The University Of Sheffield.

A bioinformatic exploration of the intersection between DNA damage, neurological disease and unique features of the cerebellum

Jacob Parker

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Molecular Biology and Biotechnology
Department/School of Science

Submission Date: 30/09/2019

Acknowledgements

Firstly, thank you to my supervisors Ian and Sherif who have supported me throughout this PhD. To Ian who has endured all sorts of office interruptions and questions, your patience has not gone unnoticed. To Sherif, I am very grateful for the part time work that has enabled me to not go penniless for the last few months. The Sudbery and El-Khamisy labs also deserve a measure of gratitude for their general good company and help over the years. Thanks also must go to my parents, who supported me financially when my funding ran out and have allowed me to live with them during the write up of this thesis, and to my family in general for their love and encouragement. I am indebted to Emily for her constant kindness, willingness to put up with long hours in the lab and long absences in general, meals after taxing days, encouraging cards and patience in general throughout this whole process. Your character is a constant source of aspiration. I am also incredibly grateful for the friends who have helped me through these four years. The Elgin five, whose friendships have shown me what true community is like, and particularly Luke Saunders, who has ever been a source of honest conversion and great laughs. Thanks also to my intellectual sparring partners Samuel Parker, James Thackery and Ben Hawkins, whom I can always rely upon to expand my mind and raise my spirits. Long may we sharpen each other. My community over at St. Thomas Crookes also deserves commendation for their emotional and spiritual support over these years.

This thesis is dedicated to the ground of being

"Oh, what am I to think

Of what the writing of a thousand lifetimes

Could not explain

If all the forest trees were pens

And all the oceans, ink?"

Declaration

I, the author, confirm that the Thesis is my own work. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not been previously been presented for an award at this, or any other, university.

Abstract

The problem of how a germline mutation present in every cell of the body can have tissue specific effects is often a complex one. However, tissue specific diseases arising from mutations in housekeeping genes that one would presume to be integral for the survival of the majority of cells represent interesting, and often unsolved cases of tissue specificity. This is all the more striking when the case in question affects a tissue but not tissues that happen to be extremely similar. Enter the autosomal recessive cerebellar ataxias (ARCAs), a group of neurodegenerative diseases whose major clinical similarity is that they all involve atrophy of the cerebellum but strangely, not the cerebral tissues. Many of the ARCAs are caused by mutations in proteins that are involved in DNA repair, particularly single strand break repair, functions one would intuitively think to be relevant for every cell. Despite much work characterising the ARCAs and the mechanisms behind their pathologies, there has yet to be a unifying theory underpinning their tissue specificity. This thesis seeks to explore the underlying mechanisms that may account for the sensitivity of the cerebellum to defects in DNA repair. In doing so, we describe several features with respect to which the cerebellum is unique relative to the cerebrum. These include lower expression of mitochondrially-associated genes, more genes with a higher rate of mapping mismatches, a higher rate of RNA editing and an increased load of germline variant calls. We also explore the role of the structural protein NuMA in the DNA damage response, and find a set of genes whose upregulation upon cellular exposure to oxidative damage is dependent upon NuMA, that NuMA is enriched at promoter regions, particularly for specific categories of genes, and that loss of NuMA leads to an increase in DNA damage at promoters.

Contents

Table of Figures

Abbreviations

4sU – 4-thiouridine

4sU-seq – 4-thiouridine sequencing

8-oxoG – 8-oxoguanine

53BP1 – Tumor Protein P53 Binding Protein 1

A – Adenine

AF – Allele frequency

AID – Activation induced cytidine deaminase

ALS – Amyotrophic lateral sclerosis

AMP – Adenosine monophosphate

AOA1/2/4 – Ataxia with oculomotor apraxia 1/2/4

APE1/2/APEX1/2 – Apurinic/Apyrimidinic endodeoxyribonuclease 1/2

APOBEC – Apolipoprotein B mRNA editing enzyme catalytic subunit

AP-seq – Apurinic site sequencing

APTX – Aprataxin

ARCA – Autosomal recessive cerebellar ataxia

ARP2/3 – Actin-related protein 2/3

AT – Ataxia telangiectasia

AT-LD – Ataxia telangiectasia like-disorder

ATM – Ataxia telangiectasia mutated

ATP – Adenosine triphosphate

ATR – Ataxia telangiectasia and Rad3-related protein

BER – Base excision repair

BLISS – Breaks labelling in situ and sequencing

BRCA1 – Breast cancer type 1 susceptibility protein/Fanconi anaemia, complementation group S

BS – BrainSpan

C – Cytosine

CB/CER – Cerebellum

CGAT – Computational Genomics Analysis Tools

CHE – Cerebellar hemisphere

ChIP – Chromatin immunoprecipitation

ChIP-seq – Chromatin immunoprecipitation sequencing

CS – Cockayne syndrome

DDR – DNA damage response

DE – Differentially expressed

DMEM – Dulbecco's Modified Eagle Medium

DNA – Deoxyribonucleic acid

DNA-seq – DNA sequencing

DNA polβ – DNA polymerase β

DRDA-ARCA – DNA damage repair defect associated-autosomal recessive cerebellar ataxia

DSB – Double strand break

DSBR – Double strand break repair

FANCD1 – Breast cancer type 2 susceptibility protein/Fanconi anemia group D1 protein

FCX – Frontal cortex

FDR – False discovery rate

FEN1 – Flap endonuclease 1

FI1G – Fragile intron one containing gene

FIG – Fragile intron gene

FPG – Fragile promoter gene

G – Guanine

GATK – Genome analysis tool kit

GG-repair – Global genome repair

GO – Gene ontology

GTEx - Genotype-tissue expression

H2AX/γ-H2AX – H2A histone family member X/phosphorylated-H2A histone family member X

H3K4me2 – Histone H3 lysine 4 dimethylation

H3K9ac – Histone H3 lysine 9 acetylation

H3K9me3 - Histone H3 lysine 9 trimethylation

H3K36me3 – Histone H3 lysine 36 trimethylation

H_2O_2 – Hydrogen peroxide

HGPS – Hutchinson Gilford progeria syndrome

HMGB1 – High Mobility Group Box 1

HR – Homologous recombination

IERG – Immediate early response gene

IgG – Immunoglobulin G

IPC – Inferior parietal cortex

Kb/KB – Kilobase

KD – Knockdown

KO – Knockout

LAD – Lamina-associated domain

LogFC – Log fold-change

M/SAR – Matrix/Scaffold attachment region

MADA – mandibuloacral dysplasia type A

MCSZ – Microcephaly with seizures

MEF – Mouse embryonic fibroblast

MEM – Minimum essential medium eagle

MM – Mismatched

MMR – Mismatch rate

Mt/mt – Mitochondrial

mtDNA – Mitochondrial DNA

mtTOP1 – Mitochondrial topoisomerase 1

N/A – Not avilable

NAD – Nicotinamide adenine dinucleotide

NBS – Nijmegen breakage syndrome

NBS1 – Nijmegen breakage syndrome 1

NBS-LD – Nijmegen breakage syndrome-like disorder

NER – Nucleotide excision repair

NHEJ – Non-homologous end joining

NRF1/2 – Nuclear factor, erythroid 2 Like 1/2

NRG – NuMA regulated gene

N.S/ns – Non-significant

NuMA – Nuclear Mitotic Apparatus Protein 1

OFC – Orbital frontal cortex

OR – Odds ratio

PAR – Poly(ADP-Ribose)

PARP/PARP1/2 – Poly(ADP-ribose) polymerase/Poly(ADP-ribose) polymerase ½

PCNA – Proliferating Cell Nuclear Antigen

PCR – Polymerase chain reaction

PDSS1 – Decaprenyl diphosphate synthase subunit 1

PNKP – Polynucleotide kinase 3'-phosphatase

Pol II – RNA polymerase II

PR – Pausing ratio/index

PVC – Primary visual cortex

qPCR – Quantitative polymerase chain reaction

rDNA – ribosomal DNA

RER – RNA editing rate

RNA – Ribonucleic acid

RNAi – RNA interference

RNA-seq – RNA sequencing

RNaseH1 – Ribonuclease H1

RPA1 – Replication Protein A1

ROS – Reactive oxygen species

RS-SCID – Radiosensitive severe combined immunodeficiency

SCA1 - Spinocerebellar ataxia 1

SCAN1 – Spinocerebellar ataxia with peripheral neuropathy 1

SDH – Succinate dehydrogenase

SETX – Senataxin

shRNA – short hairpin RNA

SNF2h - Sucrose nonfermenting protein 2 homolog

SNP – Single nucleotide polymorphism

SNV – Single nucleotide variant

SSB – Single strand break

SSBR – Single strand break repair

ssDNA – Single stranded DNA

T – Thymine

TC-repair – Transcription-coupled repair

TDP1 - Tyrosyl-DNA phosphodiesterase 1

TEC – Temporal cortex

TOP1 – Topoisomerase 1

TOP1cc - Topoisomerase 1 cleavage complex

TPM – Transcripts per million

tRNA – transfer RNA

TSS – Transcription start site

TTD – Trichothiodystrophy

TTS – Transcription termination site

UTR – Untranslated region

U - Uracil

UV – Ultraviolet

WES – Whole exome sequencing

WT – Wild type

XP – Xeroderma pigmentosum

XRCC1/4 – X-Ray repair cross complementing 1/4

XLF – XRCC4-like factor

1. Chapter 1 – Introduction

1.1 Dysfunctional DNA Damage Repair and the Brain

1.1.1 Sources of DNA damage

It is estimated that a typical mammalian cell is subject to tens of thousands of DNA lesions per day (Reviewed in Hoeijmakers, 2009). Some of these are exogenous, such as ionising radiation and UV light, but DNA is also damaged as a result of internal factors. Endogenous sources of DNA damage can be chemical attack via reactive oxygen species generated during metabolism or alkylating agents, amongst others (Reviewed in Barnes & Lindahl, 2004). These lesions can then interfere with normal cellular processes leading to the exacerbation of the initial damage. For example, the collision of replication forks with DNA adducts can lead to formation of double strand breaks (DSBs) (Ensminger et al., 2014). Additionally, the repair pathways that resolve some of these lesions, such as the base excision repair (BER) pathway, themselves involve the formation of single stranded DNA breaks (SSBs) as part of the repair process (Caldecott, 2008). Ribose contamination of DNA also represents a major contributor to the total load of endogenous DNA damage, as does the aberrant activity of enzymes that process or repair DNA (Ahel et al., 2006; Pourquier et al., 1997, 1999; Reijns et al., 2012), both of which will be discussed in detail later in this introduction. If left unchecked, these abnormalities within DNA can negatively impact transcription, replication and in turn, viability; it is therefore important for cells to have various systems in place for effectively dealing with the wide range of possible DNA lesions (Bendixen et al., 1990; Kathe et al., 2004; Kuzminov, 2001).

1.1.2 Perturbations in DNA damage Repair: Cancer or Neurological Disease?

Interestingly, mutations in proteins that process or repair DNA often lead to cancer, immunological deficiency or neurological disease and sometimes a combination of these. This would at first seem to be contradictory; neurodegeneration and neurodevelopment defects are diseases of cell death and cancer is cells acquiring abnormal longevity and levels of growth. However, this difference arises from the fact that abnormalities in distinct repair pathways will differentially affect disparate cell types. Mutations in dividing cells may occur in oncogenes, which can enable them to bypass cell cycle checkpoints or prevent apoptosis.

Crucially, there are fewer limits to becoming cancerous placed upon dividing cells compared to quiescent cells, as more barriers to uncontrolled replication need to be overcome - there is even evidence that forcing neurons to re-enter the cell cycle results in their subsequent death (Herrup and Busser, 1995) (Reviewed in Herrup, 2004). Why some types of DNA damage repair defects cause neurological defects and not cancer is less well understood. Multiple general explanations have been put forward for this phenomenon, such as the fact that neural cells are non-cycling and so unable to use homologous recombination as a form of repair and that the brain has a high rate of oxidative metabolism, consuming 20% of the body's oxygen (DNA damage and neurodegeneration reviewed in Madabhushi et al. 2014). However, whether these genuinely contribute to the sensitivity of neural cells in specific neurological disorders associated with defects in DNA damage repair has proven difficult to elucidate. Furthermore, it is sometimes unclear which pathways cause cancer, which lead to neurodevelopmental abnormalities and which result to progressive neurological disorders that manifest postnatally, or any combination thereof. Maintaining genome integrity is a complex process, with some proteins involved in several repair pathways and others providing complementary or redundant functions, sometimes resulting in variable clinical features for dysfunctional proteins in the same pathway, or even different mutations in the same protein. As such, it is not always easy to parse which loss of function is causing which phenotype. However, some patterns have emerged, which will be discussed in the following sections.

1.1.3 The Consequences of Defective Double Strand Break Repair

Double strand break repair (DSBR) consist of two pathways: homologous recombination (HR) and non-homologous end joining (NHEJ). Homologous recombination is the process by which repair occurs by using the sister chromatid as a template for filling in the gap created by the double strand break. As such, HR cannot occur in non-dividing cells, as these cells do not replicate their DNA and so do not have a homologous chromosome with which to perform HR (Reviewed Wright et al., 2018). NHEJ on the other hand is when the two sides of the break are directly ligated together, after the removal of lesions or mismatched bases from the ends and filling in of lost DNA if such steps are necessary. Therefore, NHEJ does not have the same requirement for DNA replication as HR, and so is the mechanism of DSB repair used in non-dividing cells or replicative cells that are in G1 and so do not have a sister

chromatid available for HR (Reviewed in Pannunzio et al., 2018). Perturbations in DSB repair can both lead to increased susceptibility to cancer and neurological disease, primarily neurodevelopmental defects. Abrogation of HR is generally fatal because it is required to deal with the DSBs formed during the rapid cell proliferation that the neural progenitor cells initially go through prior to differentiation (Orii et al., 2006). However, some mutations that leave enough residual protein activity to enable development are viable, but have severe phenotypic consequences, often neurological, as in the case of mutations in FANCD1/BRCA2 (Alter et al., 2007). As these cells begin to differentiate and move into G0/G1, NHEJ becomes the mechanism for DSB repair as a sister chromatid is no longer available for HR. Because of this, complete loss of function mutations in NHEJ also causes embryonic lethality, although hypomorphic mutations in several NHEJ proteins have been reported, such as DNA Ligase IV, Artemis and XLF/Cernunnos - all of which manifest in a disease termed radiosensitive severe combined immunodeficiency (RS-SCID). Mutations in both HR and NHEJ, when viable, often result in microcephaly as a common clinical feature due to the loss of progenitor cells during development. (Alter et al., 2007; Buck et al., 2006; Driscoll et al., 2001; Noordzij et al., 2003; Orii et al., 2006). However, this is not always the case, which could be due to partial functional redundancy amongst proteins acting in the same or similar pathways, differences in protein activity level between hypomorphic mutations or perhaps disparities in the relative importance of proteins to a given repair pathway.

Mutations in the signalling proteins that coordinate DSBR also lead to neurological dysfunction, although the pathology is more variable depending on the protein involved compared to the mutations occurring in factors directly involved with HR or NHEJ. The MRN complex is an important player in DSB repair signalling, being involved in both lesion detection and the initiation of downstream responses (Reviewed in Lamarche et al., 2010), and hypomorphic mutations in any of its three constituent proteins, MRE11, NBS1 and RAD50, have been identified in patients suffering from neurological disease. Mutations in MRE11 manifest in ataxia telangiectasia like-disorder (AT-LD), NBS1 in Nijmegen breakage syndrome (NBS) and RAD50 in Nijmegen breakage syndrome-like disorder (NBS-LD). Interestingly, whilst NBS and NBS-LD have microcephaly as a common feature, and so are likely pathological at a neurodevelopmental level, AT-LD recapitulates the neurodegenerative phenotype seen in ataxia telangiectasia (AT) (Carney et al., 1998;

Matsuura et al., 1998; Stewart et al., 1999; Varon et al., 1998; Waltes et al., 2009). AT itself is caused by mutations in the serine/threonine kinase ataxia –telangiectasia mutated (ATM) (Savitsky et al., 1995), which both detects DSBs and is activated by the MRN complex, upon which it auto-phosphorylates and coordinates downstream DNA repair, cell cycle checkpoints and, if necessary, apoptosis (Reviewed in Ambrose and Gatti, 2013; McKinnon, 2012). AT, NBS and NBS-LD all involve increased susceptibility to cancer, whereas AT-LD does not, highlighting the diversity in outcomes of mutations even within the same protein complex (Chun and Gatti, 2004; Digweed and Sperling, 2004; Stewart et al., 1999; Waltes et al., 2009).

1.1.3.1 The Pathology of defects in Single Strand Lesion Repair

Mutations in proteins involved with single stranded lesion repair, which consists of nucleotide excision repair (NER), single strand break repair (SSBR) and base excision repair (BER) are less associated with neurodevelopmental abnormalities than DSBR defects, although there are exceptions. Rather, defective function, if it involves neural pathology at all, most often result in neurodegeneration. Non-neural tissues do not degenerate in the same fashion, although some single strand lesion repair defects are linked to increased rates of cancer. The fact this degenerative phenotype is restricted to non-cycling neurons and does not generally affect neurodevelopment could mean alternative pathways for the repair of single stranded DNA breaks and adducts are available in cycling cells. Neural progenitor cells would therefore be able to repair the damage upon the loss of a repair factor, whereas quiescent neurons, lacking these redundant repair pathways, would die, resulting in neurodegeneration. However, if single strand break repair (SSBR) defects were generally as neurodevelopmentally toxic per se as DSBR defects, then we would expect to see massive neural progenitor loss upon differentiation when the cells stop cycling, as with perturbations in NHEJ. The fact that microcephaly occurs only with defects in very specific forms of single strand lesion repair, whilst not disproving the presence of alternative modes of repair in cycling cells, implies that single stranded lesions are less immediately dangerous to neurodevelopment, but rather accrue over time resulting in the progressive loss of neurons. These cannot be readily replaced as in most non-neural tissues, hence neurodegeneration occurs (Reviewed in Rulten and Caldecott, 2013). To further understand how mutations in different single strand lesion repair pathways lead to disparate

neurological diseases, the mechanisms of this type of repair and the pathologies that result from their dysfunction will be further discussed.

1.1.3.2 Nucleotide Excision Repair and Neurological Disease

Nucleotide excision repair is considered part of single strand lesion repair, but is often discussed separately from the specific single strand break repair pathway and base excision repair. This distinction is warranted by the fact that dysfunction within the NER pathway, involved in the removal of bulky lesions, photo-products and a subset of oxidative damage, (De Boer and Hoeijmakers, 2000; Osterod et al., 2002; Thorslund et al., 2005) does result in some diseases that can include neurodevelopmental defects, e.g. microcephaly, as part of their pathology, such as Xeroderma Pigmentosum (XP) and Trichothiodystrophy (TTD) (Faghri et al., 2008; States et al., 1998). Both of these diseases arise from mutations in the proteins of the XP complementation group, XPA-G (Bootsma and Hoeijmakers, 1991; Stefanini et al., 1986). Another NER-linked disease caused by mutations in certain proteins of the XP group and the two proteins of the Cockayne syndrome complementation group, CSA and CSB, is Cockayne syndrome (CS). CS can involve both progressive neurodegeneration and neurodevelopmental defects, but like TTD, and in contrast to XP, it does not involve increased cancer susceptibility. (Reviewed in De Boer and Hoeijmakers, 2000; Nance and Berry, 1992). An interesting phenomenon is seen depending on whether the specific XP group protein mutated is involved in transcription-coupled repair (TC-repair) or repair that occurs independently of transcription, global genome repair (GG-repair). Most of the XP complementation group are involved in both GG and TC-repair, and dysfunction in these proteins often leads to neurological pathology, which occurs in around a quarter of total XP patients. However, patients with mutations in XP-C, which is involved only in GG-NER, less commonly present with abnormalities in brain structure or function, implying that TC-repair is particularly important in the brain (Anttinen et al., 2008; Reviewed in Digiovanna and Kraemer, 2012; Reviewed in Iyama and Wilson, 2013; Mimaki et al., 1986). The fact that XP and CS variably involve both microcephaly and neurodegeneration suggest that TC-NER is key in both neurodevelopment and maintaining genome stability in the developed brain. Indeed, it has been suggested that due to the absence of replication-associated damage in the non-cycling neurons, transcription, being the major form of DNA

processing in these cells, is the major contributor to their total load of DNA damage (Lodato et al., 2015).

1.1.4 Single Strand Break Repair and Neurodegeneration

The general model for SSBR is as follows: SSBs are detected by PARP which activates and recruits several SSBR factors to the site of the break through modifying itself and other targets with poly-ADP-ribose (PAR) (Reviewed in Krishnakumar and Kraus, 2010). XRCC1 is a key protein dependent upon PARP for its accumulation at the SSB site (El-Khamisy et al., 2003). It in turn recruits or stabilises other proteins (Caldecott et al., 1996; Kubota et al., 1996; Vidal et al., 2001; Whitehouse et al., 2001). Repair then proceeds, first end-processing, using different sets of repair factors depending on which adduct is blocking further repair, followed by gap filling and ligation (Reviewed in Caldecott, 2008) (**Fig.1**). SSBs are many times more abundant than DSBs, as evidenced by analysis of breaks arising from oxidative DNA damage (Bradley and Kohn, 1979). If unrepaired, these SSBs represent a threat to genome stability and cellular homeostasis. They can block transcription, cause DSBs through colliding with a replication fork and also have the potential to deplete the cell of NAD+ and ATP due to excessive activation of PARP (Bendixen et al., 1990; Kathe et al., 2004; Kuzminov, 2001; Nagele, 1995; Zhou and Doetsch, 1994). As for their formation, SSBs can arise during Base excision repair (BER), which is often referred to under the category of single strand break repair, as SSBs are a direct intermediate of BER (Reviewed in Hegde et al., 2008). SSBs also arise through the dysfunctional activity of otherwise normal DNA processing enzymes, which can happen if they encounter an abnormal structure in the DNA. For example, Topoisomerase 1, a protein involved in the resolution of DNA supercoils, is known to become trapped on the DNA if it binds close some form of chemical adduct (Pourquier et al., 1997, 1999). Ribonucleotide contamination of DNA, which some research indicates is the most abundant DNA lesion, can generate stable SSBs via their repair due to the premature processing of transient intermediate SSBs by ligase III (Ahel et al., 2006).

As mentioned previously, neurodegeneration rather than neurodevelopmental disease is the primary hallmark of neurological diseases caused by perturbations in SSBR. A particularly interesting feature of many SSB repair defects is that they often involve atrophy of the cerebellum rather than the cerebrum. Mutations in XRCC1 were recently found to

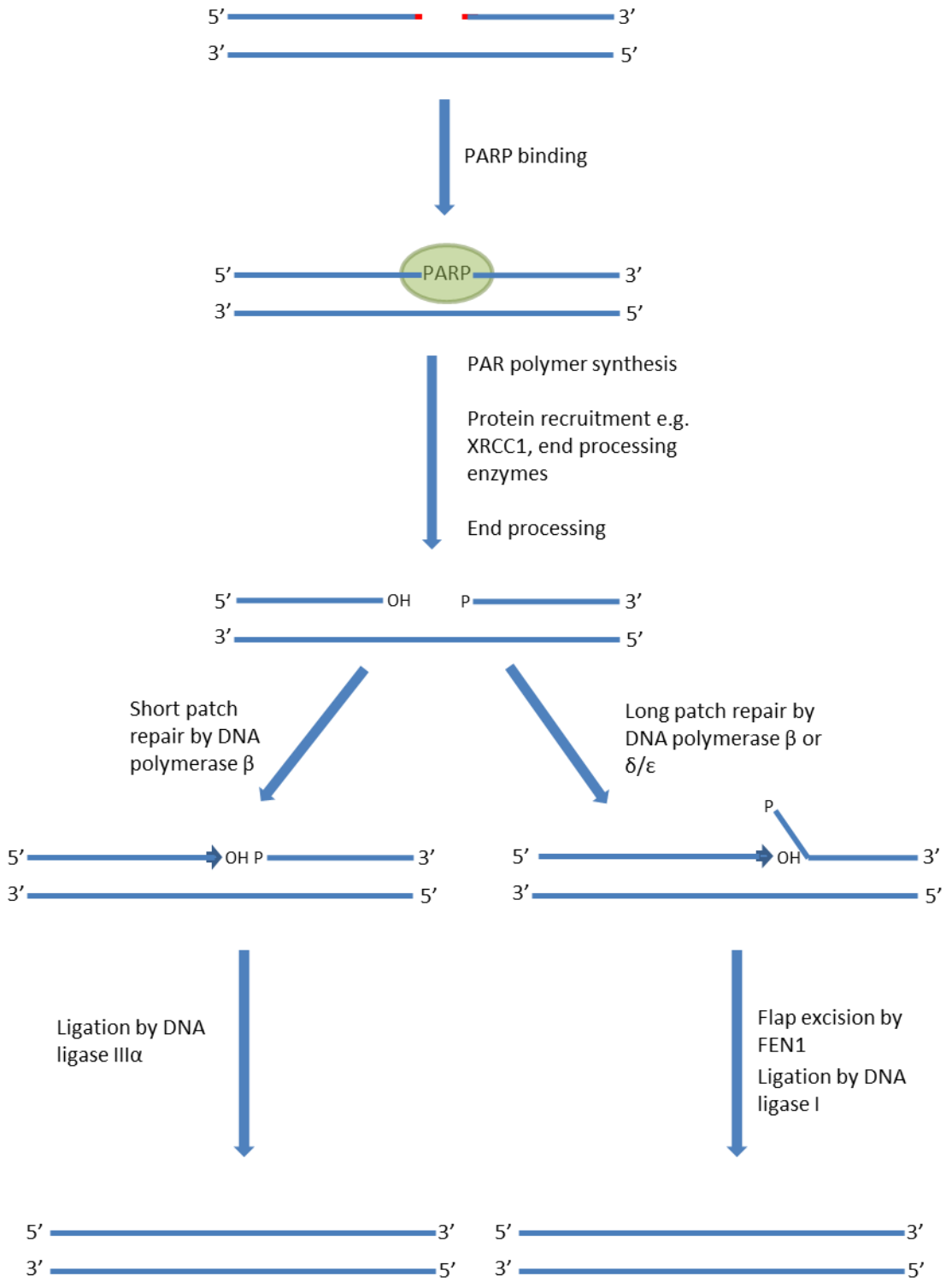


Fig.1.1 A simple overview of the single strand break repair pathway. Single strand breaks are detected by PARP1 which recruits other proteins, most notably XRCC1, to the site of the break. This in turn recruits and stabilises other proteins at the break site, including end processing enzymes. Appropriate end processing then proceeds, followed by either short patch repair for single nucleotide gap filling via DNA polymerase β and ligase III or long patch repair for the filling of larger gaps via DNA polymerase β or δ/ε, FEN1 and ligase I (adapted from Caldecott, 2007)

lead to cerebellar ataxia and oculormotor apraxia, the latter also being a repeating pattern seen amongst several diseases caused by SSBR defects. The mutation itself leads to hyperactivation of PARP1, and so potentially involves the previously discussed depletion of cellular NAD+ and ADP as part of its pathology (Hoch et al., 2017; O'Connor et al., 2018). TOP1-linked DNA breaks are cleaved from the DNA by TDP1 and the abortive activity of ligase III is counteracted by aprataxin (APTX) – mutations in both of these proteins lead to neurological diseases, spinocerebellar ataxia with peripheral neuropathy 1 (SCAN1) and ataxia with oculomotor apraxia 1 (AOA1), which again manifest in progressive neurodegeneration, particularly of the cerebellum (Ahel et al., 2006; Date et al., 2001; Fukuhara et al., 1995; Moreira et al., 2001; Pouliot et al., 1999; Takashima et al., 2002; Yang et al., 1996). PNKP, a protein whose role it is to generate the correct end moieties required for ligation is involved in both SSBR and DSBR (Reviewed in Dumitrache and McKinnon, 2017). It is of note that different mutations in this protein cause two distinct neurological disorders, one neurodevelopmental – microcephaly with seizures (MCSZ), and one involving cerebellar atrophy, ataxia with oculormotor apraxia 4 (AOA4) (Bras et al., 2015; Shen et al., 2010). It is tempting to suggest that these two disease with different causative mutations reflect the inability of PNKP to carry out its role in DSBR in the case of MCSZ and in SSBR for AOA4, as this would fit the general pattern for the neurodegenerative and neurodevelopmental diseases, although this has not been confirmed. All in all, it appears that SSB repair is integral for the maintenance of developed neurons, particularly of those in the cerebellum

1.2 The Autosomal Recessive Cerebellar Ataxias

1.2.1 Introduction to the Autosomal Recessive Cerebellar Ataxias

As discussed briefly, a portion of the aforementioned diseases affect only certain subsets of neural cells or tissues. A prime example of such a set of diseases are the autosomal recessive cerebellar ataxias (ARCAs). Although diverse in their range of clinical phenotypes, the ARCAs can be divided into two major groups on the basis of the system disrupted; one set arising from mutations in proteins that repair or process DNA, and the others being caused by metabolic defects. Many amongst the former group of ARCAs (henceforth referred to as the DNA damage repair defect associated- or DRDA-ARCAs) involve

degeneration of the cerebellum but not the cerebrum. Some of these diseases have already been mentioned (A-T, SCAN1, AOA1, AOA4, XRCC1 mutated), and the majority of the causative proteins are involved in SSBR in one form or another. A notable exception to this is Senataxin, an RNA-DNA helicase implicated in the resolution of R-loops, mutations in which cause ataxia with oculomotor apraxia 2 (Moreira et al., 2004; Yuce and West, 2013). However, R loops are known to have a detrimental effect on genome stability if left unchecked, and so the pathology of AOA2 likely still involves DNA damage (Reviewed in Skourti-Stathaki and Proudfoot, 2014). Several other DRDA-ARCAs are involved in DSBR , such as PNKP and ATM, but these proteins are either known to be involved in or have established links to SSBR (Alagoz et al., 2013; Dong and Tomkinson, 2006; PNKP reviewed in Dumitrache and McKinnon, 2017; Khoronenkova and Dianov, 2015). However, the picture is far from clear. The molecular basis of the cerebellar degeneration phenotype of this set of diseases has proven difficult to unravel. Mutations in other SSBR proteins do not result in cerebellar degeneration, and even defects in the same pathway as the DRDA-ARCA-mutated proteins can lead to neurological disorders that do not involve atrophy of the cerebellum (Reviewed in El-Khamisy, 2011) (**Fig.1.2**). In order to further elucidate the pathology of the DRDA-ARCAs and the similarities and differences between them, this section of the introduction will consist of a brief overview of the architecture of the cerebellum followed by a discussion of three of these diseases in detail: spinocerebellar ataxia with peripheral neuropathy 1, ataxia with oculomotor apraxia and ataxia telangiectasia (For a comprehesive review of the autsomal recessive cerebellar ataxias, see Fogel and Perlman, 2007).

1.2.2 The Architecture of the Cerebellum

The cerebellum as a tissue differs from the cerebrum in a number of striking ways. One immediately obvious visual contrast is that the cerebellum is structurally distinct from the cerebrum. Whereas the cerebral cortex is folded in a broad irregular fashion, the cerebellar cortex on the other hand has an extremely high level of gyrification leading to a regular structure of finely spaced folds. Within each fold, known as a folium, again there is a uniformly organised cellular architecture, unique to the cerebellum, whereby each folium is arranged into layers. The bottom layer, known as the nuclear layer, is made up predominantly of granule cells but also contains interneurons. The middle layer, the Purkinje layer, is a narrow one cell thick strip that houses the cell bodies of Purkinje cells ,

and the cell-poor top layer, the molecular layer, contains the dendrites of Purkinje cells and the axons of granule cells (cerebellar anatomy reviewed in Voogd and Glickstein, 1998). These two types of cell, granule and Purkinje, are considered the two most important sets of cells in the cerebellum, specialised in different fashions for their specific roles. The smaller granule cell is the most common type of cell in the cerebellum and the most abundant neuron in the entire human brain; indeed, it is estimated that 50-80% of all human neurons are cerebellar granule cells. Alternatively, Purkinje cells, being restricted to a single cell layer, are much fewer in number than granule cells with an estimate of the ratio of Purkinje to granule cells reaching as high as 1:2991 (Lange, 1975). Purkinje cells are also much larger than granule cells, having long branching dendrites that extend up through the molecular layer. Whilst being less abundant, Purkinje cells appear to be the major type of cell affected in cerebellar ataxias, with various reports of Purkinje cell loss or dysfunction occurring across many different diseases (Kemp et al., 2016; Sugawara et al., 2007; Xia et al., 2013). They are also exclusively found in the cerebellum, another key difference between the cerebellum and the cerebrum that may be part of the cerebellum's increased susceptibility to degeneration upon the loss of function of various housekeeping proteins.

Aside from neurons, the cerebellum also contains astrocytes and other glial cells. However, whilst in the cerebral cortex these types of cell outnumber neurons at an estimated ratio of 1:3.76, in the cerebellum the reverse is true and neurons dominate, being approximately 4.35 times more abundant than non-neurons. The skew towards neurons in the cerebellum is in fact due to the extremely high numbers of granule cells in the cerebellum rather than a reduced density of astrocytes but nonetheless, these contrasting ratios between the cerebellum and the cerebrum may lead to differences between the two tissues (Azevedo et al., 2009). One area of divergence may be in metabolic processes because of differences in the preferred means of energy production between neurons and astrocytes. Astrocytes are highly glycolytic, and on top of energy production the process of glycolysis is utilised to generate lactate, which is then extruded into the extracellular space for uptake and use by neurons. Conversely, neurons primarily rely on oxidative metabolism, although there is evidence that they can utilise glycolysis when bursts of energy are required (brain metabolism reviewed in Bélanger et al., 2011; Díaz-García and Yellen, 2019; Magistretti and Allaman, 2015). The predominance of neurons over astrocytes in the

cerebellum means that the primary mode of metabolism is oxidative, and so it is likely that this produces more reactive oxygen species and oxidative stress than in the cerebrum. There is also evidence that astrocytes play an important role in protection from oxidative stress in the brain through the production of antioxidants such as glutathione, and neurons themselves appear to rely upon astrocytes as the source of glutathione precursors for their own synthesis of the compound (Dringen et al., 1999). Astrocytes also recycle the oxidised form of the ROS scavenger ascorbic acid and release it back into the extracellular space to be utilised again by neurons (Korcok et al., 2003; Siushansian et al., 1997)(the role of astrocytes in protection from oxidative stress is reviewed in Bélanger and Magistretti, 2009). Therefore the cerebellum may be exposed to a potential double-hit of higher levels of oxidative stress and reduced protection from reactive oxygen species. This proposed increase in exposure to oxidative damage, a specialised microanatomy and unique cellular constitution could each play a part or even act in concert in sensitizing the cerebellum to defects in DNA damage repair associated with the pathology of the DRDA-ARCAs.

1.2.3 DRDA-ARCA Case Studies

1.2.3.1 Spinocerebellar Ataxia with Peripheral Neuropathy 1

Spinocerebellar ataxia with axonal neuropathy, or SCAN1, is an autosomal recessive cerebellar ataxia with an age of onset in the early teens, characterised by cerebellar atrophy, peripheral axonal sensorimotor neuropathy and distal muscle weakness. Its genetic basis is the H493R mutation in tyrosyl DNA-phosphodiesterase 1 (TDP1), whose cellular role is to remove a variety of adducts from DNA termini in order to facilitate further repair (Interthal et al., 2005). The best characterised function of TDP1 in this capacity is the release of trapped topoisomerase 1 cleavage complexes (TOP1cc) from the DNA (Pouliot et al., 1999; Takashima et al., 2002; Yang et al., 1996). Topoisomerase I is a member of the topoisomerase family of enzymes which mediate topological changes in DNA in order to allow its efficient processing by other parts of the cellular machinery. Topoisomerase I itself is involved in the resolution of DNA supercoils formed during replication and transcription, although recent research has revealed TOP1 to function in the suppression and removal of RNA-DNA hybrids (Williams et al., 2013). Its role in relaxing supercoils is however the one most pertinent in this context. It does this by nicking the DNA, rotating around it, and then

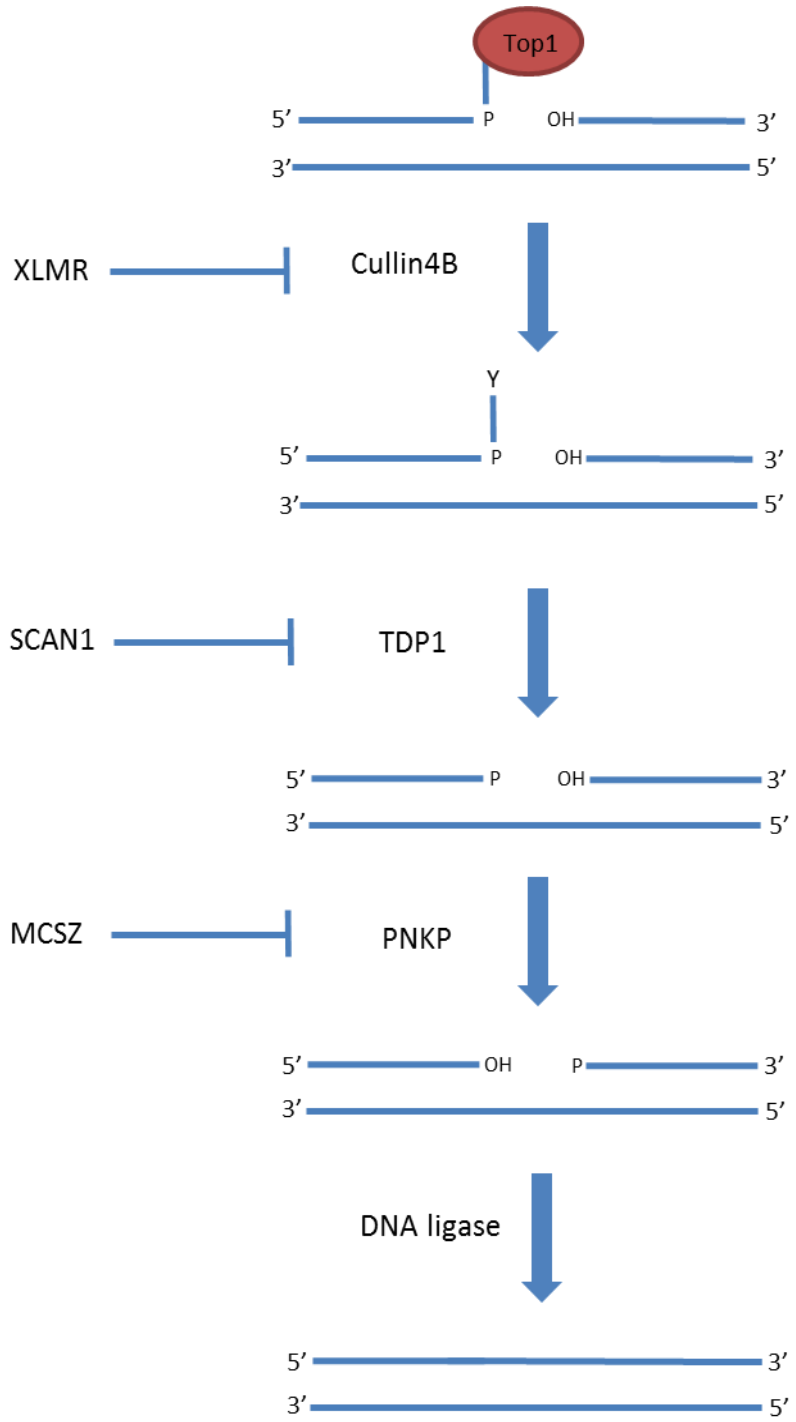


Fig.1.2 Mutations with the same pathway lead to different neurological disorders. When TOP1 becomes trapped on the DNA, the first step is proteasomal degradation of the enzyme, which cullin4B is implicated in. Mutations in this protein lead to X-linked mental retardation (XLMR) (Zou et al., 2007). The following step, cleavage of the phosphotyrosyl bond linking the remaining protein to the DNA, is mediated by TDP1, which when mutated leads to spinocerebellar ataxia with axonal neuropathy (SCAN1)(Takashima et al., 2002). Subsequently, the 3'P and 5'OH are converted to the correct end moieties required for ligation, 5'P and 3'OH. This end processing step is carried out by PNKP, mutations in which result in microcephaly and seizures (MCSZ)(Shen et al., 2010). The final step is the sealing of the nick by DNA ligase. (Adapted from El-Khamisy, 2011)

resealing the nick (Roles and mechansims of action of topoisomerases reviewed in Chen et al., 2013). However, sometimes during this process, TOP1 can become trapped on the DNA, for example, if it nicks close to another DNA lesion (Pourquier et al., 1997, 1999). These trapped TOP1ccs can interfere with transcription elongation and be converted into double stranded DNA breaks if they collide with replication or transcription machinery, thereby representing a major threat to genome stability (Bendixen et al., 1990; Hsiang et al., 1989; Tsao et al., 1993; Wu and Liu, 1997). One known pathway for the repair of these stalled TOP1s the excision of a proteolytically degraded form of this trapped TOP1cc by TDP1 through the breaking of a phospho-tyrosyl bond that links the remaining TOP1 to the DNA (El-Khamisy et al., 2007; Interthal et al., 2001; Pouliot et al., 1999; Yang et al., 1996). However, the TDP1 H493R variant has a significantly reduced rate of catalysis, extending the half-life of the normally transient TDP1-DNA reaction intermediate to around 13 minutes, allowing them to persist in the genome and negatively impact replication and transcription. This theory concerning the molecular basis of SCAN1, dubbed the TDP1 neomorph model, is that the trapped TOP1-DNA complex is replaced by a trapped TDP1-DNA complex, which the cell has no efficient way to repair (Interthal et al. 2005; Hirano et al. 2007). The true situation likely involves both unrepaired trapped TOP1 cleavage complexes and trapped TDP1-DNA complexes formed upon attempted repair having a combinatorial effect on upon genome stability (**Fig.1.3**). However, TOP1 is involved in many different processes, so it is feasible that the trapping of TOP1 on the DNA or it's replacement by TDP1 at those sites could have wide ranging effects on normal cellular function outside of its impact on genome stability.

