, M. R., Kaneva, R., Zhang, H.-W., Lu, Y.-J., Park, J. Y., Cooney, K. A., Muir, K. R., Leongamornlert, D. A., Saunders, E., Tymrakiewicz, M., Mahmud, N., Guy, M., Govindasami, K., O'Brien, L. T., Wilkinson, R. A., Hall, A. L., Sawyer, E. J., Dadaev, T., Morrison, J., Dearnaley, D. P., Horwich, A., Huddart, R. A., Khoo, V. S., Parker, C. C., Van As, N., Woodhouse, C. J., Thompson, A., Dudderidge, T., Ogden, C., Cooper, C. S., Lophatonanon, A., Southey, M. C., Hopper, J. L., English, D., Virtamo, J., Le Marchand, L., Campa, D., Kaaks, R., Lindstrom, S., Diver, W. R., Gapstur, S., Yeager, M., Cox, A., Stern, M. C., Corral, R., Aly, M., Isaacs, W., Adolfsson, J., Xu, J., Zheng, S. L., Wahlfors, T., Taari, K., Kujala, P., Klarskov, P., Nordestgaard, B. G., Røder, M. A., Frikke-Schmidt, R., Bojesen, S. E., FitzGerald, L. M., Kolb, S., Kwon, E. M., Karyadi, D. M., Orntoft, T. F., Borre, M., Rinckleb, A., Luedeke, M., Herkommer, K., Meyer, A., Serth, J., Marthick, J. R., Patterson, B., Wokolorczyk, D., Spurdle, A., Lose, F., McDonnell, S. K., Joshi, A. D., Shahabi, A., Pinto, P., Santos, J., Ray, A., Sellers, T. A., Lin, H.-Y., Stephenson, R. A., Teerlink, C., Muller, H., Rothenbacher, D., Tsuchiya, N., Narita, S., Cao, G.-W., Slavov, C., Mitev, V., UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology, UK ProtecT Study Collaborators, Australian Prostate Cancer Bioresource, PRACTICAL Consortium, Chanock, S., Gronberg, H., Haiman, C. A., Kraft, P., Easton, D. F., and Eeles, R. A. (2013). A meta-analysis of genome-wide association studies to identify prostate cancer susceptibility loci associated with aggressive and non-aggressive disease. *Hum. Mol. Genet.*, 22(2):408–415.

An, J. J., Gharami, K., Liao, G.-Y., Woo, N. H., Lau, A. G., Vanevski, F., Torre, E. R., Jones, K. R., Feng, Y., Lu, B., and Xu, B. (2008). Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell*, 134(1):175–187.

Ananth, P., Goldsmith, G., and Yathindra, N. (2013). An innate twist between crick's wobble and Watson-Crick base pairs. *RNA*, 19(8):1038–1053.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M., and Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Andreone, B. J., Chow, B. W., Tata, A., Lacoste, B., Ben-Zvi, A., Bullock, K., Deik, A. A., Ginty, D. D., Clish, C. B., and Gu, C. (2017). Blood-Brain barrier permeability is regulated by lipid Transport-Dependent suppression of Caveolae-Mediated transcytosis. *Neuron*, 94(3):581–594.e5.

Andrews, R. J., Roche, J., and Moss, W. N. (2018). ScanFold: an approach for genome-wide discovery of local RNA structural elements-applications to zika virus and HIV. *PeerJ*, 6(e6136):e6136.

Andrews, R. J., Rouse, W. B., O'Leary, C. A., Booher, N. J., and Moss, W. N. (2022). ScanFold 2.0: a rapid approach for identifying potential structured RNA targets in genomes and transcriptomes. *PeerJ*, 10:e14361.

Aranda-Orgilles, B., Saldaña-Meyer, R., Wang, E., Trompouki, E., Fassl, A., Lau, S., Mullenders, J., Rocha, P. P., Raviram, R., Guillamot, M., Sánchez-Díaz, M., Wang, K., Kayembe, C., Zhang, N., Amoasii, L., Choudhuri, A., Skok, J. A., Schober, M., Reinberg, D., Sicinski, P., Schrewe, H., Tsirigos, A., Zon, L. I., and Aifantis, I. (2016). MED12 regulates HSC-Specific enhancers independently of mediator kinase activity to control hematopoiesis. *Cell Stem Cell*, 19(6):784–799.

Atchison, L., Ghias, A., Wilkinson, F., Bonini, N., and Atchison, M. L. (2003). Transcription factor YY1 functions as a PcG protein in vivo. *EMBO J.*, 22(6):1347–1358.

Babak, T., Blencowe, B. J., and Hughes, T. R. (2007). Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*, 8(1):33.

Back, S. and Manfredi, J. J. (2021). Knockdown of target genes by siRNA in vitro. *Methods Mol. Biol.*, 2267:159–163.

Bal, E., Kumar, R., Hadigol, M., Holmes, A. B., Hilton, L. K., Loh, J. W., Dreval, K., Wong, J. C. H., Vlasevska, S., Corinaldesi, C., Soni, R. K., Basso, K., Morin, R. D., Khiabanian, H., Pasqualucci, L., and Dalla-Favera, R. (2022). Super-enhancer hypermutation alters oncogene expression in B cell lymphoma. *Nature*, pages 1–8.

Banani, S. F., Rice, A. M., Peeples, W. B., Lin, Y., Jain, S., Parker, R., and Rosen, M. K. (2016). Compositional control of Phase-Separated cellular bodies. *Cell*, 166(3).

Banerjee, P. R., Milin, A. N., Moosa, M. M., Onuchic, P. L., and Deniz, A. A. (2017). Reentrant phase transition drives dynamic substructure formation in ribonucleoprotein droplets. *Angew. Chem. Int. Ed Engl.*, 56(38):11354–11359.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, 33(3):729–740.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297.

Bäumer, D., Talbot, K., and Turner, M. R. (2014). Advances in motor neurone disease. *J. R. Soc. Med.*, 107(1):14–21.

Bava, F.-A., Eliscovich, C., Ferreira, P. G., Miñana, B., Ben-Dov, C., Guigó, R., Valcárcel, J., and Méndez, R. (2013). CPEB1 coordinates alternative 3'-UTR formation with translational regulation. *Nature*, 495(7439):121–125.

Beagan, J. A., Pastuzyn, E. D., Fernandez, L. R., Guo, M. H., Feng, K., Titus, K. R., Chandrashekar, H., Shepherd, J. D., and Phillips-Cremins, J. E. (2020). Three-dimensional genome restructuring across timescales of activity-induced neuronal gene expression. *Nat. Neurosci.*, 23(6):707–717.

Beltran, M., Tavares, M., Justin, N., Khandelwal, G., Ambrose, J., Foster, B. M., Worlock, K. B., Tvardovskiy, A., Kunzelmann, S., Herrero, J., Bartke, T., Gamblin, S. J., Wilson, J. R., and Jenner, R. G. (2019). G-tract RNA removes polycomb repressive complex 2 from genes. *Nat. Struct. Mol. Biol.*, 26(10):899–909.

Beltran, M., Yates, C. M., Skalska, L., Dawson, M., Reis, F. P., Viiri, K., Fisher, C. L., Sibley, C. R., Foster, B. M., Bartke, T., Ule, J., and Jenner, R. G. (2016). The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res.*, 26(7):896–907.

Bender, M. A., Ragoczy, T., Lee, J., Byron, R., Telling, A., Dean, A., and Groudine, M. (2012). The hypersensitive sites of the murine $\beta$-globin locus control region act independently to affect nuclear localization and transcriptional elongation. *Blood*, 119(16):3820–3827.

Bergman, D. T., Jones, T. R., Liu, V., Ray, J., Jagoda, E., Siraj, L., Kang, H. Y., Nasser, J., Kane, M., Rios, A., Nguyen, T. H., Grossman, S. R., Fulco, C. P., Lander, E. S., and Engreitz, J. M. (2022). Compatibility rules of human enhancer and promoter sequences. *Nature*, 607(7917):176–184.

Berlivet, S., Moussette, S., Ouimet, M., Verlaan, D. J., Koka, V., Al Tuwaijri, A., Kwan, T., Sinnett, D., Pastinen, T., and Naumova, A. K. (2012). Interaction between genetic and epigenetic variation defines gene expression patterns at the asthma-associated locus 17q12-q21 in lymphoblastoid cell lines. *Hum. Genet.*, 131(7):1161–1171.

Besse, F. and Ephrussi, A. (2008). Translational control of localized mRNAs: restricting protein synthesis in space and time. *Nat. Rev. Mol. Cell Biol.*, 9(12):971–980.

Bettens, K., Brouwers, N., Engelborghs, S., Van Miegroet, H., De Deyn, P. P., Theuns, J., Sleegers, K., and Van Broeckhoven, C. (2009). APP and BACE1 miRNA genetic variability has no major role in risk for alzheimer disease. *Hum. Mutat.*, 30(8):1207–1213.

Bhat-Nakshatri, P., Song, E.-K., Collins, N. R., Uversky, V. N., Dunker, A. K., O'Malley, B. W., Geistlinger, T. R., Carroll, J. S., Brown, M., and Nakshatri, H. (2013). Interplay between estrogen receptor and AKT in estradiol-induced alternative splicing. *BMC Med. Genomics*, 6:21.

Biedler, J. L., Roffler-Tarlov, S., Schachner, M., and Freedman, L. S. (1978). Multiple neurotransmitter synthesis by human neuroblastoma cell lines and clones. *Cancer Res.*, 38(11 Pt 1):3751–3757.

Birnbaum, R. Y., Clowney, E. J., Agamy, O., Kim, M. J., Zhao, J., Yamanaka, T., Pappalardo, Z., Clarke, S. L., Wenger, A. M., Nguyen, L., Gurrieri, F., Everman, D. B., Schwartz, C. E., Birk, O. S., Bejerano, G., Lomvardas, S., and Ahituv, N. (2012). Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res.*, 22(6):1059–1068.

Blayney, J., Francis, H., Camellato, B., Mitchell, L., Stolper, R., Boeke, J., Higgs, D., and Kassouf, M. (2022). Super-enhancers require a combination of classical enhancers and novel facilitator elements to drive high levels of gene expression.

Blažeković, A., Jerčić, K. G., and Borovečki, F. (2021). SNCA 3' UTR genetic variants in patients with parkinson's disease. *Biomolecules*, 11(12).

Boeynaems, S., Holehouse, A. S., Weinhardt, V., Kovacs, D., Van Lindt, J., Larabell, C., Van Den Bosch, L., Das, R., Tompa, P. S., Pappu, R. V., and Gitler, A. D. (2019). Spontaneous driving forces give rise to protein-RNA condensates with coexisting phases and complex material properties. *Proc. Natl. Acad. Sci. U. S. A.*, 116(16):7889–7898.

Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., Li, C. H., Shrinivas, K., Manteiga, J. C., Hannett, N. M., Abraham, B. J., Afeyan, L. K., Guo, Y. E., Rimel, J. K., Fant, C. B., Schuijers, J., Lee, T. I., Taatjes, D. J., and Young, R. A. (2018). Transcription factors activate genes through the Phase-Separation capacity of their activation domains. *Cell*, 175(7).

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., Tanay, A., and Cavalli, G. (2017). Multiscale 3D genome rewiring during mouse neural development. *Cell*, 171(3):557–572.e24.

Bonner, J. T. (1988). *The Evolution of Complexity by Means of Natural Selection*. Princeton University Press.

Bonnet, E., Wuyts, J., Rouzé, P., and Van de Peer, Y. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–2917.

Borsari, B., Villegas-Mirón, P., Pérez-Lluch, S., Turpin, I., Laayouni, H., Segarra-Casas, A., Bertranpetit, J., Guigó, R., and Acosta, S. (2021). Enhancers with tissue-specific activity are enriched in intronic regions. *Genome Res.*, 31(8):1325–1336.

Bose, D. A., Donahue, G., Reinberg, D., Shiekhattar, R., Bonasio, R., and Berger, S. L. (2017). RNA binding to CBP stimulates histone acetylation and transcription. *Cell*, 168(1-2):135–149.e22.

Bothma, J. P., Garcia, H. G., Ng, S., Perry, M. W., Gregor, T., and Levine, M. (2015). Enhancer additivity and non-additivity are determined by enhancer strength in the drosophila embryo. *Elife*, 4.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322.

Brennan, C. M. and Steitz, J. A. (2001). HuR and mRNA stability. *Cell. Mol. Life Sci.*, 58(2):266–277.

Britten, R. J. and Davidson, E. H. (1969). Gene regulation for higher cells: a theory. *Science*, 165(3891):349–357.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 10(12):1213–1218.

Bysani, M., Agren, R., Davegårdh, C., Volkov, P., Rönn, T., Unneberg, P., Bacos, K., and Ling, C. (2019). ATAC-seq reveals alterations in open chromatin in pancreatic islets from subjects with type 2 diabetes. *Sci. Rep.*, 9(1):7785.

Caldas, P., Luz, M., Baseggio, S., Andrade, R., Sobral, D., and Grosso, A. R. (2024). Transcription readthrough is prevalent in healthy human tissues and associated with inherent genomic features. *Commun. Biol.*, 7(1):100.

Calingasan, N. Y., Chen, J., Kiaei, M., and Beal, M. F. (2005). Beta-amyloid 42 accumulation in the lumbar spinal cord motor neurons of amyotrophic lateral sclerosis patients. *Neurobiol. Dis.*, 19(1-2):340–347.

Calo, E. and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, 49(5):825–837.

Caretti, G., Di Padova, M., Micales, B., Lyons, G. E., and Sartorelli, V. (2004). The polycomb ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation. *Genes Dev.*, 18(21):2627–2638.

Carlson, M. and Maintainer, B. P. (2015). TxDb. hsapiens. UCSC. hg19. knowngene. *Bioconductor*.

Carullo, N. V. N. and Day, J. J. (2019). Genomic enhancers in brain health and disease. *Genes*, 10(1).

Carullo, N. V. N., Phillips, Iii, R. A., Simon, R. C., Soto, S. A. R., Hinds, J. E., Salisbury, A. J., Revanna, J. S., Bunner, K. D., Ianov, L., Sultan, F. A., Savell, K. E., Gersbach, C. A., and Day, J. J. (2020). Enhancer RNAs predict enhancer-gene regulatory links and are critical for enhancer function in neuronal systems. *Nucleic Acids Res.*, 48(17):9550–9570.

Castello, A., Fischer, B., Frese, C. K., Horos, R., Alleaume, A.-M., Foehr, S., Curk, T., Krijgsveld, J., and Hentze, M. W. (2016). Comprehensive identification of RNA-Binding domains in human cells. *Mol. Cell*, 63(4):696–710.

Cedar, H. and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.*, 10(5):295–304.

Chang, K. Y., Varani, G., Bhattacharya, S., Choi, H., and McClain, W. H. (1999). Correlation of deformability at a tRNA recognition site and aminoacylation specificity. *Proc. Natl. Acad. Sci. U. S. A.*, 96(21):11764–11769.

Chang, P., Torres, J., Lewis, R. A., Mowry, K. L., Houliston, E., and King, M. L. (2004). Localization of RNAs to the mitochondrial cloud in xenopus oocytes through entrapment and association with endoplasmic reticulum. *Mol. Biol. Cell*, 15(10):4669–4681.

Chapuis, J., Flaig, A., Grenier-Boley, B., Eysert, F., Pottiez, V., Deloison, G., Vandeputte, A., Ayral, A.-M., Mendes, T., Desai, S., Goate, A. M., Kauwe, J. S. K., Leroux, F., Herledan, A., Demiautte, F., Bauer, C., Checler, F., Petersen, R. C., Blennow, K., Zetterberg, H., Minthon, L., Van Deerlin, V. M., Lee, V. M.-Y., Shaw, L. M., Trojanowski, J. Q., Albert, M., Moghekar, A., O'Brien, R., Peskind, E. R., Malmanche, N., Schellenberg, G. D., Dourlen, P., Song, O.-R., Cruchaga, C., Amouyel, P., Deprez, B., Brodin, P., Lambert, J.-C., and ADGC, Alzheimer's Disease Neuroimaging Initiative (2017). Genome-wide, high-content siRNA screening identifies the alzheimer's genetic risk factor FERMT2 as a major modulator of APP metabolism. *Acta Neuropathol.*, 133(6):955–966.

Chartier-Harlin, M.-C., Kachergus, J., Roumier, C., Mouroux, V., Douay, X., Lincoln, S., Levecque, C., Larvor, L., Andrieux, J., Hulihan, M., Waucquier, N., Defebvre, L., Amouyel, P., Farrer, M., and Destée, A. (2004). $\alpha$-synuclein locus duplication as a cause of familial parkinson's disease. *Lancet*, 364(9440):1167–1169.

Chen, C.-Y., Chen, S.-T., Juan, H.-F., and Huang, H.-C. (2012a). Lengthening of 3'UTR increases with morphological complexity in animal evolution. *Bioinformatics*, 28(24):3178–3181.

Chen, C. Y., Gherzi, R., Ong, S. E., Chan, E. L., Raijmakers, R., Pruijn, G. J., Stoecklin, G., Moroni, C., Mann, M., and Karin, M. (2001). AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell*, 107(4):451–464.

Chen, C.-Y. A. and Shyu, A.-B. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem. Sci.*, 20(11):465–470.

Chen, H., Xiao, J., Shao, T., Wang, L., Bai, J., Lin, X., Ding, N., Qu, Y., Tian, Y., Chen, X., Liu, H., Liu, H., Xu, J., and Li, X. (2019). Landscape of enhancer-enhancer cooperative regulation during human cardiac commitment. *Mol. Ther. Nucleic Acids*, 17:840–851.

Chen, L. H., Kao, P. Y. P., Fan, Y. H., Ho, D. T. Y., Chan, C. S. Y., Yik, P. Y., Ha, J. C. T., Chu, L. W., and Song, Y.-Q. (2012b). Polymorphisms of CR1, CLU and PICALM confer susceptibility of alzheimer's disease in a southern chinese population. *Neurobiol. Aging*, 33(1):210.e1–7.

Chen, Z., Snetkova, V., Bower, G., Jacinto, S., Clock, B., Dizehchi, A., Barozzi, I., Mannion, B. J., Alcaina-Caro, A., Lopez-Rios, J., Dickel, D. E., Visel, A., Pennacchio, L. A., and Kvon, E. Z. (2024). Increased enhancer-promoter interactions during developmental enhancer activation in mammals. *Nat. Genet.*, 56(4):675–685.

Cheng, L. C., Zheng, D., Baljinnyam, E., Sun, F., Ogami, K., Yeung, P. L., Hoque, M., Lu, C.-W., Manley, J. L., and Tian, B. (2020). Widespread transcript shortening through alternative polyadenylation in secretory cell differentiation. *Nat. Commun.*, 11(1):3182.

Chia, M., Li, C., Marques, S., Pelechano, V., Luscombe, N. M., and van Werven, F. J. (2021). High-resolution analysis of cell-state transitions in yeast suggests widespread transcriptional tuning by alternative starts. *Genome Biol.*, 22(1):34.

Chiò, A., Calvo, A., Mazzini, L., Cantello, R., Mora, G., Moglia, C., Corrado, L., D'Alfonso, S., Majounie, E., Renton, A., Pisano, F., Ossola, I., Brunetti, M., Traynor, B. J., Restagno, G., and PARALS (2012). Extensive genetics of ALS: a population-based study in italy. *Neurology*, 79(19):1983–1989.

Chipev, C. C. and Wolffe, A. P. (1992). Chromosomal organization of xenopus laevis oocyte and somatic 5S rRNA genes in vivo. *Mol. Cell. Biol.*, 12(1):45–55.

Cho, W. K., Spille, J. H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, I. I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, 361(6400).

Chow, B. W. and Gu, C. (2015). The molecular constituents of the blood-brain barrier. *Trends Neurosci.*, 38(10):598–608.

Chow, B. W. and Gu, C. (2017). Gradual suppression of transcytosis governs functional Blood-Retinal barrier formation. *Neuron*, 93(6):1325–1333.e3.

Chun, S., Casparino, A., Patsopoulos, N. A., Croteau-Chonka, D. C., Raby, B. A., De Jager, P. L., Sunyaev, S. R., and Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.*, 49(4):600–605.

Cichewicz, M. A., Kiran, M., Przanowska, R. K., Sobierajska, E., Shibata, Y., and Dutta, A. (2018). MUNC, an enhancer RNA upstream from the MYOD gene, induces

a subgroup of myogenic transcripts in trans independently of MyoD. *Mol. Cell. Biol.*, 38(20).

Cifuentes-Rojas, C., Hernandez, A. J., Sarma, K., and Lee, J. T. (2014). Regulatory interactions between RNA and polycomb repressive complex 2. *Mol. Cell*, 55(2):171–185.

Cirillo, L. A., Lin, F. R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K. S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell*, 9(2):279–289.

Clote, P., Ferré, F., Kranakis, E., and Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591.

Conlon, E. G., Lu, L., Sharma, A., Yamazaki, T., Tang, T., Shneider, N. A., and Manley, J. L. (2016). The C9ORF72 GGGGCC expansion forms RNA g-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. *Elife*, 5.