1.2.3.2 The importance of aprataxin in the resolution of abortive ligation events

Various missense or truncating mutations in the protein Aprataxin (APTX) have been shown to cause AOA1, a childhood onset ataxia involving cerebellar atrophy, peripheral neuropathy, distal amyotrophy, oculomotor apraxia and mental impairment (Date et al., 2001; Fukuhara et al., 1995; Moreira et al., 2001). Like TDP1, APTX is an end processing enzyme, and serves to cleave 5' AMP moieties from DNA termini (Ahel et al., 2006). These 5'AMP-DNA complexes are actually a requisite for the ligation step of DNA repair in mammalian cells, but if a DNA ligase adenylates the 5' terminus at a site of damage prior to correct end processing for the generation of ligatable ends, then resealing of the DNA nick

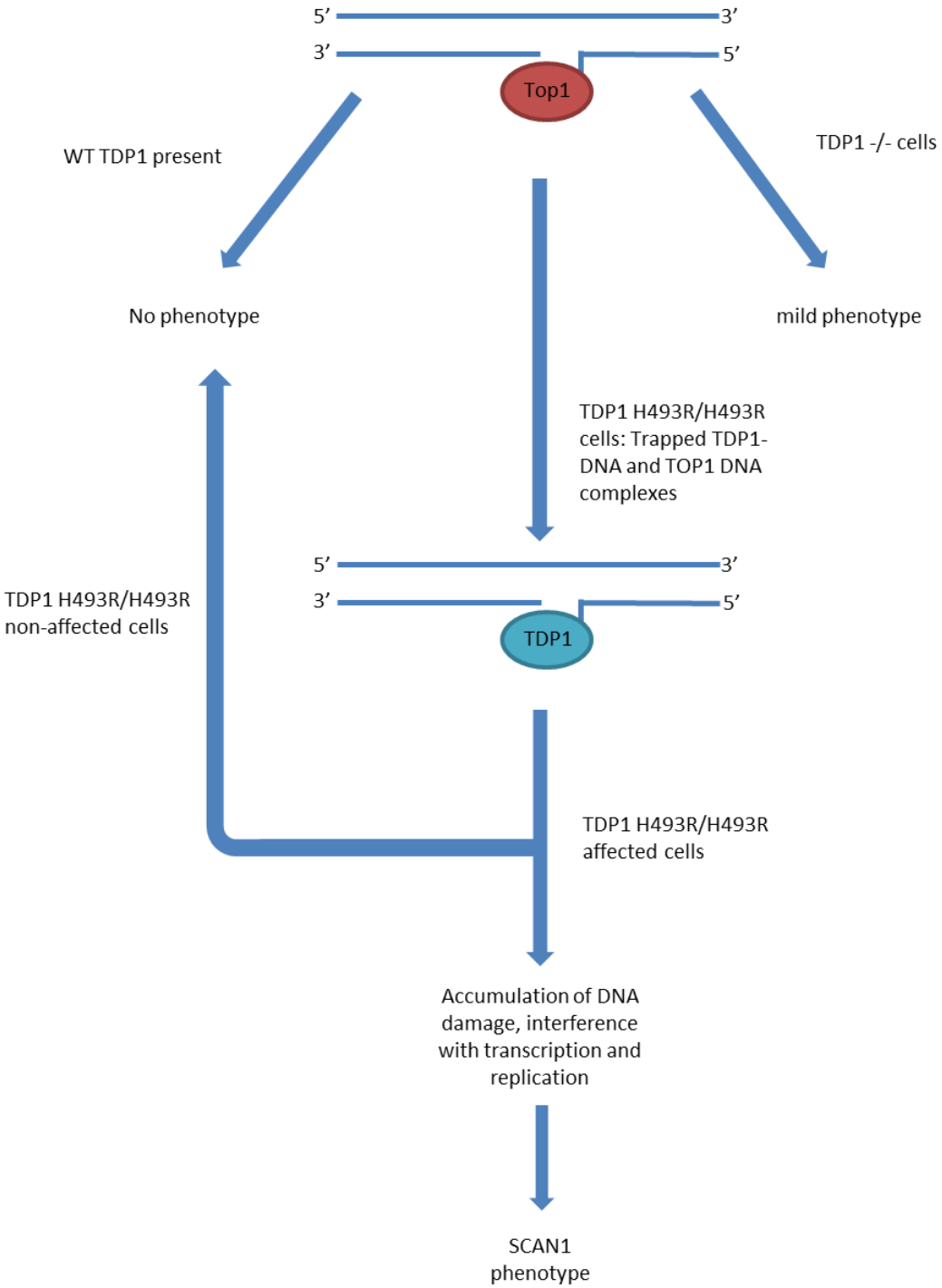


Fig.1.3 Model outlining the proposed theory as to SCAN1 pathogenesis. Trapped TOP1 cleavage complexes that arise in nuclear and mitochondrial DNA are repaired by TDP1 in WT cells and by alternative non-TDP1 dependent pathways in TDP -/- cells. In cells homozygous for the TDP1 H493R mutation, both TDP1-DNA and TOP1-DNA complexes arise. In non-affected cells, there may be repair of these lesions by pathways not present in affected cells, and/or non-affected cells could be more tolerant to mitochondrial dysfunction. In those cell types affected, there may be accumulation of TDP/TOP1-DNA complexes in both nuclear and mtDNA, and/or an increased sensitivity to loss of mitochondrial function. This could ultimately lead to interference with transcription, replication and cellular function, resulting in cell death and the SCAN1 phenotype (Adapted from Hirano et al. 2007)

cannot occur. As AMP is a chemically stable lesion, it must be actively removed from the DNA before further repair can occur (El-Khamisy et al., 2009; Harris et al., 2009; Rass et al., 2007) (For a review of eukayotic DNA ligases and their mechanisms of action, see Ellenberger & Tomkinson 2008). It has recently been shown that these intermediates brought about by abortive ligation events occur often during the removal of ribonucleotides from the genome. *In Vitro* experiments have demonstrated that when a nick is generated at an RNA-DNA junction, in many cases DNA ligase adenylates the 5'RNA, halting further repair. Supporting experiments in yeast revealed that strains which incorporated greater number of ribonucleotides into their genome due to a specific DNA polymerase ε mutation and were deficient for the yeast homolog of APTX, Hnt3, were impaired in their growth relative to those strains with either of the single mutations. Therefore, it has been postulated that the major role of APTX is the resolution of 5'AMP-RNA adducts at nicked RNA-DNA junctions (Tumbale et al., 2014) (**Fig.1.4**). If this is indeed true, the potential importance of APTX in the maintenance of genome stability is put in perspective by data from RNaseH1 KO mice, which estimates that a ribonucleotide is incorporated every 7.6Kb by mouse replicative polymerases, equating to over 1 million ribonucleotides per replicating cell (Reijns et al., 2012). In the absence of APTX it is proposed that attempted excision of RNAs from the genome generates stable 5'AMP-RNA adducts, which accumulate and interfere with vital processes, resulting in cell death and disease (Tumbale et al., 2014)

1.2.3.3 ATM and the DNA damage response

The early childhood onset disease ataxia-telangiectasia is associated with a broad range of clinical phenotypes, including atrophy of the cerebellum, immunological deficiency, ocular telangiectasias - widening of small blood vessels in the eyes - and apraxia – loss of fine motor control - to name but a few. A-T is also distinct from many other DRDA-ARCAs in that it also involves increased susceptibility to cancer (For a review of the complex A-T phenotype, see Chun & Gatti 2004). Concordant with this diverse spectrum of ailments observed in A-T patients, the protein inactivated in A-T, the serine/threonine kinase ataxia – telangiectasia mutated (ATM), has been implicated in the coordination of a wide array of cellular processes (For an overview of the cellular roles of ATM, see see Ambrose & Gatti 2013; McKinnon 2012). The role ATM is best known for, and probably most relevant to the neurodegenerative phenotype of A-T, is the coordination of various DNA damage responses.

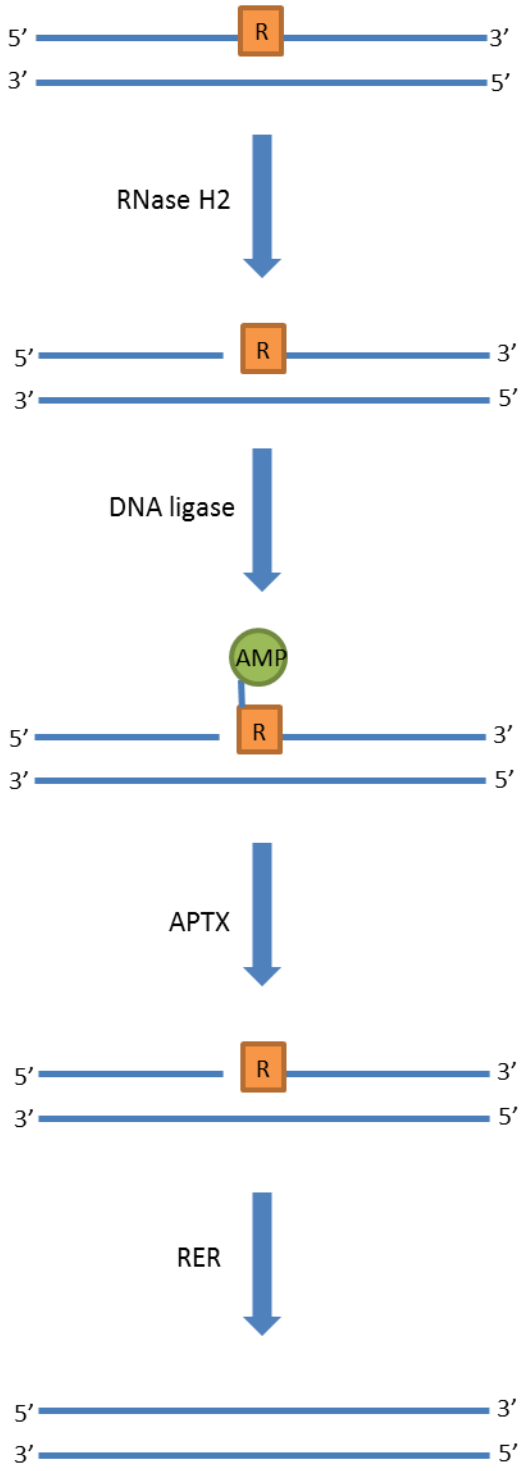


Fig.1.4 Aprataxin cleaves AMP from RNA at nicked RNA-DNA junctions. The removal of ribonucleotides from the genome through ribonucleotide excision repair (RER) involves RNaseH2 nicking 5' of the RNA generating a nicked RNA-DNA junction (Rydberg and Game, 2002). The RNA at the junction is susceptible to adenylation as a result of attempted ligation before the ribonucleotide has been removed. This group is actively removed by APTX, allowing RER to continue as normal and excise the ribonucleotide from the genome.

ATM is activated subsequent to identification of double strand breaks (DSB) by the DNA damage sensing MRN complex, but research has also shown ATM activity in response to oxidative damage independent of DSBs, and alluded to its involvement in the BER pathway (Carson et al., 2003; Chou et al., 2015; Dong and Tomkinson, 2006; Guo et al., 2010; Uziel et al., 2003). Upon activation, ATM through its kinase activity is able to phosphorylate and activate various downstream factors to initiate and coordinate an efficient DNA damage response (Reviewed in McKinnon 2012).

1.2.4 DRDA-ARCAs and Mitochondrial Dysfunction

1.2.4.1 The Brain, Mitochondria and the DRDA-ARCAs

The above sections show how several of the DRDA-ARCA-mutated proteins function in the repair of nuclear DNA, but research over recent years has implicated many of these same proteins in mtDNA maintenance. Additionally, some DRDA-ARCAs are brought about by mutations in proteins exclusive to the mitochondria, such mitochondrial DNA helicases and polymerases. Mitochondria lack the full complement of DNA-repair proteins utilised in the upkeep of nuclear DNA, and are subject to higher levels of oxidative damage, themselves being the source of reactive oxygen species (ROS), and so is it plausible that mtDNA may be more affected than nuclear DNA by the loss of a given repair enzyme (Reviewed in Sykora, Wilson, & Bohr, 2012) (Hudson et al., 1998). Given that the brain metabolises so much of the body's oxygen, it is also possible that neural cells in particular are sensitive to any perturbations in mitochondrial activity. Research into the mitochondrial functions of various DRDA-ARCA-mutated proteins and mitochondrial dysfunction in their respective DRDA-ARCAs is helping to determine the extent to which disruption of mitochondrial function may play a part in the pathology of these diseases and their tissue specificity (For an overview of the ARCAs, see Fogel & Perlman 2007). In order to further explore the pathology of the DRDA-ARCAs in relation to mitochondria, the intersection between the three previously discussed diseases and mitochondrial dysfunction will be discussed.

1.2.4.2 TDP1 and mtDNA repair

TDP1 has been shown to localize to the mitochondria in human cells where it is thought to function in the repair of mitochondrial DNA (mtDNA), although an import mechanism

remains to be determined (Das et al., 2010; Fam et al., 2013). Mitochondria also possess their own specific isoform of TOP1 (mtTOP1) and, like the nuclear isoform, this version of the protein can also become trapped on the DNA during catalysis (Zhang and Pommier, 2008; Zhang et al., 2001). Therefore, it is likely that one of the major functions of TDP1 in mitochondria is the resolution of these lesions, and there is very strong evidence to support this. Chiang et al. recently showed that cells depleted of TDP1 accumulate many times for mitochondrial TOP1 cleavage complexes. This load of trapped mtTOP1 increased even further upon overexpression of a mutant form of mtTOP1 unable to relegate the ends of nick it forms upon binding to DNA, and in this fashion traps itself in an unresolved cleavage complex. Expression of this mutant mtTOP1 also results in a compensatory upregulation of TDP1, another strong indicator that TDP1 is required their removal from the DNA (Chiang et al., 2017). TDP1 does indeed appear resolve mtTOP1-mtDNA complexes, and it may be the case that mitochondria have a greater requirement for TDP1 than the nucleus. There are several reasons underlying this proposal. To start with, it has been demonstrated that oxidative lesions in close proximity to a TOP1 cleavage site effectively inhibit the completion of the reaction and prevent resealing of the DNA (Pourquier et al., 1999). If the same is true of mtTOP1, which acts upon DNA subject to higher levels of oxidative damage relative to nuclear DNA, it is possible that mtDNA accumulates proportionally more trapped TOP1-DNA cleavage complexes, increasing the need for TDP1. This may be particularly relevant, as mitochondria lack several of the endonucleases that have been proposed to mediate a TDP1 independent mechanism of TOP1-cleavage complex repair. Also thought to be absent, or attenuated, are the double strand break repair pathways required post endonuclease activity (Liu et al., 2002; Sykora et al., 2012). Supporting this hypothesis is the observation that various human cells, both neural and peripheral, contain high levels of cytoplasmic TDP1 (Fam et al., 2013; Hirano et al., 2007). Upon induction of oxidative stress via menadione sodium bisulphite and H_2O_2 in cultured human fibroblasts, the existing pool of cellular TDP1 was shown to shift toward cytoplasmic and mitochondrial localisation (Fam et al., 2013). Investigations into mtDNA damage and repair rate through a PCR-based approach in mouse embryonic fibroblasts (MEFs) demonstrated that upon exposure to H_2O_2, MEFs in which TDP1 was absent accrued more DNA damage and were retarded in its repair (Das et al., 2010). Lack of TDP1 also has marked effects upon mitochondrial transcription. Upon TDP1 knockdown in T-REx 293 human cells, a 50% reduction in the levels of a subset of

mitochondrial transcripts was observed. The same group also observed that TDP1 depletion affects mitochondrial bioenergetics, with knockdown cells presenting with 25% decreases in oxygen consumption rate and spare respiratory capacity. Analysis of the products of free radical attack in *Tdp*$^{-/-}$ chicken cells revealed these cells accrued many times more DNA based carbon radicals than both controls and those cells treated with human TDP1 (Chiang et al., 2017). This neatly demonstrates both mitochondrial dysfunction, the increase in the products of free radical attack indicative of elevated ROS production which is in turn a marker of abnormal mitochondrial function, and also that disruption of mitochondrial function has knock on the nuclear DNA. However, whether all this is physiologically relevant to SCAN1 remains to be seen, as *TDP1*$^{-/-}$ mice fail to manifest the full range of SCAN1 symptoms, and there are conflicting reports as to the hallmarks they do present with (Hirano et al., 2007; Katyal et al., 2007). A synthesis of the TDP1 neomorph model and the data regarding the mitochondrial role of TDP1 leads to a speculative scenario whereby TDP1H493R resolves trapped mtTOP1-DNA complexes that accumulate relatively rapidly due to high levels of oxidative stress, but it itself becomes trapped on the DNA. This TDP1 trapping paired with increased levels of trapped mtTOP1ccs subsequently interferes with mitochondrial transcription, replication and ultimately, function and viability.

1.2.4.3 Dual role for APTX in the maintenance of mitochondrial function?

Immunofluorescence experiments in neural-like cells indicate that APTX, like TDP1, also localises to mitochondria, supported by the identification of an APTX isoform with a putative mitochondrial targeting sequence. Further research by the same group indicates that APTX KD cells have increased levels of reactive oxygen species, reduced levels of the mitochondrial enzyme citrate synthase and reduced mtDNA copy number. Additionally, mtDNA was assessed to accumulate 0.7 more lesions per kb in APTX deficient cells compared to controls, whereas nuclear DNA in the KD cells accrued only 0.1 more lesions relatively, as determined by qPCR. Of note is that fact that these mitochondrial dysfunction phenotypes were not recapitulated in lymphoblast cell lines derived from AOA1 patients (Sykora et al., 2011). When Akbari et al. investigated the efficiency of nuclear and mitochondrial extracts from one of these lymphoblast cell lines to repair 5'AMP-DNA adducts *in vitro*, they found that treatment with mitochondrial extracts amassed more reaction intermediates and led to slower repair than incubation with either nuclear extract

or mitochondrial extracts from WT cells. This indicates that non-resolution of the 5'AMP moiety resulting from abortive ligation events is more common in the mitochondria without APTX than in the nucleus, which the group suggested was due to alternative pathways for 5'AMP repair functional in the nucleus, but absent or attenuated in the mitochondria. However, it was reported that the AOA1 lymphoblasts were subject to lower rates of mtDNA damage compared to other cells lacking APTX, raising the question of whether the availability of secondary repair mechanisms varied between cell types (Akbari et al., 2015). In a separate study, APTX-mutant and APTX-KD cells presented with reduced respiratory chain capacity and reduced succinate dehydrogenase (SDH) and coenzyme Q_{10} (CoQ_{10}) levels. This was posited to be due to the absence of APTX leading to a depletion in APE1, a protein known to function in concert with APTX and involved in both base excision repair and regulation of gene expression in response to oxidative stress. This APE1 depletion resulted in the downregulation of proteins downstream of APE1 in the gene regulatory pathway, including NRF2, which is known to mediate expression of SDH and CoQ_{10}. This model has APTX playing a role in the regulation of mitochondrial function independent of its proposed position in the maintenance of mtDNA integrity (Garcia-Diaz et al., 2015). Conflicting reports across different cell lines and studies makes it difficult to say whether any mitochondrial dysfunction in AOA1 is due the inability of APTX to carry out one task or the other. However, it is not difficult to imagine that loss of both synergistically impacts mitochondrial function negatively, and it is premature to dismiss one hypothesis entirely, especially as it is known that ribonucleotides can be incorporated into the mitochondrial genome and their removal may involve the formation of adenylation-susceptible RNA-DNA junctions (**Fig.1.5**) (Kasiviswanathan and Copeland, 2011; Yang et al., 2002).

1.2.4.4 Coordination of mitochondrial homeostasis and mtDNA repair by ATM

Although ATM is not thought to localise to the mitochondria, there have been varying reports of mitochondrial dysfunction in both ATM null and A-T patient cells. In a study utilising A-T patient derived lymphoblastoid cells, Ambrose et al. reported that although the number of mitochondria in WT and A-T cells was similar, ATM deficient cells harboured mitochondria presenting with aberrant structural organisation, lower membrane potential and decreased respiration rates. In addition, there was increased levels of expression for various nuclear encoded oxidative damage response genes whose products were exclusively

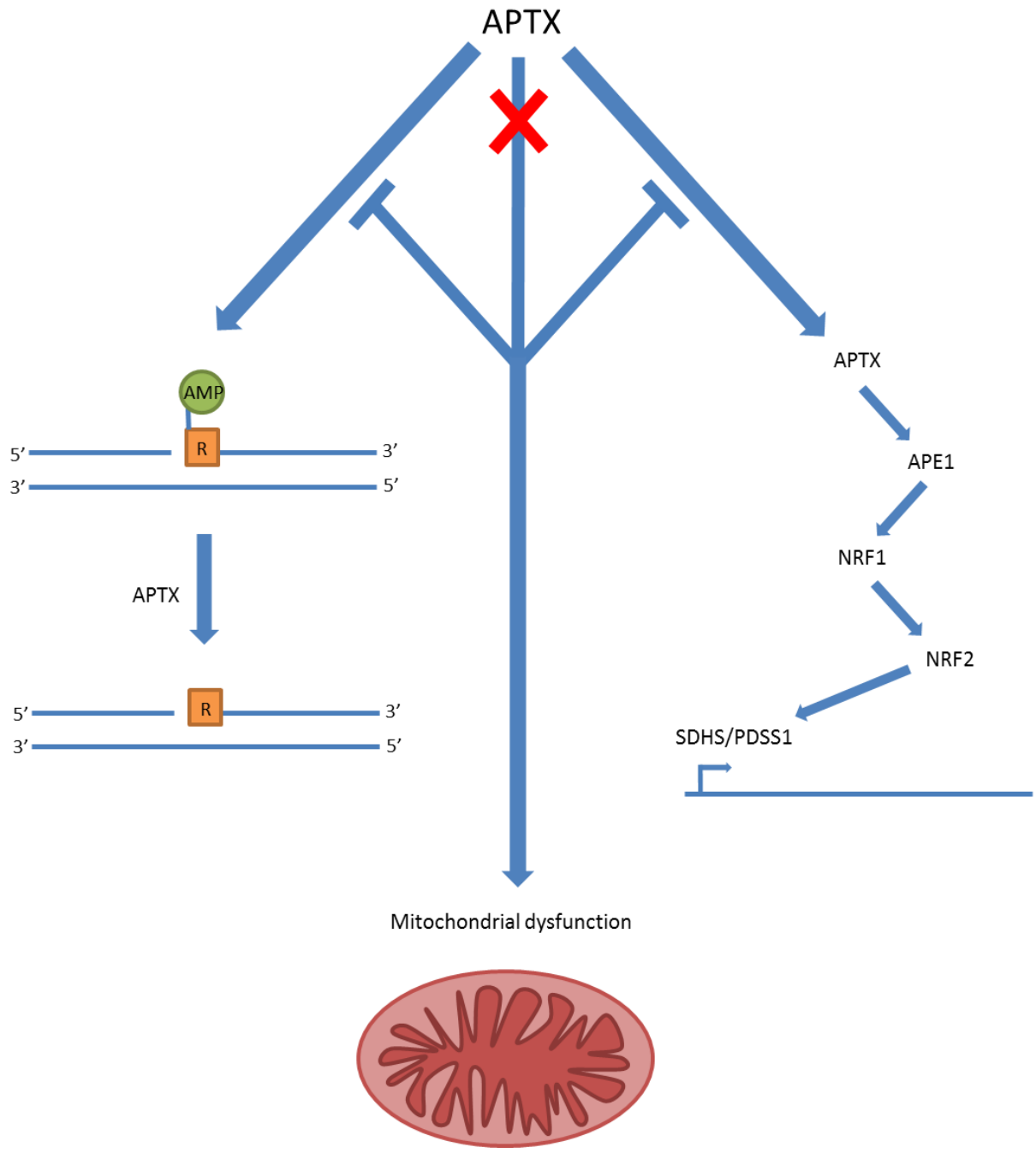


Fig.1.5 Model of how lack of aprataxin may lead to mitochondrial dysfunction. Aprataxin may perform its already characterised nuclear role in mitochondria, namely, cleaving AMP from DNA and nicked RNA-DNA junctions. Additionally, evidence suggests that APTX plays in role in coordinating the expression of succinate dehydrogenase synthase (SDHS) and Decaprenyl-diphosphate synthase subunit 1 (PDSS1), the first committed enzyme in CoQ_{10} biosynthesis. In the absence of APTX these functions will be disrupted, leading to mitochondrial dysfunction which potentially contributes to AOA1 pathogenesis.

mitochondrial, such as mtTOP1 and DNA polymerase γ, a phenomenon proposed to be part of a mitochondrial compensatory mechanism to cope with higher rates mtDNA damage (Ambrose et al., 2007). However, another group published data demonstrating that early passage, mouse derived ATM null fibroblasts had unusual increases in mitochondrial mass and mitochondrial respiration associated with elevated ROS levels and decreased electron transport chain complex I activity, which ran counter to previous research on the number of mitochondria in human cells lacking ATM. The hypothesis put forward to account for the abnormal mitochondrial mass was defective mitophagy in the ATM null cells, leading to the persistence of dysfunctional mitochondria (Valentin-Vega et al., 2012). A further study in A-T human fibroblasts reported findings that corroborated prior research with respect to increases in ROS, but defects in the removal of mitochondria were not found. Importantly, cells deficient in ATM activity accrued around 4 times more mtDNA damage relative to the WT cells, were slowed in the repair of oxidative damage to mtDNA and showed a global 50% reduction in ligase III levels (Sharma et al., 2014). This finding is of particular note, because previous research shows that whilst ligase I and III are redundant for the repair of nuclear DNA, ligase III is required for mtDNA repair and maintenance. Furthermore, when *Lig3* was knocked out specifically in the nervous systems of mice using *Nestin-cre*, the mice recapitulated the cerebellar atrophy observed in many DRD-ARCAs, albeit in an extremely accelerated form, with none of the mice surviving past 20 days post birth (Gao et al., 2011; Simsek et al., 2011). Currently, it is not clear which aspects of mitochondrial dysfunction observed in A-T and ATM null cells may be caused by defects in the coordination mtDNA repair and which result from the loss of other potential roles of ATM in mitochondrial homeostasis, such as regulating mitophagy (**Fig.1.6**). As it stands, much more work will be needed to confirm and distinguish the various mitochondrially-associated roles posited for ATM and the effects of ATM loss on the relevant pathways.

1.2.5 DRDA-ARCAs – A Summary

Although the DRDA-ARCA mutated proteins discussed throughout this introduction are implicated in both nuclear and mitochondrial roles, it is important to mention that some DRD-ARCAs are brought about by mutations in nuclear encoded proteins that have an exclusively mitochondrial function. Mutations in the mitochondrial DNA helicase twinkle and the catalytic subunit of DNA polymerase γ are known to cause infantile onset

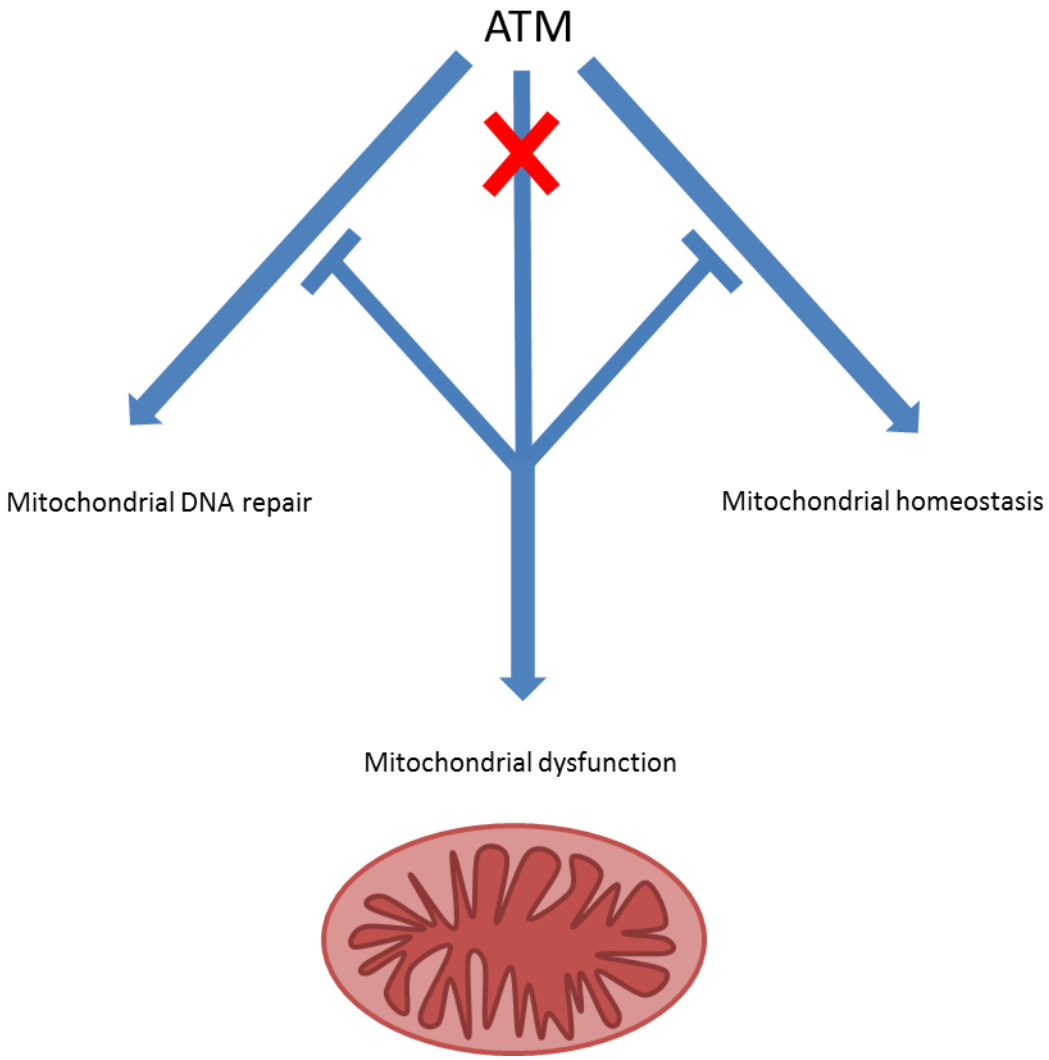


Fig.1.6 Model of how lack of ATM may lead to mitochondrial dysfunction. ATM is implicated in the coordination of mtDNA repair, specifically through the activation of DNA ligase III but other ATM targets involved in DNA repair localise to the mitochondria. There is also evidence that ATM may be involved in mitochondrial homeostasis by, for example, playing a role in mitophagy. Combinatorial loss of these ATM-mediated functions may lead to mitochondrial dysfunction, which could be involved in A-T pathology.

spinocerebellar ataxia and mitochondrial recessive ataxia syndrome respectively, providing further evidence connecting the pathology of this class of diseases to mitochondrial dysfunction (Hakonen et al., 2005; Nikali et al., 2005; Rantamäki et al., 2001). One of the major challenges with regards to the DRDA-ARCAs discussed here is determining the relative contribution of defects in nuclear DNA repair and abnormal mitochondrial function to disease pathology. Therefore, studies which allow the delineation of nuclear and mitochondrial abnormalities and their downstream effects will be invaluable in aiding our understanding of these diseases. Another potential issue for the study of these diseases is the fact that it is years before some of them manifest themselves in their human patients, and so short term studies in KO cells may not reveal certain critical, pathogenesis related features. Indeed, it may be this problem which has led to mouse models of various DRDA-ARCAs to not present with the expected phenotype; the animals may simply die before the molecular pathology progresses sufficiently far enough for the relevant clinical features present themselves. A common theme amongst studies investigating cells in which DRDA-ARCA mutated proteins are deficient is discrepancies with regards to results between distinct cell types. Based on published research, it is hard to say whether this lack of homogeneity in results is due to differences in experimental conditions or cell-specific differences in the requirement of the protein in question, but nevertheless, it would be beneficial to investigate which reported observations hold true in post-mitotic cells from the tissue affected in these disorders, the cerebellum.

This relates to one of greatest mysteries surrounding the DRDA-ARCAs – why do they manifest in atrophy of the cerebellum but not the cerebrum? If mitochondrial dysfunction does genuinely contribute to DRDA-ARCA pathology, then are there intrinsic differences between the mitochondria in the cerebral tissues and the cerebellum in terms of variables such as expression of mitochondrial genes, number of mitochondria and respiratory capacity? Research analysing relative mtDNA content across all the brain tissues in mice suggests that the cerebellum has the lowest number of mitochondria amongst all the tissues studied. However, the experiment failed to distinguish between different cell types, which may have had confounding effects, and these results remain to be confirmed in humans (Fuke et al., 2011). Ultimately, the question now isn't if loss of these proteins has a negative effect upon mitochondrial function, but rather whether the impairment of mitochondria

through deficiencies in these proteins contributes to DRDA-ARCA pathology, and if so, to what extent compared to the misregulation of nuclear DNA repair.

1.3 The Landscape of DNA Damage

1.3.1 Mutational Spectra and Regiospecific DNA damage

It is known that when certain types of DNA damage occur they result in specific base changes which can change the genome sequence in cells derived from the mother cell, thereby resulting in permanent somatic mutations. If distinct types of cells are subject to differential challenges to their DNA, then this can generate abundances or depletions in specific base changes, causing unique patterns of mutation to emerge between various cell types. The given pattern of mutation a cell has is referred to as its mutational spectrum. We are able to interpret these mutational spectra because the specific base changes brought about by different common forms of chemical attack and DNA damage are well documented. Cytosine deamination generates uracil, normally only found in RNA, and if this goes unrepaired and the DNA is replicated, this can result in a C:G→T:A transition (Duncan and Weiss, 1982; Impellizzeri et al., 1991). Deamination of cytosine is many times more likely to occur on single stranded DNA, both because there is no complementary strand protection and cytosines on single stranded DNA are substrates for deamination by activation induced cytidine deaminase (AID) or members of the APOBEC family of enzymes (Bransteitter et al., 2003; Fu et al., 2015; Chemical damage of cytosine reviewed in Nabel et al., 2012; Petersen-Mahrt and Neuberger, 2003). Therefore R loops, because they displace a strand of DNA from the duplex in order to form the RNA:DNA hybrid, may render the DNA particularly susceptible to this form of damage and subsequent base transition (R-loops reviewed in Skourti-Stathaki, 2014). This generation of a piece of single stranded DNA is also true of replication forks and transcription bubbles – this is part of a repeating pattern, perfectly normal cellular processes required for cell viability often render the cell susceptible or contribute to DNA damage. A striking example of this is ROS, by-products of cellular metabolism, generating oxidative damage of DNA. Although oxidative damage can lead to abasic sites and strand breaks, the major lesion arising from this form of chemical attack is 8-oxoguanine (8-oxoG) (Reviewed in Dizdaroglu, 1991). 8-oxoG resulting from guanine already incorporated into the DNA can base pair with adenine, potentially

generating downstream C:G→A:T transversions (Cheng et al., 1992; Thomas et al., 1997). Additionally, guanine nucleotides not incorporated into DNA can become oxidised, and these free 8-oxoGs can become mis-incorporated during replication, able to base pair with both C and A. If through this spontaneous incorporation it base pairs with adenine, 8-oxoG can generate A:T→C:G transversions (Colussi et al., 2002; Tajiri et al., 1995). In this way, it may be the case that mutations in proteins that process free 8-oxoG to prevent its accumulation in the DNA and those that excise 8-oxoG from the DNA itself lead to different mutational spectra: a enrichment of A:T→C:G and C:G→A:T transversions respectively. Alkylation of DNA is also associated with a specific change in the DNA, as it can generate O_6-methyl-guanine which can base pair with either C or T, giving the potential for a G:C→A:T transition (Aquilina et al., 1992). The presence of highly specific proteins designed to prevent and excise these different base mismatches within DNA attests to their detrimental effects if left unrepaired (For a comprehensive review of common forms of DNA damage and the mutations they cause, see Barnes and Lindahl, 2004).

Another notable feature of DNA damage is that it does not occur uniformly across the genome – that is, different regions are often more susceptible to certain types of DNA damage. On a spatial organisation level, it has been determined that whether DNA was found in lamina-associated domains (LADs) was positively correlated with levels of 8-oxoG damage, perhaps because these regions are located at the nuclear periphery and so are more likely to come into contact with damaging agents, or the LADs themselves may restrict repair (Yoshihara et al., 2014). Patterns have also emerged based on certain qualitative features of genes. Long genes with low to moderate expression have been found to harbour more 8-oxoGs and also γ-H2AX modifications, which are indicative of DSBs. Additionally, areas around origins of replication within this subset of long genes significantly overlapped with 8-oxoGs. It has been proposed that this is because the transcription of these longer genes takes more than one cell cycle to transcribe, which inevitably results in collisions between the transcription machinery and replication forks. These clashes can in turn generate single stranded DNA, which has increased susceptibility to chemical attack, hence the preponderance of oxidative DNA damage in these regions (Amente et al., 2019). Paused genes are another category that has been associated with DNA damage. Recent research utilising BLISS (Breaks Labelling In Situ and Sequencing) to map endogenous DSBs has shown

that at promoters, 5' splice sites and active enhancers, the release of paused RNA polymerase II (pol II) promotes the formation of DSBs. Whilst they found pause release to be the major predictor, they also found that gene length and topoisomerase occupancy were also determinants (Dellino et al., 2019). This work also illustrates another point – that even specific regions within genes can be sensitive to damage to greater or lesser extents. It has been demonstrated in cancer cells, (so escaping the confounding effects of purifying selection) that introns accumulate relatively more mutations than exons. This is because of the targeted recruitment of mismatch repair to exons, which is suggested to be due to differential H3K36me3 occupancy between introns and exons – indeed, there was found to be a strong negative correlation between the exon to intron ratio of H3K36me3 and level of exonic mutations (Frigola et al., 2017). In a separate study which mapped 8-oxoG and apurinic sites across the genome, damage hotspots were similarly found. A general increase damage in open chromatin (H3K9ac,H3K4me2) compared to heterochromatin (H3K9me3) was found, and intergenic regions again more susceptible than promoters and gene bodies. Interestingly, transcription levels did not seem to have on effect on the accumulation of damage. Surprising levels of variability were observed across gene bodies depending on the feature damage was aggregated across. Promoters and transcription start sites (TSSs) has the lowest relative damage enrichment of all the features assayed, being depleted in damage compared the average. Exons, UTRs and the TTS likewise showed reduced damage, whereas compared the rest of the gene body, introns were enriched for damage, but still below the levels for intergenic regions. Repetitive elements and transposons were found to be very highly enriched for damage, although no definite explanation for this was put forward (Poetsch et al., 2018).

Finally, there is the phenomenon of cell-type specific mutations. Because cell types differ in terms of physical environment and internal processes, they can be subject to different forms of DNA damage and so have unique mutational spectra. Adult stem cells in the small intestine and colon accumulate primarily deaminated methyl-cytosines, proposed to be linked to their high rate of replication. Liver cells are associated with a pattern of mutation not currently known, but this may be linked to formaldehyde-induced DNA damage, as experiments in mice have shown that the liver is adversely affected and hepatocytes can become malignant in the absence of formaldehyde-damage repair

proteins, demonstrating the need for organ specific repair of these lesions (Blokzijl et al., 2016; Pontel et al., 2015). Neurons are unique amongst cells because of their long lifespan and non-replicative status, so it has been theorised that as transcription is the major form of DNA metabolism that occurs in these cells, most mutations in neurons will be associated with this process. Analysis of single cell sequencing data has revealed that methyl-C→T substitutions are the most common type of mutation in developed neurons, and in foetal brains single nucleotide variants correlate with various markers of active transcription, adding weight to the above hypothesis (Lodato et al., 2015). All in all, research has continuously revealed the DNA damage landscape differs greatly depending on genomic location, the feature in question and the cellular environment, and all these should be taken into account when considering mutational spectra.

1.3.2 Somatic Mutations in the brain and their Detection

Mutations can fall within three categories: germline mutations passed from parent to child and present in all cells, somatic mosaic mutations arising during early development in progenitor cells or in adult stem cells and so present in a subset of cells or tissues, and finally somatic mutations that occur in differentiated, non-dividing cells and so are cell private. In a fashion reminiscent of way the neuron and non-neuronal cells are differentially effected by perturbations in DNA repair, somatic mutations in non-neuronal cells can lead to cancer, whereas their arising in neurons is being increasingly linked to various neurodegenerative disorders (Reviewed in Verheijen et al., 2018). As neurons are non-dividing, any somatic mosaicism present must have arisen during development, derived from a dividing progenitor cell. In fact, the rate of mutagenesis for neurons has been found to be greatest during this developmental phase, probably due to the fact these cells are highly replicative (Bae et al., 2018). In developed neurons, mutations are enriched at sites displaying markers of active transcription and in coding exons, as well as showing a template strand bias (Lodato et al., 2015). This is most likely because these cells are non-dividing, meaning transcription is the major form of DNA processing that occurs in these cells. This means that mutations arising from damage linked to transcription are likely relatively overrepresented in neurons compared to dividing cells, where base changes arising from replication associated damage will also make up a portion of mutation events. Additionally, transcribed genes are regions of open chromatin, meaning they are more susceptible to

chemical attack. This line of reasoning is supported by the previously discussed genome wide apurinic and 8-oxoG site mapping, which found an enrichment of damage in open compared to closed chromatin (Poetsch et al., 2018). It is also worth noting that the majority of somatic mutations in adult brains as determined by single cell sequencing are methyl-C→T substitutions, the result of methyl-C deamination (Lodato et al., 2015). As most somatic mutations in neurons appear to be associated with transcription and deamination is many times more likely to occur on ssDNA, it may be the case that transcriptionally associated R-loops are a major driver of mutations in the developed brain (Chemical damage of cytosine reviewed in Nabel et al., 2012). As some inherited neurodegenerative disorders involve premature ageing, and age is a risk factor for several notable neurodegenerative diseases, it is interesting to note that aged brains are more abundant in DNA damage markers. More than this, single cell sequencing has revealed that the total load of somatic mutations in the form of single nucleotide variants increases more or less linearly with age. Whilst they found that C→T substitutions accounted for most the mutations they found, the total proportion of these mutations compared to the whole fell as brains aged. This can be hypothesised to be due to the breakdown of cellular function, so that while transcriptionally-associated mutations may be the main factor at play in younger brains, as the brain ages dysfunction in other processes creates an increase in a different set of base changes. For example, T→C mutations particularly increased with age, suggested to be linked to the oxidation of fatty acids. Analysis of mutational signatures (sets of specific base changes thought to represent different underlying mechanisms or sources of mutation) revealed three distinct profiles. The first signature consisted primarily of C→Ts and T→C and was reminiscent of a “clock-like” or steady state (uniform with time) signature identified in cancer cells by another research group. Another mutational signature comprising mainly C→Ts did not show a strong relationship with age and was therefore proposed to be largely developmental in origin. This fingerprint also showed tissue specificity, as it increased slightly with age in the dendate gyrus of the hippocampus but not in the prefrontal cortex. This was proposed to be due to differences in neurogenesis between the two brain regions, highlighting that even subsections of the same organ can differ in their mutational landscape arising from small differences in development. The same study also analysed mutational spectra across CS and XP patients, diseases both arising from defects in NER. Brains from these patients were statistically enriched for a third set of mutations relative to

normal brains. This signature contained a comparatively high proportion of C→A mutations, indicative of oxidative damage, and also increased slightly with age in normal brains, highlighting that neurodegeneration may represent an acceleration of normal brain ageing. Differences between the disease states were also observed: CS but not XP brains were significantly enriched in the second signature linked to brain development (Lodato et al., 2018). A key discovery within this body of work was that genes related to neural function were overrepresented in terms of mutation load, and several genes that confer neurological disease when mutated in the germline were found to harbour SNVs in individual neurons. This considered in the light of the other findings has led to a model of mutation acquisition in the brain, dubbed the "use it and lose it" hypothesis. This states that as most mutations in the adult brain are associated with transcription, those genes most highly transcribed and thus likely important to neural function will have an increased disposition to the acquisition of mutations, leading to their subsequent downregulation or non-expression (Lodato et al., 2015). Corroborating evidence from another study has identified a group of genes that fall in expression after forty years of age in the frontal cortex and shown that the promoters of these age-downregulated genes are enriched for damage, as assayed by levels of 8-oxoG (Lu et al., 2004). Taken in aggregate, research into somatic brain mutations is revealing the complexity of this phenomenon, which shows variation with age, brain region, disease, transcriptional status and potentially other unaccounted for variables.