Connally, N. J., Nazeen, S., Lee, D., Shi, H., Stamatoyannopoulos, J., Chun, S., Cotsapas, C., Cassa, C. A., and Sunyaev, S. R. (2022). The missing link between genetic association and regulatory function. *Elife*, 11.

Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L., and Myers, R. M. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.*, 16(1):1–10.

Cooper-Knock, J., Zhang, S., Kenna, K. P., Moll, T., Franklin, J. P., Allen, S., Nezhad, H. G., Iacoangeli, A., Yacovzada, N. Y., Eitan, C., Hornstein, E., Elhaik, E., Celadova, P., Bose, D., Farhan, S., Fishilevich, S., Lancet, D., Morrison, K. E., Shaw, C. E., Al-Chalabi, A., Project MinE ALS Sequencing Consortium, Veldink, J. H., Kirby, J., Snyder, M. P., and Shaw, P. J. (2021). Rare variant burden analysis within enhancers identifies CAV1 as an ALS risk gene. *Cell Rep.*, 34(5):108730.

Copeland, C. A., Han, B., Tiwari, A., Austin, E. D., Loyd, J. E., West, J. D., and Kenworthy, A. K. (2017). A disease-associated frameshift mutation in caveolin-1 disrupts caveolae formation and function through introduction of a de novo ER retention signal. *Mol. Biol. Cell*, 28(22):3095–3111.

Core, L. and Adelman, K. (2019). Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes Dev.*, 33(15-16):960–982.

Cowley, M., Wood, A. J., Böhm, S., Schulz, R., and Oakey, R. J. (2012). Epigenetic control of alternative mRNA processing at the imprinted Herc3/Nap1l5 locus. *Nucleic Acids Res.*, 40(18):8917–8926.

Cramer, P. (2019). Organization and regulation of gene transcription. *Nature*, 573(7772):45–54.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.*, 107(50):21931–21936.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Crick, F. H. (1966). Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, 19(2):548–555.

Danko, C. G., Hyland, S. L., Core, L. J., Martins, A. L., Waters, C. T., Lee, H. W., Cheung, V. G., Kraus, W. L., Lis, J. T., and Siepel, A. (2015). Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods*, 12(5):433–438.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.*, 8(5):e1000384.

De Strooper, B. (2010). Proteases and proteolysis in alzheimer disease: a multifactorial view on the disease process. *Physiol. Rev.*, 90(2):465–494.

Decker, C. J., Burke, J. M., Mulvaney, P. K., and Parker, R. (2022). RNA is required for the integrity of multiple nuclear and cytoplasmic membrane-less RNP granules. *EMBO J.*, 41(9):e110137.

DeJesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., Nicholson, A. M., Finch, N. A., Flynn, H., Adamson, J., Kouri, N., Wojtas, A., Sengdy, P., Hsiung, G.-Y. R., Karydas, A., Seeley, W. W., Josephs, K. A., Coppola, G., Geschwind, D. H., Wszolek, Z. K., Feldman, H., Knopman, D. S., Petersen, R. C., Miller, B. L., Dickson, D. W., Boylan, K. B., Graff-Radford, N. R., and Rademakers, R. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*, 72(2):245–256.

Del Gatto-Konczak, F., Bourgeois, C. F., Le Guiner, C., Kister, L., Gesnel, M. C., Stévenin, J., and Breathnach, R. (2000). The RNA-binding protein TIA-1 is a novel mammalian splicing regulator acting through intron sequences adjacent to a 5' splice site. *Mol. Cell. Biol.*, 20(17):6287–6299.

Delay, C., Calon, F., Mathews, P., and Hébert, S. S. (2011). Alzheimer-specific variants in the 3'UTR of amyloid precursor protein affect microRNA function. *Mol. Neurodegener.*, 6:70.

Deng, W. and Blobel, G. A. (2017). Detecting long-range enhancer-promoter interactions by quantitative chromosome conformation capture. *Methods Mol. Biol.*, 1468:51–62.

Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C., Sriram, S., Chen, R., Rohl, C. A., Johnson, J. M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, 22(6):1173–1183.

Di Cerbo, V., Mohn, F., Ryan, D. P., Montellier, E., Kacem, S., Tropberger, P., Kallis, E., Holzner, M., Hoerner, L., Feldmann, A., Richter, F. M., Bannister, A. J., Mittler, G., Michaelis, J., Khochbin, S., Feil, R., Schuebeler, D., Owen-Hughes, T., Daujat, S., and Schneider, R. (2014). Acetylation of histone H3 at lysine 64 regulates nucleosome dynamics and facilitates transcription. *Elife*, 3:e01632.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A., and Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., and Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–108.

Doench, J. G. (2018). Am I ready for CRISPR? a user's guide to genetic screens. *Nat. Rev. Genet.*, 19(2):67–80.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., Sorek, R.,

and Rechavi, G. (2012). Topology of the human and mouse m6a RNA methylomes revealed by m6a-seq. *Nature*, 485(7397):201–206.

Donati, B., Lorenzini, E., and Ciarrocchi, A. (2018). BRD4 and cancer: going beyond transcriptional regulation. *Mol. Cancer*, 17(1):164.

Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603.

Drobot, B., Iglesias-Artola, J. M., Le Vay, K., Mayr, V., Kar, M., Kreysing, M., Mutschler, H., and Tang, T.-Y. D. (2018). Compartmentalised RNA catalysis in membrane-free coacervate protocells. *Nat. Commun.*, 9(1):3643.

Du, M., Stitzinger, S. H., Spille, J.-H., Cho, W.-K., Lee, C., Hijaz, M., Quintana, A., and Cissé, I. I. (2024). Direct observation of a condensate effect on super-enhancer controlled gene bursting. *Cell*, 187(2):331–344.e17.

Díaz-Muñoz, M. D., Kiselev, V. Y., Le Novère, N., Curk, T., Ule, J., and Turner, M. (2017). Tia1 dependent regulation of mRNA subcellular location and translation controls p53 expression in B cells. *Nat. Commun.*, 8(1):530.

Eberharter, A. and Becker, P. B. (2002). Histone acetylation: a switch between repressive and permissive chromatin. second in review series on chromatin dynamics: Second in review series on chromatin dynamics. *EMBO Rep.*, 3(3):224–229.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210.

Egawa, J., Zemljic-Harpf, A., Mandyam, C. D., Niesman, I. R., Lysenko, L. V., Kleschevnikov, A. M., Roth, D. M., Patel, H. H., Patel, P. M., and Head, B. P. (2018). Neuron-Targeted caveolin-1 promotes ultrastructural and functional hippocampal synaptic plasticity. *Cereb. Cortex*, 28(9):3255–3266.

El Mendili, M.-M., Cohen-Adad, J., Pelegrini-Issac, M., Rossignol, S., Morizot-Koutlidis, R., Marchand-Pauvert, V., Iglesias, C., Sangari, S., Katz, R., Lehericy, S., Benali, H., and Pradat, P.-F. (2014). Multi-parametric spinal cord MRI as potential progression marker in amyotrophic lateral sclerosis. *PLoS One*, 9(4):e95516.

El Mendili, M. M., Querin, G., Bede, P., and Pradat, P.-F. (2019). Spinal cord imaging in amyotrophic lateral sclerosis: Historical concepts-novel techniques. *Front. Neurol.*, 10:350.

Encinas, M., Iglesias, M., Liu, Y., Wang, H., Muhaisen, A., Ceña, V., Gallego, C., and Comella, J. X. (2000). Sequential treatment of SH-SY5Y cells with retinoic acid and brain-derived neurotrophic factor gives rise to fully differentiated, neurotrophic factor-dependent, human neuron-like cells. *J. Neurochem.*, 75(3):991–1003.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., and Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, 539(7629):452–455.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49.

Errede, M., Girolamo, F., Ferrara, G., Strippoli, M., Morando, S., Boldrin, V., Rizzi, M., Uccelli, A., Perris, R., Bendotti, C., Salmona, M., Roncali, L., and Virgintino, D. (2012). Blood-brain barrier alterations in the cerebral cortex in experimental autoimmune encephalomyelitis. *J. Neuropathol. Exp. Neurol.*, 71(10):840–854.

Fagerström, S., Påhlman, S., Gestblom, C., and Nånberg, E. (1996). Protein kinase C-epsilon is implicated in neurite outgrowth in differentiating human neuroblastoma cells. *Cell Growth Differ.*, 7(6):775–785.

Fallmann, J., Sedlyarov, V., Tanzer, A., Kovarik, P., and Hofacker, I. L. (2016). ARE-site2: an enhanced database for the comprehensive investigation of AU/GU/U-rich elements. *Nucleic Acids Res.*, 44(D1):D90–5.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y. A., Plessy, C., Vitezic, M., Severin, J., Semple, C. A., Ishizu, Y., Young, R. S., Francescatto, M., Alam, I., Albanese, D., Altschuler, G. M., Arakawa, T., Archer, J. A. C., Arner, P., Babina, M., Rennie, S., Balwierz, P. J., Beckhouse, A. G., Pradhan-Bhatt, S., Blake, J. A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Burroughs, A. M., Califano, A., Cannistraci, C. V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H. C., Dalla, E., Davis, C. A., Detmar, M., Diehl, A. D., Dohi, T., Drabløs, F., Edge, A. S. B., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson,

M. C., Faulkner, G. J., Favorov, A. V., Fisher, M. E., Frith, M. C., Fujita, R., Fukuda, S., Furlanello, C., Furino, M., Furusawa, J.-I., Geijtenbeek, T. B., Gibson, A. P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T. J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K. J., Ho Sui, S. J., Hofmann, O. M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B. R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A. S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y. I., Kawashima, T., Kempfle, J. S., Kenna, T. J., Kere, J., Khachigian, L. M., Kitamura, T., Klinken, S. P., Knox, A. J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A. T., Laros, J. F. J., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-Sim, A., Manabe, R.-I., Mar, J. C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D. A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C. L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Noma, S., Noazaki, T., Ogishima, S., Ohkura, N., Ohimiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D. A., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J. G. D., Rackham, O. J. L., Ramilowski, J. A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M. B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E. A., Schulze-Tanzil, G. G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J. W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R. K., 't Hoen, P. A. C., Tagami, M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyodo, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L. M., Verado, R., Vijayan, D., Vorontsov, I. E., Wasserman, W. W., Watanabe, S., Wells, C. A., Winteringham, L. N., Wolvetang, E., Wood, E. J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S. E., Zhang, P. G., Zhao, X., Zucchelli, S., Summers, K. M., Suzuki, H., Daub, C. O., Kawai, J., Heutink, P., Hide, W., Freeman, T. C., Lenhard, B., Bajic, V. B., Taylor, M. S., Makeev, V. J., Sandelin, A., Hume, D. A., Carninci, P., and Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo,

V. K., Epstein, C. B., Daly, M. J., Hafler, D. A., and Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343.

Favorov, A., Mularoni, L., Cope, L. M., Medvedeva, Y., Mironov, A. A., Makeev, V. J., and Wheelan, S. J. (2012). Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput. Biol.*, 8(5):e1002529.

Ferrante, R. J., Browne, S. E., Shinobu, L. A., Bowling, A. C., Baik, M. J., MacGarvey, U., Kowall, N. W., Brown, Jr, R. H., and Beal, M. F. (1997). Evidence of increased oxidative damage in both sporadic and familial amyotrophic lateral sclerosis. *J. Neurochem.*, 69(5):2064–2074.

Ferrari, K. J., Scelfo, A., Jammula, S., Cuomo, A., Barozzi, I., Stützer, A., Fischle, W., Bonaldi, T., and Pasini, D. (2014). Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. *Mol. Cell*, 53(1):49–62.

Finn, E. H. and Misteli, T. (2021). A high-throughput DNA FISH protocol to visualize genome regions in human cells. *STAR Protoc.*, 2(3):100741.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada, Y., Raychaudhuri, S., Daly, M. J., Patterson, N., Neale, B. M., and Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, 47(11):1228–1235.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D., and Cohen, D. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, 2017.

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L., and Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, 529(7584):110–114.

Flavell, S. W., Cowan, C. W., Kim, T.-K., Greer, P. L., Lin, Y., Paradis, S., Griffith, E. C., Hu, L. S., Chen, C., and Greenberg, M. E. (2006). Activity-dependent regulation of MEF2 transcription factors suppresses excitatory synapse number. *Science*, 311(5763):1008–1012.

Floor, S. N. and Doudna, J. A. (2016). Tunable protein synthesis by transcript isoforms in human cells. *Elife*, 5.

Forrest, K. M. and Gavis, E. R. (2003). Live imaging of endogenous RNA reveals a diffusion and entrapment mechanism for nanos mRNA localization in drosophila. *Curr. Biol.*, 13(14):1159–1168.

Frade, J. M. and Ovejero-Benito, M. C. (2015). Neuronal cell cycle: the neuron itself and its circumstances. *Cell Cycle*, 14(5):712–720.

Francesconi, A., Kumari, R., and Zukin, R. S. (2009). Regulation of group I metabotropic glutamate receptor trafficking and signaling by the caveolar/lipid raft pathway. *J. Neurosci.*, 29(11):3590–3602.

Francis, N. J., Kingston, R. E., and Woodcock, C. L. (2004). Chromatin compaction by a polycomb group protein complex. *Science*, 306(5701):1574–1577.

Frankel, E. A., Bevilacqua, P. C., and Keating, C. D. (2016). Polyamine/Nucleotide coacervates provide strong compartmentalization of mg2+, nucleotides, and RNA. *Langmuir*, 32(8):2041–2049.

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466(7305):490–493.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L., and Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, 47(D1):D766–D773.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19(1):92–105.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. (2016). Formation of chromosomal domains by loop extrusion. *Cell Rep.*, 15(9):2038–2049.

Fujimoto, T., Kogo, H., Nomura, R., and Une, T. (2000). Isoforms of caveolin-1 and caveolar structure. *J. Cell Sci.*, 113 Pt 19:3509–3517.

Fujinaga, K., Irwin, D., Huang, Y., Taube, R., Kurosu, T., and Peterlin, B. M. (2004). Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Mol. Cell. Biol.*, 24(2):787–795.

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer control of transcriptional bursting. *Cell*, 166(2):358–368.

Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., and Engreitz, J. M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, 354(6313):769–773.

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., and Engreitz, J. M. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.*, 51(12):1664–1669.

Fulle, S. and Gohlke, H. (2008). Analyzing the flexibility of RNA structures by constraint counting. *Biophys. J.*, 94(11):4202–4219.

Förch, P., Puig, O., Kedersha, N., Martínez, C., Granneman, S., Séraphin, B., Anderson, P., and Valcárcel, J. (2000). The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing. *Mol. Cell*, 6(5):1089–1098.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., GTEx Consortium, Nicolae, D. L., Cox, N. J., and Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, 47(9):1091–1098.

Ganser, L. R., Kelly, M. L., Herschlag, D., and Al-Hashimi, H. M. (2019). The roles of structural dynamics in the cellular functions of RNAs. *Nat. Rev. Mol. Cell Biol.*, 20(8):474–489.

Gao, P., Xia, J.-H., Sipeky, C., Dong, X.-M., Zhang, Q., Yang, Y., Zhang, P., Cruz, S. P., Zhang, K., Zhu, J., Lee, H.-M., Suleman, S., Giannareas, N., Liu, S., PRACTICAL Consortium, Tammela, T. L. J., Auvinen, A., Wang, X., Huang, Q., Wang, L., Manninen, A., Vaarala, M. H., Wang, L., Schleutker, J., and Wei, G.-H. (2018). Biology and clinical implications of the 19q13 aggressive prostate cancer susceptibility locus. *Cell*, 174(3):576–589.e18.

García-Mauriño, S. M., Rivero-Rodríguez, F., Velázquez-Cruz, A., Hernández-Vellisca, M., Díaz-Quintana, A., De la Rosa, M. A., and Díaz-Moreno, I. (2017). RNA binding protein regulation and Cross-Talk in the control of AU-rich mRNA fate. *Front Mol Biosci*, 4:71.

Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., and Shendure, J. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 176(1-2):377–390.e19.

Gaudreault, S. B., Dea, D., and Poirier, J. (2004). Increased caveolin-1 expression in alzheimer's disease brain. *Neurobiol. Aging*, 25(6):753–759.

Genova, M. L., Pich, M. M., Bernacchia, A., Bianchi, C., Biondi, A., Bovina, C., Falasca, A. I., Formiggini, G., Castelli, G. P., and Lenaz, G. (2004). The mitochondrial production of reactive oxygen species in relation to aging and pathology. *Ann. N. Y. Acad. Sci.*, 1011:86–100.

Georges, S. A., Kraus, W. L., Luger, K., Nyborg, J. K., and Laybourn, P. J. (2002). P300-mediated tax transactivation from recombinant chromatin: Histone tail deletion mimics coactivator function. *Mol. Cell. Biol.*, 22(1):127–137.

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E. E. M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 512(7512):96–100.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, 10(5):e1004383.

Gil, N. and Ulitsky, I. (2018). Production of spliced long noncoding RNAs specifies regions with increased enhancer activity. *Cell Syst*, 7(5):537–547.e3.

Gilbert, W., Marchionni, M., and McKnight, G. (1986). On the antiquity of introns. *Cell*, 46(2):151–153.

Gillies, S. D., Morrison, S. L., Oi, V. T., and Tonegawa, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell*, 33(3):717–728.

Ginno, P. A., Lim, Y. W., Lott, P. L., Korf, I., and Chédin, F. (2013). GC skew at the 5' and 3' ends of human genes links r-loop formation to epigenetic regulation and transcription termination. *Genome Res.*, 23(10):1590–1600.

Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I., and Chédin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell*, 45(6):814–825.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274(5287):546, 563–7.

Gong, X., Tang, X., Wiedmann, M., Wang, X., Peng, J., Zheng, D., Blair, L. A. C., Marshall, J., and Mao, Z. (2003). Cdk5-mediated inhibition of the protective effects of transcription factor MEF2 in neurotoxicity-induced apoptosis. *Neuron*, 38(1):33–46.

Gorbovytska, V., Kim, S.-K., Kuybu, F., Götze, M., Um, D., Kang, K., Pittroff, A., Brennecke, T., Schneider, L.-M., Leitner, A., Kim, T.-K., and Kuhn, C.-D. (2022). Enhancer RNAs stimulate pol II pause release by harnessing multivalent interactions to NELF. *Nat. Commun.*, 13(1):2429.

Gossett, L. A., Kelvin, D. J., Sternberg, E. A., and Olson, E. N. (1989). A new myocyte-specific enhancer-binding factor that recognizes a conserved element associated with multiple muscle-specific genes. *Mol. Cell. Biol.*, 9(11):5022–5033.

Goutas, A., Outskouni, Z., Papathanasiou, I., Satra, M., Koliakos, G., and Trachana, V. (2021). Dysregulation of caveolin-1 phosphorylation and nuclear translocation is associated with senescence onset. *Cells*, 10(11).

Gray, L. C., Hughes, T. R., and van den Berg, C. W. (2010). Binding of human antigen R (HuR) to an AU-rich element (ARE) in the 3'untranslated region (3'UTR) reduces the expression of decay accelerating factor (DAF). *Mol. Immunol.*, 47(16):2545–2551.

Greenway, M. J., Andersen, P. M., Russ, C., Ennis, S., Cashman, S., Donaghy, C., Patterson, V., Swingler, R., Kieran, D., Prehn, J., Morrison, K. E., Green, A., Acharya, K. R., Brown, Jr, R. H., and Hardiman, O. (2006). ANG mutations segregate with familial and 'sporadic' amyotrophic lateral sclerosis. *Nat. Genet.*, 38(4):411–413.

Gregersen, L. H., Mitter, R., and Svejstrup, J. Q. (2020). Using TTchem-seq for profiling nascent transcription and measuring transcript elongation. *Nat. Protoc.*, 15(2):604–627.

Groebe, D. R. and Uhlenbeck, O. C. (1988). Characterization of RNA hairpin loop stability. *Nucleic Acids Res.*, 16(24):11725–11735.

Gross, D. S. and Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.*, 57:159–197.

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., Macura, K., Bläss, G., Kellis, M., Werber, M., and Herrmann, B. G. (2013). The tissue-specific lncRNA fendrr is an essential regulator of heart and body wall development in the mouse. *Dev. Cell*, 24(2):206–214.

Gschwind, A. R., Mualim, K. S., Karbalayghareh, A., Sheth, M. U., Dey, K. K., Jagoda, E., Nurtdinov, R. N., Xi, W., Tan, A. S., Jones, H., Ma, X. R., Yao, D., Nasser, J., Avsec, Ž., James, B. T., Shamim, M. S., Durand, N. C., Rao, S. S. P., Mahajan, R., Doughty, B. R., Andreeva, K., Ulirsch, J. C., Fan, K., Perez, E. M., Nguyen, T. C., Kelley, D. R., Finucane, H. K., Moore, J. E., Weng, Z., Kellis, M., Bassik, M. C., Price, A. L., Beer, M. A., Guigó, R., Stamatoyannopoulos, J. A., Lieberman Aiden, E., Greenleaf, W. J., Leslie, C. S., Steinmetz, L. M., Kundaje, A., and Engreitz, J. M. (2023). An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv*.

GTEx Consortium (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330.

Guo, X.-H., Shao, L., Li, J., Mao, Y.-J., Zhang, L.-L., and Song, Q. (2016). Association of TLR-4 regulatory variants (rs41426344 and rs7873784) with rheumatoid arthritis in a chinese population. *Eur. J. Inflam.*, 14(2):118–123.

Guo, Y. E., Manteiga, J. C., Henninger, J. E., Sabari, B. R., Dall'Agnese, A., Hannett, N. M., Spille, J. H., Afeyan, L. K., Zamudio, A. V., Shrinivas, K., Abraham, B. J., Boija, A., Decker, T. M., Rimel, J. K., Fant, C. B., Lee, T. I., Cisse, I. I., Sharp, P. A., Taatjes, D. J., and Young, R. A. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature*, 572(7770).

Guschin, D. Y., Waite, A. J., Katibah, G. E., Miller, J. C., Holmes, M. C., and Rebar, E. J. (2010). A rapid and general assay for monitoring endogenous gene modification. In Mackay, J. P. and Segal, D. J., editors, *Engineered Zinc Finger Proteins: Methods and Protocols*, pages 247–256. Humana Press, Totowa, NJ.

Gusel'nikova, V. V. and Korzhevskiy, D. E. (2015). NeuN as a neuronal nuclear antigen and neuron differentiation marker. *Acta Naturae*, 7(2):42–47.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., Raitakari, O. T., Kuusisto, J., Laakso, M., Price, A. L., Pajukanta, P., and Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, 48(3):245–252.

Guvenek, A., Shin, J., De Filippis, L., Zheng, D., Wang, W., Pang, Z. P., and Tian, B. (2022). Neuronal cells display distinct stability controls of alternative polyadenylation mRNA isoforms, long Non-Coding RNAs, and mitochondrial RNAs. *Front. Genet.*, 13:840369.

Guvenek, A. and Tian, B. (2018). Analysis of alternative cleavage and polyadenylation in mature and differentiating neurons using RNA-seq data. *Quant Biol*, 6(3):253–266.

Ha, T.-Y., Choi, Y. R., Noh, H. R., Cha, S.-H., Kim, J.-B., and Park, S. M. (2021). Age-related increase in caveolin-1 expression facilitates cell-to-cell transmission of $\alpha$-synuclein in neurons. *Mol. Brain*, 14(1):122.

Haarhuis, J. H. I., van der Weide, R. H., Blomen, V. A., Yáñez-Cuna, J. O., Amendola, M., van Ruiten, M. S., Krijger, P. H. L., Teunissen, H., Medema, R. H., van Steensel, B., Brummelkamp, T. R., de Wit, E., and Rowland, B. D. (2017). The cohesin release factor WAPL restricts chromatin loop extension. *Cell*, 169(4):693–707.e14.

Hah, N., Danko, C. G., Core, L., Waterfall, J. J., Siepel, A., Lis, J. T., and Kraus, W. L. (2011). A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, 145(4):622–634.

Hah, N., Murakami, S., Nagari, A., Danko, C. G., and Kraus, W. L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.*, 23(8):1210–1223.

Hahn, M. W. and Wray, G. A. (2002). The g-value paradox. *Evol. Dev.*, 4(2):73–75.

Halees, A. S., El-Badrawi, R., and Khabar, K. S. A. (2008). ARED organism: expansion of ARED reveals AU-rich element cluster variations between human and mouse. *Nucleic Acids Res.*, 36(Database issue):D137–40.

Hardwick, L. J. A. and Philpott, A. (2014). Nervous decision-making: to divide or differentiate. *Trends Genet.*, 30(6):254–261.

Harrison, L. J. and Bose, D. (2022). Enhancer rnas step forward: New insights into enhancer function. *Development*, 149(16):dev200398.

Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., Branco Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Chougule, K., Cockburn, A., Davidson, C., De Silva, N. H., Dodiya, K., Donaldson, S., El Houdaigui, B., Naboulsi, T. E., Fatima, R., Giron, C. G., Genez, T., Grigoriadis, D., Ghattaoraya, G. S., Martinez, J. G., Gurbich, T. A., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Lodha,

D., Marques-Coelho, D., Maslen, G., Merino, G. A., Mirabueno, L. P., Mushtaq, A., Hossain, S. N., Ogeh, D. N., Sakthivel, M. P., Parker, A., Perry, M., Piližota, I., Poppleton, D., Prosovetskaia, I., Raj, S., Pérez-Silva, J. G., Salam, A. I. A., Saraf, S., Saraiva-Agostinho, N., Sheppard, D., Sinha, S., Sipos, B., Sitnik, V., Stark, W., Steed, E., Suner, M.-M., Surapaneni, L., Sutinen, K., Tricomi, F. F., Urbina-Gómez, D., Veidenberg, A., Walsh, T. A., Ware, D., Wass, E., Willhoft, N. L., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G. R., Keatley, J., Loveland, J. E., Moore, B., Mudge, J. M., Naamati, G., Tate, J., Trevanion, S. J., Winterbottom, A., Frankish, A., Hunt, S. E., Cunningham, F., Dyer, S., Finn, R. D., Martin, F. J., and Yates, A. D. (2024). Ensembl 2024. *Nucleic Acids Res.*, 52(D1):D891–D899.

Hatta, M. and Cirillo, L. A. (2007). Chromatin opening and stable perturbation of core histone:DNA contacts by FoxO1. *J. Biol. Chem.*, 282(49):35583–35593.

Hattori, C., Asai, M., Onishi, H., Sasagawa, N., Hashimoto, Y., Saido, T. C., Maruyama, K., Mizutani, S., and Ishiura, S. (2006). BACE1 interacts with lipid raft proteins. *J. Neurosci. Res.*, 84(4):912–917.

Hau, A. M., Gupta, S., Leivo, M. Z., Nakashima, K., Macias, J., Zhou, W., Hodge, A., Wulfkuhle, J., Conkright, B., Bhuvaneshwar, K., Rao, S., Madhavan, S., Petricoin, 3rd, E. F., and Hansel, D. E. (2019). Dynamic regulation of caveolin-1 phosphorylation and caveolae formation by mammalian target of rapamycin complex 2 in bladder cancer cells. *Am. J. Pathol.*, 189(9):1846–1862.

Hay, D., Hughes, J. R., Babbs, C., Davies, J. O. J., Graham, B. J., Hanssen, L., Kassouf, M. T., Marieke Oudelaar, A. M., Sharpe, J. A., Suciu, M. C., Telenius, J., Williams, R., Rode, C., Li, P.-S., Pennacchio, L. A., Sloane-Stanley, J. A., Ayyub, H., Butler, S., Sauka-Spengler, T., Gibbons, R. J., Smith, A. J. H., Wood, W. G., and Higgs, D. R. (2016). Genetic dissection of the $\alpha$-globin super-enhancer in vivo. *Nat. Genet.*, 48(8):895–903.

He, C., Sidoli, S., Warneford-Thomson, R., Tatomer, D. C., Wilusz, J. E., Garcia, B. A., and Bonasio, R. (2016). High-Resolution mapping of RNA-Binding regions in the nuclear proteome of embryonic stem cells. *Mol. Cell*, 64(2):416–430.

He, J., Ye, J., Cai, Y., Riquelme, C., Liu, J. O., Liu, X., Han, A., and Chen, L. (2011). Structure of p300 bound to MEF2 on DNA reveals a mechanism of enhanceosome assembly. *Nucleic Acids Res.*, 39(10):4464–4474.

Head, B. P., Hu, Y., Finley, J. C., Saldana, M. D., Bonds, J. A., Miyanohara, A., Niesman, I. R., Ali, S. S., Murray, F., Insel, P. A., Roth, D. M., Patel, H. H., and Patel,

P. M. (2011). Neuron-targeted caveolin-1 protein enhances signaling and promotes arborization of primary neurons. *J. Biol. Chem.*, 286(38):33310–33321.

Head, B. P., Patel, H. H., Tsutsumi, Y. M., Hu, Y., Mejia, T., Mora, R. C., Insel, P. A., Roth, D. M., Drummond, J. C., and Patel, P. M. (2008). Caveolin-1 expression is essential for N-methyl-D-aspartate receptor-mediated src and extracellular signal-regulated kinase 1/2 activation and protection of primary neurons from ischemic cell death. *FASEB J.*, 22(3):828–840.

Head, B. P., Peart, J. N., Panneerselvam, M., Yokoyama, T., Pearn, M. L., Niesman, I. R., Bonds, J. A., Schilling, J. M., Miyanohara, A., Headrick, J., Ali, S. S., Roth, D. M., Patel, P. M., and Patel, H. H. (2010). Loss of caveolin-1 accelerates neurodegeneration and aging. *PLoS One*, 5(12):e15697.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, 39(3):311–318.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589.

Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.*, 16(3):144–154.

Henninger, J. E., Oksuz, O., Shrinivas, K., Sagi, I., LeRoy, G., Zheng, M. M., Andrews, J. O., Zamudio, A. V., Lazaris, C., Hannett, N. M., Lee, T. I., Sharp, P. A., Cissé, I. I., Chakraborty, A. K., and Young, R. A. (2021). RNA-Mediated feedback control of transcriptional condensates. *Cell*, 184(1):207–225.e24.

Hentze, M. W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.*, 19(5):327–341.

Herzog, V. A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T. R., Wlotzka, W., von Haeseler, A., Zuber, J., and Ameres, S. L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods*, 14(12):1198–1204.

Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., and Gersbach, C. A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.*, 33(5):510–517.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.*, 106(23):9362–9367.

Hitz, B. C., Lee, J.-W., Jolanki, O., Kagda, M. S., Graham, K., Sud, P., Gabdank, I., Seth Strattan, J., Sloan, C. A., Dreszer, T., Rowe, L. D., Podduturi, N. R., Malladi, V. S., Chan, E. T., Davidson, J. M., Ho, M., Miyasato, S., Simison, M., Tanaka, F., Luo, Y., Whaling, I., Hong, E. L., Lee, B. T., Sandstrom, R., Rynes, E., Nelson, J., Nishida, A., Ingersoll, A., Buckley, M., Frerker, M., Kim, D. S., Boley, N., Trout, D., Dobin, A., Rahmanian, S., Wyman, D., Balderrama-Gutierrez, G., Reese, F., Durand, N. C., Dudchenko, O., Weisz, D., Rao, S. S. P., Blackburn, A., Gkountaroulis, D., Sadr, M., Olshansky, M., Eliaz, Y., Nguyen, D., Bochkov, I., Shamim, M. S., Mahajan, R., Aiden, E., Gingeras, T., Heath, S., Hirst, M., James Kent, W., Kundaje, A., Mortazavi, A., Wold, B., and Michael Cherry, J. (2023). The ENCODE uniform analysis pipelines.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., and Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–947.

Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., and Sharp, P. A. (2017). A phase separation model for transcriptional control. *Cell*, 169(1):13–23.

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., Reddy, J., Borges-Rivera, D., Lee, T. I., Jaenisch, R., Porteus, M. H., Dekker, J., and Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280):1454–1458.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125(2):167–188.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J. Y., Yehia, G., and Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Methods*, 10(2):133–139.

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.*, 99(6):1245–1260.

Hosokawa, H., Ungerbäck, J., Wang, X., Matsumoto, M., Nakayama, K. I., Cohen, S. M., Tanaka, T., and Rothenberg, E. V. (2018). Transcription factor PU.1 represses and activates gene expression in early T cells by redirecting partner transcription factor binding. *Immunity*, 48(6):1119–1134.e7.

Hou, Y., Guo, Y., Dong, S., and Yang, T. (2021). Novel roles of g-quadruplexes on enhancers in human chromatin.

Hsieh, C.-L., Fei, T., Chen, Y., Li, T., Gao, Y., Wang, X., Sun, T., Sweeney, C. J., Lee, G.-S. M., Chen, S., Balk, S. P., Liu, X. S., Brown, M., and Kantoff, P. W. (2014). Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proc. Natl. Acad. Sci. U. S. A.*, 111(20):7319–7324.

Hu, D., Garruss, A. S., Gao, X., Morgan, M. A., Cook, M., Smith, E. R., and Shilatifard, A. (2013). The mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.*, 20(9):1093–1097.

Hu, G., Schones, D. E., Cui, K., Ybarra, R., Northrup, D., Tang, Q., Gattinoni, L., Restifo, N. P., Huang, S., and Zhao, K. (2011). Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res.*, 21(10):1650–1658.

Hua, J. T., Ahmed, M., Guo, H., Zhang, Y., Chen, S., Soares, F., Lu, J., Zhou, S., Wang, M., Li, H., Larson, N. B., McDonnell, S. K., Patel, P. S., Liang, Y., Yao, C. Q., van der Kwast, T., Lupien, M., Feng, F. Y., Zoubeidi, A., Tsao, M.-S., Thibodeau, S. N., Boutros, P. C., and He, H. H. (2018). Risk SNP-Mediated Promoter-Enhancer switching drives prostate cancer through lncRNA PCAT19. *Cell*, 174(3):564–575.e18.

Huang, W. H., Nie, L. H., Zhang, L. J., Jing, L. P., Dong, F., Wang, M., Zhang, N., Liu, Y., Zhang, B. H., Chen, C., Lin, H. S., Wei, X. C., Yang, G., and Jing, C. X. (2015). Association of TLR2 and TLR4 non-missense single nucleotide polymorphisms with type 2 diabetes risk in a southern chinese population: a case-control study. *Genet. Mol. Res.*, 14(3):8694–8705.

Huang, Z., Du, G., Huang, X., Han, L., Han, X., Xu, B., Zhang, Y., Yu, M., Qin, Y., Xia, Y., Wang, X., and Lu, C. (2018). The enhancer RNA lnc-SLC4A1-1 epigenetically

regulates unexplained recurrent pregnancy loss (URPL) by activating CXCL8 and NF-kB pathway. *EBioMedicine*, 38:162–170.

Hujoel, M. L. A., Gazal, S., Hormozdiari, F., van de Geijn, B., and Price, A. L. (2019). Disease heritability enrichment of regulatory elements is concentrated in elements with ancient sequence age and conserved function across species. *Am. J. Hum. Genet.*, 104(4):611–624.

Humphreys, D. T., Westman, B. J., Martin, D. I. K., and Preiss, T. (2005). MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4e/cap and poly(a) tail function. *Proc. Natl. Acad. Sci. U. S. A.*, 102(47):16961–16966.

Hung, T.-C., Kingsley, D. M., and Boettiger, A. N. (2024). Boundary stacking interactions enable cross-TAD enhancer-promoter communication during limb development. *Nat. Genet.*, 56(2):306–314.

Ibáñez, P., Lesage, S., Janin, S., Lohmann, E., Durif, F., Destée, A., Bonnet, A.-M., Brefel-Courbon, C., Heath, S., Zelenika, D., Agid, Y., Dürr, A., Brice, A., and French Parkinson's Disease Genetics Study Group (2009). Alpha-synuclein gene rearrangements in dominantly inherited parkinsonism: frequency, phenotype, and mechanisms. *Arch. Neurol.*, 66(1):102–108.

Ibarra, I. L., Hollmann, N. M., Klaus, B., Augsten, S., Velten, B., Hennig, J., and Zaugg, J. B. (2020). Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Nat. Commun.*, 11(1):124.

Ibn-Salem, J., Muro, E. M., and Andrade-Navarro, M. A. (2017). Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic Acids Res.*, 45(1):81–91.

Ikezu, T., Trapp, B. D., Song, K. S., Schlegel, A., Lisanti, M. P., and Okamoto, T. (1998). Caveolae, plasma membrane microdomains for alpha-secretase-mediated processing of the amyloid precursor protein. *J. Biol. Chem.*, 273(17):10485–10495.

Isgrò, M. A., Bottoni, P., and Scatena, R. (2015). Neuron-specific enolase as a biomarker: Biochemical and clinical aspects. *Adv. Exp. Med. Biol.*, 867:125–143.

Jablonski, K. P., Carron, L., Mozziconacci, J., Forné, T., Hütt, M.-T., and Lesne, A. (2022). Contribution of 3D genome topological domains to genetic risk of cancers: a genome-wide computational study. *Hum. Genomics*, 16(1):2.

Jackson, D. A., Iborra, F. J., Manders, E. M., and Cook, P. R. (1998). Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. *Mol. Biol. Cell*, 9(6):1523–1536.

Jain, A. and Vale, R. D. (2017). RNA phase transitions in repeat expansion disorders. *Nature*, 546(7657):243–247.

Jeong, S. and Stein, A. (1994). Micrococcal nuclease digestion of nuclei reveals extended nucleosome ladders having anomalous DNA lengths for chromatin assembled on non-replicating plasmids in transfected cells. *Nucleic Acids Res.*, 22(3):370–375.

Jerber, J., Seaton, D. D., Cuomo, A. S. E., Kumasaka, N., Haldane, J., Steer, J., Patel, M., Pearce, D., Andersson, M., Bonder, M. J., Mountjoy, E., Ghoussaini, M., Lancaster, M. A., HipSci Consortium, Marioni, J. C., Merkle, F. T., Gaffney, D. J., and Stegle, O. (2021). Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.*, 53(3):304–312.

Ji, S., Yang, Z., Gozali, L., Kenney, T., Kocabas, A., Jinsook Park, C., and Hynes, M. (2021). Distinct expression of select and transcriptome-wide isolated 3'UTRs suggests critical roles in development and transition states. *PLoS One*, 16(5):e0250669.

Ji, X., Kong, J., and Liebhaber, S. A. (2011a). An RNA-protein complex links enhanced nuclear 3' processing with cytoplasmic mRNA stabilization. *EMBO J.*, 30(13):2622–2633.

Ji, Z., Lee, J. Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U. S. A.*, 106(17):7028–7033.

Ji, Z., Luo, W., Li, W., Hoque, M., Pan, Z., Zhao, Y., and Tian, B. (2011b). Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol. Syst. Biol.*, 7:534.

Jiang, Y., Krantz, S., Qin, X., Li, S., Gunasekara, H., Kim, Y.-M., Zimnicka, A., Bae, M., Ma, K., Toth, P. T., Hu, Y., Shajahan-Haq, A. N., Patel, H. H., Gentile, S., Bonini, M. G., Rehman, J., Liu, Y., and Minshall, R. D. (2022). Caveolin-1 controls mitochondrial damage and ROS production by regulating fission - fusion dynamics and mitophagy. *Redox Biol*, 52:102304.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502.

Juven-Gershon, T., Hsu, J.-Y., and Kadonaga, J. T. (2008). Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes Dev.*, 22(20):2823–2830.

Kaczmarek Michaels, K., Mohd Mostafa, S., Ruiz Capella, J., and Moore, C. L. (2020). Regulation of alternative polyadenylation in the yeast saccharomyces cerevisiae by histone H3K4 and H3K36 methyltransferases. *Nucleic Acids Res.*, 48(10):5407–5425.

Kadener, S., Fededa, J. P., Rosbash, M., and Kornblihtt, A. R. (2002). Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. *Proc. Natl. Acad. Sci. U. S. A.*, 99(12):8185–8190.

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., and Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435.

Kaikkonen, M. U. and Adelman, K. (2018). Emerging roles of Non-Coding RNA transcription. *Trends Biochem. Sci.*, 43(9):654–667.

Kaikkonen, M. U., Spann, N. J., Heinz, S., Romanoski, C. E., Allison, K. A., Stender, J. D., Chun, H. B., Tough, D. F., Prinjha, R. K., Benner, C., and Glass, C. K. (2013). Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol. Cell*, 51(3):310–325.

Kaneko, S., Son, J., Shen, S. S., Reinberg, D., and Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat. Struct. Mol. Biol.*, 20(11):1258–1264.

Kapoor, A., Hsu, W.-M., Wang, B.-J., Wu, G.-H., Lin, T.-Y., Lee, S.-J., Yen, C.-T., Liang, S.-M., and Liao, Y.-F. (2010). Caveolin-1 regulates $\gamma$-secretase-mediated A$\beta$PP processing by modulating spatial distribution of $\gamma$-secretase in membrane. *J. Alzheimers. Dis.*, 22(2):423–442.

Karch, C. M., Jeng, A. T., Nowotny, P., Cady, J., Cruchaga, C., and Goate, A. M. (2012). Expression of novel alzheimer's disease risk genes in control and alzheimer's disease brains. *PLoS One*, 7(11):e50976.

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A. E., Ristolainen, H., Hänninen, U. A., Cajuso, T., Kondelin, J., Tanskanen, T., Mecklin, J.-P., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Kaasinen, E., Kilpivaara, O., Tuupanen, S., Enge, M., Taipale, J., and Aaltonen, L. A. (2015). CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.*, 47(7):818–821.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, 7(12):1009–1015.

Kessler, S., Minoux, M., Joshi, O., Ben Zouari, Y., Ducret, S., Ross, F., Vilain, N., Salvi, A., Wolff, J., Kohler, H., Stadler, M. B., and Rijli, F. M. (2023). A multiple super-enhancer region establishes inter-TAD interactions and controls hoxa function in cranial neural crest. *Nat. Commun.*, 14(1):3242.

Kim, F., Pham, M., Luttrell, I., Bannerman, D. D., Tupper, J., Thaler, J., Hawn, T. R., Raines, E. W., and Schwartz, M. W. (2007). Toll-like receptor-4 mediates vascular inflammation and insulin resistance in diet-induced obesity. *Circ. Res.*, 100(11):1589–1596.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., and Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187.