The detection of somatic mutations has been an evolving process. Somatic variant callers have mainly focussed on cancer variant calling, in which a tumour sample could be compared to normal tissue in order to find cancer specific mutations. For calling mutations in the brain this is not optimal, as calling is always relative to a reference tissue. Much of the work discussed in this subchapter utilised single cell DNA sequencing to call mutations, useful for discerning between somatic mosaic and cell private mutations (Lodato et al., 2015, 2018). A third option that has been utilised recently has been the calling of variants from RNA-seq, either bulk or single cell. The advantage of this is obvious; it allows the analysis of mutational load, spectra and expression levels from a single experiment. However, there are significant challenges with this approach. One is sequence coverage – it is very difficult to call variant reliably from lowly expressed genes, a problem not encountered in whole exome sequencing. Benchmarking of a somatic variant caller able to

call from both DNA and RNA-seq, MuTect, showed that only 55% of all transcripts had the power to detect alleles at an average allele frequency (AF) of 0.32, and a smaller 33% had power to detect the alleles that were called in DNA. Where sufficient coverage was available, the results were more promising, with an 82% detection rate for corresponding sites in the DNA as opposed to 27% across all applicable sites at the same AF cut-off of 0.32 (Yizhak et al., 2019). Another issue is RNA editing, a post-transcriptional modification of RNA whereby adenines can be deaminated to form inosine, which are picked up as guanines by the sequencing machinery (RNA editing reviewed in Eisenberg and Levanon, 2018). This means that any A→G mutation picked up from variant calling of RNA-seq data could in reality be an RNA editing event. These two factors make somatic variant calling on RNA-seq data that hasn't been specially pre-treated in very unreliable. Further variant calling analyses on RNA-seq data revealed whilst DNA-seq calling identified 75,388 variants, a far higher number of 359,982 were called in RNA-seq. 65% of these DNA mutations were not recaptured in RNA and 92% of RNA mutations did not turn up in DNA, indicating a high false discovery rate (Yizhak et al., 2019). A separate analysis bore even worse results: only 6.6% of all WES variants called as true by MuTect were present in variants called in the corresponding RNA-seq. However, this study did find that this improved for mutations most relevant to cellular physiology, with variants in coding regions captured at a frequency of 15.9%, and mutations affecting function at 17.2%. It was concluded that calling from RNA-seq data was useful in conjunction with WES, as it might allow further power to detect important mutations (Coudray et al., 2018). These data strongly suggest that RNA-seq variant calling is best used in tandem with DNA-seq, and if not, that it requires much pre-filtering and treatment to acquire any sort of useful accuracy. This strict prefiltering step has been attempted with some promising results. The RNA-MuTect pipeline was developed with these shortcomings in mind and applies a rigorous system of, filtering, checks and balancing before a mutation is called as true. These include remapping putative variant containing reads with a different mapper and calling the variants again, only keeping double called mutations, filtering out RNA editing by cross-referencing variants with databases of such phenomena, removing common variants with an AF of 5% or greater found in the ExAc database, filtering out non-coding regions and pseudogenes and discarding variants determined to be sequencing errors, amongst others. When applied to their dataset, this pipeline filtered out 93% of the RNA-seq called variants, but of the 89% of mutations for

which power was available to detect in the DNA-seq data, 90% were recapitulated in DNA. Most of the RNA-only mutations were C→T mutations, suggested to possibly represent a rarer form of RNA-editing less prevalent in event databases. This massive filtering removed 2,511 variants that were also found in DNA, roughly 10% of the post-filtering set, but the sensitivity was still around 0.7 and the precision approximately 0.9. Similar results were given when the pipeline was tested on an independent set of data, and after testing on a set of normal (non-cancer) tissues, the pipeline was applied to the Genotype-Tissue Expression (GTEx) data. Ultimately, whilst not as reliable as calling from DNA-seq , this extensive quality control drastically improved the reliability of RNA-seq variant calling and enabled the physiologically interesting discoveries in their subsequent analyses, showing that this mode of mutation discovery can give relevant results (Yizhak et al., 2019).

1.4 The Link between DNA Damage Repair and Structural Proteins

Whilst not an immediately obvious association, a relatively new body of research has begun to explore the relationship between cell structure and DNA repair. The majority of this work comes from the study of lamins, components of the nuclear lamina. This is a matrix of proteins that sits below the nuclear envelope and is associated with DNA at specific sites known as LADs (Guelen et al., 2008). There is evidence that this spatial arrangement of the genome afforded by lamins plays an important role in a whole host of vital cellular processes, including DNA repair. This is supported by the fact that mutations in the *LMNA* gene that encodes the A-type lamins, (lamin A and C) through alternative transcripts are associated with cancer as well as a variety of degenerative diseases, collectively termed laminopathies, including several dystrophies, Charcot-Marie-Tooth disorder type 2 – a peripheral neuropathy and premature ageing syndromes (For in-depth reviews of nuclear lamins, see Gonzalo, 2014; Leeuw et al., 2018). The evidence linking mutations in *LMNA* to defective DNA repair is numerous. Cells from patients suffering from the laminopathies mandibuloacral dysplasia type A (MADA) and Hutchinson Gilford Progeria Syndrome (HGPS) present with general markers of genome instability, such as chromosomal aberrations, increased DNA damage and elevated levels of γH2AX foci (Liu et al., 2005; Masi et al., 2008). *Lmna−/−* mouse embryonic fibroblasts showed similar markers (though not elevated levels of γH2AX) as well as a global reduction in the levels of the DSBR protein 53BP1, important for mediating NHEJ and restricting HR, and consequently presented with NHEJ defects

(Gonzalez-suarez et al., 2009; Redwood et al., 2011). Experiments in MCF7 cells in which lamins A and C were depleted also demonstrated that despite again seeing a general reduction of 53BP1 levels, HR was also affected, showing an overall 40% reduction in activity, revealed to be due to downregulation of BRCA1 and RAD1 (Redwood et al., 2011). Specific *LMNA* mutations known to cause disease led to a reduction in γH2AX foci and mis-localisation of ataxia telangiectasia-mutated and Rad3-related protein (ATR), another DSBR factor (Manju et al., 2006). Further work in mammalian cells has shown that lamin A interacts with H2AX and γH2AX, an association that increases upon DNA damage, and lamin A/C depletion affects the spatial stability of DNA repair foci, induces sensitivity to replication stress and affects the recruitment of repair factors to sites of replicative DNA damage (Mahen et al., 2013; Singh et al., 2013). Although the precise mechanisms by which lamin A/C interface with DNA repair are not clearly understood, these data suggest that through their spatial organisation of the genome the lamins affect the expression and recruitment of repair factors, potentially acting as anchoring sites for stabilisation of other proteins involved in the DNA damage response (Mahen et al., 2013; Singh et al., 2013). However, this link between cellular structure and the DDR is not just limited to lamins. Studies across a range of models implicate actin and the actin binding proteins ARP2/3 in DSB repair. In *Xenopus* egg extracts, β-actin and the actin binding subunit of ARP2/3 were recruited to sites of DNA damage in a manner dependent upon the activity of ATM and ATR. Actin foci in the nucleus increased upon treatment with a DNA damaging agent and showed ARP2/3 dependent clustering together over time. These actin foci also colocalised with foci corresponding to the DNA damage response proteins RPA32 and RAD51, and inactivation of ARP2/3 led to reduced levels of both resection and RPA32/RAD51 foci. These results amongst others have led to a model whereby ARP2/3 promotes the polymerisation of actin at sites undergoing HR, which in turn leads to actin-directed clustering of these sites which somehow promotes effective DSB repair (Schrank et al., 2018). A similar study in *Drosophila* implicated actin in the movement of heterochromatic DSBs away from their normally repressive environment in order to enable optimal repair (Agostino et al., 2018). A third structural protein, the intermediate filament vimentin, binds to DNA secondary structures and is found within nuclear matrix attachment regions (MARs or scaffold attachment regions, SARs) (Tolstonog et al., 2001). These are specific sequences in the genome that connect to the nuclear matrix and are linked to DNA repair, particularly NHEJ, an interesting

observation when it is also considered that vimentin itself is a target of the NHEJ protein DNA-dependent protein kinase (Kotula et al., 2013; Mauldin et al., 2002; The concept of S/MARs is reviewed in Roberge and Gasser, 1992). Whilst in many cases the specifics are unclear, it is safe to say a firm relationship between DNA damage repair and cellular structure has been established, and that this represents an exciting new arm of DNA damage research.

1.5 Aims and Objectives

This introduction has aimed to explore to relationship between DNA damage and the brain, and the problem of why mutations in DNA damage repair proteins, particularly single strand break repair factors, lead to such a cerebellar specific phenotype. Throughout this discussion of currently published work, there is a distinct lack of a hypothesis by which this cerebellar sensitivity can be explained. Therefore, this thesis aims to investigate why the cerebellum is so sensitive to loss of function in DNA repair housekeeping genes. As many DRDA-ARCA diseases are relatively rare and being able to acquire to genomic data from the brains of these patients post-mortem is highly unlikely, in this thesis we have taken a different approach. Using our current knowledge about what drives the pathology of the DRDA-ARCAs, we can look for differences in such features between wild type cerebellum and cerebrum. The reason behind this is that for the cerebellum to be so specifically affected by perturbations in DNA repair, there must be intrinsic differences between the cerebellum and the cerebrum that lead to this divergence in sensitivity. Differential gene expression is an obvious analysis that will help to explore what makes the cerebellum unique and how this may relate to its disease sensitivity. The involvement of mitochondria in DRDA-ARCA pathology has been extensively discussed, so looking at mitochondrial differences between the cerebellum and the cerebrum will therefore form an important part of this work. We will also investigate whether the cerebellum is in some way already sensitive to DNA damage by attempting to measure the basal rate of mutation between tissues. In this we can leverage cutting edge understanding of how mutations accrue in the brain, both during development and in the mature tissue, in order to assess any disparities in either total mutational load or the mutational spectrum. All of this work will be performed using publicly available data, which can represent a challenge, as the ideal type of data for a given analysis is not always available. Here we will utilise the published best-

practice methods discussed in this introduction for performing perhaps unconventional analyses, such as calling mutations from RNA-seq data. Finally, this project aims to advance the research field linking structural proteins to DNA repair through a genomic analysis of the structural protein NuMA, and briefly explore how this could be relevant to DNA damage induced neurological disease.

2. Chapter 2 – Differing Patterns of Gene Expression Between the Cerebellum and Cerebrum

2.1 Introduction

The distinguishing clinical feature all DRDA-ARCAs have in common is degeneration of the cerebellum but not the cerebrum. This indicates that there must be something different about cerebellar and cerebral tissue even in healthy individuals, such that a difference in disease sensitivity can arise in the presence of dysfunctional DNA repair. Recent clustering of human tissues using multidimensional scaling of the GTEx RNA-seq showed the cerebellum as a highly distinct group relative to the cerebral tissues (Melé et al., 2015). There are several plausible contributing factors to this cerebellar uniqueness which may be linked to its susceptibility to loss of DNA repair housekeeping genes. One candidate is the fact that the cerebellum has an extended period of maturation compared to the cerebrum, and this involves rapid cell division (El-Khamisy, 2011). As previously discussed, replication stress during the expansion of progenitor cells is a major cause of DNA damage, testified to by the observation that mutations in the proteins that mitigate this often result in embryonic lethality or neurodevelopmental problems. Whilst this extended period of growth may distinguish it from most cerebral tissues developmentally, it could also mean that the fully matured cerebellum contains more somatic mutations, and as per the well characterised link between DNA damage, mutations and ageing, mean that the cerebellum as a tissue ages faster than the cerebrum (Barzilai et al., 2017). Additionally, an already existing increased somatic mutation burden may mean that upon loss of DNA repair proteins, cerebellar DNA is already a step along in terms of the negative effects of mutation acquisition, and therefore is sensitised to dysfunction in DNA repair. Another possibility is differences in mitochondria between the cerebellum and the cerebrum. The effects of the loss of DRDA-ARCA causing proteins on mitochondrial function has already been broadly described in the introduction, making it an ideal candidate for a source of difference that could explain cerebellar sensitivity. Experiments in mice have shown that the cerebellum has the lowest amount of mitochondrial DNA out of all brain tissues, which if true in humans could represent a compelling case for the specificity of the DRDA-ARCAs (Fuke et al., 2011). Simply put, the mutations that lead to cerebellar ataxias are known to cause mitochondrial dysfunction in some capacity, and so if the cerebellum has fewer mitochondria,

mitochondrial loss will represent a greater proportion of the total number of mitochondria, meaning the cerebellum is more likely to be impacted negatively by mitochondrial dysfunction. This may be why mutations in mitochondrial specific proteins also cause ARCAs or other diseases that involves cerebellar atrophy. A striking example of this is mutations in gene *Ataxin 1* leading to the dominant neurodegenerative disorder spinocerebellar ataxia type 1 (SCA1). In mouse models of this disease, the mutant protein affects HMGB1, a protein involved in the restructuring of DNA and found to be involved in the repair of mitochondrial DNA, by reducing its accumulation in neural mitochondria. Loss of mitochondria HMGB1 in turn causes increases DNA damage and loss of Purkinje cells, the major type of cell type thought to be lost in cerebellar ataxias (Ito et al., 2015). Although SCA1 is not a DRDA-ARCA, the principle remains the same: it appears that the cerebellum is extremely susceptible to perturbations in mitochondria function for a reason as of yet unkown, and this could be the source of the specificity of the ARCAs. A third, less substantiated possibility is that the mature cerebellum is subject to a greater basal rate of DNA damage relative to the tissues of the cerebrum. This could be due to more endogenous sources of DNA damage, such as ribonucleotide contamination, transcription-associated DNA damage or reactive oxygen species – although this would seem to conflict with reports of less mitochondrial in the cerebellum. Alternatively, the cerebellum may have lower basal expression of all or certain DNA damage repair factors, potentially because it is subject to less damage from specific sources, e.g. ROS if there are indeed less mitochondria in the cerebellum, transcribing and translating unnecessary amounts of proteins that deal with oxidative damage represents an energy cost. Hypothetically, this could mean that the cerebellum is sensitive to loss of one of these proteins, as there are fewer other DNA repair proteins that could act in a compensatory manner, although upregulation of compensatory proteins is known phenomenon. Whatever the case may be, the cerebellum and the cerebrum are two clearly distinct areas of the brain, and differences between them that may have a bearing on their sensitivity to DNA repair housekeeping gene loss of function are not well characterised. Therefore, shedding light on these potentially relevant differences across normally functioning cerebellum and cerebrum is the key aim of this chapter.

2.2 Results

2.2.1 Microarray data recapitulates cerebellar uniqueness with regards to gene expression

A starting point to examine differences between wild-type cerebellum and cerebrum was to analyse gene expression data. Therefore, an exon microarray summarised to genes from the BrainSpan consortium Developmental Transcriptome Dataset, a dataset consisting of microarray data from different brain tissues from a group of healthy donors, represented an ideal starting point. After filtering for samples from individuals within the 18-40 age range, this dataset consisted of 107 brain samples, 5 of which were cerebellar in origin. Differential expression analyses were performed comparing the cerebellum and the orbital frontal cortex, temporal cortex, primary visual cortex and parietal cortex in turn, these tissues being chosen in order to achieve good coverage of the range of tissue present in the cerebrum. A differential expression analysis where all the non-cerebellar brain tissues were pooled and compared to the cerebellar samples was also carried out to investigate differences on a total cerebrum-cerebellum level.

Once a set of differentially expressed genes had been determined for each analysis, overlap analyses were carried out in order to investigate what proportion of the genes determined to be differentially expressed in one analysis were also registered as being differentially expressed in another analysis (**Fig.2.1)**. The results showed that a large proportion of the genes determined to be significantly differentially expressed in one analysis were also differentially expressed in another analysis, as the middle column, the number of genes in common between the two analyses, represents a large fraction of the numbers on either side, the number of genes found to be differentially expressed in each analysis. This indicates that a large proportion of the genes that are differentially expressed between the different cerebral tissues and the cerebellum are the same genes.

Next, read counts for all the genes that were differentially expressed in at least one analysis were used to construct a heatmap. Hierarchical biclustering was applied to cluster similar samples and similar genes together. The samples clustered into four distinct groups, based on tissue they were from: cerebellum, striatum, mediodorsal nucleus of thalamus,

Tissue 1	No. of genes DE between tissue 1 and the CB	Overlap between the sets of DE genes	No. of genes DE between tissue 2 and the CB	Tissue 2
OFC	2395	2096	2392	TEC
OFC	2395	1841	2166	PVC
OFC	2395	1972	2222	IPC
OFC	2395	2043	2244	ALL
TEC	2392	1823	2166	PVC
TEC	2392	1972	2222	IPC
TEC	2392	2051	2244	ALL
PVC	2166	1809	2222	IPC
PVC	2166	1838	2244	ALL
IPC	2222	1918	2244	ALL

Fig.2.1 Table showing every possible pairwise comparison between sets of genes, each set having been determined to be differentially expressed in the cerebellum relative to a different cerebral tissue for the BrainSpan microarray data, in order to determine the number of genes in common between sets. The left and rightmost columns represent the cerebral tissues the cerebellum is being compared with, their abbreviations being: OFC = orbital frontal cortex, TEC = temporal cortex, PVC = primary visual cortex, IPC = inferior parietal cortex, ALL = read counts averaged across all non-cerebellar brain tissues in the dataset. The second columns in from the left and right contain the number of genes differentially expressed between the cerebellum and Tissue 1 and Tissue 2 respectively, and the middle column represents the number of genes shared between these two sets of genes. Differential expression determined using linear model fitting.

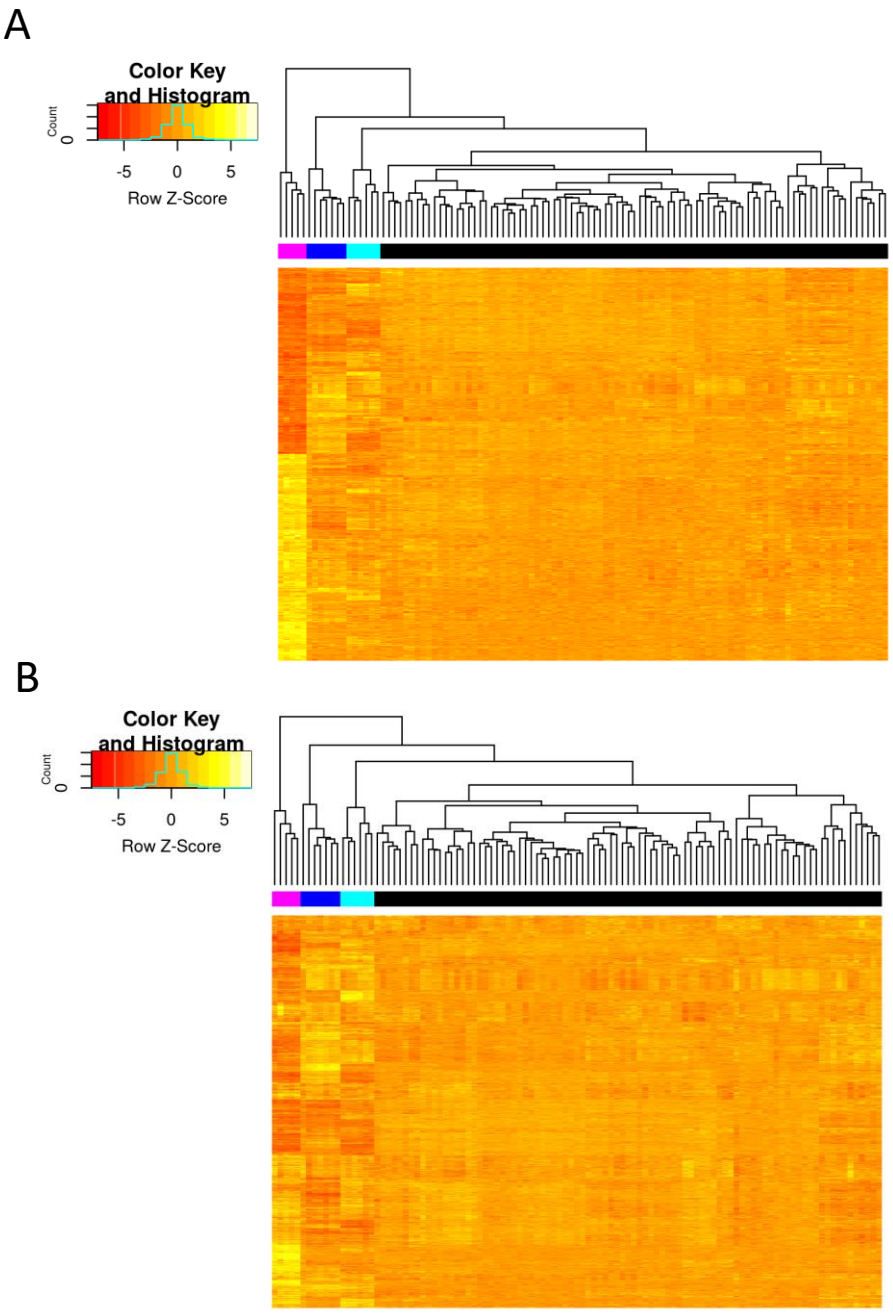


Fig.2.2 Heat maps providing a visual representation of how the samples in the BrainSpan microarray data cluster together according to gene expression. (a) Heat map plotted using genes that show differential expression in the cerebellum relative to cerebral tissues. (b) Heat map plotted using the 3000 genes with the highest variance. In both Heat maps, the three most distinct clusters of samples represent, moving from left to right: Cerebellum samples, striatum samples and mediodorsal nucleus of thalamus samples. Differential expression determined using linear model fitting.

and then the rest of the tissues (**Fig.2.2a**) – although this was unsurprising, as we selected out genes determined by our analyses to be differentially expressed in the cerebellum. Using this heatmap, three lists of potential genes of interest were identified, those that showed downregulation only in the cerebellum, genes which showed upregulation in the cerebellum and genes which showed downregulation in the cerebellum. Separate GO analyses on these sets of genes was carried out, but no categories relevant to our investigation were identified. Finally, in order to see if the cerebellum was truly distinct from the cerebral tissues in this dataset, as had been reported previously, the 3000 most variant genes in the dataset were clustered (**Fig.2.2b**). The four distinct groups from **Fig.2.2a** were recapitulated in this heatmap, identifying the cerebellum as distinct from the cerebral tissues with regards to gene expression.

As mitochondria are a potential key difference between the cerebellum and the cerebrum as regards the pathology of the DRDA-ARCAs, the next step was to look at patterns of expression for differentially expressed mitochondrial genes. However, this set of microarray data did not contain the mitochondrially-encoded genes, and so we were restricted to looking at the nuclear-encoded mitochondrial genes. To investigate mitochondrial expression, all the differentially expressed nuclear-encoded mitochondrial genes from each of the differential expression comparisons were identified and the frequency of log fold changes in gene expression between the cerebellum and each cerebral tissue for these genes was assessed. Across all of the analyses, a greater number were found to be upregulated rather than downregulated in the cerebellum (**Fig.2.3a-e**). To validate these results, the spread of logFCs for mitochondrial genes differentially expressed between the cerebellum and each tissue in the BrainSpan microarray dataset not yet analysed was likewise analysed. Most of these showed a spread of logFCs that matched the previous findings. To see whether this upregulation of mitochondrial genes in the cerebellum compared to the cerebrum was significant, the enrichment of differentially expressed mitochondrial genes amongst genes upregulated in the cerebellum compared to those downregulated relative to the pooled cerebral tissues was carried out using the Fisher's test. The test revealed that mitochondrial genes were significantly enriched in upregulated genes, shown by a significant depletion in downregulated genes with an odds ratio of 0.86 (**Fig.2.4**). To further investigate the specificity of the logFC frequency skews, the spread

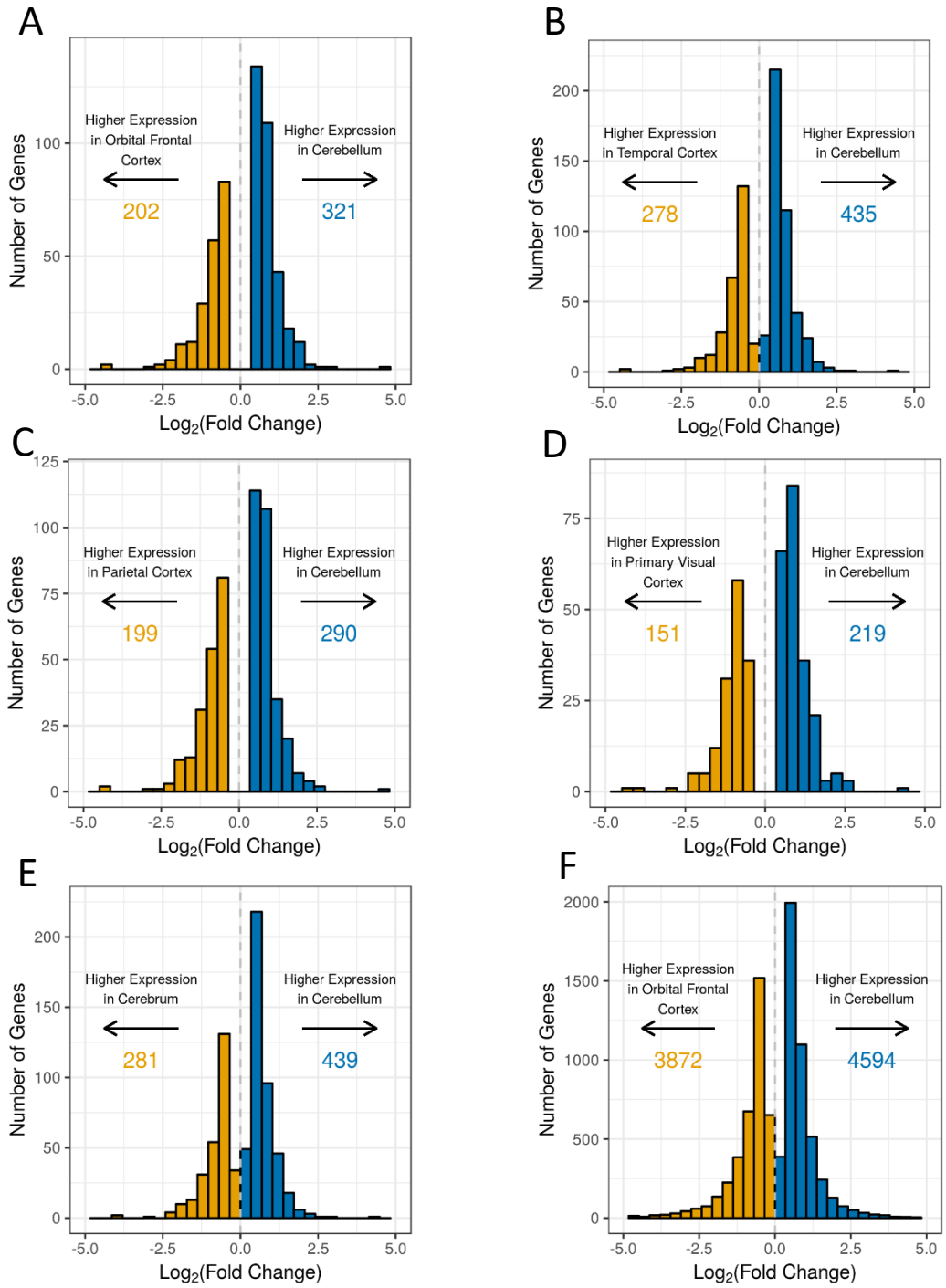


Fig.2.3 Histograms plotted using data from the BrainSpan microarray displaying the frequency of logFCs for nuclear-encoded mitochondrial genes differentially expressed between the cerebellum and (a) the orbital frontal cortex (b) the temporal cortex (c) the parietal cortex (d) the primary visual cortex and (e) all non-cerebellar brain tissues in the dataset. (f) Shows the spread of logFCs for all the genes differentially expressed between the cerebellum and all the non-cerebellar tissues. Differential expression and logFCs determined using linear model fitting.

of logFCs for all the differentially expressed genes was plotted. The logFCs of the differentially expressed genes also showed a skew towards upregulation, although this skew does not seem as strong (**Fig.2.3f**). This indicated that the observed skew within this set of data may just be a phenomenon observed amongst differentially expressed genes and is not necessarily mitochondrial specific.

2.2.2 Analysis of brain RNA-seq data does not recapture differences in mitochondrial gene expression observed in microarray data

The most interesting finding from the analysis of the BrainSpan microarray data was that the mitochondrial genes differentially expressed between the cerebellum and cerebral tissues showed a skew towards upregulation. Therefore, this part of the previous analysis was repeated using the Genotype-Tissue Expression (GTEx) dataset V6 from The Common Fund, which contains RNA-seq data for a wide range of tissues from healthy individuals. After applying quality filters, the dataset contained 351 brain RNA-seq samples, 81 of which came from the cerebellum. This dataset included read counts for the mitochondrially-encoded genes, as well as the nuclear-encoded mitochondrial genes. However, for the purposes of the initial analysis investigating the spread of logFCs for the differentially expressed mitochondrial genes, these two sets of genes were treated as one. Because there are so many more nuclear-encoded mitochondrial genes than mitochondrially-encoded genes, the addition of the mitochondrially-encoded genes into this group would be highly unlikely to affect the overall skew of logFCs.

Genes differentially expressed between the cerebellum and each of the other brain tissues were identified and histograms of the LogFCs of mitochondrial genes were plotted (**Fig.2.5a-d**). Each of these showed the same pattern as for the merged comparison of all non-cerebellar tissue to the cerebellum (**Fig.2.5e**). However, the resulting spread of logFCs was opposite to that seen for the BrainSpan microarray data analyses, namely, there were more differentially expressed mitochondrial genes that showed downregulation rather than upregulation (**Fig.2.3e** and **Fig.2.5e**). In contrast to the microarray data, in this case there was no distinct trend for all differentially expressed genes. (**Fig.2.5f**). Therefore, it appears that the skew towards downregulation in this dataset is specific to mitochondrial genes. Additionally, Fisher's tests showed that differentially expressed mitochondrial genes are significantly enriched amongst genes downregulated in the cerebellum relative to the

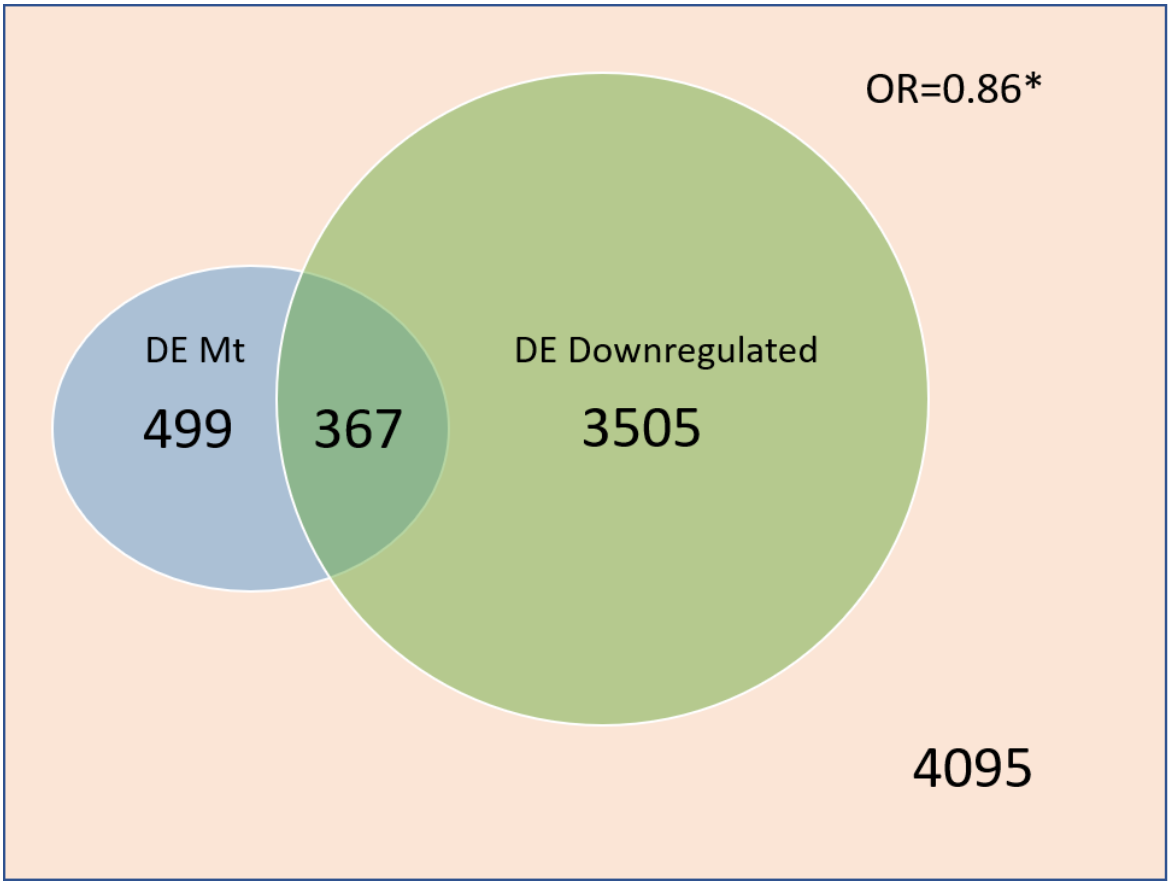


Fig.2.4 Venn diagram showing, for the BrainSpan microarray data, overlap of genes differentially downregulated in the cerebellum relative to pooled cerebral tissues with mitochondrial genes differentially expressed in the same comparison. The background gene set is differentially upregulated genes. Significance and odds ratio determined by Fisher's exact test. Differential expression determined using linear model fitting. DE = Differentially expressed, Mt = Mitochondrial, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

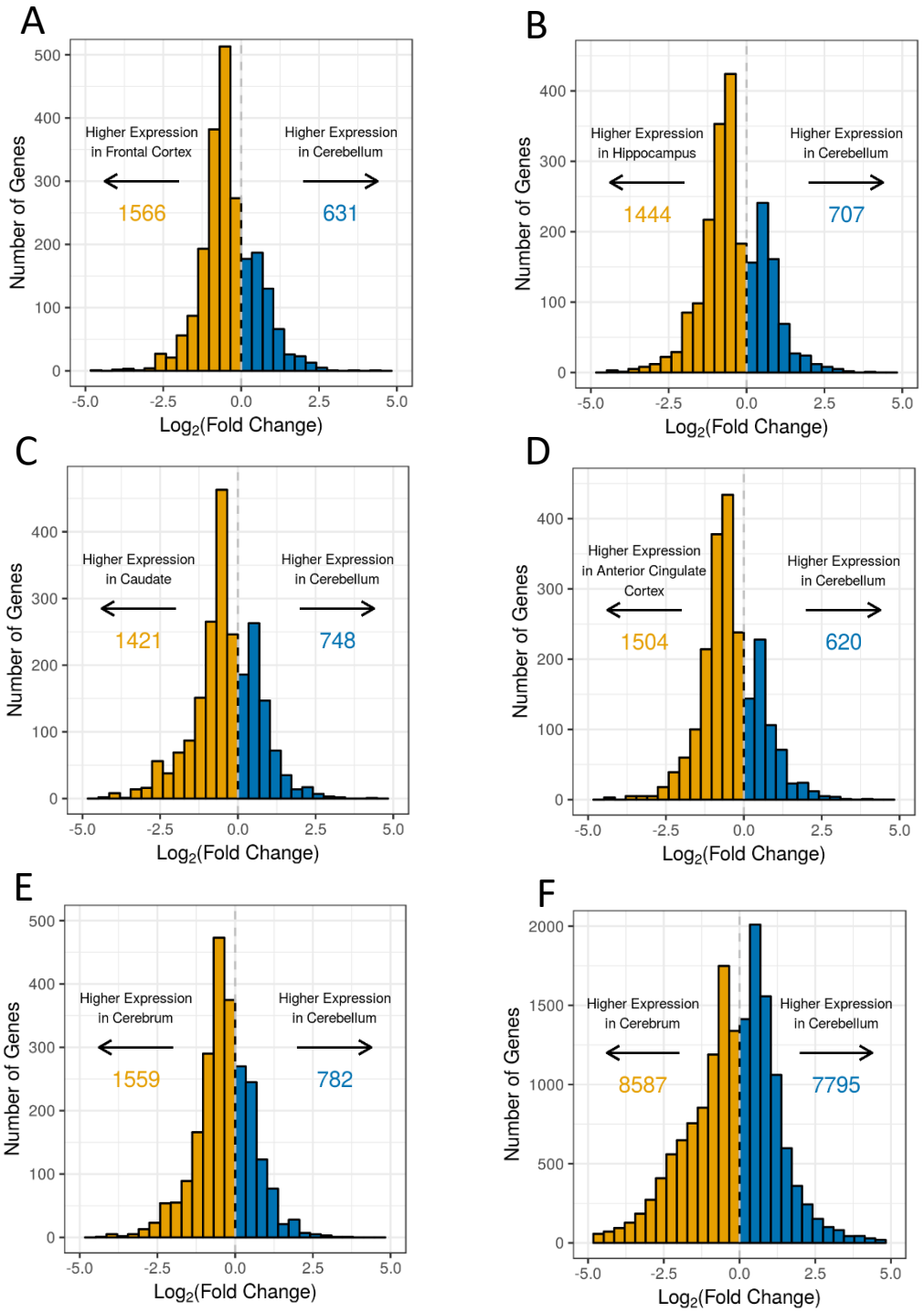


Fig.2.5 Histograms plotted using data from the GTEx v6 RNA-seq dataset displaying the frequency of logFCs for mitochondrial genes (both nuclear and mitochondrially-encoded) differentially expressed between the cerebellum and (a) the frontal cortex (b) the hippocampus (c) the caudate (d) the anterior cingulate cortex and (e) all non-cerebellar brain tissues in the dataset. (f) Shows the spread of logFCs for all the genes differentially expressed between the cerebellum and all the non-cerebellar tissues. Differential expression and logFCs determined using linear model fitting.

pooled cerebral tissues, when the comparison set was genes differentially upregulated in the cerebellum (**Fig.2.6**). The OR was 1.6, meaning the enrichment is stronger than the depletion of differentially expressed mitochondrial genes from cerebellar downregulated genes in the BrainSpan microarray dataset, which had an OR of 0.86. Finally, the spread of logFCs for the mitochondrially-encoded genes in the cerebellum relative to the frontal cortex and relative to all the non-cerebellar brain tissues was analysed. All of the mitochondrially encoded genes present in the GTEx dataset showed downregulation in the cerebellum in both instances (**Fig.2.7a-b**).