Kim, T.-K. and Shiekhattar, R. (2016). Diverse regulatory interactions of long noncoding RNAs. *Curr. Opin. Genet. Dev.*, 36:73–82.

Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S. E., Hamel, A. R., Viñuela, A., Roberts, A. L., Mangul, S., Wen, X., Wang, G., Barbeira, A. N., Garrido-Martín, D., Nadel, B. B., Zou, Y., Bonazzola, R., Quan, J., Brown, A., Martinez-Perez, A., Soria, J. M., GTEx Consortium, Getz, G., Dermitzakis, E. T., Small, K. S., Stephens, M., Xi, H. S., Im, H. K., Guigó, R., Segrè, A. V., Stranger, B. E., Ardlie, K. G., and Lappalainen, T. (2020). Cell type-specific genetic regulation of gene expression across human tissues. *Science*, 369(6509).

King, H. W. and Klose, R. J. (2017). The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. *Elife*, 6.

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116.

Klattenhoff, C. A., Scheuermann, J. C., Surface, L. E., Bradley, R. K., Fields, P. A., Steinhauser, M. L., Ding, H., Butty, V. L., Torrey, L., Haas, S., Abo, R., Tabebordbar, M., Lee, R. T., Burge, C. B., and Boyer, L. A. (2013). Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell*, 152(3):570–583.

Klein, J. C., Keith, A., Agarwal, V., Durham, T., and Shendure, J. (2018). Functional characterization of enhancer evolution in the primate lineage. *Genome Biol.*, 19(1):99.

Knowland, D., Arac, A., Sekiguchi, K. J., Hsu, M., Lutz, S. E., Perrino, J., Steinberg, G. K., Barres, B. A., Nimmerjahn, A., and Agalliu, D. (2014). Stepwise recruitment of transcellular and paracellular pathways underlies blood-brain barrier breakdown in stroke. *Neuron*, 82(3):603–617.

Kocabas, A., Duarte, T., Kumar, S., and Hynes, M. A. (2015). Widespread differential expression of coding region and 3' UTR sequences in neurons and other tissues. *Neuron*, 88(6):1149–1156.

Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T. K., Zacarias-Cabeza, J., Spicuglia, S., de la Chapelle, A. L., Heidemann, M., Hintermair, C., Eick, D., Gut, I., Ferrier, P., and Andrau, J.-C. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.*, 18(8):956–963.

Kogo, H. and Fujimoto, T. (2000). Caveolin-1 isoforms are encoded by distinct mRNAs. identification of mouse caveolin-1 mRNA variants caused by alternative transcription initiation and splicing. *FEBS Lett.*, 465(2-3):119–123.

Kolz, M., Baumert, J., Müller, M., Khuseyinova, N., Klopp, N., Thorand, B., Meisinger, C., Herder, C., Koenig, W., and Illig, T. (2008). Association between variations in the TLR4gene and incident type 2 diabetes is modified by the ratio of total cholesterol to HDL-cholesterol. *BMC Med. Genet.*, 9(1):1–12.

Kornberg, R. D. (2005). Mediator and the mechanism of transcriptional activation. *Trends Biochem. Sci.*, 30(5):235–239.

Korneev, K. V., Sviriaeva, E. N., Mitkin, N. A., Gorbacheva, A. M., Uvarova, A. N., Ustiugova, A. S., Polanovsky, O. L., Kulakovskiy, I. V., Afanasyeva, M. A., Schwartz, A. M., and Kuprash, D. V. (2020). Minor C allele of the SNP rs7873784 associated with rheumatoid arthritis and type-2 diabetes mellitus binds PU.1 and enhances TLR4 expression. *Biochim. Biophys. Acta Mol. Basis Dis.*, 1866(3):165626.

Kosicki, M., Tomberg, K., and Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.*, 36(8):765–771.

Kovalevich, J. and Langford, D. (2013). Considerations for the use of SH-SY5Y neuroblastoma cells in neurobiology. *Methods Mol. Biol.*, 1078:9–21.

Kristjánsdóttir, K., Dziubek, A., Kang, H. M., and Kwak, H. (2020). Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. *Nat. Commun.*, 11(1):5963.

Krohn, L., Wu, R. Y. J., Heilbron, K., Ruskey, J. A., Laurent, S. B., Blauwendraat, C., Alam, A., Arnulf, I., Hu, M. T. M., Dauvilliers, Y., Högl, B., Toft, M., Bjørnarå, K. A., Stefani, A., Holzknecht, E., Monaca, C. C., Abril, B., Plazzi, G., Antelmi, E., Ferini-Strambi, L., Young, P., Heidbreder, A., Cochen De Cock, V., Mollenhauer, B., Sixel-Döring, F., Trenkwalder, C., Sonka, K., Kemlink, D., Figorilli, M., Puligheddu,

M., Dijkstra, F., Viaene, M., Oertel, W., Toffoli, M., Gigli, G. L., Valente, M., Gagnon, J.-F., Nalls, M. A., Singleton, A. B., 23andMe Research Team, Desautels, A., Montplaisir, J. Y., Cannon, P., Ross, O. A., Boeve, B. F., Dupré, N., Fon, E. A., Postuma, R. B., Pihlstrøm, L., Rouleau, G. A., and Gan-Or, Z. (2020). Fine-Mapping of SNCA in rapid eye movement sleep behavior disorder and overt synucleinopathies. *Ann. Neurol.*, 87(4):584–598.

Krumbiegel, M., Pasutto, F., Mardin, C. Y., Weisschuh, N., Paoli, D., Gramer, E., Zenkel, M., Weber, B. H. F., Kruse, F. E., Schlötzer-Schrehardt, U., and Reis, A. (2009). Exploring functional candidate genes for genetic association in german patients with pseudoexfoliation syndrome and pseudoexfoliation glaucoma. *Invest. Ophthalmol. Vis. Sci.*, 50(6):2796–2801.

Kruys, V., Marinx, O., Shaw, G., Deschamps, J., and Huez, G. (1989). Translational blockade imposed by cytokine-derived UA-rich sequences. *Science*, 245(4920):852–855.

Kühn, U., Gündel, M., Knoth, A., Kerwitz, Y., Rüdel, S., and Wahle, E. (2009). Poly(A) tail length is controlled by the nuclear poly(a)-binding protein regulating the interaction between poly(a) polymerase and the cleavage and polyadenylation specificity factor. *J. Biol. Chem.*, 284(34):22803–22814.

Kundu, P., Fabian, M. R., Sonenberg, N., Bhattacharyya, S. N., and Filipowicz, W. (2012). HuR protein attenuates miRNA-mediated repression by promoting miRISC dissociation from the target RNA. *Nucleic Acids Res.*, 40(11):5088–5100.

Kuznetsov, S. V., Ren, C.-C., Woodson, S. A., and Ansari, A. (2008). Loop dependence of the stability and dynamics of nucleic acid hairpins. *Nucleic Acids Res.*, 36(4):1098–1112.

Kuznetsov, V. A., Bondarenko, V., Wongsurawat, T., Yenamandra, S. P., and Jenjaroenpun, P. (2018). Toward predictive r-loop computational biology: genome-scale prediction of r-loops reveals their association with complex promoter structures, g-quadruplexes and transcriptionally active enhancers. *Nucleic Acids Res.*, 46(15):7566–7585.

Kvon, E. Z., Waymack, R., Gad, M., and Wunderlich, Z. (2021). Enhancer redundancy in development and disease. *Nat. Rev. Genet.*, 22(5):324–336.

Kwon, B., Fansler, M. M., Patel, N. D., Lee, J., Ma, W., and Mayr, C. (2022). Enhancers regulate 3' end processing activity to control expression of alternative 3'UTR isoforms. *Nat. Commun.*, 13(1):2709.

Lagier-Tourenne, C., Polymenidou, M., Hutt, K. R., Vu, A. Q., Baughn, M., Huelga, S. C., Clutario, K. M., Ling, S.-C., Liang, T. Y., Mazur, C., Wancewicz, E., Kim,

A. S., Watt, A., Freier, S., Hicks, G. G., Donohue, J. P., Shiue, L., Bennett, C. F., Ravits, J., Cleveland, D. W., and Yeo, G. W. (2012). Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat. Neurosci.*, 15(11):1488–1497.

Lai, F., Gardini, A., Zhang, A., and Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. *Nature*, 525(7569):399–403.

Lai, F., Orom, U. A., Cesaroni, M., Beringer, M., Taatjes, D. J., Blobel, G. A., and Shiekhattar, R. (2013). Activating RNAs associate with mediator to enhance chromatin architecture and transcription. *Nature*, 494(7438):497–501.

Lai, W. S., Carballo, E., Thorn, J. M., Kennington, E. A., and Blackshear, P. J. (2000). Interactions of CCCH zinc finger proteins with mRNA. binding of tristetraprolin-related zinc finger proteins to au-rich elements and destabilization of mRNA. *J. Biol. Chem.*, 275(23):17827–17837.

Lam, E. W.-F., Brosens, J. J., Gomes, A. R., and Koo, C.-Y. (2013). Forkhead box proteins: tuning forks for transcriptional harmony. *Nat. Rev. Cancer*, 13(7):482–495.

Lambert, J.-C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O., Zelenika, D., Bullido, M. J., Tavernier, B., Letenneur, L., Bettens, K., Berr, C., Pasquier, F., Fiévet, N., Barberger-Gateau, P., Engelborghs, S., De Deyn, P., Mateo, I., Franck, A., Helisalmi, S., Porcellini, E., Hanon, O., European Alzheimer's Disease Initiative Investigators, de Pancorbo, M. M., Lendon, C., Dufouil, C., Jaillard, C., Leveillard, T., Alvarez, V., Bosco, P., Mancuso, M., Panza, F., Nacmias, B., Bossù, P., Piccardi, P., Annoni, G., Seripa, D., Galimberti, D., Hannequin, D., Licastro, F., Soininen, H., Ritchie, K., Blanché, H., Dartigues, J.-F., Tzourio, C., Gut, I., Van Broeckhoven, C., Alpérovitch, A., Lathrop, M., and Amouyel, P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with alzheimer's disease. *Nat. Genet.*, 41(10):1094–1099.

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., DeStafano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thorton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram, M. A., Zelenika, D., Vardarajan, B. N., Kamatani, Y., Lin, C. F., Gerrish, A., Schmidt, H., Kunkle, B., Dunstan, M. L., Ruiz, A., Bihoreau, M. T., Choi, S. H., Reitz, C., Pasquier, F., Cruchaga, C., Craig, D., Amin, N., Berr, C., Lopez, O. L., De Jager, P. L., Deramecourt, V., Johnston, J. A., Evans, D., Lovestone, S., Letenneur, L., Morón, F. J., Rubinsztein, D. C., Eiriksdottir, G., Sleegers, K., Goate, A. M., Fiévet, N., Huentelman, M. W., Gill, M., Brown, K., Kamboh, M. I., Keller, L., Barberger-Gateau, P., McGuiness, B., Larson, E. B., Green, R., Myers, A. J., Dufouil, C.,

Todd, S., Wallon, D., Love, S., Rogaeva, E., Gallacher, J., St George-Hyslop, P., Clarimon, J., Lleo, A., Bayer, A., Tsuang, D. W., Yu, L., Tsolaki, M., Bossù, P., Spalletta, G., Proitsi, P., Collinge, J., Sorbi, S., Sanchez-Garcia, F., Fox, N. C., Hardy, J., Deniz Naranjo, M. C., Bosco, P., Clarke, R., Brayne, C., Galimberti, D., Mancuso, M., Matthews, F., European Alzheimer's Disease Initiative (EADI), Genetic and Environmental Risk in Alzheimer's Disease, Alzheimer's Disease Genetic Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology, Moebus, S., Mecocci, P., Del Zompo, M., Maier, W., Hampel, H., Pilotto, A., Bullido, M., Panza, F., Caffarra, P., Nacmias, B., Gilbert, J. R., Mayhaus, M., Lannefelt, L., Hakonarson, H., Pichler, S., Carrasquillo, M. M., Ingelsson, M., Beekly, D., Alvarez, V., Zou, F., Valladares, O., Younkin, S. G., Coto, E., Hamilton-Nelson, K. L., Gu, W., Razquin, C., Pastor, P., Mateo, I., Owen, M. J., Faber, K. M., Jonsson, P. V., Combarros, O., O'Donovan, M. C., Cantwell, L. B., Soininen, H., Blacker, D., Mead, S., Mosley, Jr, T. H., Bennett, D. A., Harris, T. B., Fratiglioni, L., Holmes, C., de Bruijn, R. F., Passmore, P., Montine, T. J., Bettens, K., Rotter, J. I., Brice, A., Morgan, K., Foroud, T. M., Kukull, W. A., Hannequin, D., Powell, J. F., Nalls, M. A., Ritchie, K., Lunetta, K. L., Kauwe, J. S., Boerwinkle, E., Riemenschneider, M., Boada, M., Hiltuenen, M., Martin, E. R., Schmidt, R., Rujescu, D., Wang, L. S., Dartigues, J. F., Mayeux, R., Tzourio, C., Hofman, A., Nöthen, M. M., Graff, C., Psaty, B. M., Jones, L., Haines, J. L., Holmans, P. A., Lathrop, M., Pericak-Vance, M. A., Launer, L. J., Farrer, L. A., van Duijn, C. M., Van Broeckhoven, C., Moskvina, V., Seshadri, S., Williams, J., Schellenberg, G. D., and Amouyel, P. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nat. Genet.*, 45(12):1452–1458.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, 172(4):650–665.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra,

M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Landolin, J. M., Johnson, D. S., Trinklein, N. D., Aldred, S. F., Medina, C., Shulha, H., Weng, Z., and Myers, R. M. (2010). Sequence features that drive human promoter function and tissue specificity. *Genome Res.*, 20(7):890–898.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J.,

Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, 22(9):1813–1831.

Langdon, E. M., Qiu, Y., Ghanbari Niaki, A., McLaughlin, G. A., Weidmann, C. A., Gerbich, T. M., Smith, J. A., Crutchley, J. M., Termini, C. M., Weeks, K. M., Myong, S., and Gladfelter, A. S. (2018). mRNA structure determines specificity of a polyq-driven phase separation. *Science*, 360(6391):922–927.

Lasko, L. M., Jakob, C. G., Edalji, R. P., Qiu, W., Montgomery, D., Digiammarino, E. L., Hansen, T. M., Risi, R. M., Frey, R., Manaves, V., Shaw, B., Algire, M., Hessler, P., Lam, L. T., Uziel, T., Faivre, E., Ferguson, D., Buchanan, F. G., Martin, R. L., Torrent, M., Chiang, G. G., Karukurichi, K., Langston, J. W., Weinert, B. T., Choudhary, C., de Vries, P., Van Drie, J. H., McElligott, D., Kesicki, E., Marmorstein, R., Sun, C., Cole, P. A., Rosenberg, S. H., Michaelides, M. R., Lai, A., and Bromberg, K. D. (2017). Discovery of a selective catalytic p300/CBP inhibitor that targets lineage-specific tumours. *Nature*, 550(7674):128–132.

Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–1842.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, 9(8):e1003118.

Le Lay, S., Briand, N., Blouin, C. M., Chateau, D., Prado, C., Lasnier, F., Le Liepvre, X., Hajduch, E., and Dugail, I. (2010). The lipoatrophic caveolin-1 deficient mouse model reveals autophagy in mature adipocytes. *Autophagy*, 6(6):754–763.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P., and Krause, H. M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 131(1):174–187.

Lee, H., Park, D. S., Razani, B., Russell, R. G., Pestell, R. G., and Lisanti, M. P. (2002). Caveolin-1 mutations (P132L and null) and the pathogenesis of breast cancer: caveolin-1 (P132L) behaves in a dominant-negative manner and caveolin-1 (-/-) null mice show mammary epithelial cell hyperplasia. *Am. J. Pathol.*, 161(4):1357–1369.

Lee, H., Volonte, D., Galbiati, F., Iyengar, P., Lublin, D. M., Bregman, D. B., Wilson, M. T., Campos-Gonzalez, R., Bouzahzah, B., Pestell, R. G., Scherer, P. E., and Lisanti, M. P. (2000). Constitutive and growth factor-regulated phosphorylation of caveolin-1 occurs at the same site (tyr-14) in vivo: identification of a c-Src/Cav-1/Grb7 signaling cassette. *Mol. Endocrinol.*, 14(11):1750–1775.

Lee, J.-H., Wang, R., Xiong, F., Krakowiak, J., Liao, Z., Nguyen, P. T., Moroz-Omori, E. V., Shao, J., Zhu, X., Bolt, M. J., Wu, H., Singh, P. K., Bi, M., Shi, C. J., Jamal, N., Li, G., Mistry, R., Jung, S. Y., Tsai, K.-L., Ferreon, J. C., Stossi, F., Caflisch, A., Liu, Z., Mancini, M. A., and Li, W. (2021). Enhancer RNA m6a methylation facilitates transcriptional condensate formation and gene activation. *Mol. Cell*.

Lee, Y. S. and Dutta, A. (2007). The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes Dev.*, 21(9):1025–1030.

Legendre, M. and Gautheret, D. (2003). Sequence determinants in human polyadenylation site selection. *BMC Genomics*, 4(1):7.

Leontis, N. B. and Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4):499.

Leontis, N. B. and Westhof, E. (2003). Analysis of RNA motifs. *Curr. Opin. Struct. Biol.*, 13(3):300–308.

Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.*, 12(14):1725–1735.

Lettice, L. A., Horikoshi, T., Heaney, S. J. H., van Baren, M. J., van der Linde, H. C., Breedveld, G. J., Joosse, M., Akarsu, N., Oostra, B. A., Endo, N., Shibata, M., Suzuki, M., Takahashi, E., Shinka, T., Nakahori, Y., Ayusawa, D., Nakabayashi, K., Scherer, S. W., Heutink, P., Hill, R. E., and Noji, S. (2002). Disruption of a long-range cis-acting regulator for shh causes preaxial polydactyly. *Proc. Natl. Acad. Sci. U. S. A.*, 99(11):7548–7553.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.-L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K. I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M. J., Cheung, E., Liu, E., Sung, W.-K., Snyder, M., and Ruan, Y. (2012a). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98.

Li, K., Qiu, C., Sun, P., Liu, D.-C., Wu, T.-J., Wang, K., Zhou, Y.-C., Chang, X.-A., Yin, Y., Chen, F., Zhu, Y.-X., and Han, X. (2019). Ets1-mediated acetylation of FoxO1 is critical for gluconeogenesis regulation during feed-fast cycles. *Cell Rep.*, 26(11):2998–3010.e5.

Li, L., Huang, K.-L., Gao, Y., Cui, Y., Wang, G., Elrod, N. D., Li, Y., Chen, Y. E., Ji, P., Peng, F., Russell, W. K., Wagner, E. J., and Li, W. (2021). An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat. Genet.*, 53(7):994–1005.

Li, S., Seitz, R., and Lisanti, M. P. (1996). Phosphorylation of caveolin by src tyrosine kinases: THE $\alpha$-ISOFORM OF CAVEOLIN IS SELECTIVELY PHOSPHORYLATED BY v-src IN VIVO. *J. Biol. Chem.*, 271(7):3863–3868.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A. Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., Oh, S., Kim, H.-S., Glass, C. K., and Rosenfeld, M. G. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455):516–520.

Li, X. and Manley, J. L. (2005). Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell*, 122(3):365–378.

Li, Y., Sun, Y., Fu, Y., Li, M., Huang, G., Zhang, C., Liang, J., Huang, S., Shen, G., Yuan, S., Chen, L., Chen, S., and Xu, A. (2012b). Dynamic landscape of tandem 3' UTRs during zebrafish development. *Genome Res.*, 22(10):1899–1906.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., and Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604.

Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., and Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, 27(21):2380–2396.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Lidschreiber, K., Jung, L. A., von der Emde, H., Dave, K., Taipale, J., Cramer, P., and Lidschreiber, M. (2021). Transcriptionally active enhancers in human cancer cells. *Mol. Syst. Biol.*, 17(1):e9873.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny,

L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.

Lim, G. and Hohng, S. (2020). Single-molecule fluorescence studies on cotranscriptional g-quadruplex formation coupled with r-loop formation. *Nucleic Acids Res.*, 48(16):9195–9203.

Lin, B., Nasir, J., MacDonald, H., Hutchinson, G., Graham, R. K., Rommens, J. M., and Hayden, M. R. (1994). Sequence of the murine huntington disease gene: evidence for conservation, alternate splicing and polymorphism in a triplet (CCG) repeat [corrected]. *Hum. Mol. Genet.*, 3(1):85–92.

Lin, B., Rommens, J. M., Graham, R. K., Kalchman, M., MacDonald, H., Nasir, J., Delaney, A., Goldberg, Y. P., and Hayden, M. R. (1993). Differential 3' polyadenylation of the huntington disease gene results in two mRNA species with variable tissue expression. *Hum. Mol. Genet.*, 2(10):1541–1545.

Lindhorst, D. and Halfon, M. S. (2023). Reporter gene assays and chromatin-level assays define substantially non-overlapping sets of enhancer sequences. *BMC Genomics*, 24(1):17.

Liu, J. and Wang, F. (2017). Role of neuroinflammation in amyotrophic lateral sclerosis: Cellular mechanisms and therapeutic implications. *Front. Immunol.*, 8:1005.

Liu, L., Christodoulou-Vafeiadou, E., Rao, J. N., Zou, T., Xiao, L., Chung, H. K., Yang, H., Gorospe, M., Kontoyiannis, D., and Wang, J.-Y. (2014). RNA-binding protein HuR promotes growth of small intestinal mucosa by activating the wnt signaling pathway. *Mol. Biol. Cell*, 25(21):3308–3318.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. *Cell*, 165(3):535–550.

Liu, Y., Shang, G., Zhang, X., Liu, F., Zhang, C., Li, Z., Jia, J., Xu, Y., Zhang, Z., Yang, S., Zhou, B., Luan, Y., Huang, Y., Peng, Y., Han, T., He, Y., and Zheng, H. (2022). CAMTA1 gene affects the ischemia-reperfusion injury by regulating CCND1. *Front. Cell. Neurosci.*, 16:868291.

Liu, Y., Wan, X., Li, H., Chen, Y., Hu, X., Chen, H., Zhu, D., Li, C., and Zhang, Y. (2023). CTCF coordinates cell fate specification via orchestrating regulatory hubs with pioneer transcription factors. *Cell Rep.*, 42(10):113259.

Long, H. K., Osterwalder, M., Welsh, I. C., Hansen, K., Davies, J. O. J., Liu, Y. E., Koska, M., Adams, A. T., Aho, R., Arora, N., Ikeda, K., Williams, R. M., Sauka-Spengler, T., Porteus, M. H., Mohun, T., Dickel, D. E., Swigut, T., Hughes, J. R.,

Higgs, D. R., Visel, A., Selleri, L., and Wysocka, J. (2020). Loss of extreme Long-Range enhancers in human neural crest drives a craniofacial disorder. *Cell Stem Cell*, 27(5):765–783.e14.

Long, H. K., Prescott, S. L., and Wysocka, J. (2016). Ever-Changing landscapes: Transcriptional enhancers in development and evolution. *Cell*, 167(5):1170–1187.

Long, H. S., Greenaway, S., Powell, G., Mallon, A.-M., Lindgren, C. M., and Simon, M. M. (2022). Making sense of the linear genome, gene function and TADs. *Epigenetics Chromatin*, 15(1):4.

Lopes, F. M., Schröder, R., da Frota, Jr, M. L. C., Zanotto-Filho, A., Müller, C. B., Pires, A. S., Meurer, R. T., Colpo, G. D., Gelain, D. P., Kapczinski, F., Moreira, J. C. F., Fernandes, M. d. C., and Klamt, F. (2010). Comparison between proliferative and neuron-like SH-SY5Y cells as an in vitro model for parkinson disease studies. *Brain Res.*, 1337:85–94.

Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Grüning, B., Ramírez, F., and Manke, T. (2021). pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics*, 37(3):422–423.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms Mol. Biol.*, 6:26.

Lorenz, R., Wolfinger, M. T., Tanzer, A., and Hofacker, I. L. (2016). Predicting RNA secondary structures from sequence and probing data. *Methods*, 103:86–98.

Lu, H., Yu, D., Hansen, A. S., Ganguly, S., Liu, R., Heckert, A., Darzacq, X., and Zhou, Q. (2018). Phase-separation mechanism for c-terminal hyperphosphorylation of RNA polymerase II. *Nature*, 558(7709):318–323.

Lu, J., McKinsey, T. A., Zhang, C. L., and Olson, E. N. (2000). Regulation of skeletal myogenesis by association of the MEF2 transcription factor with class II histone deacetylases. *Mol. Cell*, 6(2):233–244.

Lu, X., Zhu, X., Li, Y., Liu, M., Yu, B., Wang, Y., Rao, M., Yang, H., Zhou, K., Wang, Y., Chen, Y., Chen, M., Zhuang, S., Chen, L.-F., Liu, R., and Chen, R. (2016). Multiple P-TEFbs cooperatively regulate the release of promoter-proximally paused RNA polymerase II. *Nucleic Acids Res.*, 44(14):6853–6867.

Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J.-W., Tanaka, F. Y., Adenekan, P., O'Neill, E.,

and Cherry, J. M. (2020). New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Res.*, 48(D1):D882–D889.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025.

Lutz, S. E., Smith, J. R., Kim, D. H., Olson, C. V. L., Ellefsen, K., Bates, J. M., Gandhi, S. P., and Agalliu, D. (2017). Caveolin1 is required for th1 cell infiltration, but not tight junction remodeling, at the Blood-Brain barrier in autoimmune neuroinflammation. *Cell Rep.*, 21(8):2104–2117.

Lykke-Andersen, J. and Wagner, E. (2005a). Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. *Genes Dev.*, 19(3):351–361.

Lykke-Andersen, J. and Wagner, E. (2005b). Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. *Genes Dev.*, 19(3):351–361.

Lykke-Andersen, S., Žumer, K., Molska, E. Š., Rouvière, J. O., Wu, G., Demel, C., Schwalb, B., Schmid, M., Cramer, P., and Jensen, T. H. (2021). Integrator is a genome-wide attenuator of non-productive transcription. *Mol. Cell*, 81(3):514–529.e6.

Ma, L., Gao, Z., Wu, J., Zhong, B., Xie, Y., Huang, W., and Lin, Y. (2021). Co-condensation between transcription factor and coactivator p300 modulates transcriptional bursting kinetics. *Mol. Cell*, 81(8):1682–1697.e7.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N., MacFarlane, H., Jenkins, B., Anderson, M. A., Wexler, N. S., Gusella, J. F., Bates, G. P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A.-M., Lehrach, H., Buckler, A. J., Church, D., Doucette-Stamm, L., O'Donovan, M. C., Riba-Ramirez, L., Shah, M., Stanton, V. P., Strobel, S. A., Draths, K. M., Wales, J. L., Dervan, P., Housman, D. E., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J. J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F. S., Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D., and Harper, P. S. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell*, 72(6):971–983.

Maizels, N. and Gray, L. T. (2013). The G4 genome. *PLoS Genet.*, 9(4):e1003468.

Makrides, S. C. (2003). Vectors for gene expression in mammalian cells. In *Gene Transfer and Expression in Mammalian Cells*, volume 38 of *New comprehensive biochemistry*, pages 9–26. Elsevier.

Maniatis, N. A., Chernaya, O., Shinin, V., and Minshall, R. D. (2012). Caveolins and lung function. *Adv. Exp. Med. Biol.*, 729:157–179.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Mansisidor, A. R. and Risca, V. I. (2022). Chromatin accessibility: methods, mechanisms, and biological insights. *Nucleus*, 13(1):236–276.

Mao, S.-Q., Ghanbarian, A. T., Spiegel, J., Martínez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hänsel-Hertsch, R., Tannahill, D., and Balasubramanian, S. (2018). DNA g-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol.*, 25(10):951–957.

Mao, Z., Bonni, A., Xia, F., Nadal-Vicens, M., and Greenberg, M. E. (1999). Neuronal activity-dependent cell survival mediated by transcription factor MEF2. *Science*, 286(5440):785–790.

Mariella, E., Marotta, F., Grassi, E., Gilotto, S., and Provero, P. (2019). The length of the expressed 3' UTR is an intermediate molecular phenotype linking genetic variants to complex diseases. *Front. Genet.*, 10:714.

Martinez-Outschoorn, U. E., Whitaker-Menezes, D., Lin, Z., Flomenberg, N., Howell, A., Pestell, R. G., Lisanti, M. P., and Sotgia, F. (2011). Cytokine production and inflammation drive autophagy in the tumor microenvironment: role of stromal caveolin-1 as a key regulator. *Cell Cycle*, 10(11):1784–1793.

Mashal, R. D., Koontz, J., and Sklar, J. (1995). Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases. *Nat. Genet.*, 9(2):177–183.

Mastick, C. C., Brady, M. J., and Saltiel, A. R. (1995). Insulin stimulates the tyrosine phosphorylation of caveolin. *J. Cell Biol.*, 129(6):1523–1531.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U. S. A.*, 101(19):7287–7292.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., and Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195.

Mayr, C. (2017). Regulation by 3'-untranslated regions. *Annu. Rev. Genet.*, 51:171–194.

Mayr, C. (2019). What are 3' UTRs doing? *Cold Spring Harb. Perspect. Biol.*, 11(10).

Mayr, C. and Bartel, D. P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–684.

Mayr, C., Hemann, M. T., and Bartel, D. P. (2007). Disrupting the pairing between let-7 and hmga2 enhances oncogenic transformation. *Science*, 315(5818):1576–1579.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119.

McShea, D. W. (1996). PERSPECTIVE METAZOAN COMPLEXITY AND EVOLUTION: IS THERE a TREND? *Evolution*, 50(2):477–492.

Meijlink, F., Curran, T., Miller, A. D., and Verma, I. M. (1985). Removal of a 67-base-pair sequence in the noncoding region of protooncogene fos converts it to a transforming gene. *Proc. Natl. Acad. Sci. U. S. A.*, 82(15):4987–4991.

Meng, F., Saxena, S., Liu, Y., Joshi, B., Wong, T. H., Shankar, J., Foster, L. J., Bernatchez, P., and Nabi, I. R. (2017). The phospho-caveolin-1 scaffolding domain dampens force fluctuations in focal adhesions and promotes cancer cell migration. *Mol. Biol. Cell*, 28(16):2190–2201.

Mercer, T. R., Wilhelm, D., Dinger, M. E., Soldà, G., Korbie, D. J., Glazov, E. A., Truong, V., Schwenke, M., Simons, C., Matthaei, K. I., Saint, R., Koopman, P., and Mattick, J. S. (2011). Expression of distinct RNAs from 3' untranslated regions. *Nucleic Acids Res.*, 39(6):2393–2403.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*, 149(7):1635–1646.

Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Dharia, N. V., Montgomery, P. G., Cowley, G. S., Pantel, S., Goodale, A., Lee, Y., Ali, L. D., Jiang, G., Lubonja, R., Harrington, W. F., Strickland, M., Wu, T., Hawes, D. C., Zhivich, V. A., Wyatt, M. R., Kalani, Z., Chang, J. J., Okamoto, M., Stegmaier, K., Golub, T. R., Boehm, J. S., Vazquez, F., Root, D. E., Hahn, W. C., and Tsherniak, A. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.*, 49(12):1779–1784.

Miao, L., Tang, Y., Bonneau, A. R., Chan, S. H., Kojima, M. L., Pownall, M. E., Vejnar, C. E., Gao, F., Krishnaswamy, S., Hendry, C. E., and Giraldez, A. J. (2022). The landscape of pioneer factor activity reveals the mechanisms of chromatin reprogramming and genome activation. *Mol. Cell*, 82(5):986–1002.e9.

Michael, A. K., Grand, R. S., Isbel, L., Cavadini, S., Kozicka, Z., Kempf, G., Bunker, R. D., Schenk, A. D., Graff-Meyer, A., Pathare, G. R., Weiss, J., Matsumoto, S., Burger, L., Schübeler, D., and Thomä, N. H. (2020). Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science*, 368(6498):1460–1465.

Michel, M., Demel, C., Zacher, B., Schwalb, B., Krebs, S., Blum, H., Gagneur, J., and Cramer, P. (2017). TT-seq captures enhancer landscapes immediately after T-cell stimulation. *Mol. Syst. Biol.*, 13(3):920.

Mieczkowski, J., Cook, A., Bowman, S. K., Mueller, B., Alver, B. H., Kundu, S., Deaton, A. M., Urban, J. A., Larschan, E., Park, P. J., Kingston, R. E., and Tolstorukov, M. Y. (2016). MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.*, 7:11485.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560.

Milin, A. N. and Deniz, A. A. (2018). Reentrant phase transitions and Non-Equilibrium dynamics in membraneless organelles. *Biochemistry*, 57(17):2470–2477.

Miller, J. A. and Widom, J. (2003). Collaborative competition mechanism for gene activation in vivo. *Mol. Cell. Biol.*, 23(5):1623–1632.

Mirny, L. A. (2010). Nucleosome-mediated cooperativity between transcription factors. *Proc. Natl. Acad. Sci. U. S. A.*, 107(52):22534–22539.

Mirsky, A. E. and Ris, H. (1951). The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physiol.*, 34(4):451–462.

Miska, E. A., Karlsson, C., Langley, E., Nielsen, S. J., Pines, J., and Kouzarides, T. (1999). HDAC4 deacetylase associates with and represses the MEF2 transcription factor. *EMBO J.*, 18(18):5099–5107.

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O., and Lai, E. C. (2013). Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res.*, 23(5):812–825.

Molinoff, P. B. and Axelrod, J. (1971). Biochemistry of catecholamines. *Annu. Rev. Biochem.*, 40(1):465–500.

Monk, D., Arnaud, P., Frost, J. M., Wood, A. J., Cowley, M., Martin-Trujillo, A., Guillaumet-Adkins, A., Iglesias Platas, I., Camprubi, C., Bourc'his, D., Feil, R., Moore, G. E., and Oakey, R. J. (2011). Human imprinted retrogenes exhibit non-canonical imprint chromatin signatures and reside in non-imprinted host genes. *Nucleic Acids Res.*, 39(11):4577–4586.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). A regulatory archipelago controls hox genes transcription in digits. *Cell*, 147(5):1132–1145.

Moore, G. P. (1984). The C-Value paradox. *Bioscience*, 34(7):425–429.

Moore, M. J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. *Science*, 309(5740):1514–1518.

Morales-Paytuví, F., Fajardo, A., Ruiz-Mirapeix, C., Rae, J., Tebar, F., Bosch, M., Enrich, C., Collins, B. M., Parton, R. G., and Pol, A. (2023). Early proteostasis of caveolins synchronizes trafficking, degradation, and oligomerization to prevent toxic aggregation. *J. Cell Biol.*, 222(9).

Morrison, O. and Thakur, J. (2021). Molecular complexes at euchromatin, heterochromatin and centromeric chromatin. *Int. J. Mol. Sci.*, 22(13).

Mostafavi, H., Spence, J. P., Naqvi, S., and Pritchard, J. K. (2023). Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.*, 55(11):1866–1875.

Mousavi, K., Zare, H., Dell'Orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G. L., and Sartorelli, V. (2013). ERNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol. Cell*, 51(5).

Moyo, M. B., Parker, J. B., and Chakravarti, D. (2020). Altered chromatin landscape and enhancer engagement underlie transcriptional dysregulation in MED12 mutant uterine leiomyomas. *Nat. Commun.*, 11(1):1019.

Mu, Z., Wei, W., Fair, B., Miao, J., Zhu, P., and Li, Y. I. (2021). The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biol.*, 22(1):122.

Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Ascano, Jr, M., Tuschl, T., Ohler, U., and Keene, J. D. (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell*, 43(3):327–339.

Nadel, J., Athanasiadou, R., Lemetre, C., Wijetunga, N. A., Ó Broin, P., Sato, H., Zhang, Z., Jeddeloh, J., Montagna, C., Golden, A., Seoighe, C., and Greally, J. M. (2015). RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics Chromatin*, 8:46.

Nagaswamy, U., Larios-Sanz, M., Hury, J., Collins, S., Zhang, Z., Zhao, Q., and Fox, G. E. (2002). NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res.*, 30(1):395.

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., DeStefano, A. L., Kara, E., Bras, J., Sharma, M., Schulte, C., Keller, M. F., Arepalli, S., Letson, C., Edsall, C., Stefansson, H., Liu, X., Pliner, H., Lee, J. H., Cheng, R., International Parkinson's Disease Genomics Consortium (IPDGC), Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI), 23andMe, GenePD, NeuroGenetics Research Consortium (NGRC), Hussman Institute of Human Genomics (HIHG), Ashkenazi Jewish Dataset Investigator, Cohorts for Health and Aging Research in Genetic Epidemiology (CHARGE), North American Brain Expression Consortium (NABEC), United Kingdom Brain Expression Consortium (UKBEC), Greek Parkinson's Disease Consortium, Alzheimer Genetic Analysis Group, Ikram, M. A., Ioannidis, J. P. A., Hadjigeorgiou, G. M., Bis, J. C., Martinez, M., Perlmutter, J. S., Goate, A., Marder, K., Fiske, B., Sutherland, M., Xiromerisiou, G., Myers, R. H., Clark, L. N., Stefansson, K., Hardy, J. A., Heutink, P., Chen, H., Wood, N. W., Houlden, H., Payami, H., Brice, A., Scott, W. K., Gasser, T., Bertram, L., Eriksson, N., Foroud, T., and Singleton, A. B. (2014). Large-scale meta-analysis of

genome-wide association data identifies six new risk loci for parkinson's disease. *Nat. Genet.*, 46(9):989–993.

Nardone, G., George, J., and Chirikjian, J. G. (1986). Differences in the kinetic properties of BamHI endonuclease and methylase with linear DNA substrates. *J. Biol. Chem.*, 261(26):12128–12133.

Narita, T., Ito, S., Higashijima, Y., Chu, W. K., Neumann, K., Walter, J., Satpathy, S., Liebner, T., Hamilton, W. B., Maskey, E., Prus, G., Shibata, M., Iesmantavicius, V., Brickman, J. M., Anastassiadis, K., Koseki, H., and Choudhary, C. (2021). Enhancers are activated by p300/CBP activity-dependent PIC assembly, RNAPII recruitment, and pause release. *Mol. Cell*, 81(10):2166–2182.e6.

Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., Mualim, K., Natri, H. M., Weeks, E. M., Munson, G., Kane, M., Kang, H. Y., Cui, A., Ray, J. P., Eisenhaure, T. M., Collins, R. L., Dey, K., Pfister, H., Price, A. L., Epstein, C. B., Kundaje, A., Xavier, R. J., Daly, M. J., Huang, H., Finucane, H. K., Hacohen, N., Lander, E. S., and Engreitz, J. M. (2021). Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593(7858):238–243.

Nelson, M. R., Leidal, A. M., and Smibert, C. A. (2004). Drosophila cup is an eIF4E-binding protein that functions in smaug-mediated translational repression. *EMBO J.*, 23(1):150–159–159.

Nguyen, V. T., Kiss, T., Michels, A. A., and Bensaude, O. (2001). 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature*, 414(6861):322–325.

Ni, H.-M., Williams, J. A., and Ding, W.-X. (2015). Mitochondrial dynamics and mitochondrial quality control. *Redox Biol*, 4:6–13.

Nica, A. C. and Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 368(1620):20120362.

Nigri, A., Dalla Bella, E., Ferraro, S., Medina Carrion, J. P., Demichelis, G., Bersano, E., Consonni, M., Bischof, A., Stanziano, M., Palermo, S., Lauria, G., Bruzzone, M. G., and Papinutto, N. (2023). Cervical spinal cord atrophy in amyotrophic lateral sclerosis across disease stages. *Ann. Clin. Transl. Neurol.*, 10(2):213–224.

Nishimoto, Y., Nakagawa, S., Hirose, T., Okano, H. J., Takao, M., Shibata, S., Suyama, S., Kuwako, K.-I., Imai, T., Murayama, S., Suzuki, N., and Okano, H. (2013). The long non-coding RNA nuclear-enriched abundant transcript 1_2 induces paraspeckle

formation in the motor neuron during the early phase of amyotrophic lateral sclerosis. *Mol. Brain*, 6(1):1–18.

Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E. E. M., and Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife*, 4.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385.

Nott, A., Holtman, I. R., Coufal, N. G., Schlachetzki, J. C. M., Yu, M., Hu, R., Han, C. Z., Pena, M., Xiao, J., Wu, Y., Keulen, Z., Pasillas, M. P., O'Connor, C., Nickl, C. K., Schafer, S. T., Shen, Z., Rissman, R. A., Brewer, J. B., Gosselin, D., Gonda, D. D., Levy, M. L., Rosenfeld, M. G., McVicker, G., Gage, F. H., Ren, B., and Glass, C. K. (2019). Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science*, 366(6469):1134–1139.

Ntini, E., Järvelin, A. I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P. R., Andersen, P. K., Preker, P., Valen, E., Zhao, X., Pelechano, V., Steinmetz, L. M., Sandelin, A., and Jensen, T. H. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat. Struct. Mol. Biol.*, 20(8):923–928.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., Caldas, G. V., Chen, N.-C., Cheng, H., Chin, C.-S., Chow, W., de Lima, L. G., Dishuck, P. C., Durbin, R., Dvorkina, T., Fiddes, I. T., Formenti, G., Fulton, R. S., Fungtammasan, A., Garrison, E., Grady, P. G. S., Graves-Lindsay, T. A., Hall, I. M., Hansen, N. F., Hartley, G. A., Haukness, M., Howe, K., Hunkapiller, M. W., Jain, C., Jain, M., Jarvis, E. D., Kerpedjiev, P., Kirsche, M., Kolmogorov, M., Korlach, J., Kremitzki, M., Li, H., Maduro, V. V., Marschall, T., McCartney, A. M., McDaniel, J., Miller, D. E., Mullikin, J. C., Myers, E. W., Olson, N. D., Paten, B., Peluso, P., Pevzner, P. A., Porubsky, D., Potapova, T., Rogaev, E. I., Rosenfeld, J. A., Salzberg, S. L., Schneider, V. A., Sedlazeck, F. J., Shafin, K., Shew, C. J., Shumate, A., Sims, Y., Smit, A. F. A., Soto, D. C., Sović, I., Storer, J. M., Streets, A., Sullivan, B. A., Thibaud-Nissen, F., Torrance, J., Wagner, J., Walenz, B. P., Wenger, A., Wood, J. M. D., Xiao, C., Yan, S. M., Young, A. C., Zarate, S.,

Surti, U., McCoy, R. C., Dennis, M. Y., Alexandrov, I. A., Gerton, J. L., O'Neill, R. J., Timp, W., Zook, J. M., Schatz, M. C., Eichler, E. E., Miga, K. H., and Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588):44–53.