To again further determine the specificity of the observed logFC frequency skews, the spread of logFC frequencies for all genes (not just differentially expressed ones) and then all the mitochondrial genes in both the GTEx RNA-seq data and the BrainSpan microarray data was assessed, the logFCs being calculated for the cerebellum relative to read counts averaged across all the cerebral tissues. For the microarray data, the spread of logFCs for all genes showed a skew towards downregulation, whereas in the GTEx data, there was no skew in either direction (**Fig.2.8a, b**). Interestingly, the spread of logFCs for all the mitochondrial genes for each dataset was similar to the skew seen for the differentially expressed mitochondrial genes in that dataset, towards downregulation in the GTEx data, and towards upregulation in the BrainSpan microarray data (**Fig.2.8c-d**). In order to further clarify the issue of non-agreement between these two datasets, a conservative set of mitochondrial genes that came up as being differentially expressed in the cerebellum-frontal cortex comparisons from both analyses was identified and histograms of the logFCs of these genes, one with the logFCs from the microarray data, the other with the logFCs from the GTEx RNA seq data were plotted. Both histograms matched the spread seen for the full set of differentially expressed genes for the dataset in question (**Fig.2.9a-b**). However, only 445 BrainSpan microarray genes and 439 GTEx RNA-seq genes were called as being differentially expressed in both analyses, representing 9/10ths of the mitochondrial genes called by analysis of the microarray data but only 1/5th of the mitochondrial genes called by analysis of the GTEx data. Therefore, it seems plausible that the difference between the datasets is driven by far more mitochondrial genes being called as differentially expressed in the GTEx V6 dataset. As the distribution of changes for differentially expressed mitochondrial genes in the GTEx data did not match that in the microarray analysis, we

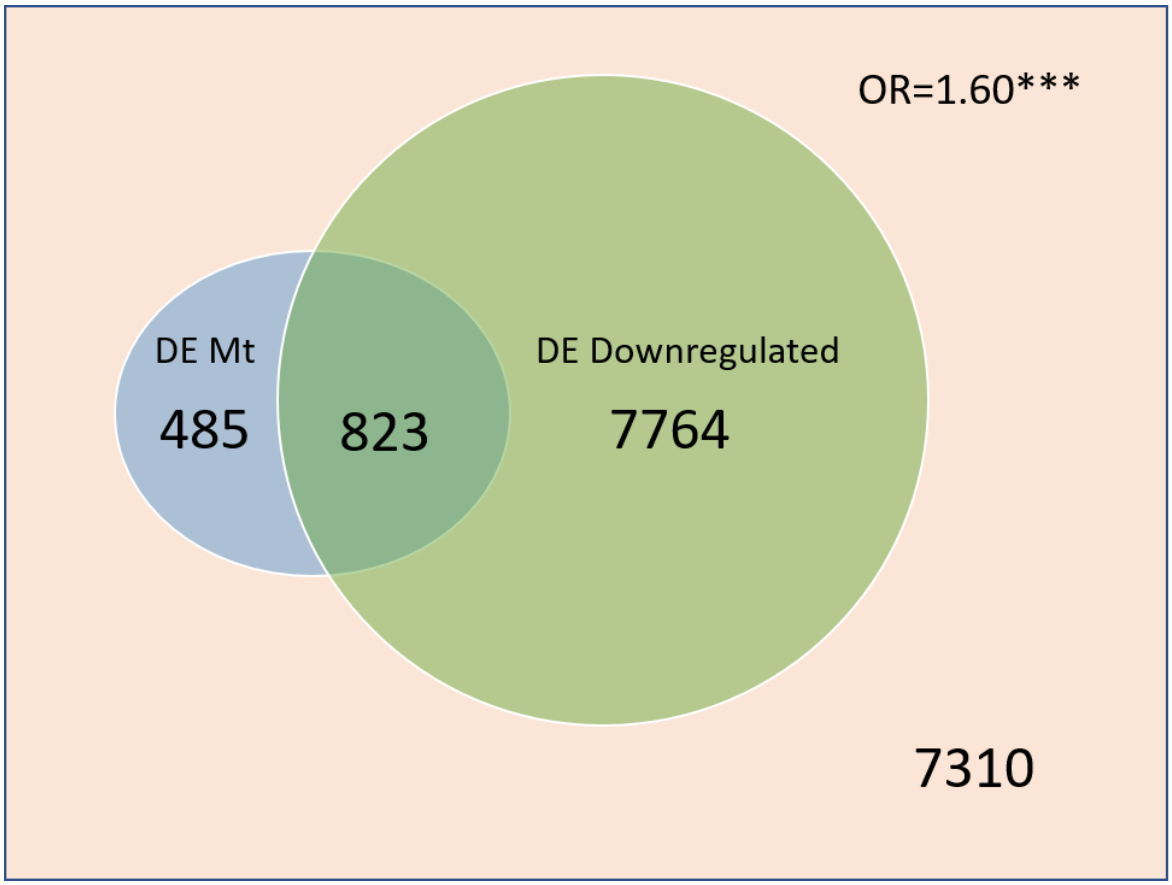


Fig.2.6 Venn diagram showing, for the GTEx V6 RNA-seq data, overlap of genes differentially downregulated in the cerebellum relative to pooled cerebral tissues with mitochondrial genes differentially expressed in the same comparison. The background gene set is differentially upregulated genes. Significance and odds ratio determined by Fisher's exact test. Differential expression determined using linear model fitting. DE = Differentially expressed, Mt = Mitochondrial , OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

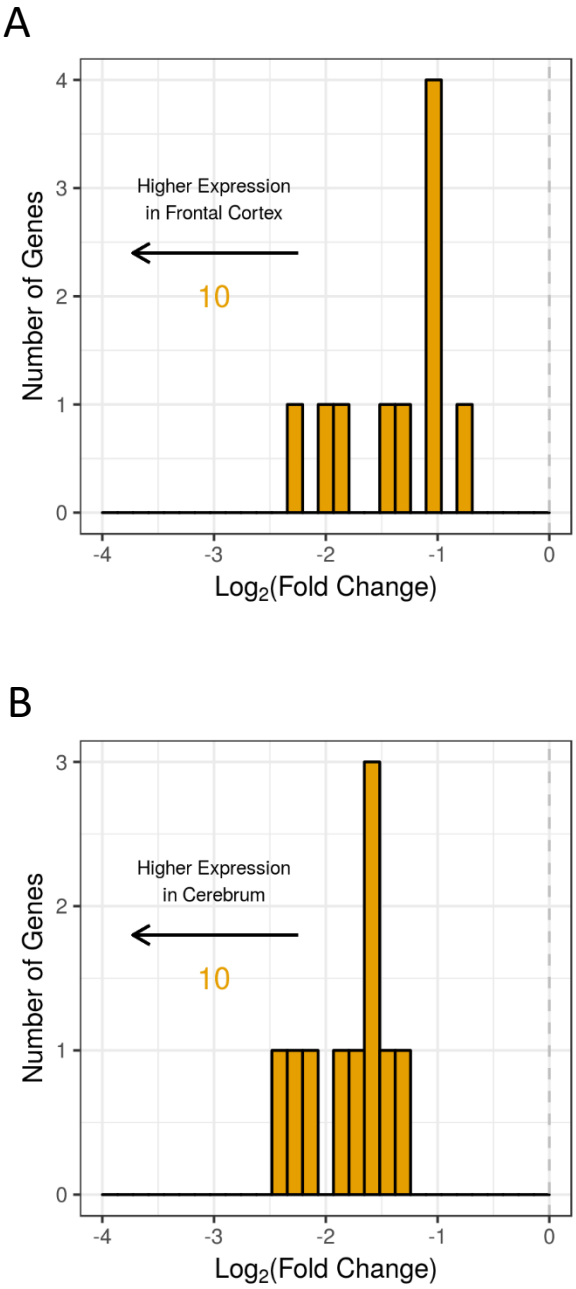


Fig.2.7 Histograms plotted using the GTEx V6 dataset displaying the frequency of log fold-changes for mitochondrially-encoded genes, differential expression and logFCs being calculated for the cerebellum relative to (a) the frontal cortex and (b) all non-cerebellar tissues in the dataset. Differential expression and logFCs determined using linear model fitting.

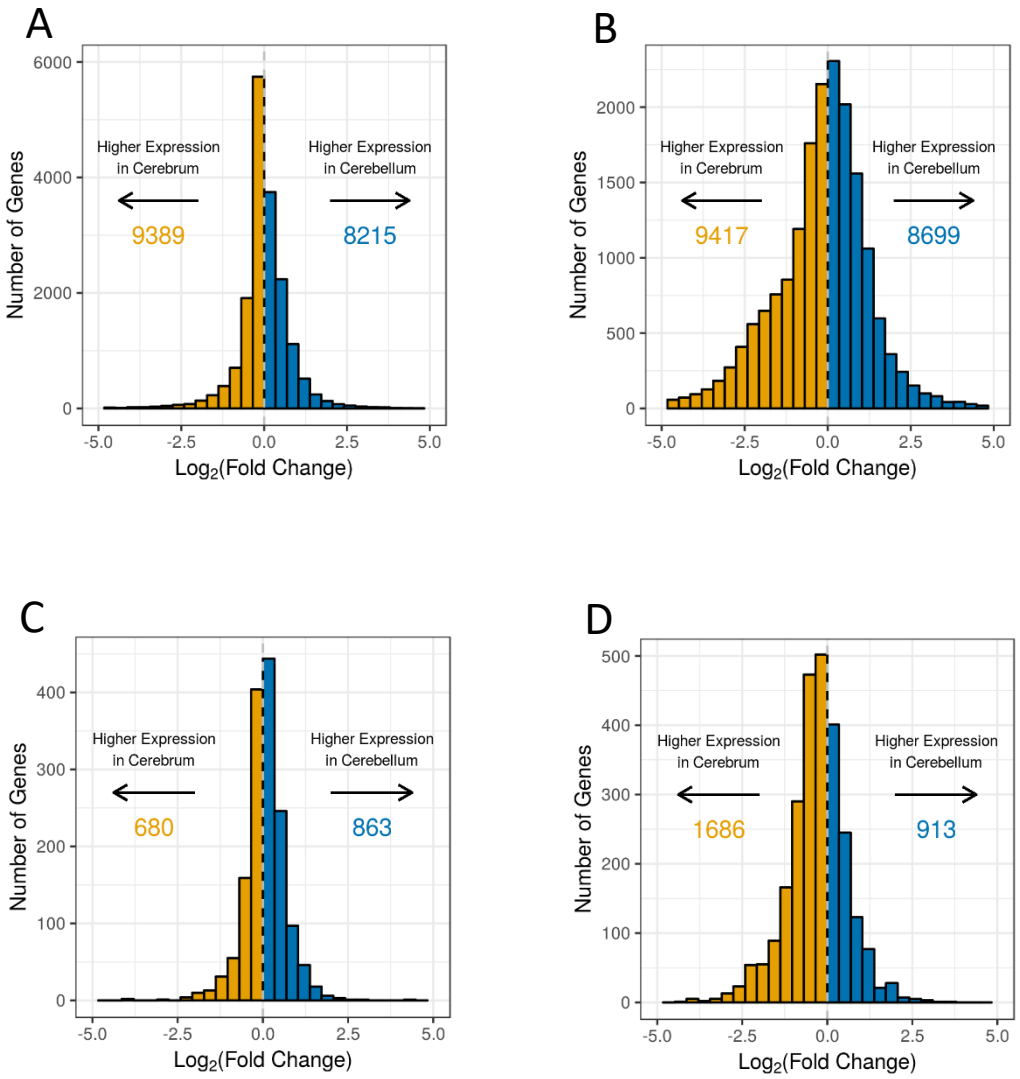


Fig.2.8 Histograms displaying the frequency of logFCs for (a) all the genes in the GTEx V6 dataset (b) all the mitochondrial genes in the GTEx V6 dataset (c) all the genes in the BrainSpan microarray dataset and (d) all the mitochondrial genes in the BrainSpan microarray dataset, logFCs being calculated for the cerebellum relative to all non-cerebellar brain tissues in each dataset. Differential expression and logFCs determined using linear model fitting.

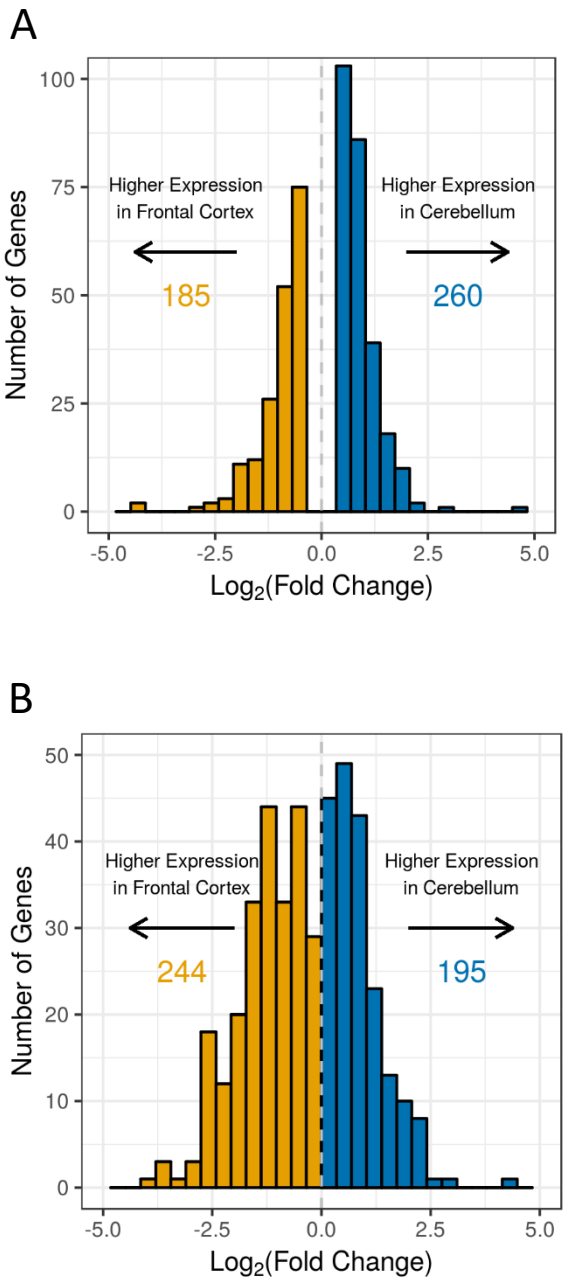


Fig.2.9 Histograms displaying the frequency of logFCs for mitochondrial genes determined to be differentially expressed between the cerebellum and the frontal cortex in both the BrainSpan microarray dataset and the GTEx V6 dataset. (a) was plotted using the BrainSpan microarray dataset and (b) was plotted using the GTEx V6 RNA-seq dataset. Differential expression and logFCs determined using linear model fitting.

examined a third set of data: the BrainSpan RNA-seq dataset. Quality filtering gave 113 brain samples including 81 cerebellar samples. The same methodology used for the analysis of the GTEx RNA-seq V6 data was applied, and the LogFCs of the mitochondrial genes differentially expressed between the cerebellum and each of the non-cerebellar brain tissues were plotted in turn. Additionally, all the mitochondrial genes, the differentially expressed genes and all the genes in the dataset were plotted, the logFCs used in this analysis being calculated with respect to all the non-cerebellar tissues. There was no clear trend across the histograms displaying the frequency of logFCs for differentially expressed mitochondrial genes, some showing a slight skew towards upregulation or downregulation, but most having no bias in the spread of logFC frequencies (**Fig.2.10a-e**), and therefore didn't agree with the results from either the GTEx RNA-seq or the BrainSpan microarray analysis. Repetition of the enrichment tests for differentially expressed mitochondrial genes gave a non-significant result, indicating lack of enrichment for this set of mitochondrial genes amongst genes either upregulated or downregulated in the cerebellum relative to the pooled cerebral tissues (**Fig.2.11**). Similarly, the spread of logFC frequencies for all the mitochondrial genes showed no bias (**Fig.2.12a**), but the spread of logFCs for all the genes and the differentially expressed genes showed a skew towards downregulation and upregulation respectively (**Fig.2.12b-c**), as seen for the equivalent histograms plotted using the BrainSpan microarray data. Although the mitochondrially-encoded genes were present in the original data, many were discarded during filtering of lowly expressed reads, and any bias in the spread of logFCs for these genes was unable to be determined.

Because none of the spreads of logFCs for differentially expressed mitochondrial genes matched between the three datasets, the correlation between the different datasets, with regards to the logFCs for each gene in common between the two datasets being correlated, was investigated. For this difference, data from the hippocampus was used, as this is a tissue present in all three datasets. As we can see, the logFCs show positive correlation across all the combinations of analyses (**Fig.2.13**). This implies that the differences in gene expression we have seen between the cerebellum and other tissues that varied between the analyses of the various datasets are not necessarily due to marked differences between the samples. The fact that the most closely correlated datasets were the GTEx RNA-seq and the BrainSpan RNA-seq data (**Fig.2.13c**) and the least correlated were the BrainSpan microarray

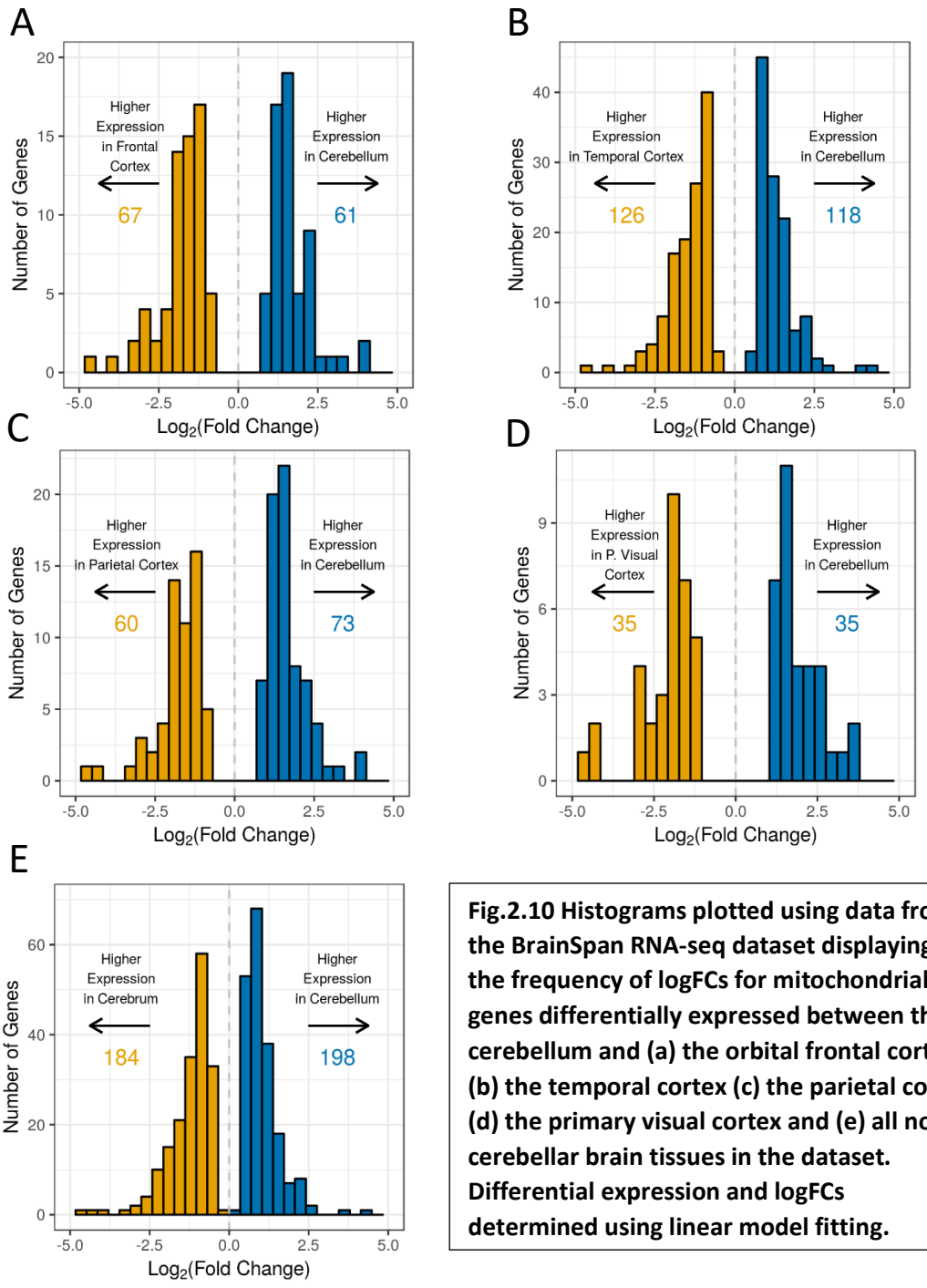


Fig.2.10 Histograms plotted using data from the BrainSpan RNA-seq dataset displaying the frequency of logFCs for mitochondrial genes differentially expressed between the cerebellum and (a) the orbital frontal cortex (b) the temporal cortex (c) the parietal cortex (d) the primary visual cortex and (e) all non-cerebellar brain tissues in the dataset. Differential expression and logFCs determined using linear model fitting.

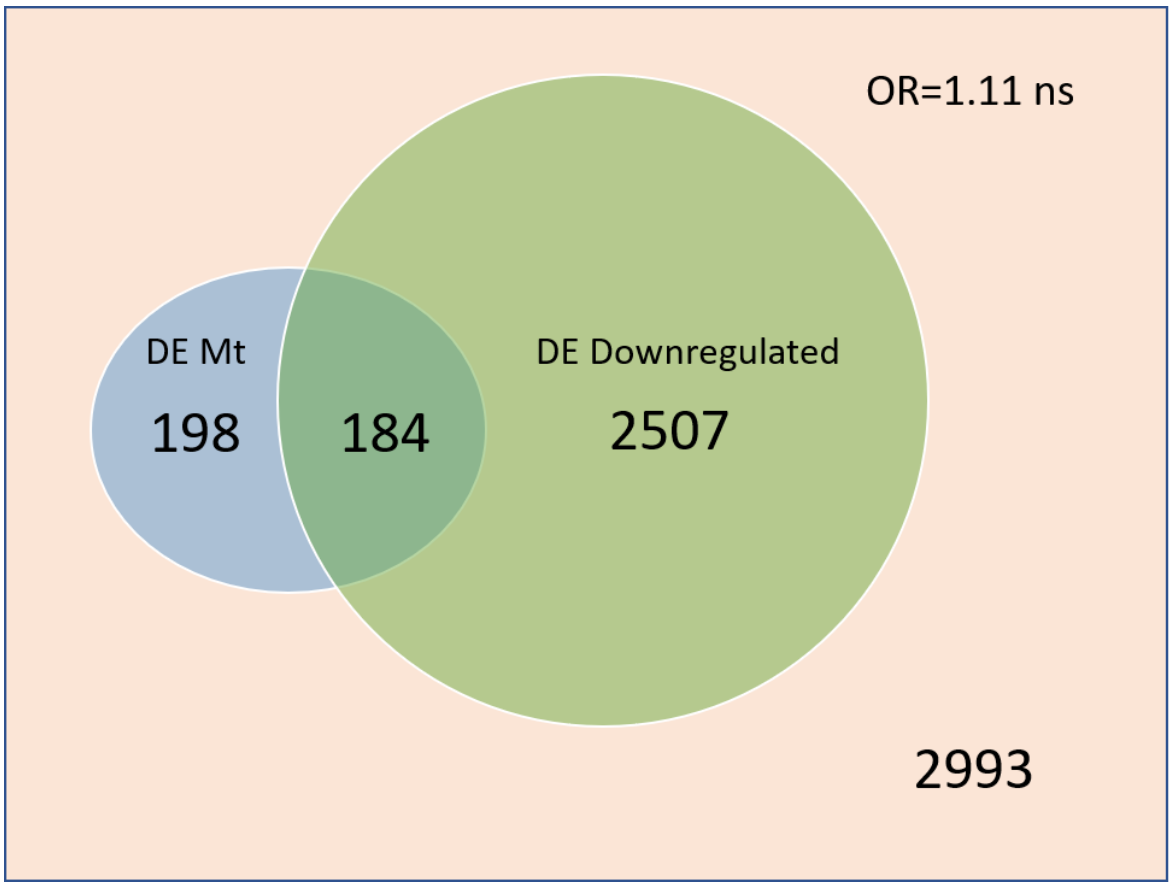


Fig.2.11 Venn diagram showing, for the BrainSpan RNA-seq data, overlap of genes differentially downregulated in the cerebellum relative to pooled cerebral tissues with mitochondrial genes differentially expressed in the same comparison. The background gene set is differentially upregulated genes. Significance and odds ratio determined by Fisher's exact test. Differential expression determined using linear model fitting. DE = Differentially expressed, Mt = Mitochondrial, OR = Odds ratio, ns = Non-significant.

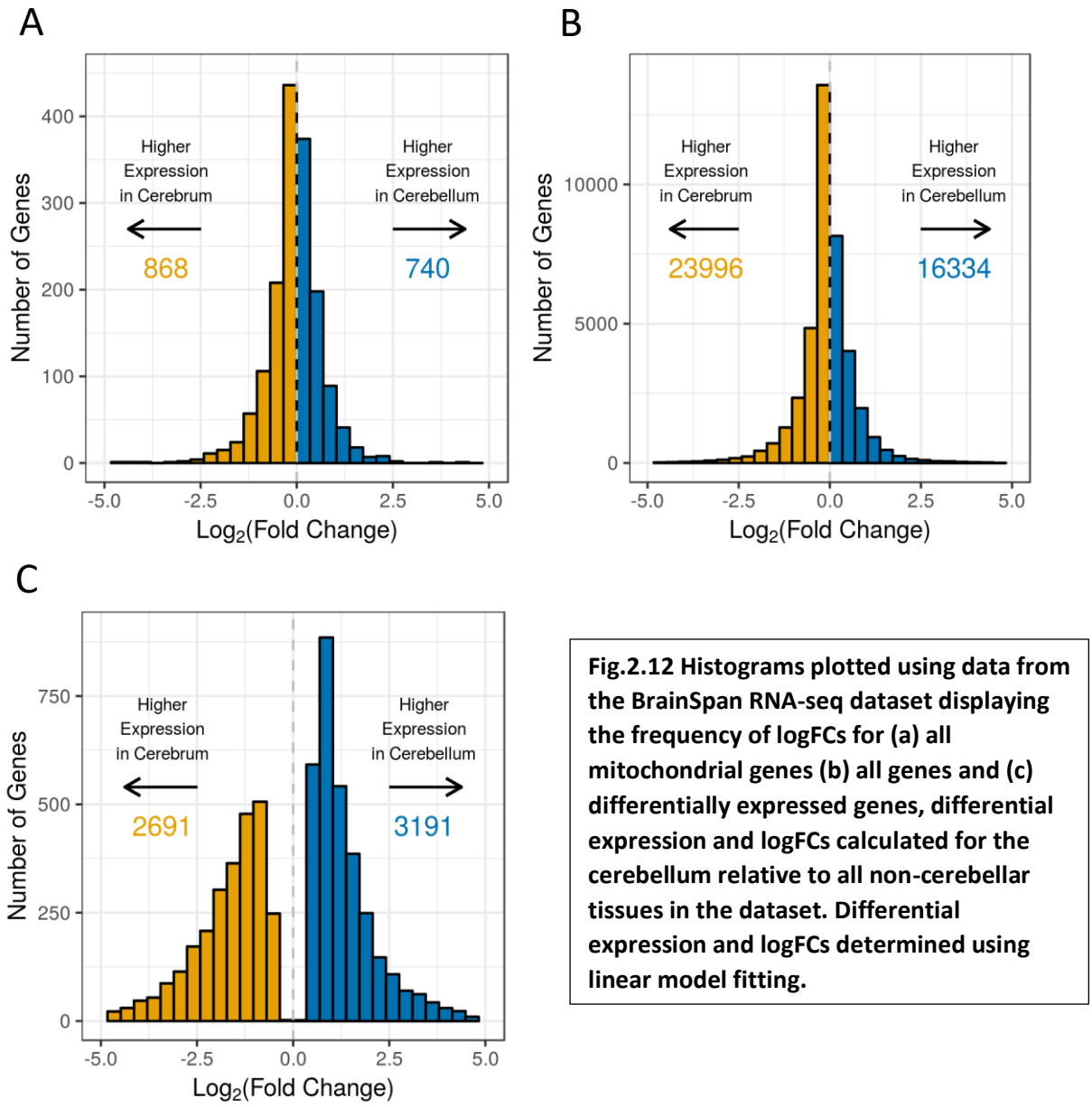


Fig.2.12 Histograms plotted using data from the BrainSpan RNA-seq dataset displaying the frequency of logFCs for (a) all mitochondrial genes (b) all genes and (c) differentially expressed genes, differential expression and logFCs calculated for the cerebellum relative to all non-cerebellar tissues in the dataset. Differential expression and logFCs determined using linear model fitting.

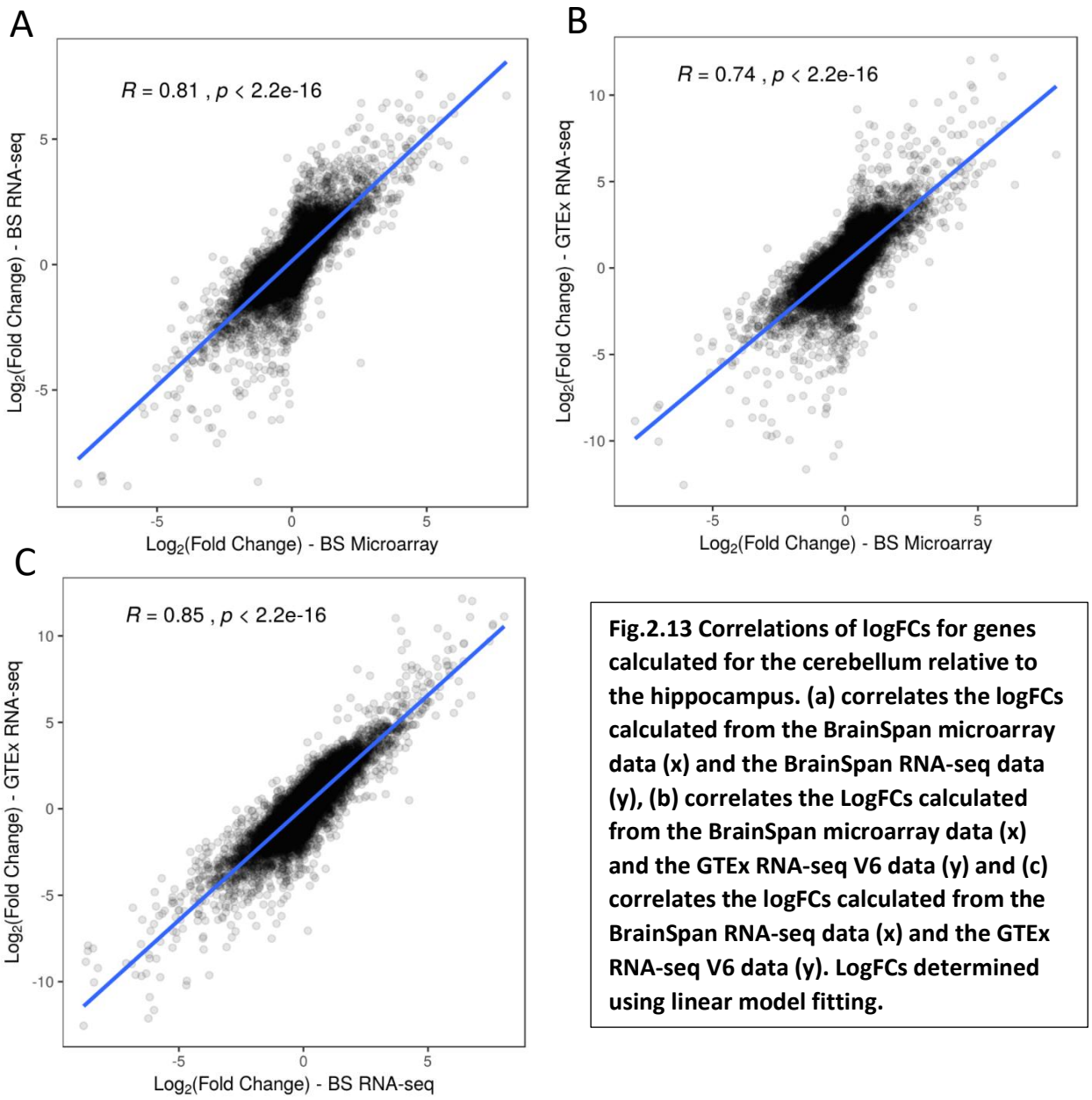


Fig.2.13 Correlations of logFCs for genes calculated for the cerebellum relative to the hippocampus. (a) correlates the logFCs calculated from the BrainSpan microarray data (x) and the BrainSpan RNA-seq data (y), (b) correlates the LogFCs calculated from the BrainSpan microarray data (x) and the GTEx RNA-seq V6 data (y) and (c) correlates the logFCs calculated from the BrainSpan RNA-seq data (x) and the GTEx RNA-seq V6 data (y). LogFCs determined using linear model fitting.

and the GTEx RNA-seq V6 data (**Fig.2.13b**), with the correlation coefficient for the BrainSpan RNA-seq and the BrainSpan microarray (**Fig.2.13a**) coming somewhere in between these two, implies that whilst samples have something to with how well the datasets correlate, the type of gene expression data is a better indicator of similarity.

2.2.3 Further RNA-seq data complements the cerebellar mitochondrial downregulation observed in the GTEx RNA-seq data

In order to try and elucidate which set of results represented genuine biology, another RNA-seq dataset was analysed. This was generated for and analysed in the paper Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS from Prudencio et al. and contains RNA-seq data for 9 matched pairs of frontal cortex and cerebellar samples derived from healthy controls. The raw data from Prudencio et al. was remapped in order to obtain read counts for the mitochondrial genes which had been filtered out in the publicly available pre-processed data. The basic method applied to the RNA-seq datasets previously analysed was repeated, this time logFCs being calculated for the cerebellum relative to the frontal cortex, comparing the results to those from the publicly available pre-processed data to make sure there were no marked differences between the two (**Fig.2.14a-b**). Differentially expressed and all mitochondrial genes showed a bias to downregulation whereas this was not seen in all genes (**Fig.2.15a-d)**. Fisher's tests to assess the enrichment of differentially expressed mitochondrial genes amongst differentially up or downregulated genes were carried out, revealing that the mitochondrial genes were highly overrepresented within those genes downregulated in the cerebellum relative to the frontal cortex with an OR of 2.19, the same direction of enrichment observed for the GTEx RNA-seq data, but with a stronger OR (**Fig.2.16**). The logFCs of all the mitochondrially-encoded genes in the dataset were then examined and it was found that, similar to the results for the same analysis of the GTEx data, all of the mitochondrially-encoded genes present in the dataset showed downregulation in the cerebellum relative to the frontal cortex (**Fig.2.15e**).

2.2.4 Remapping Prudencio et al. RNA-seq with decreasing stringency does not change the skew towards more mitochondrial genes downregulated in the cerebellum

One possibility that could explain the relative downregulation of mitochondrial genes seen across both the GTEx and Prudencio et al. RNA-seq datasets is that the mitochondrial genes

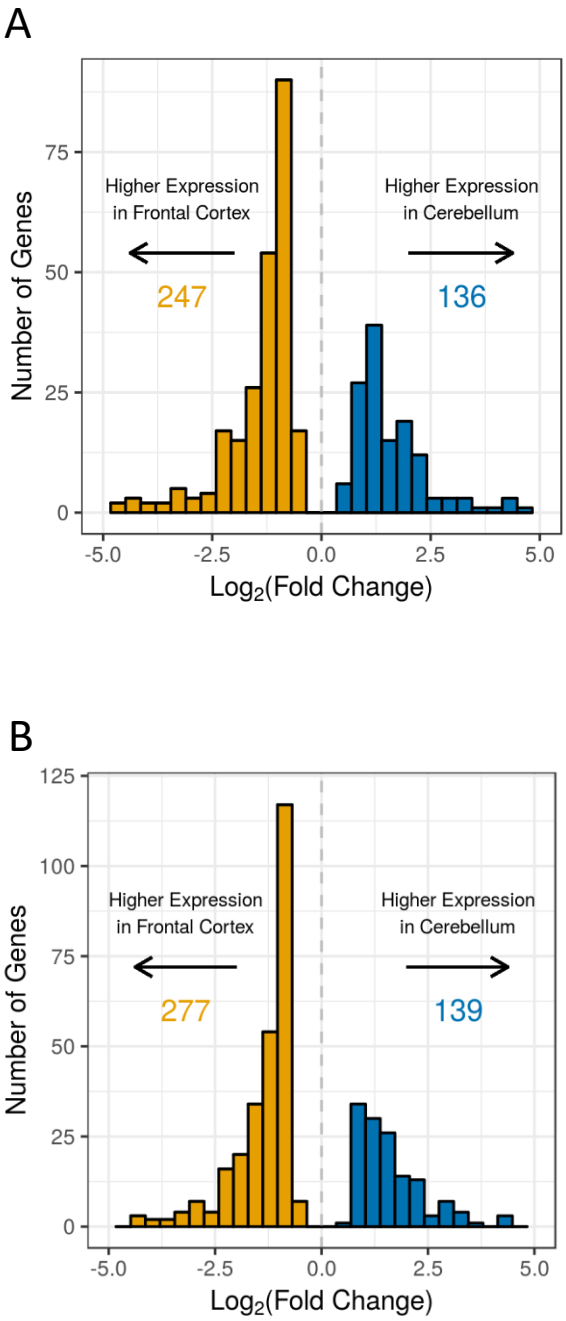


Fig.2.14 Histograms plotted using data from (a) the pre-processed, published Prudencio et al. RNA-seq dataset (b) the re-processed Prudencio et al. RNA-seq data, displaying the frequency of logFCs for differentially expressed mitochondrial genes, logFCs and differential expression being calculated for the cerebellum relative to frontal cortex. Differential expression and logFCs determined using linear model fitting.

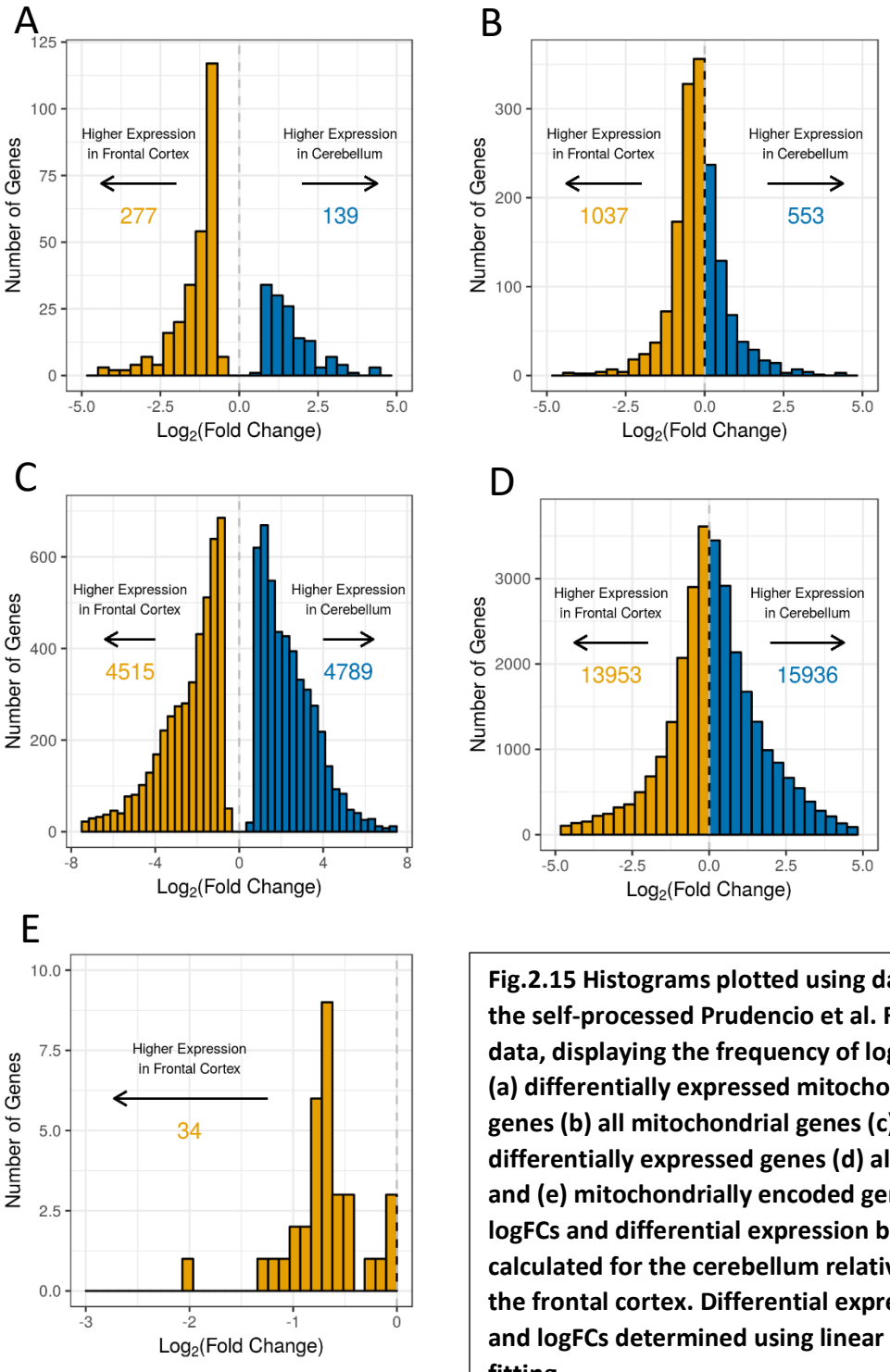


Fig.2.15 Histograms plotted using data from the self-processed Prudencio et al. RNA-seq data, displaying the frequency of logFCs for (a) differentially expressed mitochondrial genes (b) all mitochondrial genes (c) differentially expressed genes (d) all genes and (e) mitochondrially encoded genes, logFCs and differential expression being calculated for the cerebellum relative to the frontal cortex. Differential expression and logFCs determined using linear model fitting.

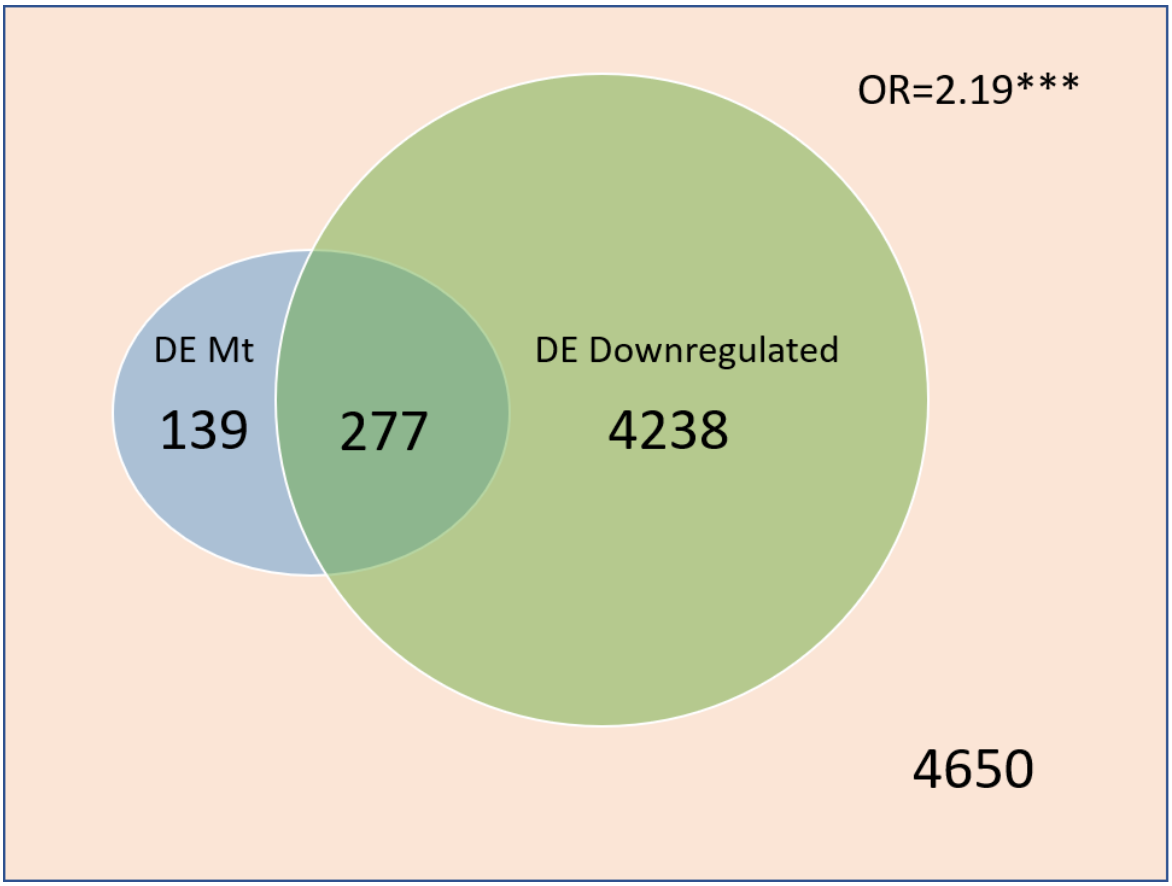


Fig.2.16 Venn diagram showing, for the Prudencio et al. healthy controls RNA-seq data, overlap of genes differentially downregulated in the cerebellum relative to the forntal cortex with mitochondrial genes differentially expressed in the same comparison. The background gene set is differentially upregulated genes. Significance and odds ratio determined by Fisher's exact test. Differential expression determined using linear model fitting. DE = Differentially expressed, Mt = Mitochondrial, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

are not mapping to the reference genome due to mismatches in their reads, perhaps due to a higher mutation rate for these genes. To test this hypothesis, we remapped the Prudencio et al. data three times with decreasing stringency, allowing for 5% of the read to be mismatched, then 10% and finally 15%. Differential expression analyses were performed on each of these different sets of read counts and histograms showing the spread of log fold-changes for differentially expressed mitochondrial genes were plotted. These showed that decreasing the stringency did not affect the skew towards cerebellar downregulation for mitochondrial genes (**Fig.2.17a-c**).