O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of MicroRNA biogenesis, mechanisms of actions, and circulation. *Front. Endocrinol.*, 9:402.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symp. Biol.*, 23:366–370.

Oksuz, O., Henninger, J. E., Warneford-Thomson, R., Zheng, M. M., Erb, H., Vancura, A., Overholt, K. J., Hawken, S. W., Banani, S. F., Lauman, R., Reich, L. N., Robertson, A. L., Hannett, N. M., Lee, T. I., Zon, L. I., Bonasio, R., and Young, R. A. (2023). Transcription factors interact with RNA to regulate genes. *Mol. Cell*, 83(14):2449–2463.e13.

Oktaba, K., Zhang, W., Lotz, T. S., Jun, D. J., Lemke, S. B., Ng, S. P., Esposito, E., Levine, M., and Hilgers, V. (2015). ELAV links paused pol II to alternative polyadenylation in the drosophila nervous system. *Mol. Cell*, 57(2):341–348.

Orgel, L. E. and Crick, F. H. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284(5757):604–607.

Osma-Garcia, I. C., Capitan-Sobrino, D., Mouysset, M., Aubert, Y., Maloudi, O., Turner, M., and Diaz-Muñoz, M. D. (2022). The splicing regulators TIA1 and TIAL1 are required for the expression of the DNA damage repair machinery during B cell lymphopoiesis. *Cell Rep.*, 41(12):111869.

Ospelt, C., Brentano, F., Rengel, Y., Stanczyk, J., Kolling, C., Tak, P. P., Gay, R. E., Gay, S., and Kyburz, D. (2008). Overexpression of toll-like receptors 3 and 4 in synovial tissue from patients with early rheumatoid arthritis: toll-like receptor expression in early and longstanding arthritis. *Arthritis Rheum.*, 58(12):3684–3692.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Dickel, D. E., Visel, A., and Pennacchio, L. A. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, 554(7691):239–243.

Oudelaar, A. M., Davies, J. O. J., Hanssen, L. L. P., Telenius, J. M., Schwessinger, R., Liu, Y., Brown, J. M., Downes, D. J., Chiariello, A. M., Bianco, S., Nicodemi, M., Buckle, V. J., Dekker, J., Higgs, D. R., and Hughes, J. R. (2018). Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nat. Genet.*, 50(12):1744–1751.

Oudelaar, A. M., Harrold, C. L., Hanssen, L. L. P., Telenius, J. M., Higgs, D. R., and Hughes, J. R. (2019). A revised model for promoter competition based on multi-way chromatin interactions at the -globin locus. *Nat. Commun.*, 10(1):5412.

Ounzain, S., Micheletti, R., Arnan, C., Plaisance, I., Cecchi, D., Schroen, B., Reverter, F., Alexanian, M., Gonzales, C., Ng, S. Y., Bussotti, G., Pezzuto, I., Notredame, C., Heymans, S., Guigó, R., Johnson, R., and Pedrazzini, T. (2015). CARMEN, a human super enhancer-associated long noncoding RNA controlling cardiac specification, differentiation and homeostasis. *J. Mol. Cell. Cardiol.*, 89(Pt A):98–112.

Padhy, B., Hayat, B., Nanda, G. G., Mohanty, P. P., and Alone, D. P. (2017). Pseudoexfoliation and alzheimer's associated CLU risk variant, rs2279590, lies within an enhancer element and regulates CLU, EPHX2 and PTK2B gene expression. *Hum. Mol. Genet.*, 26(22):4519–4529.

Padhy, B., Nanda, G. G., Chowdhury, M., Padhi, D., Rao, A., and Alone, D. P. (2014). Role of an extracellular chaperone, clusterin in the pathogenesis of pseudoexfoliation syndrome and pseudoexfoliation glaucoma. *Exp. Eye Res.*, 127:69–76.

Pak, C. W., Kosno, M., Holehouse, A. S., Padrick, S. B., Mittal, A., Ali, R., Yunus, A. A., Liu, D. R., Pappu, R. V., and Rosen, M. K. (2016). Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. *Mol. Cell*, 63(1):72–85.

Palmer, T. D., Willhoite, A. R., and Gage, F. H. (2000). Vascular niche for adult hippocampal neurogenesis. *J. Comp. Neurol.*, 425(4):479–494.

Pan, X., Papasani, M., Hao, Y., Calamito, M., Wei, F., Quinn, Iii, W. J., Basu, A., Wang, J., Hodawadekar, S., Zaprazna, K., Liu, H., Shi, Y., Allman, D., Cancro, M., and Atchison, M. L. (2013). YY1 controls Ig$\kappa$ repertoire and b-cell development, and localizes with condensin on the Ig$\kappa$ locus. *EMBO J.*, 32(8):1168–1182.

Pandit, R., Koh, W. K., Sullivan, R. K. P., Palliyaguru, T., Parton, R. G., and Götz, J. (2020). Role for caveolin-mediated transcytosis in facilitating transport of large cargoes into the brain via ultrasound. *J. Control. Release*, 327:667–675.

Panigrahi, A. K., Lonard, D. M., and O'Malley, B. W. (2023). Enhancer-promoter entanglement explains their transcriptional interdependence. *Proc. Natl. Acad. Sci. U. S. A.*, 120(4):e2216436120.

Parton, R. G. and del Pozo, M. A. (2013). Caveolae as plasma membrane sensors, protectors and organizers. *Nat. Rev. Mol. Cell Biol.*, 14(2):98–112.

Patel, P. J., Ren, Y., and Yan, Z. (2023). Epigenomic analysis of alzheimer's disease brains reveals diminished CTCF binding on genes involved in synaptic organization. *Neurobiol. Dis.*, 184:106192.

Pedchenko, T. V., Park, G. Y., Joo, M., Blackwell, T. S., and Christman, J. W. (2005). Inducible binding of PU.1 and interacting proteins to the toll-like receptor 4 promoter during endotoxemia. *Am. J. Physiol. Lung Cell. Mol. Physiol.*, 289(3):L429–37.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., Federation, A., Chao, J., Elliott, O., Liu, Z.-P., Economides, A. N., Bradner, J. E., Rabadan, R., and Basu, U. (2015). RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. *Cell*, 161(4):774–789.

Pelkmans, L. and Zerial, M. (2005). Kinase-regulated quantal assemblies and kiss-and-run recycling of caveolae. *Nature*, 436(7047):128–133.

Peng, S. S., Chen, C. A., Xu, N., and Shyu, A. (1998). RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein. *EMBO J.*, 17(12):3461–3470–3470.

Pertea, G. and Pertea, M. (2020). GFF utilities: GffRead and GffCompare. *F1000Research*, 9(304):304.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, 33(3):290–295.

Petrovic, A. G. and Polavarapu, P. L. (2008). The quadruplex-duplex structural transition of polyriboguanylic acid. *J. Phys. Chem. B*, 112(7):2245–2254.

Phillips-Cremins, J. E. and Corces, V. G. (2013). Chromatin insulators: linking genome organization to cellular function. *Mol. Cell*, 50(4):461–474.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, 94(4):559–573.

Pnueli, L., Rudnizky, S., Yosefzon, Y., and Melamed, P. (2015). RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin $\alpha$-subunit gene. *Proc. Natl. Acad. Sci. U. S. A.*, 112(14):4369–4374.

Pol, A., Luetterforst, R., Lindsay, M., Heino, S., Ikonen, E., and Parton, R. G. (2001). A caveolin dominant negative mutant associates with lipid bodies and induces intracellular cholesterol imbalance. *J. Cell Biol.*, 152(5):1057–1070.

Pollex, T., Rabinowitz, A., Gambetta, M. C., Marco-Ferreres, R., Viales, R. R., Jankowski, A., Schaub, C., and Furlong, E. E. M. (2024). Enhancer-promoter interactions become more instructive in the transition from cell-fate specification to tissue differentiation. *Nat. Genet.*, 56(4):686–696.

Pradeepa, M. M., Grimes, G. R., Kumar, Y., Olley, G., Taylor, G. C. A., Schneider, R., and Bickmore, W. A. (2016). Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.*, 48(6):681–686.

Przanowska, R. K., Weidmann, C. A., Saha, S., Cichewicz, M. A., Jensen, K. N., Przanowski, P., Irving, P. S., Janes, K. A., Guertin, M. J., Weeks, K. M., and Dutta, A. (2022). Distinct MUNC lncRNA structural domains regulate transcription of different promyogenic factors. *Cell Rep.*, 38(7):110361.

Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D. J., Pauli, A., Hacohen, N., Schier, A. F., Blackshear, P. J., Friedman, N., Amit, I., and Regev, A. (2014). High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*, 159(7):1698–1710.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283.

Rademakers, R., Eriksen, J. L., Baker, M., Robinson, T., Ahmed, Z., Lincoln, S. J., Finch, N., Rutherford, N. J., Crook, R. J., Josephs, K. A., Boeve, B. F., Knopman, D. S., Petersen, R. C., Parisi, J. E., Caselli, R. J., Wszolek, Z. K., Uitti, R. J., Feldman, H., Hutton, M. L., Mackenzie, I. R., Graff-Radford, N. R., and Dickson, D. W. (2008). Common variation in the mir-659 binding-site of GRN is a major risk factor for TDP43-positive frontotemporal dementia. *Hum. Mol. Genet.*, 17(23):3631–3642.

Radstake, T. R. D. J., Roelofs, M. F., Jenniskens, Y. M., Oppers-Walgreen, B., van Riel, P. L. C. M., Barrera, P., Joosten, L. A. B., and van den Berg, W. B. (2004). Expression of toll-like receptors 2 and 4 in rheumatoid synovial tissue and regulation by proinflammatory cytokines interleukin-12 and interleukin-18 via interferon-gamma. *Arthritis Rheum.*, 50(12):3856–3865.

Rahnamoun, H., Lee, J., Sun, Z., Lu, H., Ramsey, K. M., Komives, E. A., and Lauberth, S. M. (2018). RNAs interact with BRD4 to promote enhanced chromatin engagement and transcription activation. *Nat. Struct. Mol. Biol.*, 25(8):687–697.

Rahnamoun, H., Lu, H., Duttke, S. H., Benner, C., Glass, C. K., and Lauberth, S. M. (2017). Mutant p53 shapes the enhancer landscape of cancer cells in response to chronic immune signaling. *Nat. Commun.*, 8(1):754.

Ramsköld, D., Wang, E. T., Burge, C. B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, 5(12):e1000598.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.*, 9(1):189.

Raney, B. J., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Fischer, C., Navarro Gonzalez, J., Hickey, G., Hinrichs, A. S., Kuhn, R. M., Lee, B. T., Lee, C. M., Le Mercier, P., Miga, K. H., Nassar, L. R., Nejad, P., Paten, B., Perez, G., Schmelter, D., Speir, M. L., Wick, B. D., Zweig, A. S., Haussler, D., Kent, W. J., and Haeussler, M. (2024). The UCSC genome browser database: 2024 update. *Nucleic Acids Res.*, 52(D1):D1082–D1088.

Rao, J. N., Neumann, L., Wenzel, S., Schweimer, K., Rösch, P., and Wöhrl, B. M. (2006). Structural studies on the RNA-recognition motif of NELF e, a cellular negative transcription elongation factor involved in the regulation of HIV transcription. *Biochem. J*, 400(3):449–456.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.

Ratmeyer, L., Vinayak, R., Zhong, Y. Y., Zon, G., and Wilson, W. D. (1994). Sequence specific thermodynamic and structural properties for DNA.RNA duplexes. *Biochemistry*, 33(17):5298–5304.

Rauluseviciute, I., Riudavets-Puig, R., Blanc-Mathieu, R., Castro-Mondragon, J. A., Ferenc, K., Kumar, V., Lemma, R. B., Lucas, J., Chèneby, J., Baranasic, D., Khan, A., Fornes, O., Gundersen, S., Johansen, M., Hovig, E., Lenhard, B., Sandelin, A., Wasserman, W. W., Parcy, F., and Mathelier, A. (2024). JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 52(D1):D174–D182.

Reddy, K., Zamiri, B., Stanley, S. Y. R., Macgregor, Jr, R. B., and Pearson, C. E. (2013). The disease-associated r(GGGGCC)n repeat from the c9orf72 gene forms

tract length-dependent uni- and multimolecular RNA g-quadruplex structures. *J. Biol. Chem.*, 288(14):9860–9866.

Rehli, M., Poltorak, A., Schwarzfischer, L., Krause, S. W., Andreesen, R., and Beutler, B. (2000). PU.1 and interferon consensus sequence-binding protein regulate the myeloid expression of the human toll-like receptor 4 gene. *J. Biol. Chem.*, 275(13):9773–9781.

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., Schymick, J. C., Laaksovirta, H., van Swieten, J. C., Myllykangas, L., Kalimo, H., Paetau, A., Abramzon, Y., Remes, A. M., Kaganovich, A., Scholz, S. W., Duckworth, J., Ding, J., Harmer, D. W., Hernandez, D. G., Johnson, J. O., Mok, K., Ryten, M., Trabzuni, D., Guerreiro, R. J., Orrell, R. W., Neal, J., Murray, A., Pearson, J., Jansen, I. E., Sondervan, D., Seelaar, H., Blake, D., Young, K., Halliwell, N., Callister, J. B., Toulson, G., Richardson, A., Gerhard, A., Snowden, J., Mann, D., Neary, D., Nalls, M. A., Peuralinna, T., Jansson, L., Isoviita, V.-M., Kaivorinne, A.-L., Hölttä-Vuori, M., Ikonen, E., Sulkava, R., Benatar, M., Wuu, J., Chiò, A., Restagno, G., Borghero, G., Sabatelli, M., ITALSGEN Consortium, Heckerman, D., Rogaeva, E., Zinman, L., Rothstein, J. D., Sendtner, M., Drepper, C., Eichler, E. E., Alkan, C., Abdullaev, Z., Pack, S. D., Dutra, A., Pak, E., Hardy, J., Singleton, A., Williams, N. M., Heutink, P., Pickering-Brown, S., Morris, H. R., Tienari, P. J., and Traynor, B. J. (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, 72(2):257–268.

Replogle, J. M., Bonnar, J. L., Pogson, A. N., Liem, C. R., Maier, N. K., Ding, Y., Russell, B. J., Wang, X., Leng, K., Guna, A., Norman, T. M., Pak, R. A., Ramos, D. M., Ward, M. E., Gilbert, L. A., Kampmann, M., Weissman, J. S., and Jost, M. (2022). Maximizing CRISPRi efficacy and accessibility with dual-sgRNA libraries and optimal effectors. *Elife*, 11.

Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11:129.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, 39(17):e118.

Richter, S., Cao, H., and Rana, T. M. (2002). Specific HIV-1 TAR RNA loop sequence and functional groups are required for human cyclin T1-Tat-TAR ternary complex formation. *Biochemistry*, 41(20):6391–6397.

Ries, R. J., Zaccara, S., Klein, P., Olarerin-George, A., Namkoong, S., Pickering, B. F., Patil, D. P., Kwak, H., Lee, J. H., and Jaffrey, S. R. (2019). m6a enhances the phase separation potential of mRNA. *Nature*, 571(7765):424–428.

Riet, J., Ramos, L. R. V., Lewis, R. V., and Marins, L. F. (2017). Improving the PCR protocol to amplify a repetitive DNA sequence. *Genetics and molecular research : GMR*, 16(3).

Ringrose, L., Ehret, H., and Paro, R. (2004). Distinct contributions of histone H3 lysine 9 and 27 methylation to locus-specific stability of polycomb complexes. *Mol. Cell*, 16(4):641–653.

Rissone, P., Bizarro, C. V., and Ritort, F. (2022). Stem-loop formation drives RNA folding in mechanical unzipping experiments. *Proc. Natl. Acad. Sci. U. S. A.*, 119(3):e2025575119.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Rockman, M. V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nat. Rev. Genet.*, 7(11):862–872.

Roden, C. and Gladfelter, A. S. (2021). RNA contributions to the form and function of biomolecular condensates. *Nat. Rev. Mol. Cell Biol.*, 22(3):183–195.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, 79(1):233–269.

Rojas-Fernandez, A., Herhaus, L., Macartney, T., Lachaud, C., Hay, R. T., and Sapkota, G. P. (2015). Rapid generation of endogenously driven transcriptional reporters in cells through CRISPR/Cas9. *Sci. Rep.*, 5(1):9811.

Romo, L., Ashar-Patel, A., Pfister, E., and Aronin, N. (2017). Alterations in mRNA 3' UTR isoform abundance accompany gene expression changes in human huntington's disease brains. *Cell Rep.*, 20(13):3057–3070.

Roses, A. D. (2021). Methods for detecting structural variants in neurodegenerative disease.

Ross, R. A., Spengler, B. A., and Biedler, J. L. (1983). Coordinate morphological and biochemical interconversion of human neuroblastoma cells. *J. Natl. Cancer Inst.*, 71(4):741–747.

Roy, S., Luetterforst, R., Harding, A., Apolloni, A., Etheridge, M., Stang, E., Rolls, B., Hancock, J. F., and Parton, R. G. (1999). Dominant-negative caveolin inhibits H-Ras function by disrupting cholesterol-rich plasma membrane domains. *Nat. Cell Biol.*, 1(2):98–105.

Rudakou, U., Yu, E., Krohn, L., Ruskey, J. A., Asayesh, F., Dauvilliers, Y., Spiegelman, D., Greenbaum, L., Fahn, S., Waters, C. H., Dupré, N., Rouleau, G. A., Hassin-Baer, S., Fon, E. A., Alcalay, R. N., and Gan-Or, Z. (2021). Targeted sequencing of parkinson's disease loci genes highlights SYT11, FGF20 and other associations. *Brain*, 144(2):462–472.

Ryan, M., Heverin, M., McLaughlin, R. L., and Hardiman, O. (2019). Lifetime risk and heritability of amyotrophic lateral sclerosis. *JAMA Neurol.*, 76(11):1367–1374.

Saccon, R. A., Bunton-Stasyshyn, R. K. A., Fisher, E. M. C., and Fratta, P. (2013). Is SOD1 loss of function involved in amyotrophic lateral sclerosis? *Brain*, 136(Pt 8):2342–2358.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific shh expression and truncation of the mouse limb. *Development*, 132(4):797–803.

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., and Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U. S. A.*, 112(47):E6456–65.

Sanchez-Mut, J. V., Heyn, H., Silva, B. A., Dixsaut, L., Garcia-Esparcia, P., Vidal, E., Sayols, S., Glauser, L., Monteagudo-Sánchez, A., Perez-Tur, J., Ferrer, I., Monk, D., Schneider, B., Esteller, M., and Gräff, J. (2018). PM20D1 is a quantitative trait locus associated with alzheimer's disease. *Nat. Med.*, 24(5):598–603.

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*, 320(5883):1643–1647.

Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5(1):99.

Sandler, H., Kreth, J., Timmers, H. T. M., and Stoecklin, G. (2011). Not1 mediates recruitment of the deadenylase caf1 to mRNAs targeted for degradation by tristetraprolin. *Nucleic Acids Res.*, 39(10):4373–4386.

Sanjana, N. E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods*, 11(8):783–784.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113.

Sanz, L. A., Hartono, S. R., Lim, Y. W., Steyaert, S., Rajpurkar, A., Ginno, P. A., Xu, X., and Chédin, F. (2016). Prevalent, dynamic, and conserved R-Loop structures associate with specific epigenomic signatures in mammals. *Mol. Cell*, 63(1):167–178.

Sawada, A., Wang, S., Jian, M., Leem, J., Wackerbarth, J., Egawa, J., Schilling, J. M., Platoshyn, O., Zemljic-Harpf, A., Roth, D. M., Patel, H. H., Patel, P. M., Marsala, M., and Head, B. P. (2019). Neuron-targeted caveolin-1 improves neuromuscular function and extends survival in SOD1G93A mice. *FASEB J.*, 33(6):7545–7554.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Res.*, 22(9):1748–1759.

Schaukowitch, K., Joo, J. Y., Liu, X., Watts, J. K., Martinez, C., and Kim, T. K. (2014). Enhancer RNA facilitates NELF release from immediate early genes. *Mol. Cell*, 56(1).