2.2.5 Single Strand Break Repair Genes are downregulated in the cerebellum relative the cerebrum and frontal cortex in the GTEx RNA-seq data

Having established the GTEx and Prudencio et al. RNA-seq datasets as the most reliable sets of data, we wanted to examine a cohort of single strand break repair and DRDA-ARCA disease proteins to see how their expression differed between the cerebellum and cerebral tissues. Therefore, we looked for their differential expression in the cerebellum – frontal cortex comparison for the GTEx RNA-seq dataset. Out of the 15 genes surveyed, 13 were found be differentially expressed, and 11 of these 13 had a lower expression in the cerebellum relative to the frontal cortex (**Fig.2.18a**). In order to elucidate whether this was true for the cerebellum compared all the cerebral tissues, the analysis was repeated for genes differentially expressed between the cerebellum and the pooled cerebral tissues. The results were remarkably consistent with the cerebellum – frontal cortex comparison: the same 13 genes differentially expressed along with a 14th, APTX, but the LogFC for APTX was so small as to be negligible. Additionally, the same 11 had lower expression in the cerebellum (**Fig.2.18b**). Although for many of these genes the downregulation observed is fairly weak, DNA polβ, which is known to be important for base excision repair, was highly downregulated in the cerebellum across both comparisons.

2.3 Discussion

Out of the four sets of data analysed, only two showed the same effect: the GTEx V6 and Prudencio et al. RNA-seq data. In both of these sets of data many more mitochondrial genes (amongst both the differentially expressed mitochondrial genes and all the mitochondrial

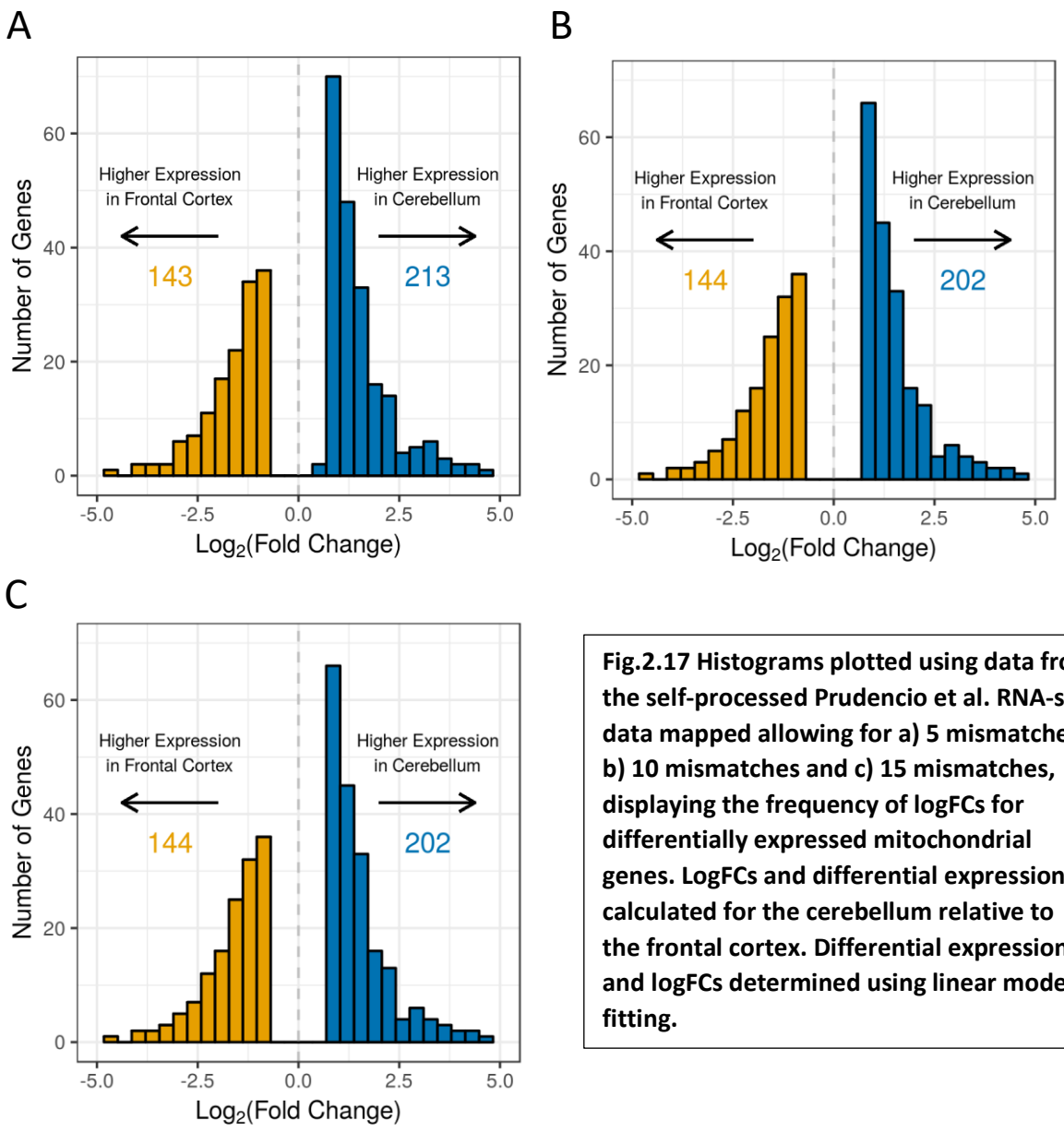


Fig.2.17 Histograms plotted using data from the self-processed Prudencio et al. RNA-seq data mapped allowing for a) 5 mismatches, b) 10 mismatches and c) 15 mismatches, displaying the frequency of logFCs for differentially expressed mitochondrial genes. LogFCs and differential expression calculated for the cerebellum relative to the frontal cortex. Differential expression and logFCs determined using linear model fitting.

A

Gene	logFC	FDR
RPA	↓ -0.303	1.64e-06
APEX1	↓ -0.15	1.78e-02
APEX2	↑ 0.363	2.99e-04
PARP1	↑ 0.336	1.01e-06
PARP2	↓ -0.423	9.12e-08
XRCC1	↓ -0.591	8.06e-12
DNApolB	↓ -2.8	7.91e-25
PCNA	↓ -0.727	2.29e-13
FEN1	↓ -0.286	1.14e-03
Ligase III	↓ -0.774	9.26e-14
TDP1	↓ -0.444	1.97e-07
ATM	↓ -1.43	3.28e-19
SETX	↓ -0.768	1.30e-15

B

Gene	logFC	FDR
RPA	↓ -0.334	3.15e-22
APEX1	↓ -0.264	1.30e-15
APEX2	↑ 0.404	1.84e-11
PARP1	↑ 0.45	1.47e-13
PARP2	↓ -0.765	1.12e-32
XRCC1	↓ -0.594	2.55e-36
DNApolB	↓ -2.89	2.94e-73
PCNA	↓ -0.603	3.61e-24
FEN1	↓ -0.779	1.98e-28
Ligase III	↓ -0.868	2.20e-46
TDP1	↓ -0.53	2.77e-25
APTX	↑ 0.0754	1.83e-02
ATM	↓ -1.03	3.76e-33
SETX	↓ -0.712	1.33e-41

Fig.2.18 Tables showing log fold-changes (logFC) in expression and Bonferroni-corrected p-values (FDR) for single strand break repair proteins differentially expressed between the cerebellum and a) the frontal cortex and b) the pooled cerebral tissues for the GTEx V6 RNA-seq dataset. If the logFC is yellow and is next to a down arrow, that gene is downregulated in the cerebellum relative to the tissue(s) of comparison. If the logFC is blue and next to an up arrow, that gene is upregulated in the cerebellum relative to the tissue(s) of comparison. Differential expression and logFCs determined using linear model fitting.

genes) show downregulation than show upregulation in the cerebellum relative to the cerebral tissue(s). Additionally, all of the mitochondrially-encoded genes show downregulation in the cerebellum. It can be tentatively suggested that the GTEx and Prudencio et al. results represent the most likely scenario with respect to differences in genes expression across the two tissues. This is because they are the only two of our datasets to agree with each other and the fact that the BrainSpan data represents two methods of examining gene expression from the same initial samples, so any issues with sample preparation would carry forward and be seen in both these datasets. Additionally, we were able to filter the GTEx data based on a number of important quality control metrics such as RNA integrity, mapping rate and duplication rate, whereas none of this data was available for the Brainspan data and as such no prefiltering was performed. Finally, the GTEx dataset contains more samples than the Brainspan data, and so outliers in gene expression are less likely to affect the overall result. The questions that arises from these analyses is what are the possible explanations for this phenomenon? First of all, the downregulation of mitochondrial genes could be a normal feature of cerebellar biology and may not represent any sort of abnormal mitochondrial function. This does not mean however that this mitochondrial downregulation effect is not associated with disease progression under conditions of increased genomic instability and mitochondrial dysfunction as seen in the DRDA-ARCAs. Alternatively, this relative downregulation be a result of there simply being less mitochondria in the cerebellum relative to the frontal cortex. However, less mitochondria does not necessarily mean a lower level of mitochondrial gene expression, as it has been shown that in disease states decreased mtDNA content does not always correlate with decreased mitochondrial transcription due to compensatory mechanisms (Barthélémy et al., 2001). Despite this caveat, the previously referenced studies in healthy mouse brains which have reported low mtDNA levels in the cerebellum relative to the other tissues of the brain adds weight to this hypothesis (Fuke et al., 2011). Another option is that if the cerebellum has a higher mutation rate relative to the cerebral tissues generally or for certain genes, it could it be due to increased mutation rates for mitochondrial genes. A proposed increased mutation rate could lead to an actual decrease in transcript levels due to mutations resulting in gene inactivation in some cells or a perceived decrease due to reads from highly mutated genes not mapping to the reference genome due to an increased number of mismatches. This second scenario appears unlikely, based on the results seen for

repeated analysis of the Prudencio et al. data with decreasing stringency when mapping. However, either of these scenarios could lead to the relative mitochondrial gene downregulation that we observe. Why the cerebellum could have a high basal mutation rate genome wide or for gene subsets? This could be due to the cerebellum's extended period of development relative to other neurological tissues, which could burden cerebellar cells with an increased mutation load (El-Khamisy, 2011). However, as the damage that results from development is replication associated, why the mitochondria would be particularly affected is not clear. It has been suggested recently that somatic mutations in brain tissues are preferentially acquired in transcriptionally active genes (due to the absence of DNA replication in the non-cycling cells), and this could lead to a "use it, lose it" model, whereby the most transcriptionally active genes acquire mutations at a higher rate and therefore lose function faster (Lodato et al., 2015). Building a hypothesis around this phenomenon could lead to a model whereby the mitochondria acquire more mutations not developmentally but in the mature brain due to the high energy demand requiring high expression of mitochondrial genes and leading to the production of more ROS, and this could be particularly pronounced in the cerebellum. In relation to this possibility, it is intriguing to note that the majority of key SSBR proteins are downregulated in the cerebellum relative to the cerebrum. Enhanced mutation rate could alternatively be a result of their being less repair proteins present in the cerebellum, or this could further contribute to the scenario described above. It is particularly worth noting that DNA polβ, which is involved in base excision repair, is highly downregulated in the cerebellum (Ray et al., 2013). If the cerebellum is deficient in base excision, then that could indeed go part of the way in explaining an increased mutational load. It should also be pointed out that some of these ideas are not mutually exclusive and a satisfying explanation may be reached by considering several or all of them as contributing factors. One potential confounding factor is that the cerebellum is made up mainly of two types of cell, granule cells and purkinje cells. The differing results observed across the datasets could be as a result of differing proportions of the cell types. Purkinje cells are generally regarded as being the cells affected in DRDA-ARCAs, and so it would be ideal to separate out these two cells types in order to isolate data from the purkinje cells. This could be done through single cell RNA-seq from the cerebellum, and so this type of sequencing data would constitute an ideal extension to this project. Additionally, qPCR validation of some of the most downregulated genes in the cerebellum

relative to the cerebrum would enable us to be confident that the results seen for the GTEx and Prudencio RNA-seq datasets were the correct ones out of all the datasets analysed.

2.4 Methods

2.4.1 Datasets

The BrainSpan microarray and RNA-seq data was obtained from the website of the BrainSpan consortium (http://www.brainspan.org) under the titles "Exon microarray summarized to genes" and "RNA-Seq Gencode v10 summarized to genes" respectively (Allen Insitute, 2010) . The GTEx RNA-seq data was obtained from the GTEx portal (http://www.gtexportal.org) under the heading "GTEx Analysis V6" (Carithers and Moore, 2015) The both the pre-processed and raw data for the healthy controls from the ALS study performed by Prudencio et al. 2015 (available under the accession GSE67196) (Prudencio et al., 2015)

2.4.2 Analysis of microarray data

To perform differential gene expression analyses on the microarray data using the Limma (Ritchie et al., 2015; Smyth, 2004) package in R studio (R Core team, 2019; R Studio Team, 2015), the data was subsetted and formatted for the tissues being compared and samples from donors within the age range 18 – 40 (as 40 was the highest age in the set), and the phenotype, genotype and expression data was combined into an expression set object for downstream analysis. The data was blocked according to the donor IDs to remove batch effects and a model matrix was generated based on the tissue and the block groups. A linear model was then fit and empirical Bayesian shrinkage applied. Differentially expressed genes were identified using a Benjamin-Hochberg procedure corrected p value cut-off of 0.05. When looking at all genes in the dataset was suitable, a p value of 1 was used instead. To build the heatmaps, batch effects were first removed from the relevant subset of data using the `removeBatchEffect` command and then the donors were clustered according to the Euclidean distance and the genes clustered according to the Pearson distance using the `hclust` function. The heatmaps were then plotted using the `heatmap.2` function. To identify mitochondrial genes, the gene ids of mitochondrial

genes were extracted from the org.Hs.eg.db annotation database package for R using GO terms for mitochondrial pulled from the GO annotation package GO.db (Carlson, 2015, 2019) Statistical enrichment and odds ratios were calculated using Fisher's exact test through the R function `fisher.test`.

2.4.3 Analysis of RNA-seq data

To perform differential gene expression analyses on the RNA-seq datasets using the edgeR package in R, the data was first subsetted and formatted for the tissues being compared and certain samples (Robinson and Smyth, 2008; Robinson et al., 2010). For analysis of the BrainSpan RNA-seq dataset, only samples from donors within the age range 18 – 40 (40 was the highest age in the set) were selected. For the GTEx V6 RNA-seq dataset, only samples that were from donors between the ages of 20 – 59 were selected. These samples were then subsetted for samples that had an RNA integrity of equal to or greater than 6, a mapping rate of over 0.8 and a duplication rate of less than 0.5. For the analysis of the healthy control RNA-seq data from Prudencio et al., all the samples that passed the quality control checks as discussed were used for the differential gene expression analyses. For all the datasets, lowly expressed genes were filtered from the datasets. Genes that didn't have a read count above 0 in at least a number of samples equal to half the size of the smallest sample group were removed e.g. if the frontal cortex samples represented the smallest sample group, and there were 8 of them, genes would need to have a read count of above zero in at least 4 samples. The library size was then re-computed, the data blocked according to the donor IDs to remove batch effects and a model matrix generated based on the tissue and the block groups as before. A generalised linear model was fit and differentially expressed genes identified using a Benjamini-Hochberg procedure corrected p-value cut-off of 0.05. Mitochondrial genes were extracted as discussed in **Analysis of microarray data.** Statistical enrichment and odds ratios were calculated using Fisher's exact test through the R function `fisher.test`.

2.4.4 Processing RNA-seq data

The processing of the raw RNA-seq data from Prudencio et al. was done using the CGAT pipelines (https://github.com/CGATOxford/CGATPipelines), specifically the quality control pipeline, the mapping pipeline and the RNA-seq differential expression pipeline. To map the

reads to the human reference genome (Ensembl 75, hg19), `hisat2` was used with default options (Kim et al., 2015a; Zerbino et al., 2018). To generate read counts, the mapped data was run through `FeatureCounts` as part of the RNA-seq differential expression pipeline from CGAT pipelines (Liao et al., 2014). The samples cereb-WT-24 and FCX-WT-95 were not included due to their poor quality.

2.4.5 Data manipulation and figures

Plots were generated using the gplot and the ggplot2 packages (Wickham, 2016). To generate the overlap table seen in **Fig.2.1**, the R package gridExtra was used (Auguie, 2017). The tables seen in **Fig.2.18** were generated using the R package formattable.

2.1.1 Tabular summary of datasets and analyses

Dataset	Experimental design/Samples utilised	Analyses performed
BrainSpan Microarray	5 cerebellar cortex samples 7 orbital frontal cortex samples 14 temporal cortex samples (pooled 7 inferolateral temporal cortex (area TEv, area 20) and 7 posterior (caudal) superior temporal cortex (area 22c) samples) 7 primary visual cortex (striate cortex, area V1/17) samples 7 posteroventral (inferior) parietal cortex samples 102 pooled cerebral samples	Differential expression analyses Category enrichment (Fisher's) test (cerebellar cortex and pooled cerebral samples) Correlation analysis of log fold-changes between datasets (cerebellar cortex and hippocampus samples)
BrainSpan RNA-seq	8 cerebellar cortex samples 8 orbital frontal cortex samples 16 temporal cortex samples (pooled 8 inferolateral temporal cortex (area TEv, area 20) and 8 posterior (caudal) superior temporal cortex (area 22c) samples) 7 primary visual cortex (striate cortex, area V1/17) samples 8 posteroventral (inferior) parietal cortex samples 114 pooled cerebral samples	Differential expression analyses Category enrichment (Fisher's) test (cerebellar cortex and pooled cerebral samples) Correlation analysis of log fold-changes between datasets (cerebellar cortex and hippocampus samples)

GTEx V6 RNA-seq	81 cerebellum samples (pooled 41 Brain – Cerebellum and 40 Brain – Cerebellar Hemisphere samples) 35 Brain – Frontal Cortex (BA9) samples 22 Brain – Hippocampus samples 28 Brain – Caudate (basal ganglia) samples 21 Brain – Anterior cingulate cortex (BA24) samples 270 pooled cerebral samples	Differential expression analyses Category enrichment (Fisher's) test (cerebellum and pooled cerebral samples) Correlation analysis of log fold-changes between datasets (cerebellum and hippocampus samples
Prudencio et al. 2015 RNA-seq (available under the GEO accession GSE67196)	9 donor matched pairs of cerebellum and frontal cortex samples	Differential expression analyses Category enrichment (Fisher's) test

3. Chapter 3 – Comparing the mutational landscape of the Cerebellum and the Cerebrum

3.1 Introduction

As discussed in the introduction to the previous chapter, there are several known inherent differences between the cerebellum and the cerebrum: namely the cerebellum's extended period of maturation relative to most cerebral tissues, it's distinct transcriptional profile and sensitivity to mitochondrial diseases (El-Khamisy, 2011; Melé et al., 2015). Also mentioned is the possibility that on top of these differences and possibly even a driver of them, the cerebellum has a mutational landscape distinct from that of the cerebrum. This could take either or both of two forms, either a higher or lower basal mutation rate or alternatively a divergent mutational spectra indicating different relative contributions from various sources of DNA damage. It may also be the case that the cerebellum is enriched relative to the cerebrum for a particular type of somatic mutation with respect to mosaic and cell private. An increased mosaic mutational load could arise due to differences in development between two tissues whereas more cell private mutations would be associated with a different set of challenges in the developed brain. The techniques for calling these varieties of mutations has been discussed in brief, but it should be noted that somatic mosaic mutations can be picked up by germline variant callers if they are sufficiently abundant. This is because germline variant callers try to identify alternative allele frequencies close to 0.5 and 1 to identify heterozygous or homozygous variants respectively, although bespoke somatic mosaic variant calling programs have been created. Cell private mutations on the other hand represent a more difficult problem because of their low abundance resulting in few supporting reads, a problem compounded when calling from RNA-seq data because of coverage issues (Coudray et al., 2018; Yizhak et al., 2019), meaning it is not possible to accurately call base changes from bulk RNA-seq data. This amongst other problems means that cell private mutations require dedicated pre-processing and calling software. Somatic variant callers also come in many different flavours to tackle the myriad issues surrounding this type of variant calling. Many are built around comparing matched tumour and normal samples, whilst others are dedicated to finding higher frequency variants in lower coverage data and vice versa. Finally, there are callers specifically built to call from RNA-seq by

attempting to address problems such as the low coverage for lowly expressed, false positive mutations around splicing junctions due to alignment errors and RNA editing. Benchmarking studies have largely concluded that variant calling from RNA-seq is best performed in conjunction with matched DNA-seq data, but can give relatively robust results when subject to extensive treatment and prefiltering (Coudray et al., 2018; Yizhak et al., 2019). Recently, there has been success in using single-cell DNA-seq to investigate the landscape of somatic mutations in the brain, a technique attractive because cell private mutations become far easier to detect with confidence (Lodato et al., 2015). Similarly, single cell RNA-seq data is starting to be used for somatic variant calling, however, many of the problems with calling from bulk RNA-seq unfortunately carry over. Whilst these programs often have the advantage of being independently benchmarked and subject to the review of the scientific community, despite tailoring to different applications there are some things these tools are not designed to do. An example relevant to this chapter is that whilst somatic variant callers are able to give an idea of the mutational spectra arising in a tissue based on the differing proportions of the base changes that occur, it is more difficult to assess differences in basal mutation rate across specific genes. This chapter aims to investigate potential differences in the mutational landscape of the cerebellum and the cerebrum, in terms of spectra and mutation load, using publicly available RNA-seq data. As such, it explores a range of ways to glean relevant information from this type of data. It also aims to further explore and expand on the cerebellar – frontal cortex differences discovered in the previous chapter, including the possibility of mitochondrial genes having a greater rate of mutation in the cerebellum and the application of the “use it and lose it” hypothesis to this and potentially other subsets of genes.

3.2 Results

3.2.1 Development of a pipeline that identifies mismatches in sequence data

One of the downsides to investigating phenomena through public data is that relevant data in an ideal format may not be available. This is a problem increasingly compounded the more specialised the area of work. When investigating the spectrum and load of mutations across different brain regions, bespoke data is not readily available. Therefore, we had to

use less optimised data to test our hypotheses. As well as the type of data, the number of samples for each brain region also needed to be taken into account. An ideal dataset with very few samples may be insufficient because it does not have sufficient power to detect the differences we are looking for, especially with respect to genome wide and cell specific differences in mutation rate. It is likely that a relevant dataset would have to come from a consortium or similar large project. RNA or DNA sequence from brains represents another problem, because it can only be acquired post-mortem it tends to be less abundant than data from other tissues. These factors combined severely restrict the pool of workable data. At the time of beginning this project, the most relevant and comprehensive set of data was the GTEx RNA-seq dataset. This collection has many advantages. For one, a wide variety of brain regions are sequenced, meaning we would be able to select the regions we determined to best suit the project. Records of quality control and phenotypic data are also kept as part of the service, meaning samples meeting specific criteria, such as age or RNA integrity, can be selected for downstream analysis. GTEx also keeps data on the variants called across all of their different samples. Finally, GTEx generally contains over 100 samples for each tissue it sequences, and this is true of all the neural tissues except the substantia nigra and the spinal cord. These factors made it an obvious candidate for our investigation. However, it is still RNA-seq data, which as discussed is not optimal for calling variants. Furthermore, although it can be done, variant callers are not designed to look at different mutation rates across genes, a question we were interested in because of the system of "use and lose it" proposed to be in effect in the mature brain. Because of this we decided to build a script and complementary pipeline that detected and recorded reference genome – RNA-seq read single nucleotide mismatches. This was done separately for each individual gene passed to the script, the base change occurring logged and the number of bases assessed for each gene likewise recorded, in order to calculate per gene mismatch rates and mutational spectra accordingly. These mismatches were intended to act as measures of mutational load. Whilst any one mismatch might be a sequencing error, it was reasoned that the error rate across the tissues being compared would be equal, so the importance is placed on a relative mismatch rate, not on any single mismatch or an accurate reading of the total number of mutations. Based on this logic any differences between the tissues being compared would have to arise from disparities in the underlying level of mutation. Because accurate downstream analysis hinged on this principle, it was important to make

sure that there were no confounding effects that could lead to inter-tissue discrepancies and false results. Important filtering steps to offset the unreliability of calling from RNA-seq data were also integrated in the pipeline or the code itself. Reads mapping to more than one position, duplicate read, unmapped reads and bases with a quality score lower than 30 were all discarded prior to mismatch calling. To remove SNPs, we took the GTEx VCF file containing all the SNPs called for each sample through whole blood exome sequencing, and each candidate mismatch was cross-referenced with the variant called for that sample within the GTEx VCF file. If a match was found, the mismatch was not recorded. The samples run through the pipeline were also subject to phenotypic criteria and quality control filters in order to minimise the chance of poor samples leading to unreliable results. This constituted a basic version of this mismatch calling pipeline, henceforth referred to as pipeline V.1. However, after working with this initial version of pipeline, we realised that it lacked a feature necessary for effective variant calling RNA from-seq data: a way to identify and filter out RNA-editing events (Yizhak et al., 2019). RNA editing of an adenosine to an inosine base is major confounding factor when calling variants from RNA-seq data. As RNA-editing is picked up by the sequencer as a G base, these events manifest as A to G base changes and are therefore picked up by our script as mismatches. Therefore, in a similar manner to SNP filtering, each potential A to G mismatch was searched for in an online database of RNA editing events called REDIPortal, and discarded from downstream analysis if a hit was found (Picardi et al., 2017). This scenario was also recorded by our script as a likely RNA editing event at this position and added to an RNA edit count. Indels were also excluded from our analyses by discarding mismatches that occur in the context of three or more mismatches in a 5 base window in this subsequent iteration of the pipeline. Alterations to how germline SNPs were excluded were also made. Instead of using the GTEx VCF file, germline variants were called from the RNA-seq using the GATK HaplotypeCaller and each candidate mismatch was cross-referenced with these variants. It was reasoned that this might better capture tissue specific mosaic mutations, and so both improve accuracy and identify an interesting subset of mutations. A final change made was to how base changes were recorded. Pipeline V.1 did not take into account the strandedness of the read when recording the base changes, leading to erroneous results. This can be seen by plotting the proportion of mismatches represented by each possible base transition and colour coding it by tissue. This creates a “mirror image” situation where each of the

transition's reverse complement in terms of base pairing had almost the same representation. This was a result that was suspiciously uniform. Therefore, a strandedness checker was implemented to check whether a read was the reverse complement of the original read, and if so, the bases from the original read were recreated through the principles of base complementarity. A modified version of this newer pipeline that only assessed mismatched falling within exons was also created. These updated pipelines will be referred to as pipeline V.2 total gene and pipeline V.2 exons for the pipeline that records mismatches across the whole gene body and the version that logs mismatched only in exons respectively. Irrespective of the version of the pipeline used, the samples run through the pipeline were also subject to phenotypic criteria and quality control filters in order to minimise the chance of poor samples leading to unreliable results. Only samples from individuals between the ages of 20 and 59 and that had a Hardy scale score of 1 or 2 were carried forward to quality control filters. The Hardy scale score is a value that indicates the manner in which an individual died. A score of 1 or 2 on the scale correspond to "1) Violent and fast death Deaths due to accident, blunt force trauma or suicide, terminal phase estimated at < 10 minutes 2) Fast death of natural causes, sudden unexpected deaths of people who had been reasonably healthy, after a terminal phase estimated at < 1 hour (with sudden death from a myocardial infarction as a model cause of death for this category)" We applied these phenotypic filters to eliminate samples from particularly elderly individuals with the age based checks and from individuals suffering from terminal disease with the Hardy score filters, both circumstances of which might involve the breakdown of normal tissue function and gene expression. As for the RNA quality checks, only samples that had an RNA integrity score of 6 or greater, a mapping rate of 0.8 or greater and a duplication rate of 0.5 or lower.

3.2.2 The cerebellum has a higher tissue mismatch rate and more genes with a higher mismatch rate than the frontal cortex for pipeline V.1 results

After building our first iteration of the mismatches calling pipeline, pipeline V.1, we decided to run GTEx RNA-seq data from the cerebellum and the frontal cortex through the pipeline. The frontal cortex was chosen as the tissue of comparison because it is highly distinct in terms of location from the cerebellum and it is the only tissue aside from the cerebellum to appear in both the GTEx and Prudencio et al. derived RNA-seq data, therefore allowing

continuity with our previous work on patterns of differential expression between the two tissues. After the phenotypic and quality control filters, 24 cerebellum and 29 frontal cortex samples were mapped, with each sample consisting roughly of 60-80 million mapped reads, and then put through the pipeline. Subsequent analysis of the mismatch matrices generated by the pipeline consisted of removing genes that had less than 10 bases and 1 mismatch recorded, and then summing the number of mismatches and bases over each sample and dividing the mismatch number by the number of bases to get a mismatch rate of average mismatches per base for each sample. T-tests assessing whether the mismatch rate had any relationship to the tissue of origin revealed that the cerebellar samples had a significantly higher average mismatch rate than the frontal cortex, 0.00266 per sequenced base compared to the cortex rate of 0.00249 (**Fig.3.1a**). In order to investigate the accumulation of mismatches in specific genes, mismatch rates were similarly calculated but on a per gene basis within each sample by dividing the total mismatches recorded for a given gene by the number of reads mapping to that gene. Linear models were then fit assessing which genes had a significantly differential mismatch rate between the cerebellum and the frontal cortex. Of the ~8500 genes which we determined to have different rates of mismatch between the two tissues, the vast majority of them had a higher rate of mismatch in the cerebellum rather than the frontal cortex, 7463 compared to 1162 respectively (**Fig.3.1b**). Base changes could not be analysed because of errors in the base recording system identified after running the data through pipeline V.1, but this was amended in pipeline V.2.

3.2.3 Genes with higher differential mismatch rates in the cerebellum and the frontal cortex are enriched for genes showing relative downregulation in the relevant tissue

Having identified two set of genes, those that have a higher mismatch rate in the cerebellum relative to the frontal cortex and vice versa, enrichment tests were carried out to see whether specific subsets of genes were enriched amongst either of these sets of genes. First of all we looked at those genes with a higher rate of mismatch in the cerebellum, and looked for over or underrepresentation of genes that we had determined to be significantly downregulated or upregulated in the cerebellum relative to the frontal cortex through analysis of the same GTEx RNA-seq data. Fisher's tests revealed that these cerebellar mismatched genes were significantly enriched for genes downregulated in the cerebellum, with a substantial OR of 2.01 and similarly depleted of those genes upregulated

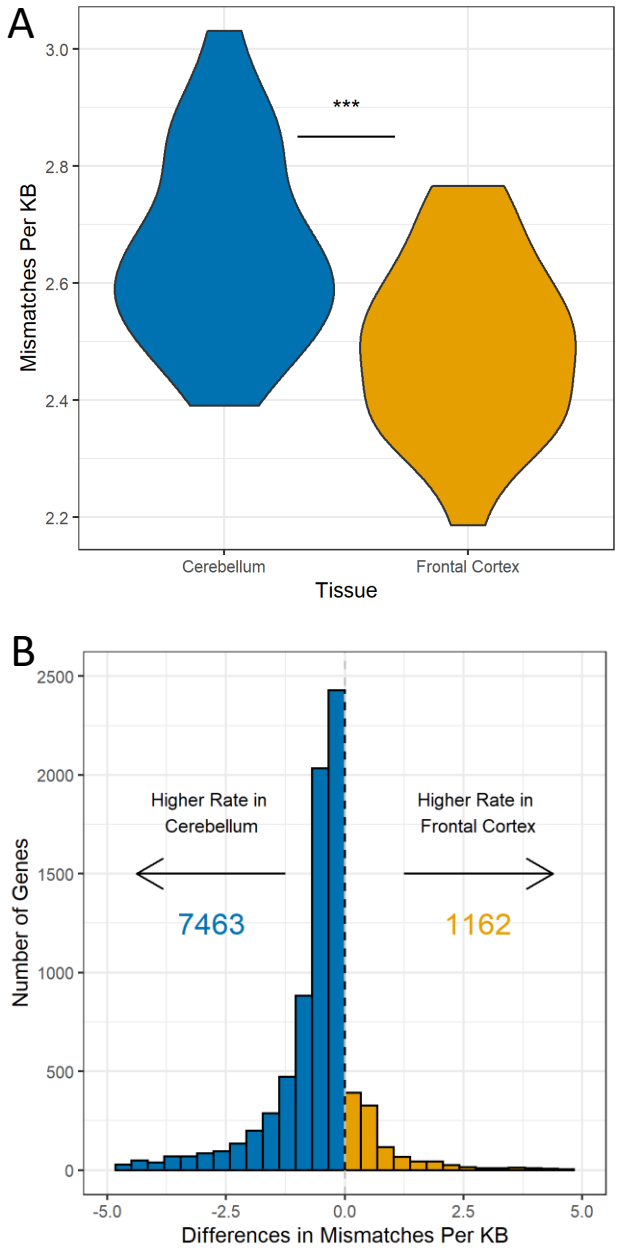
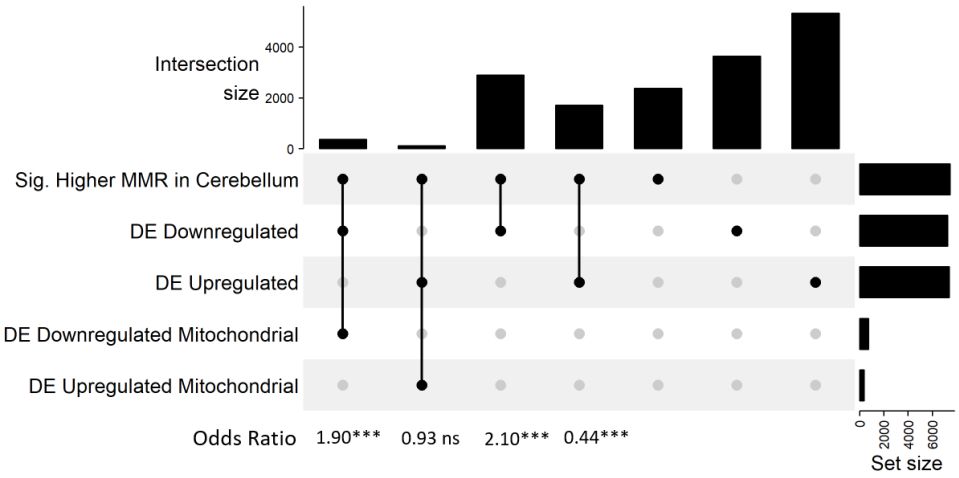


Fig.3.1a) Violin plot showing the spread of mismatches per kilobase (KB), determined using pipeline V.1 , for RNA-seq samples from the cerebellum and frontal cortex. Significance determined by t-test.

b) Histogram showing number of genes with differential rates of mismatch between the cerebellum and frontal cortex RNA-seq samples plotted against differences in mismatches per kilobase, determined using pipeline V.1 , between the two tissues. Genes with differential rates of mismatch were identified using linear model fitting. p-values : * <= 0.05. ** <=0.01. * <= 0.001**

in the cerebellum with an OR of 0.44. As differences in mitochondrial expression were a major finding in the chapter 2 of this thesis, we also decided to investigate enrichment of mitochondrial genes amongst the significantly mismatched set. The total set of mitochondrial genes was not assessed for enrichment because only 200 out of the ~1300 GO annotated mitochondrial genes were not differentially expressed, and the addition of this small group of non-differentially expressed mitochondrial genes would be unlikely to drastically change any observed over or under representation. Therefore, only the differentially expressed mitochondrial genes were tested for enrichment and they were split into those downregulated and those upregulated in the cerebellum relative to the frontal cortex. Those downregulated showed a significant enrichment within the cerebellar mismatched genes, although a slightly weaker enrichment compared to the cerebellar downregulated genes overall, with an odds ratio of 1.9, whereas the upregulated set showed a minor, non-significant depletion (**Fig.3.2a)**. Having determined enrichment/depletion for certain genes subsets within the genes with a higher rate of mismatch in the cerebellum, we wanted to assess whether this result was specific to this set of cerebellar mismatched genes. Therefore, we repeated the enrichment analyses, but instead using the set of genes that had a significantly higher rate of mismatch in the frontal cortex relative to the cerebellum. It should be noted that as the sets of genes showing differential expression in the cerebellum were calculated relative to the frontal cortex, any genes that are upregulated or downregulated in the cerebellum represent genes that show relative downregulation or upregulation respectively in the frontal cortex. Bearing this in mind, the enrichment tests for the mismatched frontal cortex genes gave the opposite result as to those for the cerebellar genes. Genes downregulated in the cerebellum and therefore relatively upregulated in the frontal cortex were significantly depleted, (OR=0.4), whereas the genes upregulated in the cerebellum/downregulated in the frontal cortex were significantly overrepresented amongst this set of genes with an odds ratio of 2.76 (**Fig.3.2b**). Cerebellar downregulated and upregulated mitochondrial genes showed weak non-significant depletion and significant enrichment (OR=1.61) respectively. Collectively, these results indicate that genes with significantly higher rates of mismatch between the cerebellum and the frontal cortex are enriched for genes downregulated/depleted for genes upregulated within the tissue those genes have a higher mismatch rate in.

A



B

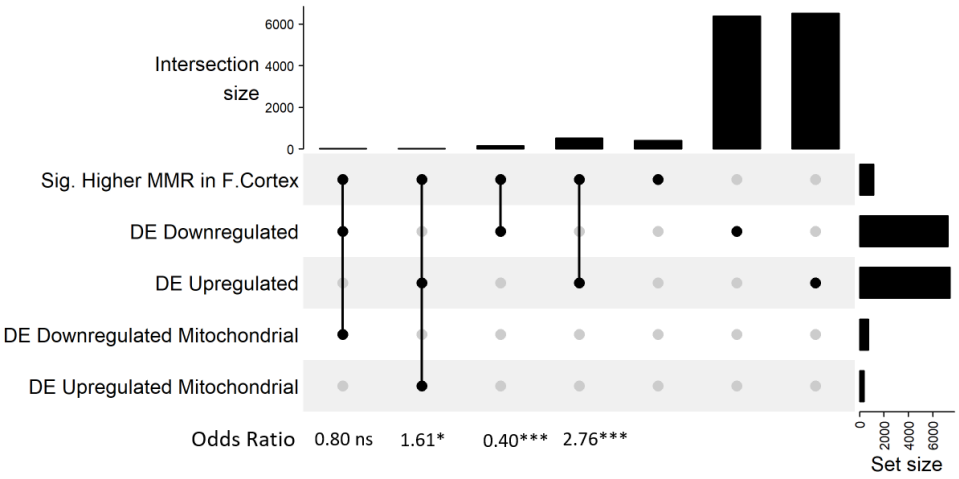


Fig.3.2 Upset plots of enrichment of specific gene categories amongst a) genes with a significantly higher mismatch rate in the cerebellum compared to the frontal cortex, and b) genes with a significantly higher mismatch rate in the frontal cortex relative to the cerebellum. Significance and Odds Ratios (OR) determined by Fisher's exact test. Mismatch rates calculated from pipeline V.1 results. Sig. = Significantly, DE = Differentially, MMR = Mismatch rate , ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

3.2.4 Assessing the efficacy of mismatch pipeline V.1

This difference in both the tissue wide and gene specific differences in mismatch rates between the cerebellum and the frontal cortex could be due to higher levels of damage in the cerebellum. However, alternative explanations must be ruled out if we are to make this claim. The tissue from the cerebellum could be of a poorer quality for any number of reasons, leading to more sequencing errors. If the cerebellum has a greater load of sequencing errors, then our previous assumption that the only source of difference between the cerebellum and the frontal cortex must be a biological one no longer applies. As part of the quality control step when selecting samples for mapping, an RNA integrity filter was already applied. However, it is possible that even above this RIN threshold there are differences between the tissues. Therefore we plotted the RNA integrities of the quality filtered cerebellar and frontal cortex samples that we had taken forward for mapping. As a point of comparison, we also included the RINs for quality filtered cerebellar hemisphere samples that we had not included in our initial run of pipeline V.1. It is important to note here that these cerebellar samples are actually intended as duplicate samples to the cerebellum samples and are not distinct from them. The difference between them is the cerebellum samples were preserved in PAXgene tissue fixative solution, whereas the cerebellar hemisphere tissue samples were taken later as close as possible to the initial site, so had a longer time period from death to preservation, and were preserved by snap freezing. The frontal cortex samples were sampled and preserved in a similar fashion to the cerebellar hemisphere samples. The PAXgene tissue fixative solution preserved counterparts to the frontal cortex samples were named Cortex and not included in this analysis. This seemly strange selection of samples for analysis came about because this information about the preservation of these samples was not disclosed to us until the writing of this thesis. Up until this time, based on the provided annotations, it was understood that that the cerebellum and the cerebellar hemisphere samples differed in terms of the area of the cerebellum the tissue sample was taken from. Because of the lateness of the true distinction between the cerebellar samples being made clear to us, there was not adequate time to repeat the analysis with the Cortex included and to limit the direct comparisons to samples preserved by the same method. These differences between tissues should be borne in mind when interpreting the results. Regardless, It was indeed found that the cerebellar samples

had a lower RIN than the cortex samples. However, the cerebellar hemisphere samples had a higher average RIN than the frontal cortex (**Fig.3.3a**). Therefore, we decided to run pipeline V.1 and the downstream statistical tests again, this time including samples for the cerebellar hemisphere. The reasoning underpinning this approach was that if the cerebellar hemisphere samples had a higher average RIN than the frontal cortex samples but were found to have a tissue rate of mismatch and more genes with a higher rate of mismatch amongst differentially mismatched genes relative to the frontal cortex, that it may be reasonably posited than differences in tissue quality above our quality thresholds do not have a major effect upon the mismatch rates. Upon running the statistical test it was found that when the two cerebellar samples were compared to each other, there was no significant difference in mismatch rates, but when either of them were compared to the frontal cortex, in both comparisons the cerebellar tissues had a higher mismatch rate (**Fig.3.3b**). Similar to the cerebellum, the cerebellar hemisphere also had many more genes with a significantly higher mismatch rate relative to the frontal cortex (**Fig.3.3c**). However, when the cerebellum and the cerebellar hemisphere were compared in this fashion, very few genes had differential rates of mismatch (**Fig.3.3d**). Based on this, we can conclude that at the very least RNA integrity is not the driving factor for this difference in mismatch rates.