Scherer, P. E., Lewis, R. Y., Volonte, D., Engelman, J. A., Galbiati, F., Couet, J., Kohtz, D. S., van Donselaar, E., Peters, P., and Lisanti, M. P. (1997). Cell-type and tissue-specific expression of caveolin-2. caveolins 1 and 2 co-localize and form a stable hetero-oligomeric complex in vivo. *J. Biol. Chem.*, 272(46):29337–29346.

Scherer, P. E., Tang, Z., Chun, M., Sargiacomo, M., Lodish, H. F., and Lisanti, M. P. (1995). Caveolin isoforms differ in their n-terminal protein sequence and subcellular distribution. identification and epitope mapping of an isoform-specific monoclonal antibody probe. *J. Biol. Chem.*, 270(27):16395–16401.

Schild, C., Claret, F. X., Wahli, W., and Wolffe, A. P. (1993). A nucleosome-dependent static loop potentiates estrogen-regulated transcription from the xenopus vitellogenin B1 promoter in vitro. *EMBO J.*, 12(2):423–433.

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, 148(1-2):335–348.

Schmiedel, B. J., Seumois, G., Samaniego-Castruita, D., Cayford, J., Schulten, V., Chavez, L., Ay, F., Sette, A., Peters, B., and Vijayanand, P. (2016). 17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells. *Nat. Commun.*, 7:13426.

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., and Ren, B. (2016). A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.*, 17(8):2042–2059.

Scholes, C., Biette, K. M., Harden, T. T., and DePace, A. H. (2019). Signal integration by shadow enhancers and enhancer duplications varies across the drosophila embryo. *Cell Rep.*, 26(9):2407–2418.e5.

Schrijvers, E. M. C., Koudstaal, P. J., Hofman, A., and Breteler, M. M. B. (2011). Plasma clusterin and the risk of alzheimer disease. *JAMA*, 305(13):1322–1326.

Schwalb, B., Michel, M., Zacher, B., Frühauf, K., Demel, C., Tresch, A., Gagneur, J., and Cramer, P. (2016). TT-seq maps the human transient transcriptome. *Science*, 352(6290):1225–1228.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C. H., Mirny, L., and Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678):51–56.

Sciortino, F., Bansil, R., Stanley, H. E., and Alstrom, P. (1993). Interference of phase separation and gelation: A zeroth-order kinetic model. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, 47(6):4615–4618.

Seffens, W. and Digby, D. (1999). mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.*, 27(7):1578–1584.

Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., Young, R. A., and Sharp, P. A. (2008). Divergent transcription from active promoters. *Science*, 322(5909):1849–1851.

Semotok, J. L., Cooperstock, R. L., Pinder, B. D., Vari, H. K., Lipshitz, H. D., and Smibert, C. A. (2005). Smaug recruits the CCR4/POP2/NOT deadenylase complex to trigger maternal transcript localization in the early drosophila embryo. *Curr. Biol.*, 15(4):284–294.

Sera, T. and Wolffe, A. P. (1998). Role of histone H1 as an architectural determinant of chromatin structure and as a specific repressor of transcription on xenopus oocyte 5S rRNA genes. *Mol. Cell. Biol.*, 18(7):3668–3680.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–472.

Shafik, A. M., Zhang, F., Guo, Z., Dai, Q., Pajdzik, K., Li, Y., Kang, Y., Yao, B., Wu, H., He, C., Allen, E. G., Duan, R., and Jin, P. (2021). N6-methyladenosine dynamics in neurodevelopment and aging, and its potential role in alzheimer's disease. *Genome Biol.*, 22(1):17.

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., and Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, 343(6166):84–87.

Shalizi, A., Gaudillière, B., Yuan, Z., Stegmüller, J., Shirogane, T., Ge, Q., Tan, Y., Schulman, B., Harper, J. W., and Bonni, A. (2006). A calcium-regulated MEF2 sumoylation switch controls postsynaptic differentiation. *Science*, 311(5763):1012–1017.

Shaw, G. and Kamen, R. (1986). A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell*, 46(5):659–667.

Shaw, P. J., Ince, P. G., Falkous, G., and Mantle, D. (1995). Oxidative damage to protein in sporadic motor neuron disease spinal cord. *Ann. Neurol.*, 38(4):691–695.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V., and Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120.

Shepard, P. J., Choi, E.-A., Lu, J., Flanagan, L. A., Hertel, K. J., and Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17(4):761–772.

Shi, H., Kokoeva, M. V., Inouye, K., Tzameli, I., Yin, H., and Flier, J. S. (2006). TLR4 links innate immunity and fatty acid-induced insulin resistance. *J. Clin. Invest.*, 116(11):3015–3025.

Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, 3rd, J. R., Frank, J., and Manley, J. L. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. *Mol. Cell*, 33(3):365–376.

Shi, Y., Lee, J. S., and Galvin, K. M. (1997). Everything you have ever wanted to know about yin yang 1... *Biochim. Biophys. Acta*, 1332(2):F49–66.

Shi, Y., Tan, S.-H., Ng, S., Zhou, J., Yang, N.-D., Koo, G.-B., McMahon, K.-A., Parton, R. G., Hill, M. M., Del Pozo, M. A., Kim, Y.-S., and Shen, H.-M. (2015). Critical role of CAV1/caveolin-1 in cell stress responses in human breast cancer cells via modulation of lysosomal function and autophagy. *Autophagy*, 11(5):769–784.

Shii, L., Song, L., Maurer, K., Zhang, Z., and Sullivan, K. E. (2017). SERPINB2 is regulated by dynamic interactions with pause-release proteins and enhancer RNAs. *Mol. Immunol.*, 88.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., and Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.*, 100(26):15776–15781.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, 15(4):272–286.

Shrinivas, K., Sabari, B. R., Coffey, E. L., Klein, I. A., Boija, A., Zamudio, A. V., Schuijers, J., Hannett, N. M., Sharp, P. A., Young, R. A., and Chakraborty, A. K. (2019). Enhancer features that drive formation of transcriptional condensates. *Mol. Cell*, 75(3):549–561.e7.

Shui, I. M., Lindström, S., Kibel, A. S., Berndt, S. I., Campa, D., Gerke, T., Penney, K. L., Albanes, D., Berg, C., Bueno-de Mesquita, H. B., Chanock, S., Crawford, E. D., Diver, W. R., Gapstur, S. M., Gaziano, J. M., Giles, G. G., Henderson, B., Hoover, R., Johansson, M., Le Marchand, L., Ma, J., Navarro, C., Overvad, K., Schumacher, F. R., Severi, G., Siddiq, A., Stampfer, M., Stevens, V. L., Travis, R. C., Trichopoulos, D., Vineis, P., Mucci, L. A., Yeager, M., Giovannucci, E., and Kraft, P. (2014). Prostate cancer (PCa) risk variants and risk of fatal PCa in the national cancer institute breast and prostate cancer cohort consortium. *Eur. Urol.*, 65(6):1069–1075.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371):74–79.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050.

Sigova, A. A., Abraham, B. J., Ji, X., Molinie, B., Hannett, N. M., Guo, Y. E., Jangi, M., Giallourakis, C. C., Sharp, P. A., and Young, R. A. (2015). Transcription factor trapping by RNA in gene regulatory elements. *Science*, 350(6263):978–981.

Simon, M., North, J. A., Shimko, J. C., Forties, R. A., Ferdinand, M. B., Manohar, M., Zhang, M., Fishel, R., Ottesen, J. J., and Poirier, M. G. (2011). Histone fold modifications control nucleosome unwrapping and disassembly. *Proc. Natl. Acad. Sci. U. S. A.*, 108(31):12711–12716.

Simón-Sánchez, J., Schulte, C., Bras, J. M., Sharma, M., Gibbs, J. R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S. W., Hernandez, D. G., Krüger, R., Federoff, M., Klein, C., Goate, A., Perlmutter, J., Bonin, M., Nalls, M. A., Illig, T., Gieger, C., Houlden, H., Steffens, M., Okun, M. S., Racette, B. A., Cookson, M. R., Foote, K. D., Fernandez, H. H., Traynor, B. J., Schreiber, S., Arepalli, S., Zonozi, R., Gwinn, K., van der Brug, M., Lopez, G., Chanock, S. J., Schatzkin, A., Park, Y., Hollenbeck, A., Gao, J., Huang, X., Wood, N. W., Lorenz, D., Deuschl, G., Chen, H., Riess, O., Hardy, J. A., Singleton, A. B., and Gasser, T. (2009). Genome-wide association study reveals genetic risk underlying parkinson's disease. *Nat. Genet.*, 41(12):1308–1312.

Singh, I., Lee, S.-H., Sperling, A. S., Samur, M. K., Tai, Y.-T., Fulciniti, M., Munshi, N. C., Mayr, C., and Leslie, C. S. (2018). Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat. Commun.*, 9(1):1716.

Skene, P. J. and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*, 6.

Smale, S. T. and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, 72:449–479.

Smibert, P., Miura, P., Westholm, J. O., Shenker, S., May, G., Duff, M. O., Zhang, D., Eads, B. D., Carlson, J., Brown, J. B., Eisman, R. C., Andrews, J., Kaufman, T., Cherbas, P., Celniker, S. E., Graveley, B. R., and Lai, E. C. (2012). Global patterns of tissue-specific alternative polyadenylation in drosophila. *Cell Rep.*, 1(3):277–289.

Smits, A. H., Ziebell, F., Joberty, G., Zinn, N., Mueller, W. F., Clauder-Münster, S., Eberhard, D., Fälth Savitski, M., Grandi, P., Jakob, P., Michon, A.-M., Sun, H., Tessmer, K., Bürckstümmer, T., Bantscheff, M., Steinmetz, L. M., Drewes, G., and

Huber, W. (2019). Biological plasticity rescues target activity in CRISPR knock outs. *Nat. Methods*, 16(11):1087–1093.

Soldner, F., Stelzer, Y., Shivalila, C. S., Abraham, B. J., Latourelle, J. C., Barrasa, M. I., Goldmann, J., Myers, R. H., Young, R. A., and Jaenisch, R. (2016). Parkinson-associated risk variant in distal enhancer of $\alpha$-synuclein modulates target gene expression. *Nature*, 533(7601):95–99.

Song, L. and Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, 2010(2):db.prot5384.

Sotiriou, S., Gibney, G., Baxevanis, A. D., and Nussbaum, R. L. (2009). A single nucleotide polymorphism in the 3'UTR of the SNCA gene encoding alpha-synuclein is a new potential susceptibility locus for parkinson disease. *Neurosci. Lett.*, 461(2):196–201.

Spain, S. L. and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Hum. Mol. Genet.*, 24(R1):R111–9.

Sparrow, D. B., Miska, E. A., Langley, E., Reynaud-Deonauth, S., Kotecha, S., Towers, N., Spohr, G., Kouzarides, T., and Mohun, T. J. (1999). MEF-2 function is modified by a novel co-repressor, MITR. *EMBO J.*, 18(18):5085–5098.

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.*, 91(6):1011–1021.

Spies, N., Nielsen, C. B., Padgett, R. A., and Burge, C. B. (2009). Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell*, 36(2):245–254.

Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9):613–626.

Spurlock, 3rd, C. F., Shaginurova, G., Tossberg, J. T., Hester, J. D., Chapman, N., Guo, Y., Crooke, 3rd, P. S., and Aune, T. M. (2017). Profiles of long noncoding RNAs in human naive and memory T cells. *J. Immunol.*, 199(2):547–558.

Srivastava, D. and Mahony, S. (2020). Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochim. Biophys. Acta Gene Regul. Mech.*, 1863(6):194443.

Stacey, S. N., Sulem, P., Jonasdottir, A., Masson, G., Gudmundsson, J., Gudbjartsson, D. F., Magnusson, O. T., Gudjonsson, S. A., Sigurgeirsson, B., Thorisdottir, K., Ragnarsson, R., Benediktsdottir, K. R., Nexø, B. A., Tjønneland, A., Overvad, K.,

Rudnai, P., Gurzau, E., Koppova, K., Hemminki, K., Corredera, C., Fuentelsaz, V., Grasa, P., Navarrete, S., Fuertes, F., García-Prats, M. D., Sanambrosio, E., Panadero, A., De Juan, A., Garcia, A., Rivera, F., Planelles, D., Soriano, V., Requena, C., Aben, K. K., van Rossum, M. M., Cremers, R. G. H. M., van Oort, I. M., van Spronsen, D.-J., Schalken, J. A., Peters, W. H. M., Helfand, B. T., Donovan, J. L., Hamdy, F. C., Badescu, D., Codreanu, O., Jinga, M., Csiki, I. E., Constantinescu, V., Badea, P., Mates, I. N., Dinu, D. E., Constantin, A., Mates, D., Kristjansdottir, S., Agnarsson, B. A., Jonsson, E., Barkardottir, R. B., Einarsson, G. V., Sigurdsson, F., Moller, P. H., Stefansson, T., Valdimarsson, T., Johannsson, O. T., Sigurdsson, H., Jonsson, T., Jonasson, J. G., Tryggvadottir, L., Rice, T., Hansen, H. M., Xiao, Y., Lachance, D. H., O Neill, B. P., Kosel, M. L., Decker, P. A., Thorleifsson, G., Johannsdottir, H., Helgadottir, H. T., Sigurdsson, A., Steinthorsdottir, V., Lindblom, A., Swedish Low-risk Colorectal Cancer Study Group, Sandler, R. S., Keku, T. O., Banasik, K., Jørgensen, T., Witte, D. R., Hansen, T., Pedersen, O., Jinga, V., Neal, D. E., Catalona, W. J., Wrensch, M., Wiencke, J., Jenkins, R. B., Nagore, E., Vogel, U., Kiemeney, L. A., Kumar, R., Mayordomo, J. I., Olafsson, J. H., Kong, A., Thorsteinsdottir, U., Rafnar, T., and Stefansson, K. (2011). A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat. Genet.*, 43(11):1098–1103.

Stasevich, T. J., Hayashi-Takanaka, Y., Sato, Y., Maehara, K., Ohkawa, Y., Sakata-Sogawa, K., Tokunaga, M., Nagase, T., Nozaki, N., McNally, J. G., and Kimura, H. (2014). Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature*.

Stephen, T. K. L., Cofresi, L. A., Quiroz, E., Owusu-Ansah, K., Ibrahim, Y., Qualls, E., Marshall, J., Li, W., Shetti, A., Bonds, J. A., Minshall, R. D., Cologna, S. M., and Lazarov, O. (2023). Caveolin-1 autonomously regulates hippocampal neurogenesis via mitochondrial dynamics. *bioRxiv*.

Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41–45.

Strober, B. J., Elorbany, R., Rhodes, K., Krishnan, N., Tayeb, K., Battle, A., and Gilad, Y. (2019). Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R., and Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.*, 101(16):6062–6067.

Svoboda, P. and Di Cara, A. (2006). Hairpin RNA: a secondary structure of primary importance. *Cell. Mol. Life Sci.*, 63(7-8):901–908.

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Res.*, 24(3):390–400.

Taatjes, D. J. (2010). The human mediator complex: a versatile, genome-wide regulator of transcription. *Trends Biochem. Sci.*, 35(6):315–322.

Taft, R. J., Pheasant, M., and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29(3):288–299.

Takayasu, Y., Takeuchi, K., Kumari, R., Bennett, M. V. L., Zukin, R. S., and Francesconi, A. (2010). Caveolin-1 knockout mice exhibit impaired induction of mGluR-dependent long-term depression at CA3-CA1 synapses. *Proc. Natl. Acad. Sci. U. S. A.*, 107(50):21778–21783.

Tanay, A., O'Donnell, A. H., Damelin, M., and Bestor, T. H. (2007). Hyperconserved CpG domains underlie polycomb-binding sites. *Proc. Natl. Acad. Sci. U. S. A.*, 104(13):5521–5526.

Taskiran, I. I., Spanier, K. I., Dickmänken, H., Kempynck, N., Pančíková, A., Ekşi, E. C., Hulselmans, G., Ismail, J. N., Theunis, K., Vandepoel, R., Christiaens, V., Mauduit, D., and Aerts, S. (2024). Cell-type-directed design of synthetic enhancers. *Nature*, 626(7997):212–220.

Taylor, G. C. A., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M. M., and Bickmore, W. A. (2013). H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Res.*, 23(12):2053–2065.

Theunissen, F., Flynn, L. L., Anderton, R. S., Mastaglia, F., Pytte, J., Jiang, L., Hodgetts, S., Burns, D. K., Saunders, A., Fletcher, S., Wilton, S. D., and Akkari, P. A. (2020). Structural variants may be a source of missing heritability in sALS. *Front. Neurosci.*, 14:47.

Thirumalai, D. and Woodson, S. A. (1996). Kinetics of folding of proteins and RNA. *Acc. Chem. Res.*, 29(9):433–439.

Thomas, Jr, C. A. (1971). The genetic organization of chromosomes. *Annu. Rev. Genet.*, 5:237–256.

Thomas, H., Feng, S., Huber, M., Loubiere, V., Vanina, D., Pitasi, M., Stark, A., and Buecker, C. (2023). Enhancer cooperativity can compensate for loss of activity over large genomic distances.

Thorn, H., Stenkula, K. G., Karlsson, M., Ortegren, U., Nystrom, F. H., Gustavsson, J., and Stralfors, P. (2003). Cell surface orifices of caveolae and localization of caveolin to the necks of caveolae in adipocytes. *Mol. Biol. Cell*, 14(10):3967–3976.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82.

Tian, B., Hu, J., Zhang, H., and Lutz, C. S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, 33(1):201–212.

Tian, B. and Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, 18(1):18–30.

Tollervey, J. R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A. L., Zupunski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C. E., and Ule, J. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.*, 14(4):452–458.

Tomashevski, A., Webster, D. R., Grammas, P., Gorospe, M., and Kruman, I. I. (2010). Cyclin-C-dependent cell-cycle entry is required for activation of non-homologous end joining DNA repair in postmitotic neurons. *Cell Death Differ.*, 17(7):1189–1198.

Trabjerg, B. B., Garton, F. C., van Rheenen, W., Fang, F., Henderson, R. D., Mortensen, P. B., Agerbo, E., and Wray, N. R. (2020). ALS in danish registries: Heritability and links to psychiatric and cardiovascular disorders. *Neurol Genet*, 6(2):e398.

Trevino, T. N., Fogel, A. B., Minshall, R., Richner, J. M., and Lutz, S. E. (2023). Caveolin-1 mediates neuroinflammation and cognitive impairment in SARS-CoV-2 infection. *bioRxiv*.

Tropberger, P., Pott, S., Keller, C., Kamieniarz-Gdula, K., Caron, M., Richter, F., Li, G., Mittler, G., Liu, E. T., Bühler, M., Margueron, R., and Schneider, R. (2013). Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer. *Cell*, 152(4):859–872.

Trotta, E. (2014). On the normalization of the minimum free energy of RNAs by sequence length. *PLoS One*, 9(11):e113380.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.*, 45(2):124–130.

Tsai, P.-F., Dell'Orso, S., Rodriguez, J., Vivanco, K. O., Ko, K.-D., Jiang, K., Juan, A. H., Sarshad, A. A., Vian, L., Tran, M., Wangsa, D., Wang, A. H., Perovanovic, J., Anastasakis, D., Ralston, E., Ried, T., Sun, H.-W., Hafner, M., Larson, D. R., and Sartorelli, V. (2018). A Muscle-Specific enhancer RNA mediates cohesin recruitment and regulates transcription in trans. *Mol. Cell*, 71(1):129–141.e8.

Turner, M. R. and Benatar, M. (2015). Ensuring continued progress in biomarkers for amyotrophic lateral sclerosis. *Muscle Nerve*, 51(1):14–18.

Tushev, G., Glock, C., Heumüller, M., Biever, A., Jovanovic, M., and Schuman, E. M. (2018). Alternative 3' UTRs modify the localization, regulatory potential, stability, and plasticity of mRNAs in neuronal compartments. *Neuron*, 98(3):495–511.e6.