We also considered the possibility that high mismatch rates at low numbers of bases could be skewing the data. This is a problem because of very lowly expressed genes. If a gene has a very low expression level, then it is possible that all the reads corresponding to that gene all come from a very small subset of cells that may for one reason or another be of lower quality. This could lead to a situation where several mismatches are present in genes for which a low number of bases are recorded due to low expression. If one tissue has a lower average expression level overall, then samples originating from that tissue may be inflated for these highly mismatched low base genes, and this could artificially inflate the mismatch rate and skew the overall rate for the tissue. Based on our previous analyses into differential patterns of gene expression across cerebellar and cerebral tissues, this should not be the case as there was no log fold-change skew when looking at either the differentially expressed genes or all genes across the GTEx RNA-seq dataset. Nevertheless, there was still a possibility that this was driving the difference between the tissues, especially as we know there are marked differences in expression between the frontal

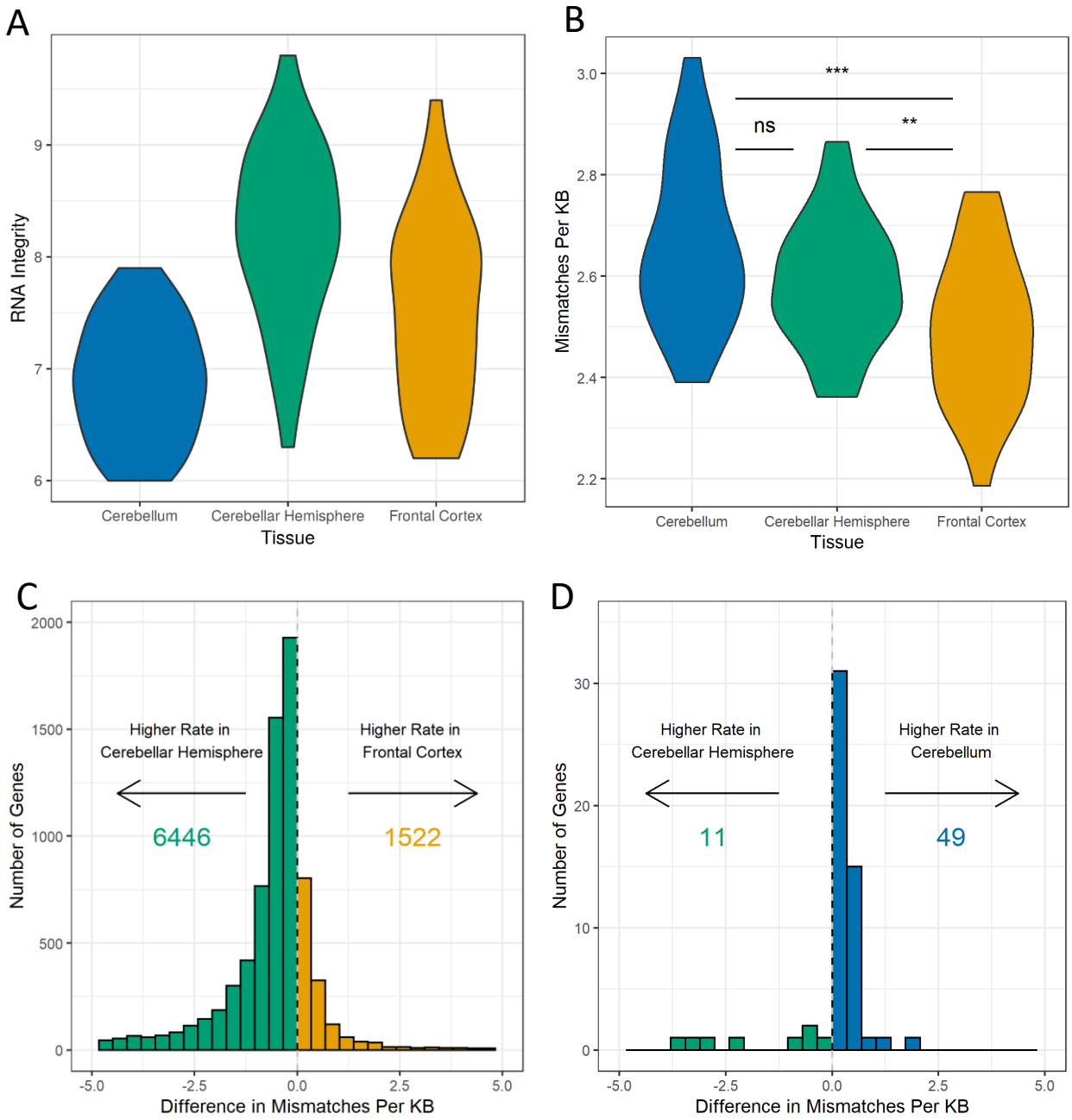


Fig.3.3a) Violin plot showing the spread of RNA integrity values for quality filtered RNA-seq samples from the cerebellum, cerebellar hemisphere and frontal cortex.

b) Violin plot showing the spread of mismatches per kilobase (KB) determined using pipeline V.1 for RNA-seq samples from the cerebellum, cerebellar hemisphere and the frontal cortex. Significance determined by t-test.

c-d) Histogram showing number of genes with differential rates of mismatch between c) the cerebellum and frontal cortex and d) the cerebellar hemisphere and frontal cortex RNA-seq samples plotted against differences in mismatches per kilobase, determined using pipeline V.1, between the two tissues. Genes with differential rates of mismatch were identified using linear model fitting.

ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

cortex and the cerebellum e.g. for mitochondrial genes. Therefore, after filtering out genes with very low bases recorded and no mismatches from the dataset, we plotted log-bases depending on whether they came from the cerebellum, the cerebellar hemisphere, or the frontal cortex (**Fig.3.4a**). The resulting graph showed that at low base levels it tended to be genes from the frontal cortex that had high mismatch rates. Therefore, if this was a major driving factor in determining the overall average number of mismatches per base for the tissue, we would expect to see the frontal cortex come out with a higher mismatch rate. To corroborate this finding, we then repeated the tissue comparison t-test on the mismatch dataset subsetted to included only genes which had over 1000 bases recorded. When this was done, the t-test still determined the difference in mismatch rates to be significant and the cerebellar tissues still had many more genes with a higher rate of mismatch relative to the frontal cortex (**Fig.3.4b,c,d**). We can therefore conclude that even if this effect could skew mismatch rates, it does not account for the higher overall rate of mismatch we see in the cerebellum relative to the frontal cortex.

Another potential issue when assessing mismatch rates from pipeline V.1 could be that the statistical models we are using to distinguish significant differences between the tissue wide and gene level mismatch rates are ill-suited to this type of data, so the results are erroneous. In order to test the reliability of the statistics employed, read simulation tests were undertaken. This involved simulating reads based on read counts present in real samples from both the cerebellum and the frontal cortex, but with a uniform error model – i.e. the same error rate in each tissue. As there are differences in expression levels between these two tissues, this simulation allows investigation into the effects of expression levels on mismatch rate, and whether expression level affects our statistical tests. If expression level does not affect the mismatch rate, then we would expect to see very little difference in overall rate between the tissues and see very few genes that show significant differential mismatch rates. In this way, the simulations can act as a negative control for our experiment. After uniform error model read simulation, and mapping, the resulting samples were put through pipeline V.1 and the relevant downstream statistical analyses performed. The results from the simulation showed that the t-tests assessing differences in mismatch rates between tissues did give a significant result, although the average rates were a lot closer together than the average rates for the real data (**Fig.3.5a**). This is likely due to the

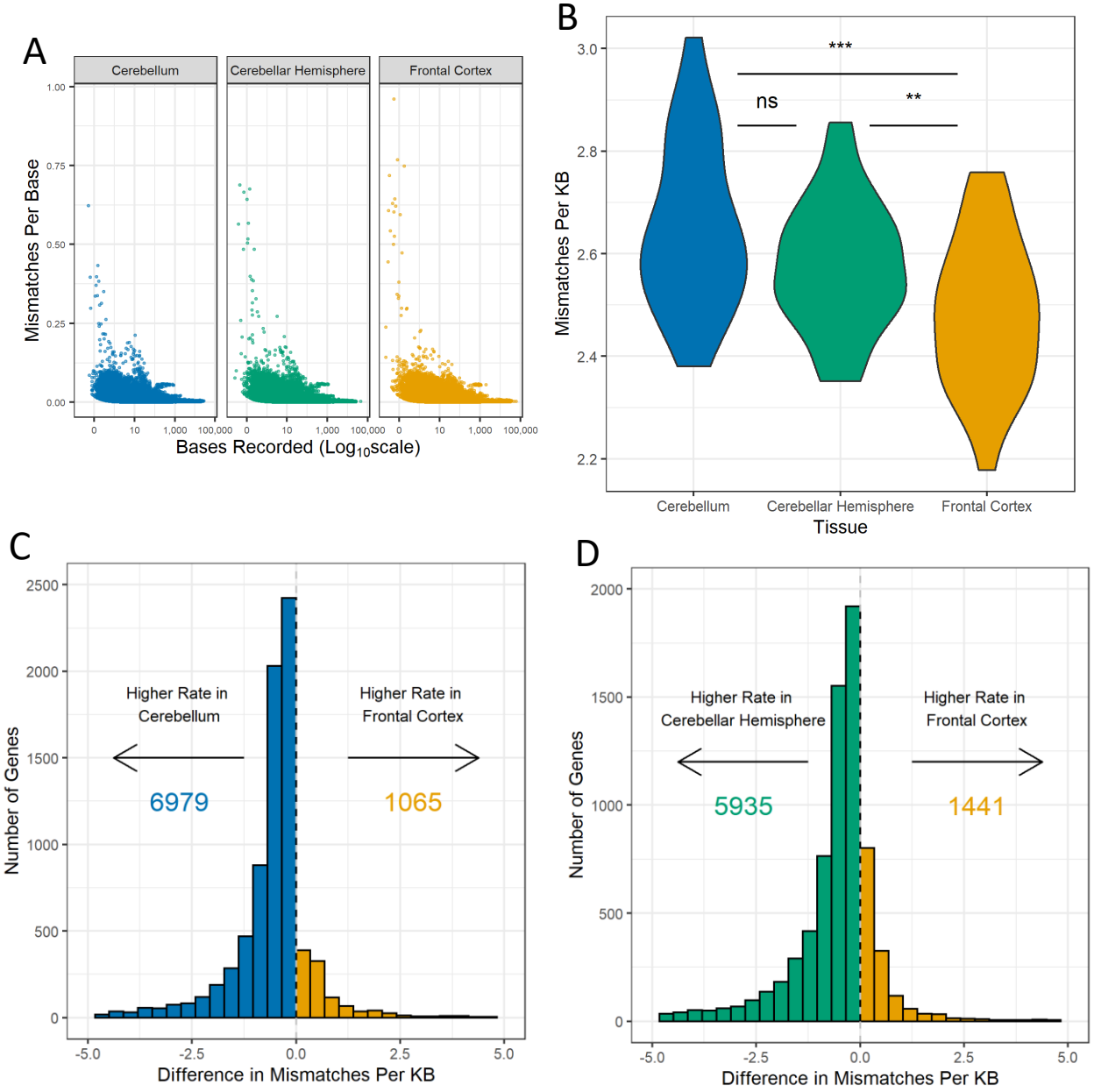


Fig.3.4a) Scatterplots showing for each gene in our analysis mismatches per base (KB), determined using pipeline V.1, plotted against bases recorded by the pipeline. Plots are split and colour coded according to whether the data comes from the cerebellum, cerebellar hemisphere or the frontal cortex.

b) Violin plot showing the spread of mismatches per kilobase (KB), determined using pipeline V.1, for genes with more than 1000 bases recorded in RNA-seq samples from the cerebellum, cerebellar hemisphere and the frontal cortex. Significance determined by t-test.

c-d) Histogram showing number of genes that have differential rates of mismatch between c) the cerebellum and frontal cortex and d) the cerebellar hemisphere and frontal cortex RNA-seq samples and have more than 1000 bases recorded plotted against differences in mismatches per kilobase, determined using pipeline V.1, between the two tissues. Genes with differential rates of mismatch were identified using linear model fitting.

ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

fact that a uniform error model produces a very low variance for the tissue wide mutation rates. However, when linear models were used to identify genes that were differentially mismatched between the simulated tissues, very few genes were called as having significantly different mismatch rates (**Fig.3.5b**). This indicates that perhaps the per-gene model is a better indicator of differential rates of damage between the tissues, although the range of the difference in the tissue wide mutation rate should also be taken into account.

Whilst investigating other potential confounding effects we came across the phenomenon of RNA-editing, and understood how that could have serious effects on mismatch rates and our results, especially if one tissue had a higher rate of RNA editing. The most common form of RNA-editing is A→I, picked up as A→ G by the sequencer. Therefore as rough way of investigating the effect of RNA editing on our results in the absence of a defined RNA editing filter, we removed all A→G transitions from our total set of mismatches for all samples and repeated our statistical tests. The t-test assessing the mismatch rates between the cerebellum and the frontal cortex retained significance but when the cerebellar hemisphere and the frontal cortex were compared, the result was non-significant, as was the cerebellum-cerebellar hemisphere comparison (**Fig.3.6a**). The per gene linear model tests showed slightly different results. When each of the cerebellar tissues were compared to the frontal cortex, fewer genes had a higher mismatch rate in the cerebellum, although the skew towards more genes with a higher mismatch rate in the cerebellum remained (**Fig.3.6b,c**). The number of genes with a higher rate in the frontal cortex showed a very small increase, although this is likely negligible. This indicates that RNA editing is potentially a major confounding effect when it comes to our mismatch analyses, and prompted the inclusion of the RNA-editing checker and logger in our script as part of pipeline V.2.

3.2.5 Pipeline V.2 results show lower mismatch rates across all tissues and reduced or loss of significance in comparisons of average mismatch rate across tissues

We ran a slightly expanded set of quality filtered GTEx RNA-seq samples analysed in pipeline V.1 through both pipeline V.2 total gene and pipeline V.2 exons. The statistical tests were performed in a similar fashion except prior to analysis a stricter filter was placed on the data, so that only genes with more than 500 bases recorded, corresponding to more than 5 reads, and at least one mismatch were put through to subsequent analyses. This was done

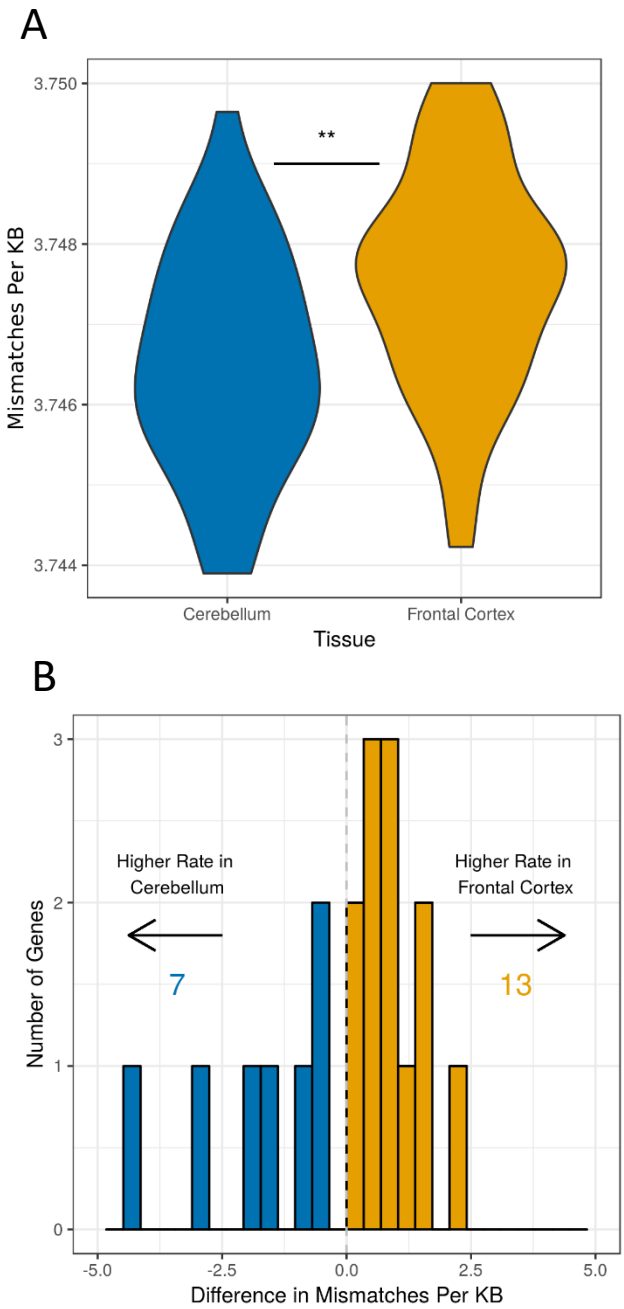


Fig.3.5a) Violin plot showing the spread of mismatches per kilobase (KB), determined using pipeline V.1, for RNA-seq samples simulated from cerebellum and frontal cortex RNA-seq data. Simulation was performed using a uniform error model. Significance determined by t-test.

b) Histogram showing number of genes with differential rates of mismatch between RNA-seq samples simulated from cerebellum and frontal cortex RNA-seq data plotted against differences in mismatches per kilobase, determined using pipeline V.1 , between the two tissues. Simulation was performed using a uniform error model. Genes with differential rates of mismatch were identified using linear model fitting.

p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

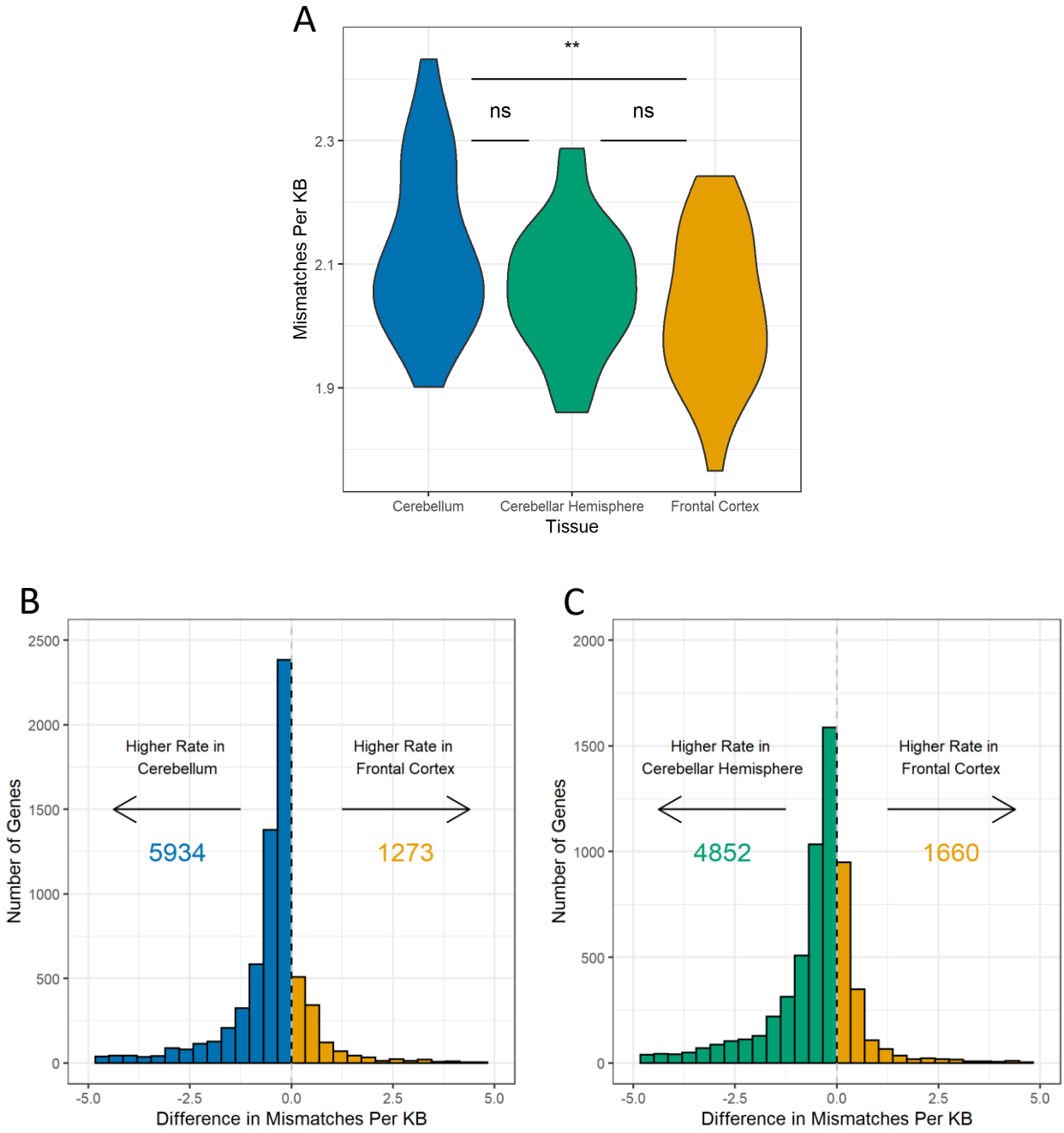


Fig.3.6 a) Violin plot showing the spread of mismatches per kilobase (KB) with A→G mismatches removed, determined using pipeline V.1, for RNA-seq samples from the cerebellum, cerebellar hemisphere and the frontal cortex. Significance determined by t-test.

b-c) Histogram showing number of genes that have differential rates of mismatch between b) the cerebellum and frontal cortex and c) cerebellar hemisphere and frontal cortex RNA-seq samples when A→G mismatches are removed plotted against differences in mismatches per kilobase, determined using pipeline V.1, between the two tissues. Genes with differential rates of mismatch were identified using linear model fitting.

ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

due to identifying the potential problem of high number of mismatches at low numbers of bases recorded, discussed in the previous section. Tissue comparisons of mismatch rates for pipeline V.2 total gene results using t-tests showed that whilst significance between the rates of the cerebellum and the frontal cortex was retained, there was no significant difference between the cerebellar hemisphere and frontal cortex rates (**Fig.3.7a**). Also of note was that the sample average mismatch rates for each tissue were much reduced compared the pipeline V.1 average rates. The cerebellum had its average rate of mismatch reduced from 2.66 to 1.98 per kb, the cerebellar hemisphere from 2.59 to 1.92 per kb and the frontal cortex from 2.49 to 1.87 per kb. These reduced rates for pipeline V.2 are even lower than the average mismatch rates for pipeline V.1 with A→Gs removed, indicating that the reduction in rate is not just due to the removal of RNA-editing events, but is likely also affected by the indel checker implemented in pipeline V.2. When the results for pipeline V.2 exons were analysed, there was no significant difference between any of the tissues in terms of average mismatch rate per sample (**Fig.3.7b**).

3.2.6 Specific Base changes are enriched between the cerebellum, cerebellar hemisphere and frontal cortex

Having obtained reliable base change data in pipeline V.2, we wanted to analyse the general mutational spectra in the brain. Looking at the changes occurring in the results from pipeline V.2 total gene (**Fig.3.8a**), the most common mutation is A→G. A→Gs do not result from the common forms of damage discussed in the introduction to this thesis, and so perhaps represent unfiltered RNA-editing events. A→Gs are closely followed by T→Cs in terms of rate, another mutation that lacks a clearly defined cause. It should be noted however, that this is the reverse complement of the A→G mutation associated with RNA editing, potentially implying that although we have corrected for strandedness when recording base changes, some changes are not being reverse complemented for an unknown reason. Three sets of mutations are all approximately tied for having the third greatest rate: G→A, T→A and T→G. G→As are a marker of unrepaired deamination. The initial C→U caused by deamination, if not repaired, can lead to a G→A on the opposing strand through replication of the mutation. However, it must be remembered that brain cells are non-cycling, so if this is the cause of these mutations it must have occurred during development or alternatively, another mechanism is responsible for the abundance of this

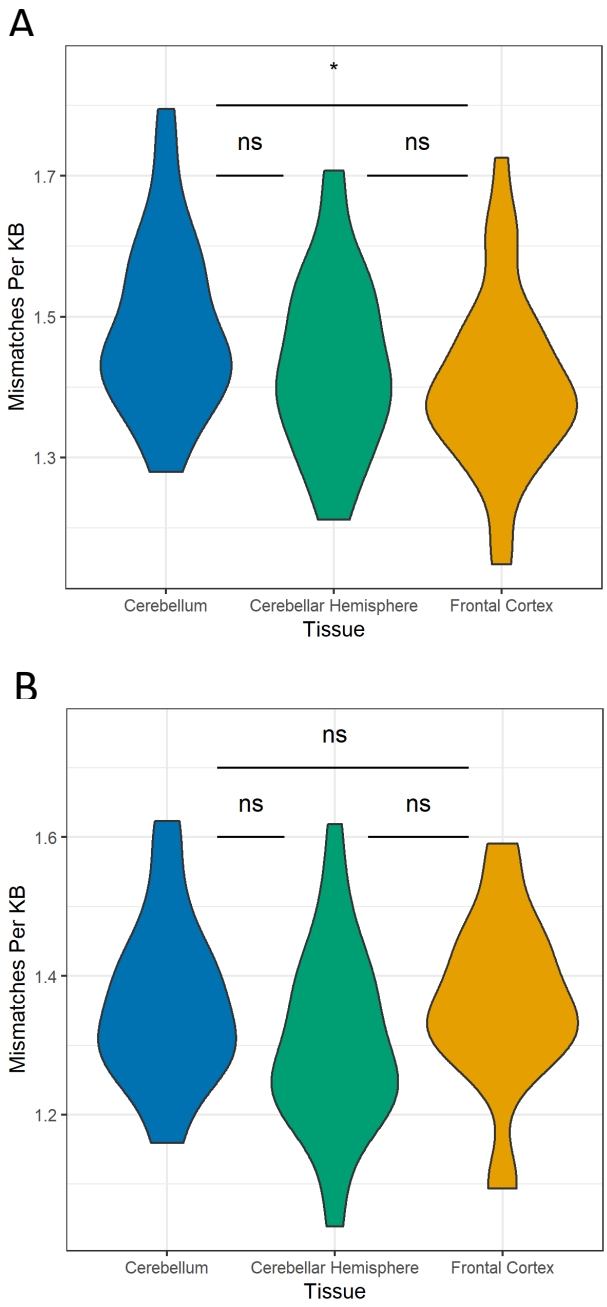
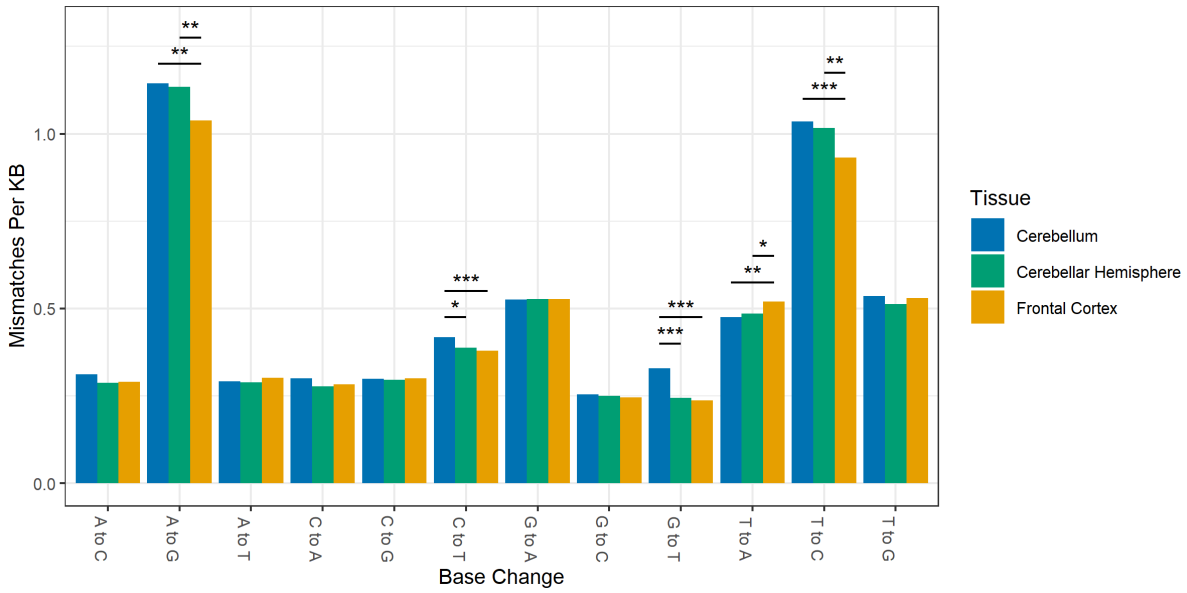


Fig.3.7 a-b) Violin plot showing the spread of mismatches per kilobase (KB), a) determined using pipeline V.2 total genes and b) determined using pipeline V.2 exons, for RNA-seq samples from the cerebellum, cerebellar hemisphere and the frontal cortex. Significance determined by t-test.

ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

A



B

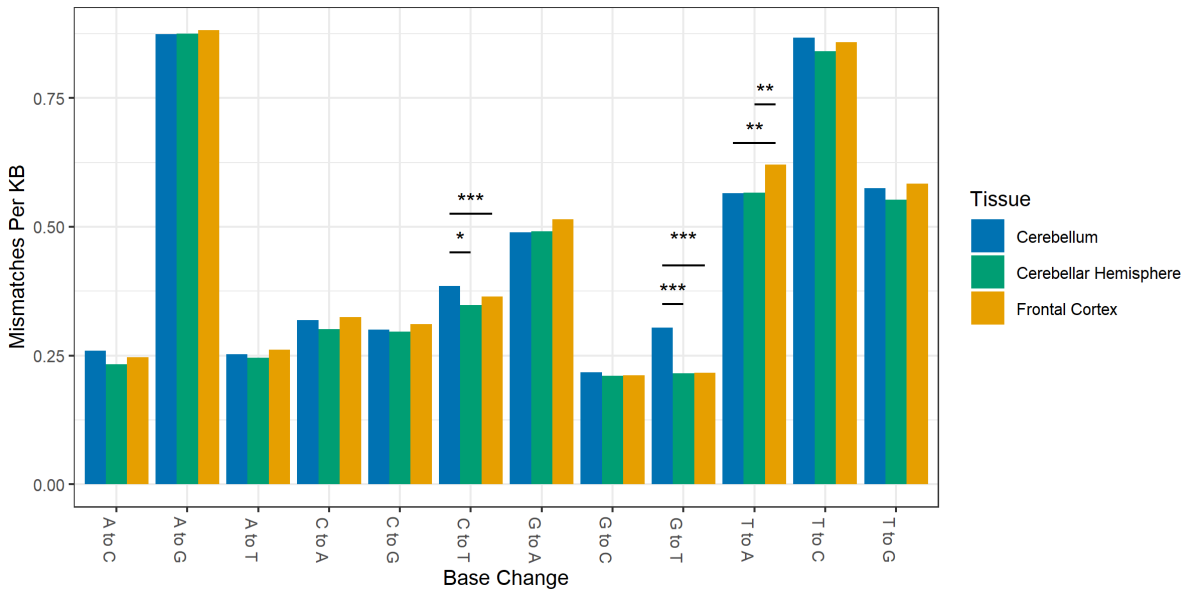


Fig.3.8 a-b) Bar chart showing average mismatches per kilobase (KB), a) determined by pipeline V.2 total genes and b) determined by pipeline V.2 exons, for each base change in the RNA-seq samples from the cerebellum, cerebellar hemisphere and the frontal cortex. Significance determined by t-test. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

transition, as will apply to mutations similarly known to arise from incorporation during replication. T→A is again not known to be caused by any common forms of damage, and the fact that no alternative explanation exists as with the A→Gs and T→Cs makes this intriguing. On the other hand, T→Gs are indicative of oxidative mutations, as free 8-oxoG can base pair with adenine during replication resulting in A:T → C:G mutations after subsequent replication. The final base change that has a rate higher than the other is C→T. C→T is indicative of alkylation mutations stabilised by replication, as the initial O_6-methyl-guanine can base pair with T, and both transient and stabilised deamination, as the C→U formed by this type of chemical attack will be picked up as a C to T by the sequencer, and downstream replication can form a stable C:G → T:A transition. C→Ts are also associated with transcriptional damage, as laid out by Lodato et al., (2015). The general pattern described here also holds true for the base changes recorded for pipeline V.2 exons, however, the A→G and T→C mutations have a lower overall rate compared to the results for pipeline V.2 total genes (**Fig.3.8b**).

We also wanted to compare these mutations across the different tissues. Statistical comparison of the rate of these base changes across the different tissues was carried out by t-tests, which revealed that in the pipeline V.2 total genes results, A→G and T→C mutations have a significantly higher rate in the cerebellum and the cerebellar hemisphere relative to the frontal cortex (**Fig.3.8a**). C→T and G→T mutations had a significantly higher rate in the cerebellum compared to both the cerebellar hemisphere and the frontal cortex, and the frontal cortex had a greater rate for T→A changes relative to both the cerebellum and cerebellar hemisphere. The results for pipeline V.2 exons were the same aside from significance between tissues for the rate of A→G and T→C mutations was lost (**Fig.3.8b**).

3.2.7 Pipeline V.2 recapitulates the patterns for genes with differential mismatch rates and category enrichment observed in pipeline V.1

Although the tissue comparisons of average mismatch rate per sample showed reduced significance for pipeline V.2, our simulations indicated that differences in mismatch rates between genes was a more effective way of looking at differential mutation rates between the tissues of interest. Therefore, we repeated the linear model fitting to identify genes with significantly different mismatch rates between the three tissues. In the cerebellum-frontal cortex comparison, slightly fewer genes had differential rates of mismatch between the two

tissues, and the number of genes that had a higher mismatch rate in the cerebellum and the frontal cortex both fell indicating that the genes for which significance was lost were not specific to one tissue (**Fig.3.9a**). The cerebellar skew observed in pipeline V.1 remained, with the cerebellum having 7179 genes with a higher mismatch rate than the frontal cortex and compared to 882 in the opposite direction. When the cerebellar hemisphere and the frontal cortex were compared, although the number of genes with a significantly higher mismatch rate in the cerebellar hemisphere decreased, the number of genes with a relatively greater rate in the frontal cortex increased by a slight amount (**Fig.3.9b**). However, again the skew for genes with higher mismatch rates was heavily toward the cerebellar tissue, with 5380 and 1841 genes with a higher mismatch rate in the cerebellar hemisphere and frontal cortex respectively. When this was repeated for pipeline V.2 exons results, there was dramatic decrease in the number of genes with differentially mismatch rates across the board. The 7179/882 genes differentially mismatched between the cerebellum and frontal cortex fell to just 83 in the cerebellum and 36 in the frontal cortex (**Fig.3.9c**). Although the numbers fell for this comparison, again, more genes had a higher mismatch rate in the cerebellum. This was not true of the exon results for the cerebellar hemisphere – frontal cortex comparison which had only 59 genes with a higher rate in the hemisphere and 126 for the frontal cortex (**Fig.3.9d**). This represents an altered skew towards more genes having a higher mismatch rate in the frontal cortex compared to the cerebellar hemisphere when restricting our analysis to exons. Across both the pipeline V.2 total gene and pipeline V.2 exons results, no or a negligibly small number of genes had significantly differential mismatch rates between the cerebellum and the cerebellar hemisphere. Next we wanted to investigate whether the patterns of gene enrichment and depletion observed for pipeline V.1 were present in the genes determined to have differential rates of mismatch by both versions of pipeline V.2. For these pipeline V.2 enrichment tests, the cerebellum and the cerebellar hemisphere were treated as the same tissue when performing linear model fitting to identify genes with differential mismatch rates, and it was the gene set resulting from this grouped linear model analysis that were tested for enrichment of categories. The genes with a higher rate of mismatch in the cerebellar tissues for pipeline V.2 total gene were strongly enriched for genes downregulated in the cerebellum with an OR of 3.4 and strongly depleted for those upregulated, OR=0.25 (**Fig.3.10a**). This patterns of enrichment and depletion is the same as that observed for pipeline V.1 but the over/underrepresentation as given by the OR is much

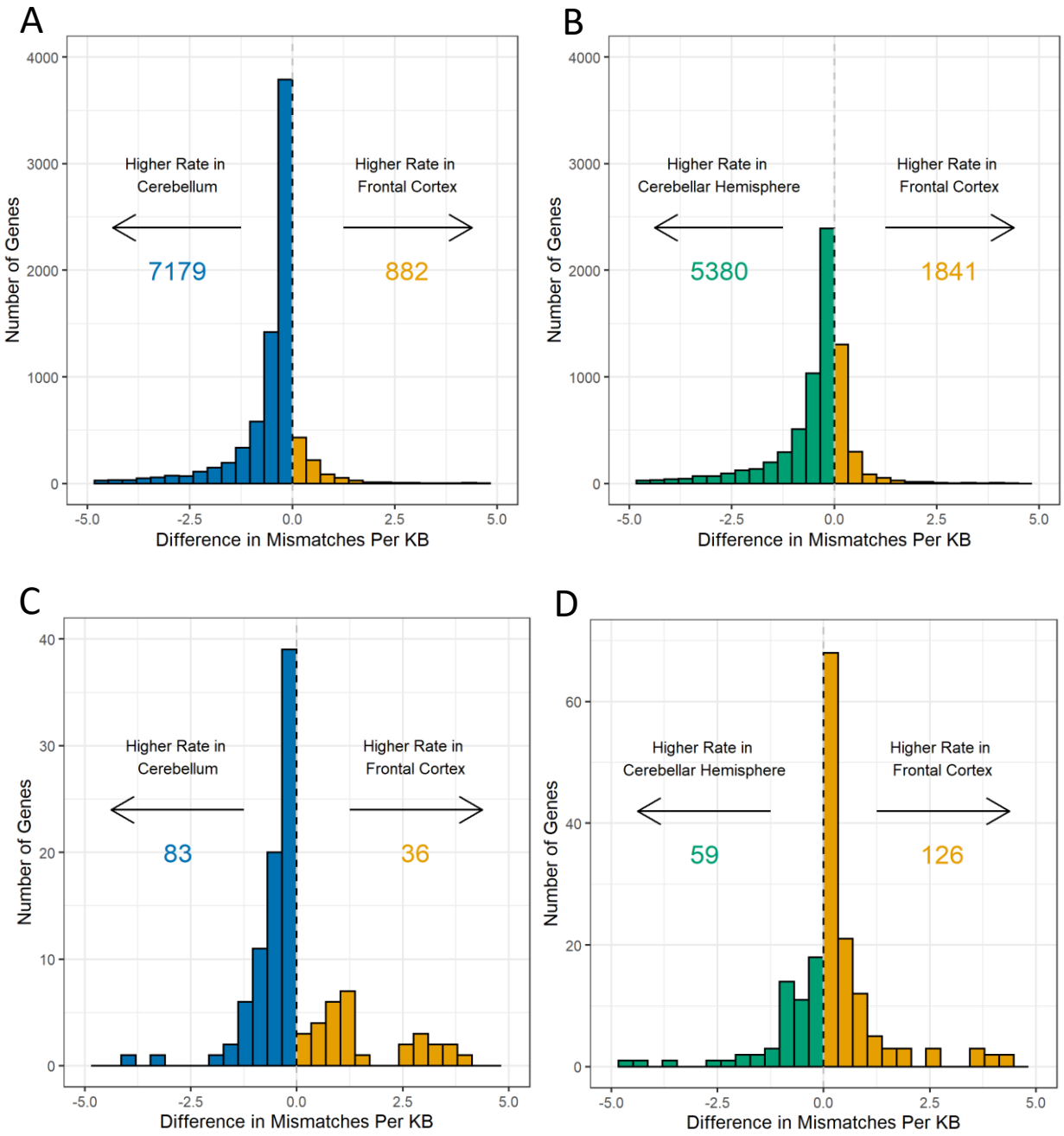
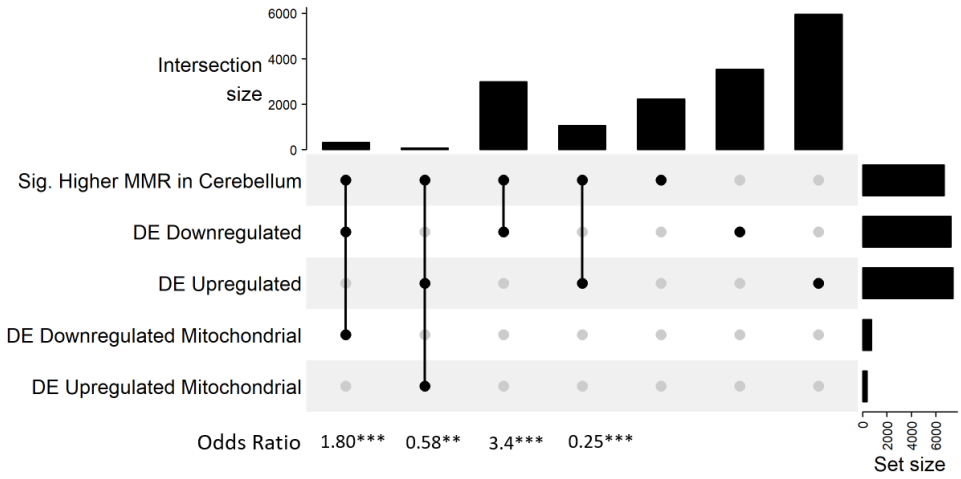


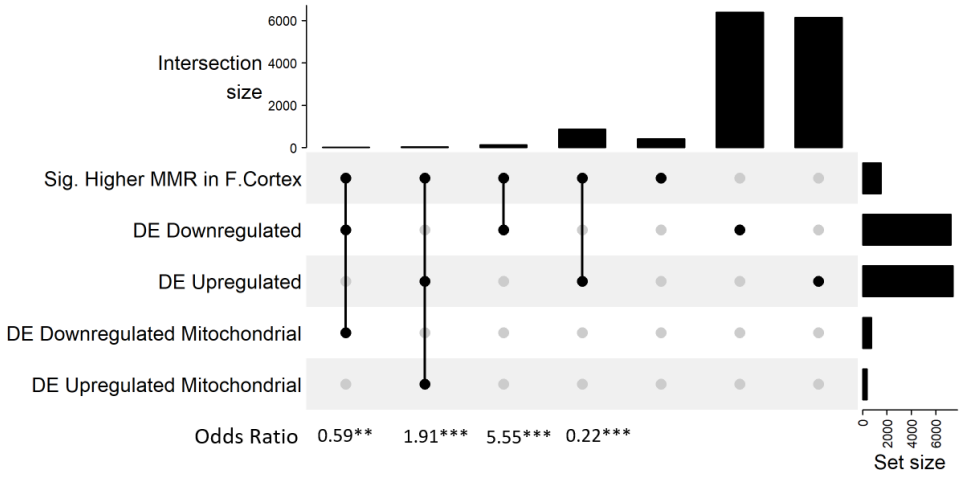
Fig.3.9a-d) Histogram showing number of genes that have differential rates of mismatch between b) the cerebellum and frontal cortex RNA-seq samples for pipeline V.2 total gene results, b) cerebellar hemisphere and frontal cortex RNA-seq samples for pipeline V.2 total gene results, c) cerebellum and frontal cortex RNA-seq samples for pipeline V.2 exon results and d) cerebellar hemisphere and frontal cortex RNA-seq samples for pipeline V.2 exon results, plotted against differences in mismatches per kilobase between the two tissues. Genes with differential rates of mismatch were identified using linear model fitting.