Ulitsky, I., Shkumatava, A., Jan, C. H., Subtelny, A. O., Koppstein, D., Bell, G. W., Sive, H., and Bartel, D. P. (2012). Extensive alternative polyadenylation during zebrafish development. *Genome Res.*, 22(10):2054–2066.

van Arensbergen, J., Pagie, L., FitzPatrick, V. D., de Haas, M., Baltissen, M. P., Comoglio, F., van der Weide, R. H., Teunissen, H., Võsa, U., Franke, L., de Wit, E., Vermeulen, M., Bussemaker, H. J., and van Steensel, B. (2019). High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.*, 51(7):1160–1169.

van Rheenen, W., Shatunov, A., Dekker, A. M., McLaughlin, R. L., Diekstra, F. P., Pulit, S. L., van der Spek, R. A. A., Võsa, U., de Jong, S., Robinson, M. R., Yang, J., Fogh, I., van Doormaal, P. T., Tazelaar, G. H. P., Koppers, M., Blokhuis, A. M., Sproviero, W., Jones, A. R., Kenna, K. P., van Eijk, K. R., Harschnitz, O., Schellevis, R. D., Brands, W. J., Medic, J., Menelaou, A., Vajda, A., Ticozzi, N., Lin, K., Rogelj, B., Vrabec, K., Ravnik-Glavač, M., Koritnik, B., Zidar, J., Leonardis, L., Grošelj, L. D., Millecamps, S., Salachas, F., Meininger, V., de Carvalho, M., Pinto, S., Mora, J. S., Rojas-García, R., Polak, M., Chandran, S., Colville, S., Swingler, R., Morrison, K. E., Shaw, P. J., Hardy, J., Orrell, R. W., Pittman, A., Sidle, K., Fratta, P., Malaspina, A., Topp, S., Petri, S., Abdulla, S., Drepper, C., Sendtner, M., Meyer, T., Ophoff, R. A., Staats, K. A., Wiedau-Pazos, M., Lomen-Hoerth, C., Van Deerlin, V. M., Trojanowski, J. Q., Elman, L., McCluskey, L., Basak, A. N., Tunca, C.,

Hamzeiy, H., Parman, Y., Meitinger, T., Lichtner, P., Radivojkov-Blagojevic, M., Andres, C. R., Maurel, C., Bensimon, G., Landwehrmeyer, B., Brice, A., Payan, C. A. M., Saker-Delye, S., Dürr, A., Wood, N. W., Tittmann, L., Lieb, W., Franke, A., Rietschel, M., Cichon, S., Nöthen, M. M., Amouyel, P., Tzourio, C., Dartigues, J.-F., Uitterlinden, A. G., Rivadeneira, F., Estrada, K., Hofman, A., Curtis, C., Blauw, H. M., van der Kooi, A. J., de Visser, M., Goris, A., Weber, M., Shaw, C. E., Smith, B. N., Pansarasa, O., Cereda, C., Del Bo, R., Comi, G. P., D'Alfonso, S., Bertolin, C., Sorarù, G., Mazzini, L., Pensato, V., Gellera, C., Tiloca, C., Ratti, A., Calvo, A., Moglia, C., Brunetti, M., Arcuti, S., Capozzo, R., Zecca, C., Lunetta, C., Penco, S., Riva, N., Padovani, A., Filosto, M., Muller, B., Stuit, R. J., PARALS Registry, SLALOM Group, SLAP Registry, FALS Sequencing Consortium, SLAGEN Consortium, NNIPPS Study Group, Blair, I., Zhang, K., McCann, E. P., Fifita, J. A., Nicholson, G. A., Rowe, D. B., Pamphlett, R., Kiernan, M. C., Grosskreutz, J., Witte, O. W., Ringer, T., Prell, T., Stubendorff, B., Kurth, I., Hübner, C. A., Leigh, P. N., Casale, F., Chio, A., Beghi, E., Pupillo, E., Tortelli, R., Logroscino, G., Powell, J., Ludolph, A. C., Weishaupt, J. H., Robberecht, W., Van Damme, P., Franke, L., Pers, T. H., Brown, R. H., Glass, J. D., Landers, J. E., Hardiman, O., Andersen, P. M., Corcia, P., Vourc'h, P., Silani, V., Wray, N. R., Visscher, P. M., de Bakker, P. I. W., van Es, M. A., Pasterkamp, R. J., Lewis, C. M., Breen, G., Al-Chalabi, A., van den Berg, L. H., and Veldink, J. H. (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.*, 48(9):1043–1048.

Venkatesh, S. and Workman, J. L. (2015). Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.*, 16(3):178–189.

Verlaan, D. J., Berlivet, S., Hunninghake, G. M., Madore, A.-M., Larivière, M., Moussette, S., Grundberg, E., Kwan, T., Ouimet, M., Ge, B., Hoberman, R., Swiatek, M., Dias, J., Lam, K. C. L., Koka, V., Harmsen, E., Soto-Quiros, M., Avila, L., Celedón, J. C., Weiss, S. T., Dewar, K., Sinnett, D., Laprise, C., Raby, B. A., Pastinen, T., and Naumova, A. K. (2009). Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. *Am. J. Hum. Genet.*, 85(3):377–393.

Vetrivel, K. S. and Thinakaran, G. (2006). Amyloidogenic processing of beta-amyloid precursor protein in intracellular compartments. *Neurology*, 66(2 Suppl 1):S69–73.

Vicens, Q. and Kieft, J. S. (2022). Thoughts on how to think (and talk) about RNA structure. *Proc. Natl. Acad. Sci. U. S. A.*, 119(17):e2112677119.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M. A., Bertelsen, M. F.,

Murchison, E. P., Flicek, P., and Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–566.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (2007). VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, 35(Database issue):D88–92.

Visser, M., Kayser, M., and Palstra, R.-J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res.*, 22(3):446–455.

Vučićević, D., Corradin, O., Ntini, E., Scacheri, P. C., and Ørom, U. A. (2015). Long ncRNA expression associates with tissue-specific enhancers. *Cell Cycle*, 14(2):253–260.

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. A., Winston, F., Buratowski, S., and Handa, H. (1998). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human spt4 and spt5 homologs. *Genes Dev.*, 12(3):343–356.

Waddington, C. H. (1957). *The strategy of the genes: A discussion of some aspects of theoretical biology*. Allen & Unwin, St Leonards, NSW, Australia.

Wahle, E. and Kühn, U. (1997). The mechanism of 3' cleavage and polyadenylation of eukaryotic pre-mRNA. *Prog. Nucleic Acid Res. Mol. Biol.*, 57:41–71.

Wai, D. C. C., Shihab, M., Low, J. K. K., and Mackay, J. P. (2016). The zinc fingers of YY1 bind single-stranded RNA with low sequence specificity. *Nucleic Acids Res.*, 44(19):9153–9165.

Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A., Glass, C. K., Rosenfeld, M. G., and Fu, X.-D. (2011a). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, 474(7351):390–394.

Wang, D., Wu, X., Jiang, G., Yang, J., Yu, Z., Yang, Y., Yang, W., Niu, X., Tang, K., and Gong, J. (2022). Systematic analysis of the effects of genetic variants on chromatin accessibility to decipher functional variants in non-coding regions. *Front. Oncol.*, 12:1035855.

Wang, G., van der Walt, J. M., Mayhew, G., Li, Y.-J., Züchner, S., Scott, W. K., Martin, E. R., and Vance, J. M. (2008a). Variation in the miRNA-433 binding site of FGF20 confers risk for parkinson disease by overexpression of alpha-synuclein. *Am. J. Hum. Genet.*, 82(2):283–289.

Wang, L., Jia, P., Wolfinger, R. D., Chen, X., and Zhao, Z. (2011b). Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics*, 98(1):1–8.

Wang, P.-Y., Rao, J. N., Zou, T., Liu, L., Xiao, L., Yu, T.-X., Turner, D. J., Gorospe, M., and Wang, J.-Y. (2010a). Post-transcriptional regulation of MEK-1 by polyamines through the RNA-binding protein HuR modulating intestinal epithelial apoptosis. *Biochem. J*, 426(3):293–306.

Wang, R., Nambiar, R., Zheng, D., and Tian, B. (2018). PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res.*, 46(D1):D315–D319.

Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M. G., Glass, C. K., and Kurokawa, R. (2008b). Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature*, 454(7200):126–130.

Wang, X., Goodrich, K. J., Gooding, A. R., Naeem, H., Archer, S., Paucek, R. D., Youmans, D. T., Cech, T. R., and Davidovich, C. (2017a). Targeting of polycomb repressive complex 2 to RNA by short repeats of consecutive guanines. *Mol. Cell*, 65(6):1056–1067.e5.

Wang, X., Paucek, R. D., Gooding, A. R., Brown, Z. Z., Ge, E. J., Muir, T. W., and Cech, T. R. (2017b). Molecular analysis of PRC2 recruitment to DNA in chromatin and its inhibition by RNA. *Nat. Struct. Mol. Biol.*, 24(12):1028–1038.

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N. M., Rot, G., Zupan, B., Curk, T., and Ule, J. (2010b). iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol.*, 8(10):e1000530.

Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Watson, M., Park, Y., and Thoreen, C. (2020). Roadblock-qPCR: A simple and inexpensive strategy for targeted measurements of mRNA stability. *RNA*.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS

catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, 42(Database issue):D1001–6.

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M., and Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.*, 13(9):R50.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis.

Wiestner, A., Tehrani, M., Chiorazzi, M., Wright, G., Gibellini, F., Nakayama, K., Liu, H., Rosenwald, A., Muller-Hermelink, H. K., Ott, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Vose, J., Armitage, J. O., Gascoyne, R. D., Connors, J. M., Campo, E., Montserrat, E., Bosch, F., Smeland, E. B., Kvaloy, S., Holte, H., Delabie, J., Fisher, R. I., Grogan, T. M., Miller, T. P., Wilson, W. H., Jaffe, E. S., and Staudt, L. M. (2007). Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. *Blood*, 109(11):4599–4606.

Wilkinson, F. H., Park, K., and Atchison, M. L. (2006). Polycomb recruitment to DNA in vivo by the YY1 REPO domain. *Proc. Natl. Acad. Sci. U. S. A.*, 103(51):19296–19301.

Wolfe, M. S. and Guénette, S. Y. (2007). APP at a glance. *J. Cell Sci.*, 120(Pt 18):3157–3161.

Wolffe, A. P. (1989). Dominant and specific repression of xenopus oocyte 5S RNA genes and satellite I DNA by histone H1. *EMBO J.*, 8(2):527–537.

Wolffe, A. P. and Guschin, D. (2000). Chromatin structural features and targets that regulate transcription. *J. Struct. Biol.*, 129(2-3):102–122.

Wood, A. J., Schulz, R., Woodfine, K., Koltowska, K., Beechey, C. V., Peters, J., Bourc'his, D., and Oakey, R. J. (2008). Regulation of alternative polyadenylation by genomic imprinting. *Genes Dev.*, 22(9):1141–1146.

Wu, C. and Huang, J. (2024). Enhancer selectivity across cell types delineates three functionally distinct enhancer-promoter regulation patterns. *BMC Genomics*, 25(1):483.

Wu, H., Deng, R., Chen, X., Wong, W. C., Chen, H., Gao, L., Nie, Y., Wu, W., and Shen, J. (2016). Caveolin-1 is critical for lymphocyte trafficking into central nervous system during experimental autoimmune encephalomyelitis. *J. Neurosci.*, 36(19):5193–5199.

Wulfridge, P. and Sarma, K. (2021). A nuclease- and bisulfite-based strategy captures strand-specific r-loops genome-wide. *Elife*, 10.

Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R. R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M. J., Walther, N., Koch, B., Kueblbeck, M., Ellenberg, J., Zuber, J., Fraser, P., and Peters, J.-M. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.*, 36(24):3573–3599.

Xia, J.-H. and Wei, G.-H. (2018). Oncogenic regulatory circuits driven by 19q13 rs11672691 underlies prostate cancer aggressiveness. *Mol Cell Oncol*, 5(6):e1516451.

Xia, Z., Donehower, L. A., Cooper, T. A., Neilson, J. R., Wheeler, D. A., Wagner, E. J., and Li, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat. Commun.*, 5:5274.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345.

Xiong, L., Wu, F., Wu, Q., Xu, L., Cheung, O. K., Kang, W., Mok, M. T., Szeto, L. L. M., Lun, C.-Y., Lung, R. W., Zhang, J., Yu, K. H., Lee, S.-D., Huang, G., Wang, C.-M., Liu, J., Yu, Z., Yu, D.-Y., Chou, J.-L., Huang, W.-H., Feng, B., Cheung, Y.-S., Lai, P. B., Tan, P., Wong, N., Chan, M. W., Huang, T. H., Yip, K. Y., Cheng, A. S., and To, K.-F. (2019). Aberrant enhancer hypomethylation contributes to hepatic carcinogenesis through global transcriptional reprogramming. *Nat. Commun.*, 10(1):335.

Xu, H., An, J. J., and Xu, B. (2017). Distinct cellular toxicity of two mutant huntingtin mRNA variants due to translation regulation. *PLoS One*, 12(5):e0177610.

Xu, J., Xu, X., Huang, D., Luo, Y., Lin, L., Bai, X., Zheng, Y., Yang, Q., Cheng, Y., Huang, A., Shi, J., Bo, X., Gu, J., and Chen, H. (2024). A comprehensive benchmarking with interpretation and operational guidance for the hierarchy of topologically associating domains. *Nat. Commun.*, 15(1):4376.

Xu, W., He, C., Kaye, E. G., Li, J., Mu, M., Nelson, G. M., Dong, L., Wang, J., Wu, F., Shi, Y. G., Adelman, K., Lan, F., Shi, Y., and Shen, H. (2022). Dynamic control of chromatin-associated m6a methylation regulates nascent RNA synthesis. *Mol. Cell*, 0(0).

Yamaguchi, Y., Inukai, N., Narita, T., Wada, T., and Handa, H. (2002). Evidence that negative elongation factor represses transcription elongation through binding to a DRB Sensitivity-Inducing Factor/RNA polymerase II complex and RNA. *Mol. Cell. Biol.*, 22(9).

Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., and Handa, H. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, 97(1).

Yan, Q., Shields, E. J., Bonasio, R., and Sarma, K. (2019). Mapping native R-Loops genome-wide using a targeted nuclease approach. *Cell Rep.*, 29(5):1369–1380.e5.

Yang, A. C., Stevens, M. Y., Chen, M. B., Lee, D. P., Stähli, D., Gate, D., Contrepois, K., Chen, W., Iram, T., Zhang, L., Vest, R. T., Chaney, A., Lehallier, B., Olsson, N., du Bois, H., Hsieh, R., Cropper, H. C., Berdnik, D., Li, L., Wang, E. Y., Traber, G. M., Bertozzi, C. R., Luo, J., Snyder, M. P., Elias, J. E., Quake, S. R., James, M. L., and Wyss-Coray, T. (2020). Physiological blood-brain transport is impaired with age by a shift in transcytosis. *Nature*, 583(7816):425–430.

Yang, H., Wei, C., Li, Q., Shou, T., Yang, Y., Xiao, C., Yu, M., Li, M., Yang, Z., Zhang, J., and Zheng, B. (2013). Association of TLR4 gene non-missense single nucleotide polymorphisms with rheumatoid arthritis in chinese han population. *Rheumatol. Int.*, 33(5):1283–1288.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42(7):565–569.

Yang, M. G., Ling, E., Cowley, C. J., Greenberg, M. E., and Vierbuchen, T. (2022). Characterization of sequence determinants of enhancer function using natural genetic variation. *Elife*, 11.

Yang, Q., She, H., Gearing, M., Colla, E., Lee, M., Shacka, J. J., and Mao, Z. (2009). Regulation of neuronal survival factor MEF2D by chaperone-mediated autophagy. *Science*, 323(5910):124–127.

Yang, Z., Zhu, Q., Luo, K., and Zhou, Q. (2001). The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature*, 414(6861):317–322.

Yerbury, J. J., Poon, S., Meehan, S., Thompson, B., Kumita, J. R., Dobson, C. M., and Wilson, M. R. (2007). The extracellular chaperone clusterin influences amyloid formation and toxicity by interacting with prefibrillar structures. *FASEB J.*, 21(10):2312–2322.

Youn, H. D., Grozinger, C. M., and Liu, J. O. (2000). Calcium regulates transcriptional repression of myocyte enhancer factor 2 by histone deacetylase 4. *J. Biol. Chem.*, 275(29):22563–22567.

Young, C. P., Hillyer, C., Hokamp, K., Fitzpatrick, D. J., Konstantinov, N. K., Welty, J. S., Ness, S. A., Werner-Washburne, M., Fleming, A. B., and Osley, M. A. (2017). Distinct histone methylation and transcription profiles are established during the development of cellular quiescence in yeast. *BMC Genomics*, 18(1):107.

Yu, T.-X., Wang, P.-Y., Rao, J. N., Zou, T., Liu, L., Xiao, L., Gorospe, M., and Wang, J.-Y. (2011). Chk2-dependent HuR phosphorylation regulates occludin mRNA translation and epithelial barrier function. *Nucleic Acids Res.*, 39(19):8472–8487.

Zaret, K. S. and Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.*, 25(21):2227–2241.

Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.*, 21(8):1273–1283.

Zhang, H., Lee, J. Y., and Tian, B. (2005). Biased alternative polyadenylation in human tissues. *Genome Biol.*, 6(12):R100.

Zhang, J., Lee, D., Dhiman, V., Jiang, P., Xu, J., McGillivray, P., Yang, H., Liu, J., Meyerson, W., Clarke, D., Gu, M., Li, S., Lou, S., Xu, J., Lochovsky, L., Ung, M., Ma, L., Yu, S., Cao, Q., Harmanci, A., Yan, K.-K., Sethi, A., Gürsoy, G., Schoenberg, M. R., Rozowsky, J., Warrell, J., Emani, P., Yang, Y. T., Galeev, T., Kong, X., Liu, S., Li, X., Krishnan, J., Feng, Y., Rivera-Mulia, J. C., Adrian, J., Broach, J. R., Bolt, M., Moran, J., Fitzgerald, D., Dileep, V., Liu, T., Mei, S., Sasaki, T., Trevilla-Garcia, C., Wang, S., Wang, Y., Zang, C., Wang, D., Klein, R. J., Snyder, M., Gilbert, D. M., Yip, K., Cheng, C., Yue, F., Liu, X. S., White, K. P., and Gerstein, M. (2020a). An integrative ENCODE resource for cancer genomics. *Nat. Commun.*, 11(1):3696.

Zhang, J., Li, C. C., Li, X., Liu, Y., Wang, Q., Zhang, G., Xiong, H., He, A., and Ai, S. (2023). Novel enhancers conferring compensatory transcriptional regulation of Nkx2-5 in heart development. *iScience*, 26(4):106509.

Zhang, S., Cooper-Knock, J., Weimer, A. K., Shi, M., Moll, T., Marshall, J. N. G., Harvey, C., Nezhad, H. G., Franklin, J., Souza, C. D. S., Ning, K., Wang, C., Li, J., Dilliott, A. A., Farhan, S., Elhaik, E., Pasniceanu, I., Livesey, M. R., Eitan, C., Hornstein, E., Kenna, K. P., Project MinE ALS Sequencing Consortium, Veldink, J. H., Ferraiuolo, L., Shaw, P. J., and Snyder, M. P. (2022). Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. *Neuron*, 110(6):992–1008.e11.

Zhang, T., Zhang, Z., Dong, Q., Xiong, J., and Zhu, B. (2020b). Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol.*, 21(1):45.

Zhang, X., Choi, P. S., Francis, J. M., Imielinski, M., Watanabe, H., Cherniack, A. D., and Meyerson, M. (2016). Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat. Genet.*, 48(2):176–182.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-seq (MACS). *Genome Biol.*, 9(9):R137.

Zhang, Z., Zong, C., Jiang, M., Hu, H., Cheng, X., Ni, J., Yi, X., Jiang, B., Tian, F., Chang, M.-W., Su, W., Zhu, L., Li, J., Xiang, X., Miao, C., Gorospe, M., de Cabo, R., Dou, Y., Ju, Z., Yang, J., Jiang, C., Yang, Z., and Wang, W. (2020c). Hepatic HuR modulates lipid homeostasis in response to high-fat diet. *Nature Communications*, 11(1):3067.

Zhao, Y., Wang, L., Ren, S., Wang, L., Blackburn, P. R., McNulty, M. S., Gao, X., Qiao, M., Vessella, R. L., Kohli, M., Zhang, J., Karnes, R. J., Tindall, D. J., Kim, Y., MacLeod, R., Ekker, S. C., Kang, T., Sun, Y., and Huang, H. (2016). Activation of P-TEFb by androgen Receptor-Regulated enhancer RNAs in Castration-Resistant prostate cancer. *Cell Rep.*, 15(3).

Zhong, X., Leontis, N., Qian, S., Itaya, A., Qi, Y., Boris-Lawrie, K., and Ding, B. (2006). Tertiary structural and functional analyses of a viroid RNA motif by isostericity matrix and mutagenesis reveal its essential role in replication. *J. Virol.*, 80(17):8566–8581.

Zhu, Y., Sun, L., Chen, Z., Whitaker, J. W., Wang, T., and Wang, W. (2013). Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res.*, 41(22):10032–10043.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, 48(5):481–487.

Zimnicka, A. M., Husain, Y. S., Shajahan, A. N., Sverdlov, M., Chaga, O., Chen, Z., Toth, P. T., Klomp, J., Karginov, A. V., Tiruppathi, C., Malik, A. B., and Minshall, R. D. (2016). Src-dependent phosphorylation of caveolin-1 tyr-14 promotes swelling and release of caveolae. *Mol. Biol. Cell*, 27(13):2090–2106.

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70.

Zuin, J., Dixon, J. R., van der Reijden, M. I. J. A., Ye, Z., Kolovos, P., Brouwer, R. W. W., van de Corput, M. P. C., van de Werken, H. J. G., Knoch, T. A., van IJcken,

W. F. J., Grosveld, F. G., Ren, B., and Wendt, K. S. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U. S. A.*, 111(3):996–1001.

Zuin, J., Roth, G., Zhan, Y., Cramard, J., Redolfi, J., Piskadlo, E., Mach, P., Kryzhanovska, M., Tihanyi, G., Kohler, H., Eder, M., Leemans, C., van Steensel, B., Meister, P., Smallwood, S., and Giorgetti, L. (2022). Nonlinear control of transcription through enhancer-promoter interactions. *Nature*, 604(7906):571–577.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9(1):133–148.