stronger for these results, which were 2.1 for the genes downregulated in the cerebellum relative to the frontal cortex and 0.44 for the genes relatively upregulated. For differentially expressed downregulated and upregulated mitochondrial genes there is a significant enrichment slighter lower than that observed for pipeline V.1, OR=1.8 compared to 1.9, and a now significant ,stronger depletion with an OR of 0.53 compared to the previous, non-significant 0.93. Again this reverse pattern is captured in the results for the genes with a higher mismatch rate in the frontal cortex. Those genes that are downregulated in the cerebellum and therefore relatively upregulated in the frontal cortex are depleted to a greater extent than they were for pipeline V.1, with an OR of 0.22 as compared to 0.4 (**Fig.3.10b**). This stronger effect also applied to the enrichment of those genes upregulated in the cerebellum/downregulated in the frontal cortex which were significantly enriched with an OR of 5.55 compared to pipeline V.1's 2.76. The same occurred with the mitochondrial genes, those showing differential downregulation and those differentially upregulated were significantly enriched with a 1.91 OR and significantly depleted with an OR of 0.59 respectively. Both of these results showed stronger ORs in their respective directions than the pipeline V.1 results. Repeating this analysis with mismatch data restricted to exons as per pipeline V.2 exons did change the results somewhat. Although the same general pattern seen across the genes showing a greater rate of mismatch in each tissue remained, there were some exceptions. First of all, as the sets of genes showing significantly differential mismatch rates were smaller, the overlaps with the categories was smaller in terms of the raw number of genes (**Fig.3.10c,d**). Secondly, although the pattern remained, most notably for those upregulated and downregulated cerebellar genes, the effect in terms of OR for most enrichments and depletions was slightly reduced. The enrichment of downregulated genes amongst the cerebellar mismatched set showed an OR of 1.73 relative to the 3.4 observed for pipeline V.2 total genes, and cerebellar upregulated genes showed a weaker depletion with the OR rising from 0.25 to 0.55 (**Fig.3.10c**). The depletion of genes downregulated in the cerebellum within the frontal cortex mismatched genes was also weakened, the OR rising from 0.22 to 0.41, and the overrepresentation of the cerebellar upregulated genes similarly reduced from 5.55 to 3.47 (**Fig.3.10d**). Some categories even had significance abrogated, as was the case for the enrichment of mitochondrial genes differentially downregulated in the cerebellum amongst genes showing a higher mismatch rate in the cerebellum, which had its OR reduced from 1.8 to a non-

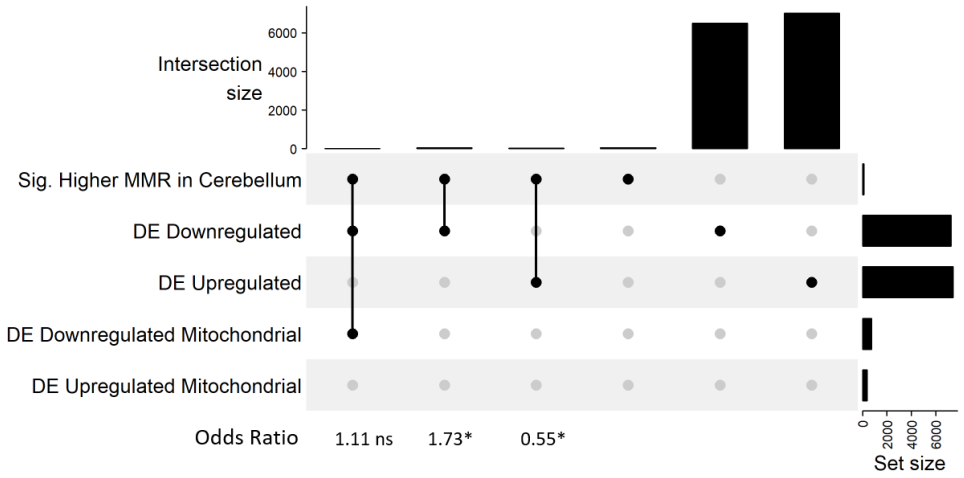
A


Intersection size
6000
4000
2000
0
Sig. Higher MMR in Cerebellum
DE Downregulated
DE Upregulated
DE Downregulated Mitochondrial
DE Upregulated Mitochondrial
Odds Ratio
1.80***
0.58**
3.4***
0.25***
0
2000
4000
6000
Set size

B


Intersection size
6000
4000
2000
0
Sig. Higher MMR in F.Cortex
DE Downregulated
DE Upregulated
DE Downregulated Mitochondrial
DE Upregulated Mitochondrial
Odds Ratio
0.59**
1.91***
5.55***
0.22***
0
2000
4000
6000
Set size

C



D

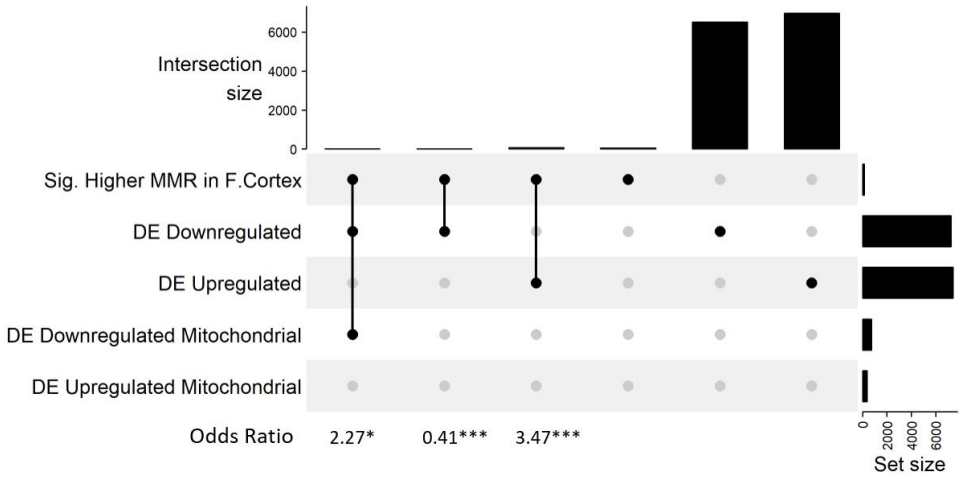
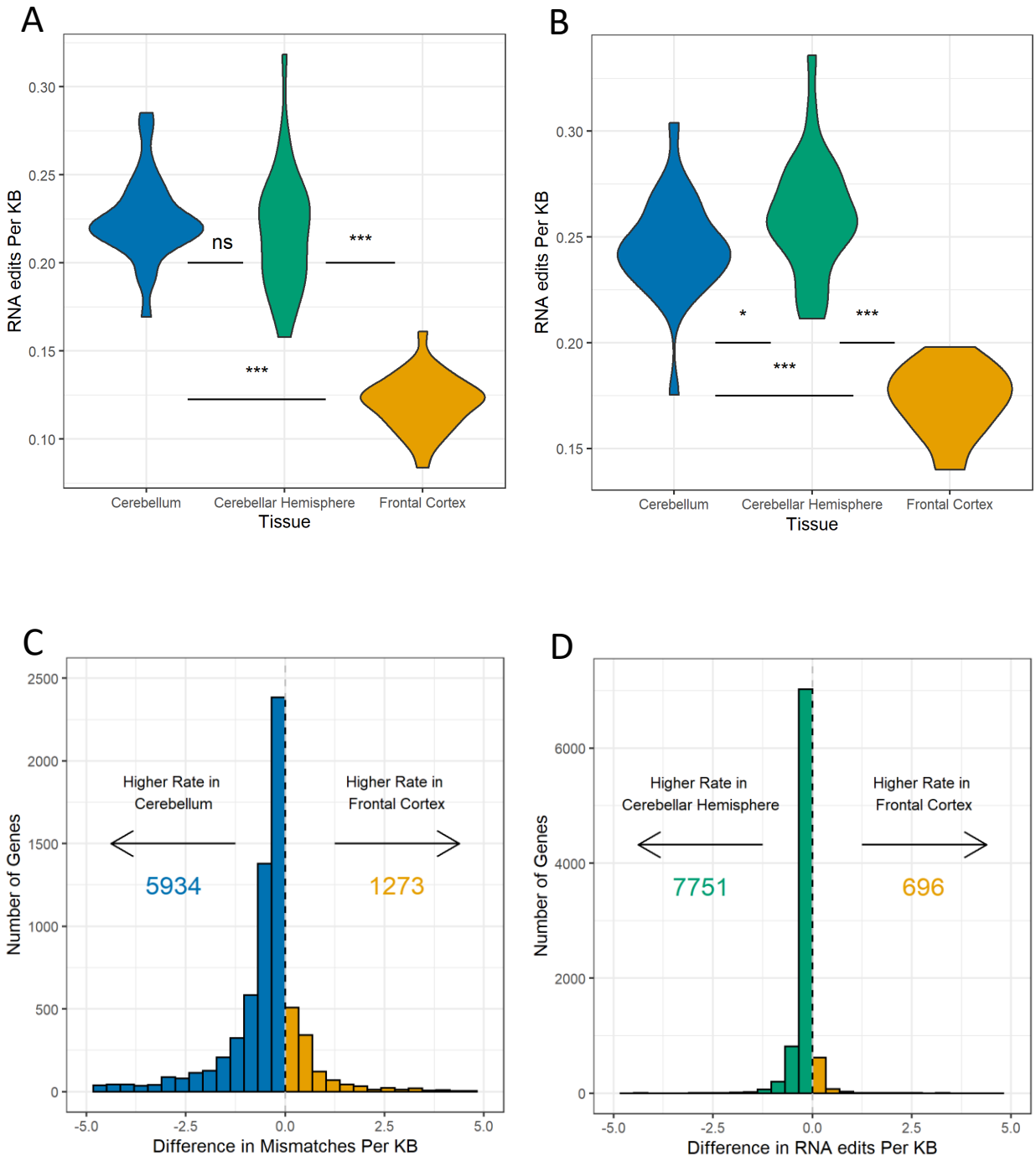


Fig.3.10a-d) Upset plots of enrichment of specific gene categories amongst a) genes with a significantly higher mismatch rate in the cerebellum compared to the frontal cortex for pipeline V.2 total genes results b) genes with a significantly higher mismatch rate in the frontal cortex relative to the cerebellum for pipeline V.2 total gene results, c) genes with a significantly higher mismatch rate in the cerebellum compared to the frontal cortex for pipeline V.2 exon results and d) genes with a significantly higher mismatch rate in the frontal cortex compared to the cerebellum for pipeline V.2 exon results. Significance and Odds Ratios (OR) determined by Fisher's exact test. Sig. = Significantly, DE = Differentially, ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

significant 1.11. Mitochondrial genes differentially upregulated in the cerebellum were not present at all in genes with a higher mismatch rate in either the cerebellum or the frontal cortex (**Fig.3.10c,d**). Finally, there was one result that did not fit the previously observed pattern. Mitochondrial genes differentially downregulated in the cerebellum relative to the frontal cortex were significantly enriched in genes with a significantly higher mismatch rate in the frontal cortex, with a strong OR of 2.27 (**Fig.3.10d**). However, in the enrichment tests for pipeline V.2 total gene this category was significantly underrepresented amongst this set of genes, as indicated by an OR of 0.59 (**Fig.3.10b**). These results show the persistence of patterns of enrichment and depletion for several gene categories amongst subsets of genes with differential mismatch rates across multiple versions of the pipeline, makes these findings more robust.

3.2.8 The cerebellum has a greater rate of RNA editing than the frontal cortex

As we had logged when a potential SNP was found within REDIportal and treated as an RNA editing event, comparisons of the rate of RNA editing events were carried out across the three different tissues in our analysis. This was done in a fashion similar to the mismatch analysis. An RNA editing rate (average RNA editing events per base) for each sample was calculated by summing all the RNA editing events and dividing this by the sum of all the recorded bases on a per sample basis. The same was done on a per gene basis to get gene specific RNA-editing rates for each sample. T-tests were then used to assess differences in rates between the selected tissues. The cerebellum and the cerebellar hemisphere both separately have a significantly higher average per sample rate of RNA editing than the frontal cortex, but there was no difference when these two cerebellar tissues were compared to each other (**Fig.3.11a**). A similar pattern was seen again when performing the analysis on the pipeline V.2 exons data, although the sample average RNA editing rates were greater across all tissues and the cerebellar hemisphere had a significantly larger rate than the cerebellum (**Fig.3.11b**). Linear model fitting to identify genes that had significantly different rates of RNA editing between the tissues revealed striking inter-tissue difference. In a fashion mimicking the results for genes with differential mismatch rates between tissues, vastly more genes had a higher rate of RNA editing in the cerebellum and the cerebellar hemisphere when these tissues were compared in turn to the frontal cortex for the pipeline V.2 total gene RNA editing data. The cerebellum – frontal cortex comparison


A
RNA edits Per KB
0.30
0.25
0.20
0.15
0.10
ns
***
***
Cerebellum
Cerebellar Hemisphere
Frontal Cortex
Tissue
B
RNA edits Per KB
0.30
0.25
0.20
0.15
*
***
***
Cerebellum
Cerebellar Hemisphere
Frontal Cortex
Tissue
C
Number of Genes
2500
2000
1500
1000
500
0
Higher Rate in Cerebellum
5934
Higher Rate in Frontal Cortex
1273
-5.0
-2.5
0.0
2.5
5.0
Difference in Mismatches Per KB
D
Number of Genes
6000
4000
2000
0
Higher Rate in Cerebellar Hemisphere
7751
Higher Rate in Frontal Cortex
696
-5.0
-2.5
0.0
2.5
5.0
Difference in RNA edits Per KB

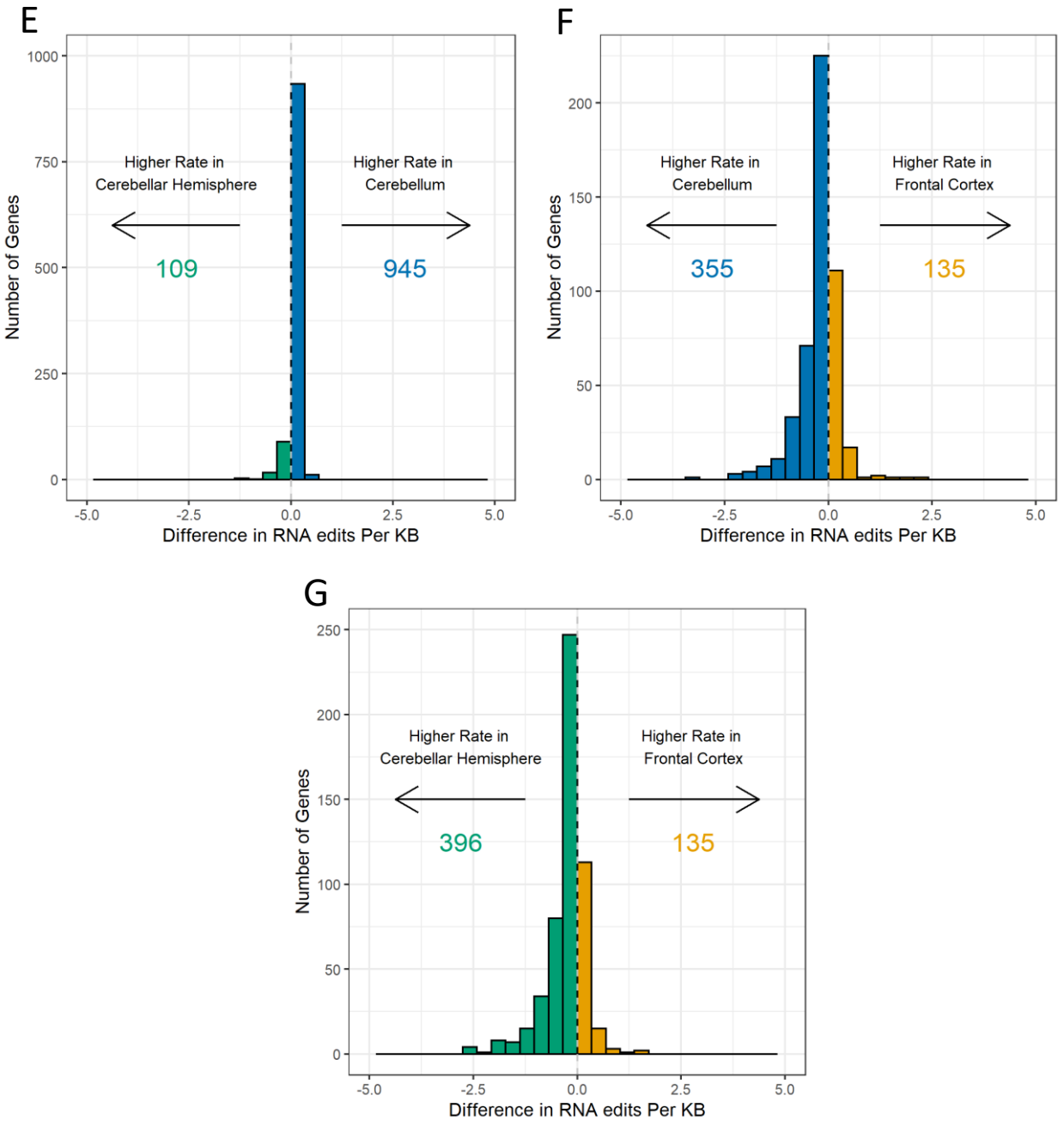


Fig.3.11a-b) Violin plots showing the spread of RNA editing events per kilobase (KB), determined from a) pipeline V.2 total gene results and b) pipeline V.2 exon results, for RNA-seq samples from the cerebellum, cerebellar hemisphere and the frontal cortex. Significance determined by t-test.

c-g) Histogram showing number of genes that have differential rates of RNA editing between c) the cerebellum and frontal cortex RNA-seq samples for pipeline V.2 total gene results, d) cerebellar hemisphere and frontal cortex RNA-seq samples for pipeline V.2 total gene results, e) cerebellum and cerebellar hemisphere RNA-seq samples for pipeline V.2 total gene results, f) cerebellum and frontal cortex RNA-seq samples for pipeline V.2 exon results and g) cerebellum and frontal cortex RNA-seq samples for pipeline V.2 total gene results, plotted against differences in RNA editing events per kilobase between the two tissues. Genes with differential rates of RNA editing were identified using linear model fitting.

ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

showed a 8137 to 727 skew in favour of the cerebellum (**Fig.3.11c**) and the cerebellar hemisphere – frontal cortex tests similarly revealed 7751 genes with a higher RNA editing rate in the cerebellar hemisphere and only 696 with a higher rate in the frontal cortex (**Fig.3.11d**). When tests were run between the cerebellum and the cerebellar hemisphere, the number of genes with a relatively different rate of RNA editing fell massively compared to the number of genes with a higher rate in each of the tissues in question when they were compared to the frontal cortex. 109 genes were differentially RNA edited to a higher degree in the cerebellar hemisphere, whereas 945 genes were more edited in the cerebellum, representing a skew towards the more genes highly RNA edited in the cerebellum in this comparison (**Fig.3.11e**). Again, in a fashion reminiscent of the per gene mismatch rate test results for the exons, when the per gene RNA editing analysis was restricted to exons, the overall number of genes showing a differential rate of RNA editing was drastically reduced. When the cerebellum and the cerebellar hemisphere were compared to the frontal cortex for this analysis, only 355 and 396 genes underwent more RNA editing in the cerebellar tissue, but in both cases the number of genes with a higher rate of RNA editing in the frontal cortex was 135, indicating the phenomenon of more genes with a greater RNA editing rate in the cerebellar tissues was still present (**Fig.3.11f,g**). The cerebellum – cerebellar hemisphere per genes RNA editing rate comparison for pipeline V.2 exons returned no genes as having a differential rate of RNA editing between the two tissues. We then analysed the sets of genes that had differential RNA editing rates between the tissues for biologically relevant GO categories using GO-seq, but nothing of immediate interest was found. Nevertheless, these results clearly show that the cerebellar tissues have a greater rate of RNA editing within both exonic and intronic portions of the RNA relative to the frontal cortex.

3.2.9 Summary of different mismatches pipeline runs and versions

In order to summarise and view the differences we see between different pipeline runs we have provided a summary table of the different runs and versions of the pipeline (**Fig.3.12**). One of the main differences is that the average mismatch rate across all the tissues falls between pipeline V.1 and pipeline V.2 total genes, then falls again in pipeline V.2 exons. In parallel with this drop in rate, fewer genes are called as differentially mismatched in V.2 whole gene compared to V.1. The drop is only slight however when compared to the drop in

Pipeline	SNP source	MMR Tests	Significant Base Changes	Enrichment Tests	RNA editing Tests
Pipeline V.1	GTEx VCF	**CER avg. MMR**: 2.66 **CHE avg. MMR**: 2.59 **FCX avg. MMR**: 2.49 **CER vs FCX***** 7463 | 1162 **CHE vs FCX*** 6446 | 1522 **CER vs CHE** ns 49 | 11	N/A	**CER MM sig. enriched:** DE down, DE down Mt **CER MM sig. depleted:** DE up **FCX MM sig. enriched:** DE up, DE up Mt **FCX MM sig. depleted:** DE down	N/A
Pipeline V.2 total gene	GATK germline variant calling	**CER avg. MMR**: 1.98 **CHE avg. MMR**: 1.92 **FCX avg. MMR**: 1.87 **CER vs FCX*** 7179 | 882 **CHE vs FCX** ns 5380 | 1841 **CER vs CHE** ns N/A	**CER vs FCX**: G to T, C to T, T to A, T to C, A to G **CHE vs FCX**: T to A, T to C, A to G **CER vs CHE**: G to T, C to T	**CER MM sig. enriched:** DE down, DE down Mt **CER MM sig. depleted:** DE up, DE up Mt **FCX MM sig. enriched:** DE up, DE up Mt **FCX MM sig. depleted:** DE down, DE down Mt	**CER avg. RER**: 0.225 **CHE avg. RER**: 0.217 **FCX avg. RER**: 0.121 **CER vs FCX***** 8137 | 727 **CHE vs FCX***** 7751 | 696 **CER vs CHE** ns 945 | 109
Pipeline V.2 total exon	GATK germline variant calling	**CER avg. MMR**: 1.39 **CHE avg. MMR**: 1.30 **FCX avg. MMR**: 1.36 **CER vs FCX** ns 83 | 36 **CHE vs FCX** ns 59 | 126 **CER vs CHE** ns N/A	**CER vs FCX**: G to T, C to T, T to A **CHE vs FCX**: T to A, **CER vs CHE**: G to T, C to T	**CER MM sig. enriched:** DE down **CER MM sig. depleted:** DE up **FCX MM sig. enriched:** DE up, DE down Mt **FCX MM sig. depleted:** DE down	**CER avg. RER**: 0.245 **CHE avg. RER**: 0.260 **FCX avg. RER**: 0.173 **CER vs FCX***** 355 | 135 **CHE vs FCX***** 396 | 135 **CER vs CHE*** N/A

Fig.3.12 Summary table of the results from different versions of the mismatches pipeline. Across the whole table CER = cerebellum, CHE = cerebellar hemisphere and FCX = frontal cortex, avg. = average, sig. = significantly, MM = mismatched, MMR = mismatch rate, RER = RNA editing rate, DE = differentially expressed, Mt = mitochondrial. The average mismatch rates in the MMR Tests column were calculated across all the samples for a given tissue. The significant differences in MMR as shown in the MMR Tests column were determined by t-tests. The numbers shown below the t-test results in this column are the number of genes with a higher mismatch rate in each tissue when the average mismatch rates of genes were compared to each other using linear model fitting. The order of the numbers reflects the order in which the tissues are written above e.g. CER vs FCX written above 7463 | 1162 indicates 7463 genes had a higher mismatch rate in the cerebellum and 1162 genes had a higher mismatch rate in the frontal cortex when the per gene mismatch rates were compared across tissues. Significant base changes as seen in the column with this title were determined by t-tests. Enrichment and depletion as seen in the Enrichment Tests column was determined by Fisher's exact test. The average RNA editing rates in the RNA Editing Tests column were calculated across all the samples for a given tissue. The significant differences in RER as shown in the RNA Editing Tests column were determined by t-tests. The numbers shown below the t-test results in this column are the number of genes with a higher RNA editing rate in each tissue when the average RNA editing rates of genes were compared to each other using linear model fitting. The order of the numbers reflects the order in which the tissues are written above e.g. CER vs FCX written above 8137 | 727 indicates 8137 genes had a higher RNA editing rate in the cerebellum and 727 genes had a higher RNA editing rate in the frontal cortex when the per gene RNA editing rates were compared across tissues. ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

the number of differentially mismatched genes when comparing V.2 total gene and exons. Additionally, the cerebellar hemisphere vs frontal cortex comparison for pipeline V.2 exons shows the frontal cortex as having more genes with a relatively higher mismatch rate, which is a big change. There are no results for the base changes for pipeline V.1 because the base recording code was not built properly, but there are a few differences between V.2 total genes and exons. These are that in V.2 total genes, A→G and T→C base changes are called as having a differential rate between both cerebellar tissues and the frontal cortex, whereas in pipeline V.2 exons, this significance is lost. The enrichment tests do show some variation across pipeline versions but the general pattern is similar: genes downregulated in the tissue that the set of genes have a higher mismatch rate in are enriched amongst the higher mismatch rate genes, but the opposite is true for upregulated genes. The difference between pipeline V.1 and V.2 total gene is that mitochondrial genes upregulated in the cerebellum are depleted from genes with a higher rate of mismatch in the cerebellum, and mitochondrial genes upregulated in the frontal cortex/downregulated in the cerebellum show the same depletion within genes with a higher mismatch rate in the frontal cortex. The change that occurs when comparing V.2 total gene and V.2 exons is that the previously assessed categories of mitochondrial genes no longer show enrichment or depletion, except for mitochondrial genes upregulated in the frontal cortex being enriched amongst genes with a higher mismatch rate in the frontal cortex. Finally, there are the RNA editing results. RNA editing events were not recorded for pipeline V.1 because the filter was built into this version of the pipeline, but one noticeable difference in the RNA editing results for pipeline V.2 total genes and V.2 exons is that the rate of RNA editing slightly increases in pipeline V.2 exons across all three tissues. However, the number of genes that show differential levels of RNA editing across the tissues falls dramatically. For the comparisons between the frontal cortex and the cerebellar tissues, the number falls from the thousands to the low hundreds, but the general trend of the cerebellar tissues having more genes with significantly higher rate of RNA editing compared to the frontal cortex remains. The final difference is the tests to identify genes showing differential levels of RNA editing between the cerebellar tissues gave results for the pipeline V.2 total gene data, where more genes had a higher rate of editing in the cerebellum compared to the cerebellar hemisphere samples, but when this test was repeated for pipeline V.2 exons, no genes were called as having a differential rate of RNA editing.

3.2.10 The cerebellum has more RNA-seq variant calls than the frontal cortex

Because it is possible for germline variant callers to pick up mosaic mutations, we decided to analyse the variants called from our cerebellar and frontal cortex samples as part of the mismatches pipeline. However, before statistical tests were performed, we thought it possible that RNA-editing events were being called as SNPs, and as our previously discussed data showed that the cerebellum had a greater rate of RNA-editing than the frontal cortex, this difference could be driving any apparent increase in SNPs in the cerebellum relative to the frontal cortex if such a result was found. Therefore, prior to our analysis, all A → G calls removed. Subsequent t-tests on this A→G filtered data revealed that both the cerebellum and the cerebellar hemisphere had significantly more SNPs than the frontal cortex (**Fig.3.13a**). When these tests were repeated for SNPs that fell within exons, the significance between the two cerebellar tissues and the frontal cortex remained (**Fig.3.13b**). Finally, differences in the number of SNPs mapping to the mitochondria were analysed but there was no significant difference between any of three tissues in terms of numbers mitochondrial variants, likely because of the very low number of mitochondrial SNPs (**Fig.3.13c**). To make sure the significant differences observed between the tissues were not a function of there being vastly more reads in samples coming from the cerebellum, the average number of reads per sample for each tissue was plotted. This revealed very little difference in the average number of mapped reads (**Fig.3.13d**, indicating that large variations in read numbers are not driving this observation and confirming a bona fide increase in the number of SNPs in the RNA-seq samples derived from the cerebellum and the cerebellar hemisphere relative to the frontal cortex samples.

3.3 Discussion

Our results, largely corroborated across multiple versions of our pipelines suggest that the cerebellum and the cerebellar hemisphere have a greater basal rate of mutation than the frontal cortex, particularly for specific sets of genes. Additionally, the cerebellar tissues appear to have more SNVs and a greater rate of RNA editing. However, many caveats and alternative explanations exist and so each set of results needs to be carefully assessed in detail.

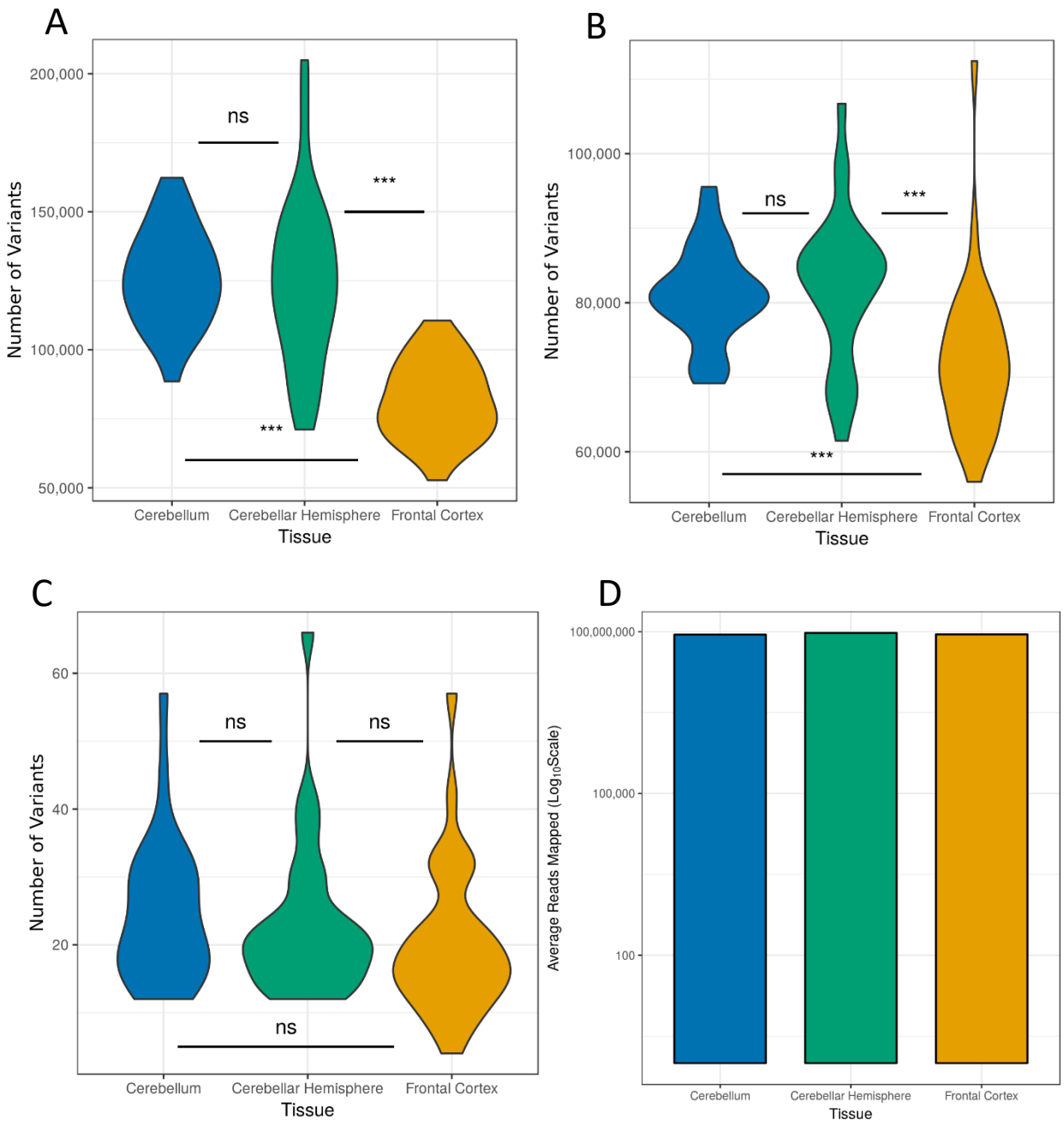


Fig.3.13a-c) Violin plots showing the spread of numbers of variants, for a) all variants b) variants within exons and c) variants within mitochondrial genes, identified by running a germline variant caller on RNA-seq samples from the cerebellum, cerebellar hemisphere and the frontal cortex.

d) Bar chart showing the average number of mapped reads for the RNA-seq samples from the cerebellum, cerebellar hemisphere and the frontal cortex

ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

The sample average mismatch rate comparisons across tissues for when RNA editing was filtered out in pipeline V.1 and pipeline V.2 total genes match each other in terms of their relation to the pipeline V.1 results – significance between the cerebellar hemisphere and the frontal cortex was lost and the cerebellum – frontal cortex comparison also lost significance. This is encouraging to a certain extent, as it shows that the filters introduced in pipeline V.2 are functioning as intended, but it also implies that much of the differences between tissues in pipeline V.1 was driven by RNA editing. Although the cerebellum – frontal cortex comparison did retain significance, it is entirely possible this is due to the differences in how the tissues were preserved. However, it should be noted that across many other results the cerebellar hemisphere follows the same pattern as seen for the cerebellum, albeit with a slightly weaker effect in terms of the numbers of genes involved. This loss was even greater for the pipeline V.2 exons results, in which there were no significant comparisons. In light of this, and the fact that our RNA-seq simulations showed that even with a uniform error model you can still get significance between tissues, these tissue level tests should be taken to be overall less reliable than the other results. On the other hand, For both pipeline V.1 with and without the A→Gs removed and pipeline V.2 total genes, many more differentially mismatched genes had a higher mismatch rate in the cerebellum and the cerebellar hemisphere than did in the frontal cortex. The fact this result has been recapitulated across multiple versions of the pipeline with different filters means it is more reliable than the tissue mismatch rate comparisons. However, this general pattern is lost when looking just at exons, as the total number of genes called as having differential rates of mismatch falls dramatically and in the cerebellar hemisphere – frontal cortex comparison, samples from both of which were preserved in the same way, there are more genes with a higher mismatch rate in the frontal than the hemisphere. It would appear therefore that at the exon level this pattern breaks down in the most accurate comparison. This indicates that many of our mismatches are not occurring in exons and instead are perhaps coming from introns in unspliced RNA. This indeed fits with recent research on mutation rates across different regions of genes, which showed that in cancer cells exons are protected from DNA damage relative to introns, proposed to be due to the presence of a specific histone marker in exons that recruits mismatch repair machinery. Also, AP-seq has determined that exons are highly depleted of oxidative damage whereas introns have a level of damage just below that of intergenic regions, which have relatively high damage

levels (Poetsch et al., 2018). Therefore it is not surprising that we see less genes differentially mismatched between tissues when restricting analysis to exons. However, it seems difficult to imagine that most of these mutations are occurring in pre-spliced introns, as it would require a large amount of the GTEx dataset to consist of such unspliced RNA. Another possibility relates to a filter we did not include within our mismatches pipeline, namely, getting rid of mismatches that occur around intron-exon junctions. Due to misalignment of reads across such junctions mismatches can occur, and it has been recommended to filter our such mismatches as unreliable when calling variants from RNA-seq data. It may be said that our indel filter may in fact detect and discard such misaligned bases. However, this is not a catch-all, and therefore, if we were to propose an important follow up to this data, it would be to implement such an intron-exon junction window mismatches filter and also some sort of end trimming of reads.

The differences in base changes are intriguing, but difficult to interpret. It was suggested that the fact that A→Gs and T→Cs are enriched in the cerebellar tissues compared to the frontal cortex, and the fact they are the base changes with the highest rate may mean they represent unfiltered RNA editing events, the A→Gs being normally recorded and the T→Cs being the RNA editing events on the reverse strand that are not being corrected to the reverse complement for some unknown reason. This hypothesis is given weight by the fact that significance between the tissues for these base changes disappears in the pipeline V.2 exon results, as research has shown that the cerebellum has an increase in RNA editing that is particularly notable in repetitive sequences, which is linked to RNA-hyperediting within UTRs and particularly relevant to this argument, introns (Tan et al., 2017; Walkley and Li, 2017). Suffice to say, when introns are removed from the analyses by only focussing on exons, the main source of difference in RNA editing between the cerebellum and the frontal cortex would be nullified. G→A, T→G and C→T mutations are markers of deamination, oxidation and alkylation induced damage respectively. In all these cases, in order for the chemical attack to cause the base change, it would have to have happened during replication (Reviewed in Barnes and Lindahl, 2004). As mentioned previously, if such chemical attack is the source of these mutations, then it must have happened during development. Although there is no difference between any of the tissues for G→A and T→G mutations, this is interesting because it implies that during development

these tissues were not exposed to differential levels of attack from these types of damaging agents. C→T and G→T mutations on the other hand are enriched in the cerebellum relative to both the frontal cortex and the cerebellar hemisphere. As well as representing replication stabilised alkylation mutations, they can also be caused by unrepaired spontaneous deamination and are also associated with transcription. However, due to the rate for C→Ts showing a significant difference between the cerebellum and cerebellar hemisphere, effectively duplicate samples preserved differently, and there being no difference between the cerebellar hemisphere and the frontal cortex, both of which were preserved by snap freezing, it is overwhelmingly likely that this difference is driven by the differing methods of preservation. This also applies to G→T mutations, which show the exact same pattern of differential rates. The strangest result from this analysis was that T→As are enriched in the frontal cortex compared to the cerebellar tissues. There is no common form of attack that causes such a base change, so the biology underlying this remains a mystery.

The patterns of over and under representation of gene categories recaptured across different analyses show that genes with a higher rate of mismatch in one tissue relative to the other are enriched for genes downregulated and depleted for genes upregulated in that tissue. This is also true for differentially expressed mitochondrial genes, however, as the number of non-differentially expressed mitochondrial genes was too small to accurately assess enrichment, we were unable to determine whether this effect was merely due to their being downregulated and upregulated, and not because they were specifically mitochondrial. The fact that this applies to genes with higher rates of mismatch in both the cerebellum and the frontal cortex means that it is obviously not a tissue specific effect. It is tempting to suggest that the enrichment of downregulated genes amongst the mismatched genes represents the “use it and lose it” hypothesis in action (Lodato et al., 2015). The downregulated genes that overlap with the genes with a higher rate of mismatch in one of the tissues may be downregulated because they are prone to mutations, which thereby cause lower levels of transcription or loss of expression altogether in some cells. However, this hypothesis hinges on those genes starting out as highly expressed, and so the fact that genes upregulated in the tissue in which the genes have a higher rate of mismatch are depleted from the mismatched genes is a potential problem with this argument. Alternatively, the observed pattern of enrichment and depletion could be a false positive

resulting from some sort of technical issue. However, we applied a far more stringent base number cut off for pipeline V.2 analyses, so it is unlikely to be caused by high numbers of mismatches and low numbers of bases, as explored when assessing the efficacy of pipeline V.1.

The observation that the cerebellum has a greater rate of RNA-editing is interesting within itself, though not unprecedented. Recent research had shown that the cerebellum is distinct from other brain tissues with respect to RNA editing, proposed to be due to higher expression of the RNA-editing enzyme ADAR2 within the cerebellum. The cerebellum also has higher levels of RNA editing than cerebral tissues, particularly at repetitive sites (Tan et al., 2017). Other research has indicated that the brain has a greater number of both tissue-specific RNA-editing events and RNA-edited tissue specific genes (Picardi et al., 2015). Additionally, across the tissues assessed it was found that the sets of genes that underwent tissue specific RNA-editing were enriched for genes involved in diseases specific to that tissue, and this was also true of the brain. Specifically, genes that only underwent RNA-editing in the brain were enriched for genes involved in neurological and neurodegenerative disorders (Picardi et al., 2015). This observation paired with the fact that the cerebellum has a greater level of RNA editing than the cerebrum raises interesting questions. If the cerebellum has a higher basal mutation rate then this could interfere with RNA-editing by changing the usually edited bases to different ones, preventing the usual A→I conversion. In a disease state associated with increased genome instability, such as manifests in the DRDA-ARCAs, the effect could be magnified and in turn could lead to dysfunction of transcriptome regulation which contributes to overall pathology. This is a very specific scenario however, and it would presumably have to occur across many cells to have any effect, so it is unlikely that this is the case. It may also be postulated that perhaps because the cerebellum has a higher level of RNA editing than the cerebrum it is much more sensitive to perturbations in RNA-editing that may come about as a result of such mutations at RNA-editing sites. There are complications with this hypothesis though, because the higher level of RNA editing in the cerebellum is mainly associated with repetitive sites found primarily in introns or UTRs (Tan et al., 2017). These repetitive sites tend to undergo what is termed hyper-editing, A → I editing of a large number of adenosines found close to each other within the same transcript (Walkley and Li, 2017). It is not known how the abrogation of a single editing site

in the context of hyper-editing would affect the expression of a given gene. Therefore, whether mutations at editing sites may have a biological effect that could promote disease states remains unresolved. It should also be noted that no known link between the DRDA-ARCAs and RNA-editing have been established as of yet. A more important current matter is the effect that this increased level of RNA editing in the cerebellum could be having on our results. Due to this phenomenon, our previous assumption that all variables relating to mutation calling from RNA-seq data, somatic mutation rate notwithstanding, were the same between the cerebellum and the frontal cortex no longer holds true. Despite the fact we have taken steps to filter out RNA editing events from our analysis, no database will be 100% accurate. Low-frequency RNA-editing events not contained within REDIportal but are nevertheless occurring at a higher level in the cerebellum could be driving this observed difference in tissue wide mutation rate between the cerebellum and the frontal cortex. The genes showing an increased rate of mismatch in cerebellum are less susceptible to this line of attack however, as they were not enriched amongst genes showing a significantly increased rate of RNA-editing in the cerebellum. Nevertheless, it is possible, if unlikely, that these genes contain many low frequency RNA editing events that are driving this difference in number of genes that have significantly higher mismatch rates between the cerebellum and the frontal cortex. Although it has been known about for a while now, RNA-editing is still an emerging field and as more research is carried out and detection techniques improve, we may yet find answers to these questions.

Our analysis of the germline variants called in both the cerebellum and the frontal cortex revealed that the cerebellum had significantly more called variants. There are several potential explanations for this. One is that because germline variant callers can pick up somatic mosaic mutations, this difference is due to there being an overabundance of somatic mosaic mutations in the cerebellum relative to the frontal cortex, resulting in an apparent increased load of germline SNPs. There may be some precedent for this line of reasoning. As mentioned earlier, the cerebellum undergoes an extended period of maturation compared to most other cerebral tissues, and this involves rapid cell proliferation (El-Khamisy, 2011). Cell proliferation requires DNA replication which can lead to DSBs due to replication stress, which can in turn introduce mutations (Reviewed in Zeman and Cimprich, 2014). If a mutation arises in a progenitor cell early on in this period of

extended development, then it will be propagated to each subsequent daughter cell and become somatic mosaic. This could be what is leading to the observed increased in germline variant calls in the cerebellum. The second potential driving factor is that genes containing SNPs have a higher expression level in the cerebellum. Whilst this is certainly a possibility, there is no context for why this might be happening. The obvious explanation would be that the cerebellum merely has a higher level of gene expression overall compared to the frontal cortex. However, this cannot be determined by looking at log fold-change skews between the cerebellum and the frontal cortex from our earlier differential expression analyses, because of the normalisations applied. The third option is that non-A→I RNA-editing not removed by our getting rid of A→G base changes is driving this difference. The other form of RNA editing that occurs in mammalian cells is C→U editing, which would be picked up as a C→T change by the sequencer (Keegan et al., 2001). An issue with this this hypothesis is that A→I RNA editing is many times more abundant than C→U editing. Currently, the pool of RNAs that are modified by C→U editing is very small, although it has been expanded since this form of RNA editing was first discovered (Keegan et al., 2001; Rosenberg et al., 2011). Therefore, whilst it is unwise to rule out C→U editing as a factor, it is unlikely to have a major effect on the number of mutations derived from germline variant calling. A final, obvious option is that this is a difference driven by some sort of non-biological difference between the samples from the two tissues. As discussed in the results section, differences in RNA integrity do not seem to be driving the disparities between tissues we observe. A difference could be arising from the technical issues associated with calling variants from RNA-seq data, but it is hard to imagine what this might be. In the absence of a strong alternative hypothesis, the idea somatic mosaicism with its grounding in the biology of cerebellar development is very appealing, even if the numbers involved in the cerebellum-frontal cortex difference do appear to be perhaps to high to be accounted for purely by this. It may be the case that the effect is being driven by a combination of these factors, and more work would need to be done to determine their relative effects, or alternatively isolate the single causative factor.

In conclusion, whilst we were not able to build outright on the work from the previous chapter, as it could not be determined whether the enrichment of mitochondrial genes amongst differentially mismatched genes was as a result of their differential up or

downregulation or because they were mitochondrial, the enrichment of downregulated genes amongst mismatched genes could be evidence for the “use it and lose it hypothesis” at play. In addition, the fact that we repeatedly call a set of genes that show differential rates of mismatch between tissues is worthy of note, and these groups of genes warrant further exploration. The possibility of somatic mosaicism leading to there being more SNVs in the cerebellar tissues is also a major finding, and could explain some of our results, as our mismatches pipeline will call both cell private and somatic mosaic mutations. The elevated levels of RNA editing in the cerebellum and the cerebellar hemisphere compared to the frontal cortex and the links between RNA editing and neurological disease warrant further research, but in this body of work it is likely these events represent a major confounding effect. Therefore, expanding on this work would ideally involve matched single cell RNA and DNA-seq from the relevant tissues in order to filter out RNA-editing events and also double call mutations to improve reliability. The mismatches pipeline should be fitted with an intron-exon junction window mismatches filter and a published somatic variant caller utilised to double check our mismatch calls. Additionally, further work should be done testing the mismatches pipeline through positive controls to give a proof of principle for the work done in the chapter.

3.4 Methods

3.4.1 Identification of mismatches

The GTEx RNA-seq dataset was filtered to exclude samples not matching the tissues of interest and those that has an RNA integrity of less than 6, an mapping rate of less than 0.8 and a duplication rate higher than 0.5 (Carithers and Moore, 2015). Additionally, it was made sure that for every frontal cortex sample that passed the filters there was either a cerebellum or cerebellar hemisphere sample from the same donor that also met the criteria, and vice versa, otherwise the sample was discarded. Samples that passed these filters were downloaded via their SRA accessions using `fastq-dump` from SRA-Toolkit. These samples were then quality assessed using `fastQC` and mapped to the human reference genome (Ensembl 85, hg38) using hisat2 (Andrews, 2015; Kim et al., 2015b; Zerbino et al., 2018). hisat2 was set to allow up to 15% of the read to be mismatched to the

reference genome to make sure no reads of sufficient quality were being excluded due to mismatches (parameters: `--ignore-quals --score-min L,0,-0.9`). After mapping, bam files were put through the mismatch pipeline, built using the CGAT pipelines framework (Sims et al., 2014). Read group names were added to each sample based on the name of the dataset using `picard AddOrReplaceReadGroups` (parameters: `RGLB=lib1, RGPL=ILLUMINA RGPU=unit1`) (Broad Institute, 2016). Duplicates were then removed from the BAM files using `Picard MarkDuplicates` in conjunction with `samtools view` using the `-F` option to remove reads with the flag 1024 (PCR or optical duplicates) (Li et al., 2009). Reads subsequently split into exon segements using Genome Analysis Toolkit (GATK) `SplitNCigarReads`, which identifies reads with Ns in their cigar string, i.e. reads spanning a splice junction and splits them apart (parameters: `-rf ReassignOneMappingQuality -RMQF 255 -RMQT 60 -U ALLOW_N_CIGAR_READS`) (McKenna et al., 2010). In pipeline V.2, if the pipeline.ini option `vcfavail` was set to 0, germline variant calling using the GATK `HaplotypeCaller` was then carried out on the bam files output by `SplitNCigarReads` (parameters: `-dontUseSoftClippedBases -stand_call_conf 20.0`). The header of the VCF file was then renamed according to the sample name as it appears in the BAM file using `bcftools reheader` (Li, 2011). After variant calling, if specified, or after running `SplitNCigarReads` if variant calling was not set in the .ini options, mismatches were then called using a custom python script built using the CGAT scripts framework. Iteration of mismatch calling was done over each gene contained within a specified GTF file, in all instances discussed in this chapter this GTF contained all genes within the hg38 Ensembl 85 annotation. Mismatches were identified by iterating over bases in each read from each gene and assessing them for non-alignment to the reference genome. Reads that were either duplicate, unmapped, had an unmapped mate or mapped to multiple locations within the genome were not iterated over. Putative mismatches were also subject to several checks and filters. In pipeline V.2 they were cross referenced with SNPs contained within either the GTEx VCF or the germline VCF for the relevant sample generated at an earlier step in the pipeline and if an alternate allele that matched the mismatched base was found that the mismatch position, the mismatch was regarded as a SNP and discarded. Mismatches were

also cross referenced with a downloaded version of the online RNA-editing database REDIPortal in pipeline V.2 (Picardi et al., 2017). Again, if an RNA editing event that matched the base change and position of the mismatch was identified, the mismatch was treated as an RNA editing event and not recorded as a true mismatch but instead added to an RNA editing counter. Mismatches thought to be part of indels were also discarded by identifying mismatches that occurred in the context of three or more mismatches in a 5 base window. Indels were recorded by counting the number of Is and Ds inside the cigar string. Both these features related to indel were implemented in pipeline V.2. Potential mismatches that fell below a base quality of less than 30 were also not recorded as real mismatches. The base change coinciding with a mismatch was also logged. If the read was on the reverse strand, then the base change was recorded as the reverse complement in pipeline V.2. pysam (https://github.com/pysam-developers/pysam) was widely used throughout the script for recovery of read and base data, iteration over reads and reading in the BAM files. FASTA files, GTF files and the REDIPortal RNA-editing BED file were read in and manipulated using `IndexedFasta`, `GTF.flat_gene_iterator`/`GTF.iterator` and `Bed.readandIndex` respectively from CGAT (Sims et al., 2014). VCF files were read in and manipulated using `pyVCF` (https://github.com/jamescasbon/PyVCF/). `IOTools` from CGAT was also used throughout the script. Pipeline V.2 exons differed from pipeline V.2 total gene in that bases not falling within exons were not logged or carried forward for further analysis as part of the mismatches counting script. `pipeline_rnaseqmismatches` and `pipeline_rnaseqmismatchesexons`, correspond to pipeline V.2 total gene and pipeline V.2 exons respectively, and their associated mismatch calling scripts are available at: https://github.com/jdparker101/pipeline_rnaseqmismatches and https://github.com/jdparker101/pipeline_rnaseqmismatches_exons.

3.4.2 Analysis of mismatch data

Analysis of the mismatch data was performed in R studio (R Core team, 2019; R Studio Team, 2015). Mismatch databases were read into R studio using RSQLite (Müller et al., 2018). Prior to any sort of statistical test, genes that had no mismatches or less than 10 bases recorded for pipeline V.1/500 bases recorded for pipeline V.2 were filtered out of the data on a per sample basis. Overall mismatch rates for each sample were calculated by

summing the number of bases assessed and the number of mismatches logged over all genes and then dividing the number of bases by the number of mismatches to generate a value representing the number of mismatches per base. Mismatch rates for each gene within a sample were calculated by dividing the number of mismatches by number of bases recorded on a per gene basis to get a mismatches per base rate for each gene. RNA editing rates on a per sample basis and a per gene basis within each sample were calculated in the same way as the analogous overall sample and per gene mismatch rates, except instead of mismatches the number of RNA editing events was divided by the number of bases. The `t.test` function in R was used to assess differences in overall mismatch rate and overall RNA editing rate between samples from different tissues. This function also gave the average mismatch rates across all the samples for both tissues. Genes with differential mismatch rates were identified through the fitting of linear models using the R function `lm`. Tests for enrichment of various gene categories within sets of genes with differential mismatch rates were performed using Fisher's tests in R using `fishers.test`. For information on how genes differentially expressed between the cerebellum and the frontal cortex were identified, see the methods section of chapter 2. To identify mitochondrial genes, the gene ids of mitochondrial genes were extracted from the org.Hs.eg.db annotation database package for R using mitochondrial GO terms pulled from the GO annotation package GO.db (Carlson, 2019b, 2019a). GO analysis was carried out using the R package GOseq (Young et al., 2010).

3.4.3 Read simulation

Reads were simulated using a custom pipeline named `pipeline_readsimulation` built using the CGAT pipelines framework. The GTEx RNA-seq samples were selected by applying the same quality control filters used to pre-select samples for the mismatch analysis (but without the requirement for donor paired cerebellum/cerebellar hemisphere – frontal cortex samples): RNA integrity >= 6, mapping rate >= 0.8 and duplication rate <= 0.5 to the GTEx RNA-seq read counts matrix (available from the GTEx portal: https://gtexportal.org/home/datasets). The R package Biostrings was then used to read in a FASTA file for hg38 and subset it for only genes present in the GTEx RNA-seq read counts matrix and re-write a new, filtered FASTA file (Pagès et al., 2017). The GTEx read counts matrix was then split up and new count matrices containing all the counts for each

individual sample were generated. From each of these sample count matrices and the previously subsetted FASTA file, a pair of simulated read containing FASTA files, simulating paired-end sequencing, was generated using the `simulate_experiment_countmat` function from the R package polyester (Frazee et al., 2015). This generates reads based in number upon the read counts for each gene in the provided counts matrix. Reads were generated using the default error model, a uniform error model, and with the default error rate, 0.005, or 1 per 200 bases. The resulting FASTA files were then converted into FASTQ files using a custom python script, `fastafastqconversion.py` built using the CGAT scripts template and utilising `Fastq`, `FastaIterator` and `IOTools` from CGAT (Sims et al., 2014). These FASTQ files were then put through the mismatches pipeline and the resulting mismatches matrix analysed in R according to the protocols laid down under the **Analysis of mismatch data** subheading of this chapter. `pipeline_readsimulation` and its associated scripts can be found here: https://github.com/jdparker101/pipeline_readsimulation.

3.4.4 Data manipulation in R and plots

Within R, data manipulation was performed using the dplyr and tidyr packages (Wickham and Henry, 2019; Wickham et al., 2019). Plots were generated using base R and the ggplot2 package (R Core team, 2019; Wickham, 2016).

3.4.5 Tabular summary of datasets and analyses

Dataset	Experimental design/Samples utilised	Analyses performed
GTEx V6 RNA-seq	Pipeline V.1 analyses: 24 Brain – Cerebellum samples, 27 Brain – Cerebellar Hemisphere samples, 29 Brain – Frontal Cortex (BA9) samples Pipeline V.2 analyses: 27 Brain – Cerebellum samples, 30 Brain – Cerebellar Hemisphere samples, 32 Brain – Frontal Cortex (BA9) samples Read simulation: 41 Brain – Cerebellum, 40 Brain – Cerebellar Hemisphere	Mismatch rate analyses (pipeline V.1, pipeline V.2 total gene and pipeline V.2 exons) Read simulation Variant calling analyses (pipeline V.2 sample set) Category enrichment (Fisher's) tests Mutational spectra analysis (pipeline V.2 sample set) RNA editing analyses (pipeline V.2 sample set)

	samples, 35 Brain – Frontal Cortex (BA9) samples	

4. A Genomic Study Into the Role of NuMA in DNA repair

4.1 Introduction

The nuclear mitotic apparatus (NuMA) is a large protein consisting of a globular head and tail linked by a coiled coil domain, first discovered nearly 40 years ago due to its intriguing cell cycle specific localisation (Yang et al., 1992). NuMA localisation was restricted to the nucleus during interphase, but upon cells entering mitosis it relocated to the cytoplasm, particularly the spindle poles (Lydersen and Pettijohn, 1980). This observation in turn led to the characterisation of NuMA's best known role in assembly of the spindle microtubules. NuMA binds to and bundles together microtubules through its C-terminal domain, which facilitates its role in concentrating and stabilising these molecules at the spindle poles and mediating their interaction with the centromere (Gaglio et al., 1995; Merdes et al., 1996, 2000). Concordant with an important role for NuMA during mitosis, disruption of NuMA function results in mitotic abnormalities and complications in the re-formation of the nucleus post-mitosis, and RNAi experiments in mice have shown NuMA to be essential for viability (Reviewed in Cleveland, 1995; Harborth et al., 2001; Kallajoki et al., 1993).

Although the spindle assembly function of NuMA is its most well characterised, the protein has also been implicated in a variety of disparate roles. NuMA has long been considered a putative component of the nuclear matrix due to its high abundance within the nucleus, its absence from non-spherical nuclei and enrichment in the insoluble fraction of the nucleus - which contains other nuclear matrix components. It also has the ability to form multi-arm oligomers, with its overexpression resulting in the formation of a nuclei-filling scaffold and presence within some nuclear filaments (Compton et al., 1992; Harborth, 1999; Lydersen and Pettijohn, 1980; Merdes and Cleveland, 1998). It has also been linked to genome organisation, a function likely related to its role as a structural protein as the nuclear matrix has been suggested to facilitate the spatial distribution of the genome as discussed in the introduction. Indeed, NuMA has been shown to bind to matrix attachment regions (MARs) *in vitro*, possibly through its S/TPXX domain - a motif found in DNA binding gene regulatory proteins, and be involved in the 3D arrangement of chromatin in human mammary epithelial cells, which in turn is related to their differentiation (Abad et al., 2007; Ludérus et al., 2012; Suzuki, 1989). Also relevant to nuclear structural organisation is that

NuMA localises to the nucleolus, where studies have indicated it mediates both the nucleolar stress response and rDNA transcription through binding to rDNA promoters and rDNA transcription proteins such as RNA polymerase I and components of B-WICH, a chromatin remodeler (Jayaraman et al., 2017). Separate from its nucleolar functions, NuMA is additionally linked to genomic stress response, where it particularly promotes the transcription of pro-cell cycle arrest genes by p53 (Endo et al., 2013; Ohata et al., 2013). Finally, most interestingly for this thesis, there is a small body of evidence connecting NuMA to DNA repair. It has been repeatedly suggested across analyses of the aforementioned NuMA functions that it being a structural protein means it acts as a scaffold for the recruitment and stabilisation of other important factors, and this has been recapitulated in investigations relating to its involvement in DNA damage response. NuMA has been demonstrated to accumulate at sites of DNA damage in a manner dependent on the activity of PARP, where it is postulated to function epistatically with the chromatin remodeler SNF2h in HR via binding to and recruitment of SNF2h to the location of the damage (Vidi et al., 2014). Silencing of NuMA also led to the depletion of several repair proteins such as BRCA1 and CtIP at the damage site, reduction of HR levels by 60%, premature loss of γH2AX foci and the absence of chromatin de-condensation at DNA breaks, a phenomenon observed in control cells (Vidi et al., 2014). A separate investigation demonstrated that the treatment of basally polarised breast epithelial cells with the DSB inducing drug bleomycin, led to the re-localisation of NuMA within the nucleus, and BLM treatment in conjunction with a NuMA KD resulted in a reduced percentage of cells with γH2AX foci (Vidi et al., 2012). The association of NuMA with MARs, themselves linked to DNA damage repair, is also relevant (Ludérus et al., 1994). NuMA has associations with cancer, with it being overexpressed in epithelial ovarian tumours, although whether this is linked to its DNA damage response or mitotic function is unclear, and whether it is a cause or response to genome instability also remains to be elucidated (Brüning-Richardson et al., 2012). Further confounding easy explanations are observations that NuMA is alternatively spliced into three main categories of isoform, long, medium and finally a short isoform that may act as a tumour suppressor (Qin et al., 2017; Wu et al., 2014). However, other research has shown that favouring the generation through splicing of a slightly shorter variant of the long isoform lacking the 14 amino acid coding exon 16 promotes cell proliferation and centrosome amplification in otherwise normal MCF10A cells. Analysis of luminal tumours demonstrated that as the

preference for the shorter long isoform increased, so did a signal for aneuploidy (Sebestyén et al., 2018).

This demonstrates the possibility of diverse roles for different isoforms of NUMA, and perhaps the wide range of functions discussed, including within the DDR, are mediated separately by a range of differentially spliced transcripts. This is one angle from which to explore the action of NUMA in the DDR, but there are many others. For example, its localisation to sites of DNA damage was shown to be PARP dependent, and PARP is primarily associated with SSB repair, yet so far only a link with HR has been established (Vidi et al., 2014). Is there therefore a potential role for NuMA within SSB repair? Furthermore, the establishment of NuMA binding to rDNA, whether directly or indirectly, and its association with MARs, raises the question as whether it binds to other regions of the genome, or even particular subsets of genes, in order to facilitate effective DDR (Jayaraman et al., 2017; Ludérus et al., 1994). There is also the interesting prospect of NuMA mediating tissue specific DDR. It has been demonstrated that the nuclear matrix in its coordination of 3D chromatin structure is important for tissue differentiation and phenotype, and this could extend to the coordination of DNA repair in a tissue specific manner (Abad, et al., 2007; Lelievre et al., 1998). As disruption of lamins is known to cause neuropathies, perhaps dysregulation of NuMA function would cause similar defects if it was not embryonically lethal (Reviewed in Gonzalo, 2014). The possibilities discussed here represent a wealth of avenues to explore in investigating how NuMA relates to the DDR.

4.2 Results

4.2.1 NuMA shows increased expression in the cerebellum relative to all other brain regions across the GTEx RNA-seq dataset

Due to the association between mutations in lamins and neuropathies, we resolved to examine expression of NuMA across different regions of the brain. For this, the GTEx RNA-seq data was utilised, due to its large variety of brain regions with high numbers of samples. Violin plots showed that NuMA expression was far higher in the cerebellum and cerebellar hemisphere than in any other brain tissue **(Fig.4.1a)**.

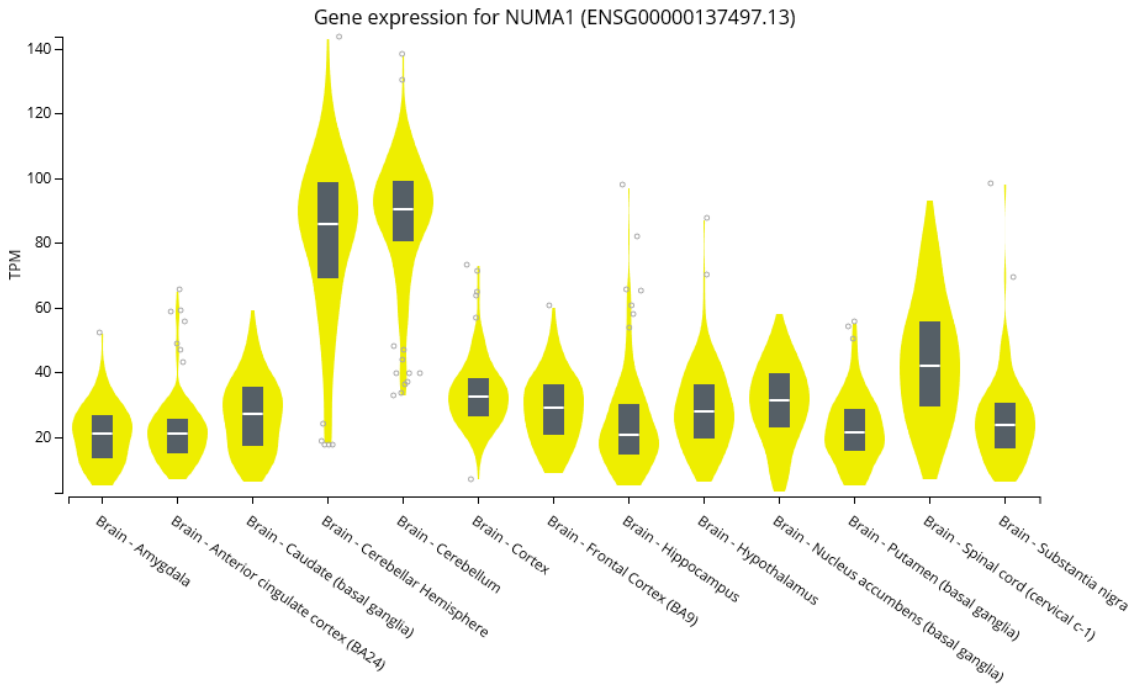


Fig.4.1 a) Violin plot showing the spread of transcripts per million (TPM) of NuMA across samples of each brain tissue contained within the GTEx RNA-seq dataset. Generated using the GTEx portal at https://gtexportal.org/home/gene/NUMA. TPM = Transcripts per million

4.2.2 Specific categories of genes are enriched amongst genes differentially expressed in H_2O_2+ WT vs H_2O_2- WT cells and H_2O_2- WT vs H_2O_2+ $NuMA_{KD}$ cells

To investigate how NuMA may impact the transcriptional landscape of the cell in the presence of DNA damage, 4sU-sequencing was carried out by Swagat Ray of the El-Khamisy of the University of Sheffield. This technique is designed for the exclusive capture and sequencing of nascent RNA. The procedure for 4sU-seq is as follows. Cells are treated with the uridine nucleotide analogue 4-thiouridine (4sU), which is then incorporated into newly transcribed RNA. Total RNA is then extracted, the 4sU containing RNA tagged with biotin, the 4sU containing, biotin-tagged RNA extracted via streptavidin, and this captured nascent RNA sent off for sequencing (Gilad et al., 2014). In this case, 4sU-seq was carried out on wildtype (WT) and NuMA doxycycline-inducible shRNA knockdown (KD) RPE-1 cells, both in the presence and absence of H_2O_2 treatment, so allowing the immediate effect on transcription brought about by the induction of damage to be determined. The set of samples consisted of two replicates per condition, with each sample corresponding to roughly 40-60 million mapped reads. Differential expression analyses revealed 4,579 and 1,696 genes differentially upregulated and downregulated, with a Benjamini-Hochberg procedure adjusted p-value value cut-off of 0.05 or less, between control and H_2O_2 treated WT cells (**Fig.4.2a**), and 1,209 upregulated and 4,996 downregulated differentially expressed genes between H_2O_2 treated WT and KD cells (**Fig.4.2b**). As NuMA has already been shown to localise to specific sites in the genome, we reasoned that it may function in facilitating the expression of certain damage response genes or perhaps protect certain categories of genes from damage. Therefore, the enrichment of differentially expressed genes for specific gene sets was examined. It was recently revealed that release of promoter proximal pausing is a strong determinant for the formation of endogenous DSBs (Dellino et al., 2019). Promoter proximal pausing is the phenomenon whereby an RNA polymerase undergoing transcription elongation stops ~50 base pairs downstream of the transcription start site, which can then be subsequently released to continue transcription. As for its purpose, pausing is thought to act as another level at which gene expression can be regulated (Promoter proximal pausing reviewed in Adelman and Lis, 2012; Li and Gilmour, 2011). Because of the link between DSB formation and pause-release, paused genes were the first category assessed. Pausing indices for genes were calculated as the ratio between promoter

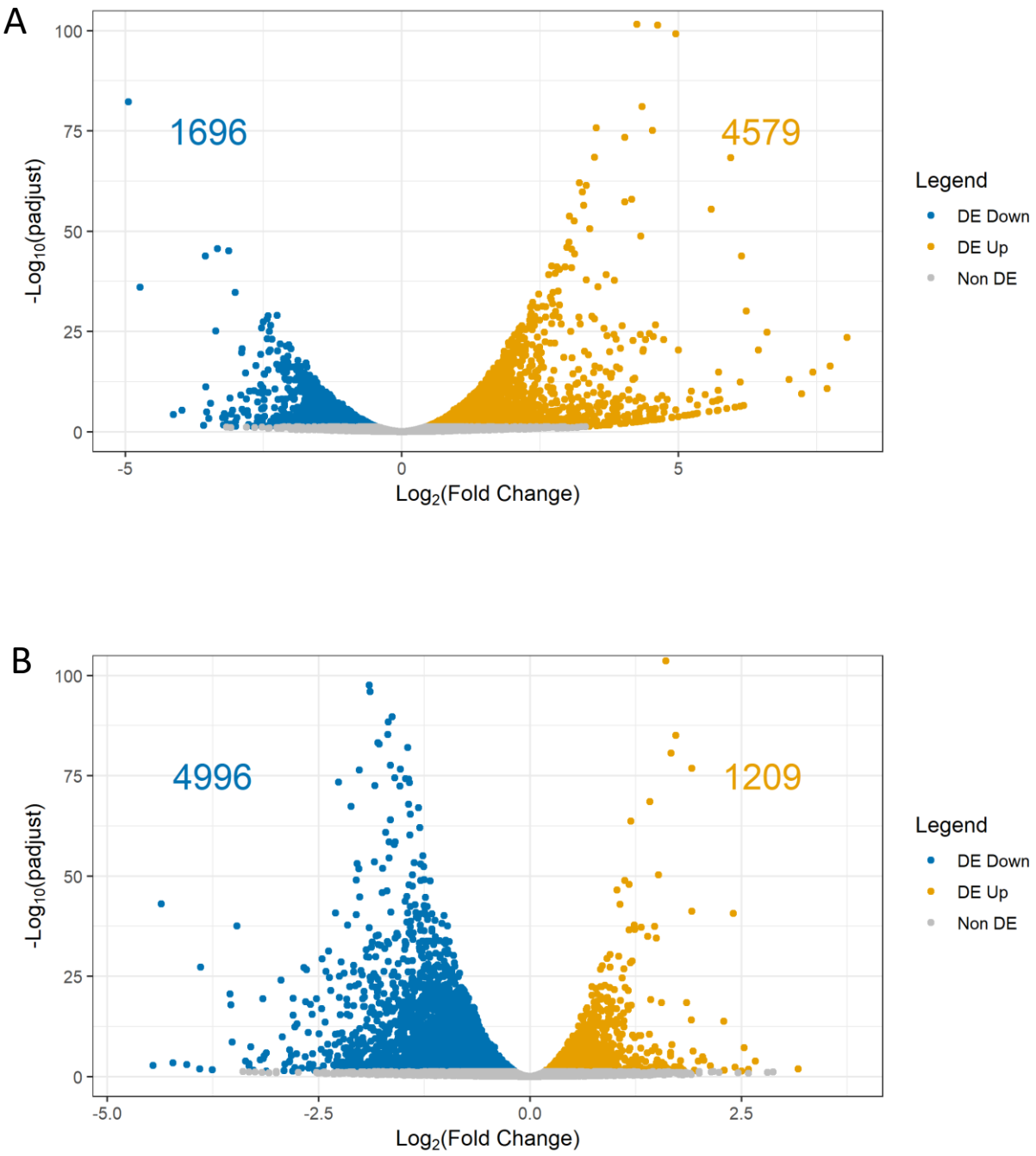


Fig.4.2 Volcano plots of -Log10 transformed adjusted p values plotted against Log2 fold change for genes differentially expressed in a) WT H_2O_2+ vs WT H_2O_2- cells and b) $NuMA_{KD}$ H_2O_2+ vs WT H_2O_2+ cells. Log2 fold changes were taken from the respective comparisons the genes were found to be differentially expressed within and are presented with respect to the first the first condition in each comparison. Cut off for differential expressed was an adjusted p-value <= 0.05. DE = Differentially expressed.

bound and gene body RNA polymerase II, determined from RNA polymerase II ChIP-seq, and paused genes were defined as those genes with a pausing index of 2 or greater. Subsequent tests revealed paused genes were significantly enriched amongst genes both differentially upregulated and downregulated upon treatment of WT cells with H_2O_2 **(Fig.4.3a,b)**. The paused genes showed a greater relative enrichment amongst the downregulated genes than the upregulated genes with an odds ratio (OR) of 1.49 compared to 1.36. When genes differentially expressed upon NuMA KD in H_2O_2 treated cells were considered for paused gene over or under-representation, no statistical enrichment was found within upregulated genes, but the downregulated genes were indeed enriched with an OR of 1.57 **(Fig.4.4a,b)**. The enrichment of immediate early response genes (IERGs) was also considered, due to NuMA being involved in the p53-mediated stress response. The enrichment of the IERGs was even more striking than the results for paused genes. For the WT H_2O_2 treatment comparison, out of the 45 genes assessed, 32 were present amongst the 4,579 upregulated genes (OR=6.13), a significant overrepresentation, and only one was found within downregulated genes, demonstrating a depletion, albeit a non-significant one **(Fig.4.5a,b)**. For both the H_2O_2 treatment NuMA KD upregulated and downregulated genes, a significant enrichment of OR=4.98 and OR=1.92 was observed respectively **(Fig.4.6a,b)**. Genes with fragile promoters (FPGs) and fragile introns (FIGs) were also of interest because their fragile makes them obvious targets for DNA repair factor. Fragile promoter and introns were defined as promoter or introns that overlapped at least one endogenous DSB. These endogenous DSBs were determined by Breaks labelling in situ and sequencing (BLISS) carried out by Dellino et al., (2019), a technique which allows the labelling of DSBs and subsequent sequencing of the adjacent DNA, thereby enabling identification of DSB sites in the genome (Yan et al., 2017). H_2O_2+ WT vs H_2O_2- WT upregulated and downregulated genes showed a very slight enrichment (OR=1.21) and a non-significant depletion respectively for genes containing fragile promoters **(Fig.4.7a,b)**. The results for fragile intron containing genes was similar apart from the enrichment was stronger (OR=1.85), as was the underrepresentation (OR=0.56), which was also statistically significant **(Fig.4.8a,b)**. As for the H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT comparison, the reverse pattern manifested itself. Across fragile promoter containing genes, there was a non-significant result for enrichment or depletion amongst upregulated genes but a slightly overrepresentation within the downregulated set of genes (OR=1.28) **(Fig.4.9a,b)**, whereas for genes containing fragile

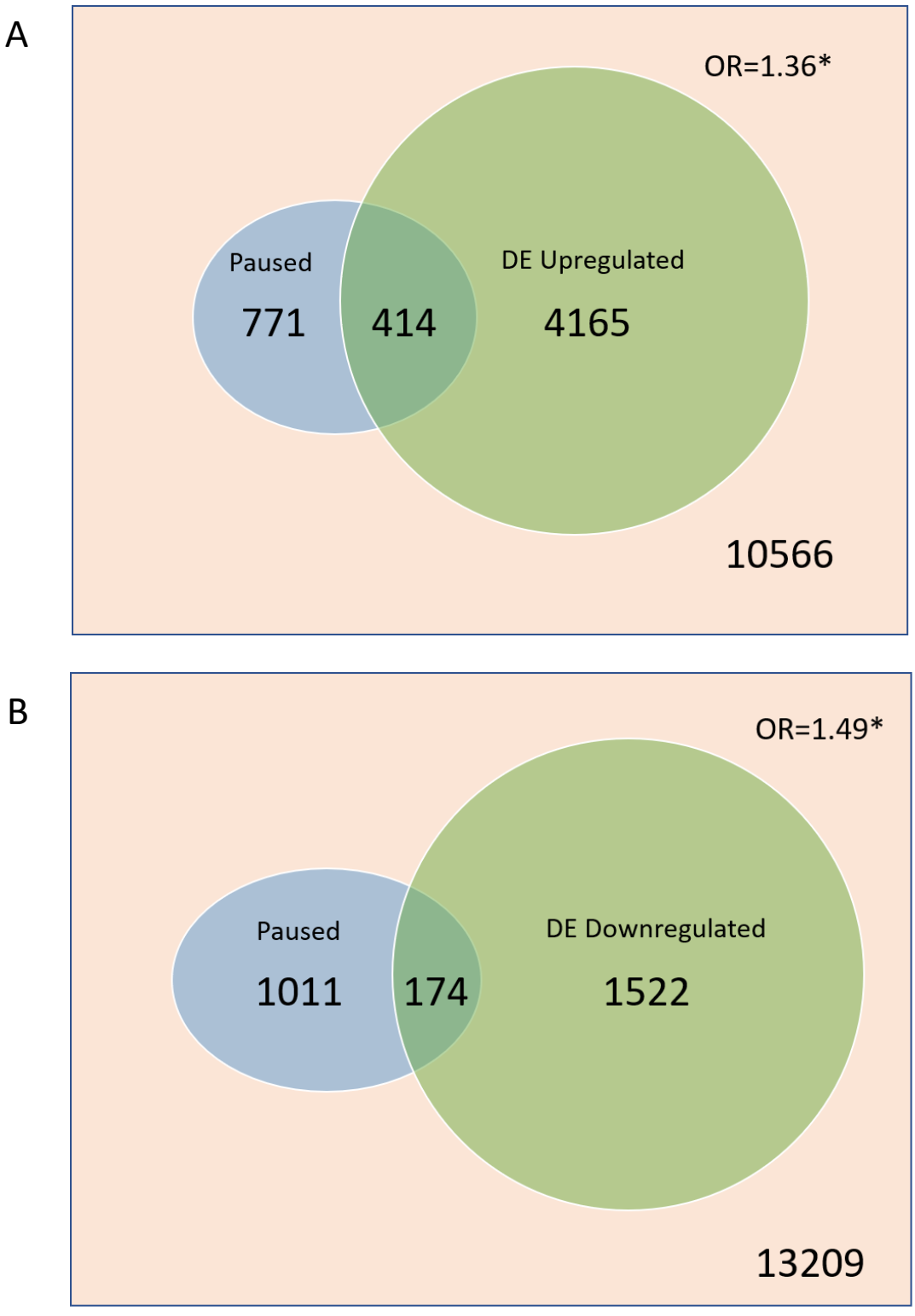


Fig.4.3 Venn diagram showing overlap of paused genes with genes a) differentially upregulated and b) differentially downregulated upon H_2O_2 treatment in WT cells relative to untreated WT cells. The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. DE = Differentially expressed, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

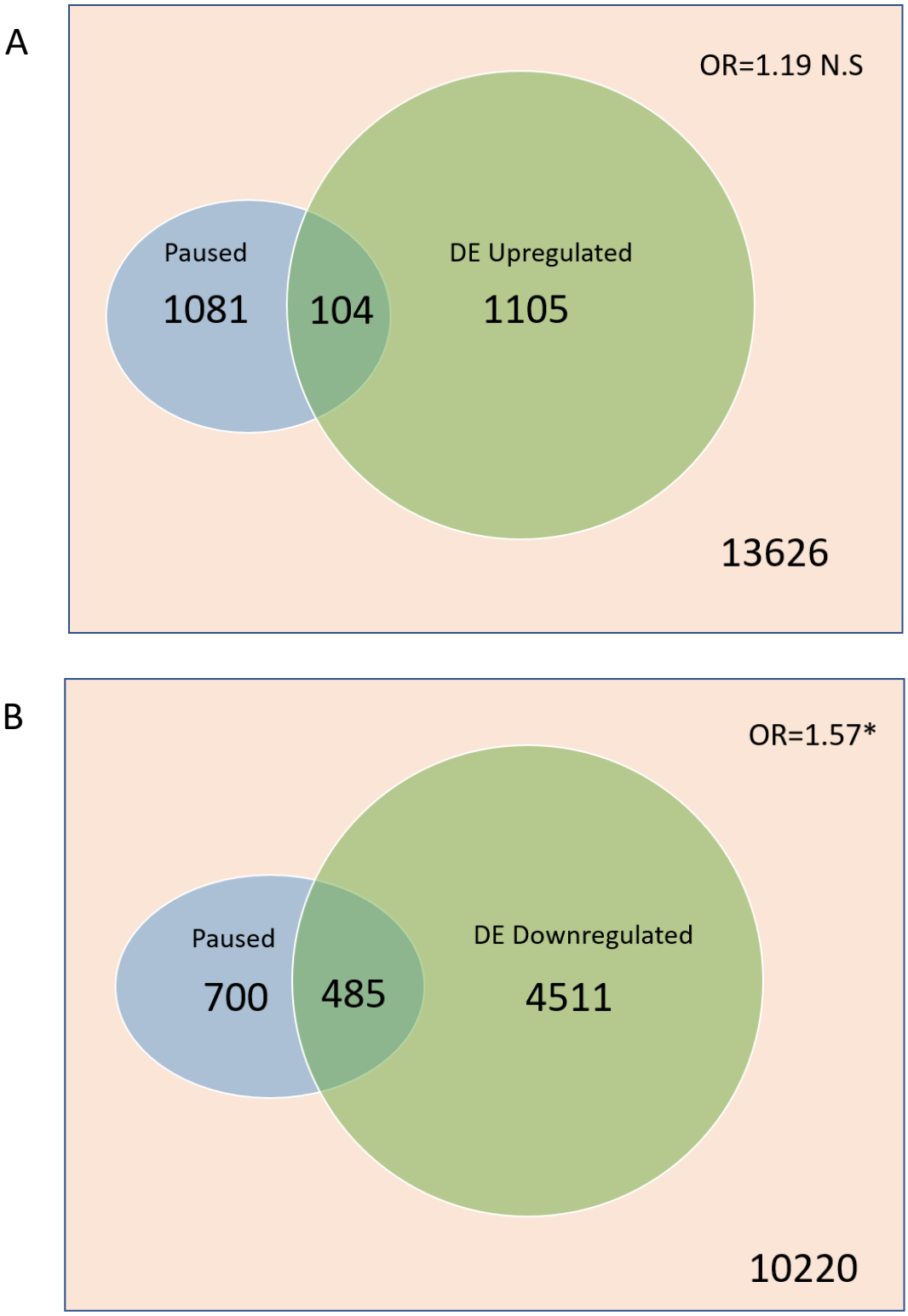


Fig.4.4 Venn diagram showing overlap of paused genes with genes a) differentially upregulated and b) differentially downregulated upon NuMA KD H_2O_2 treatment relative to H_2O_2 treated WT cells. The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. DE=Differentially expressed, N.S=Non-significant, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

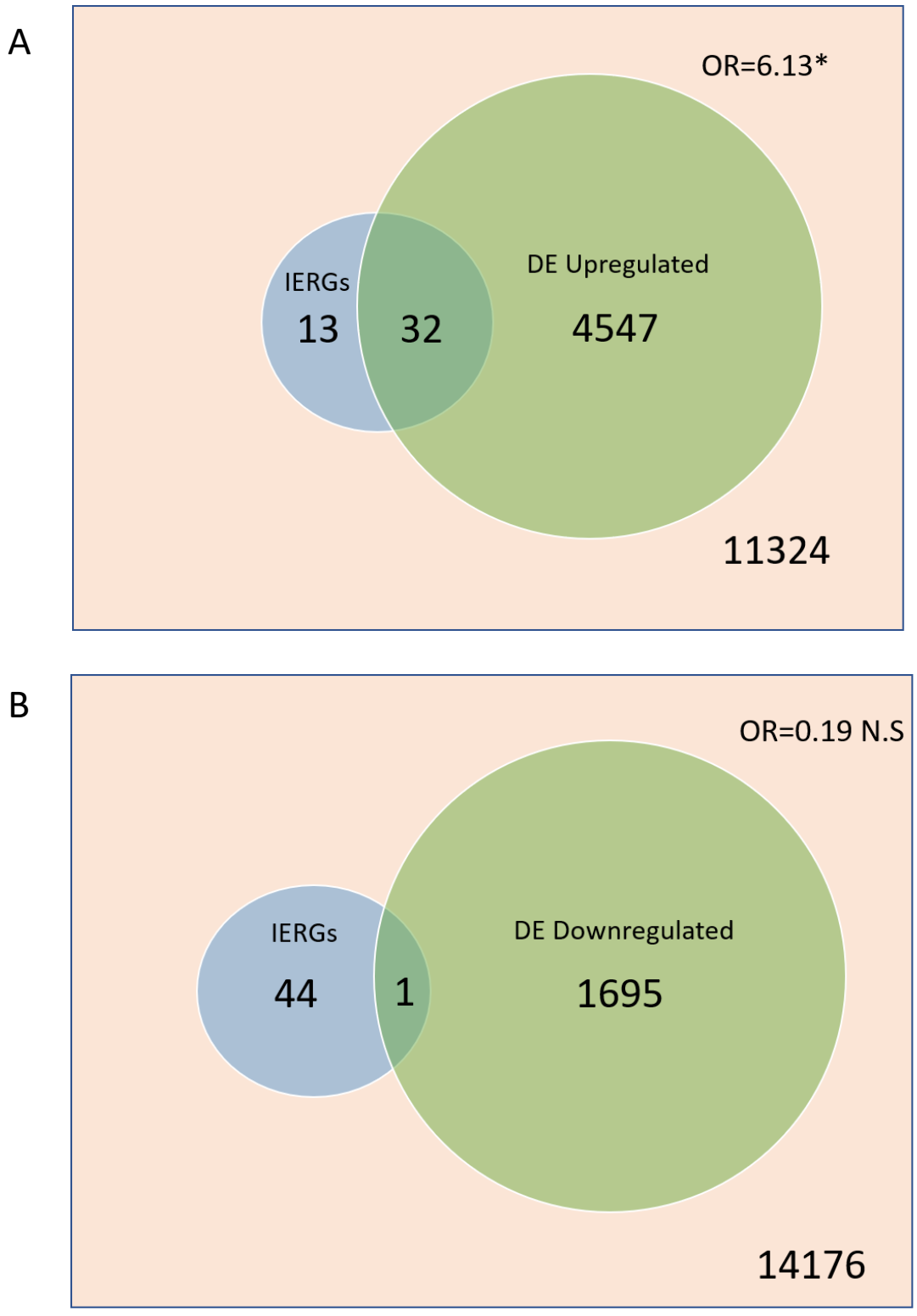


Fig.4.5 Venn diagram showing overlap of immediate early response genes (derived from Tullai et al. 2007) with genes a) differentially upregulated and b) differentially downregulated upon H_2O_2 treatment in WT cells relative to untreated WT cells. The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. DE = Differentially expressed, IERGs=Immediate early response genes, N.S = Non-significant, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

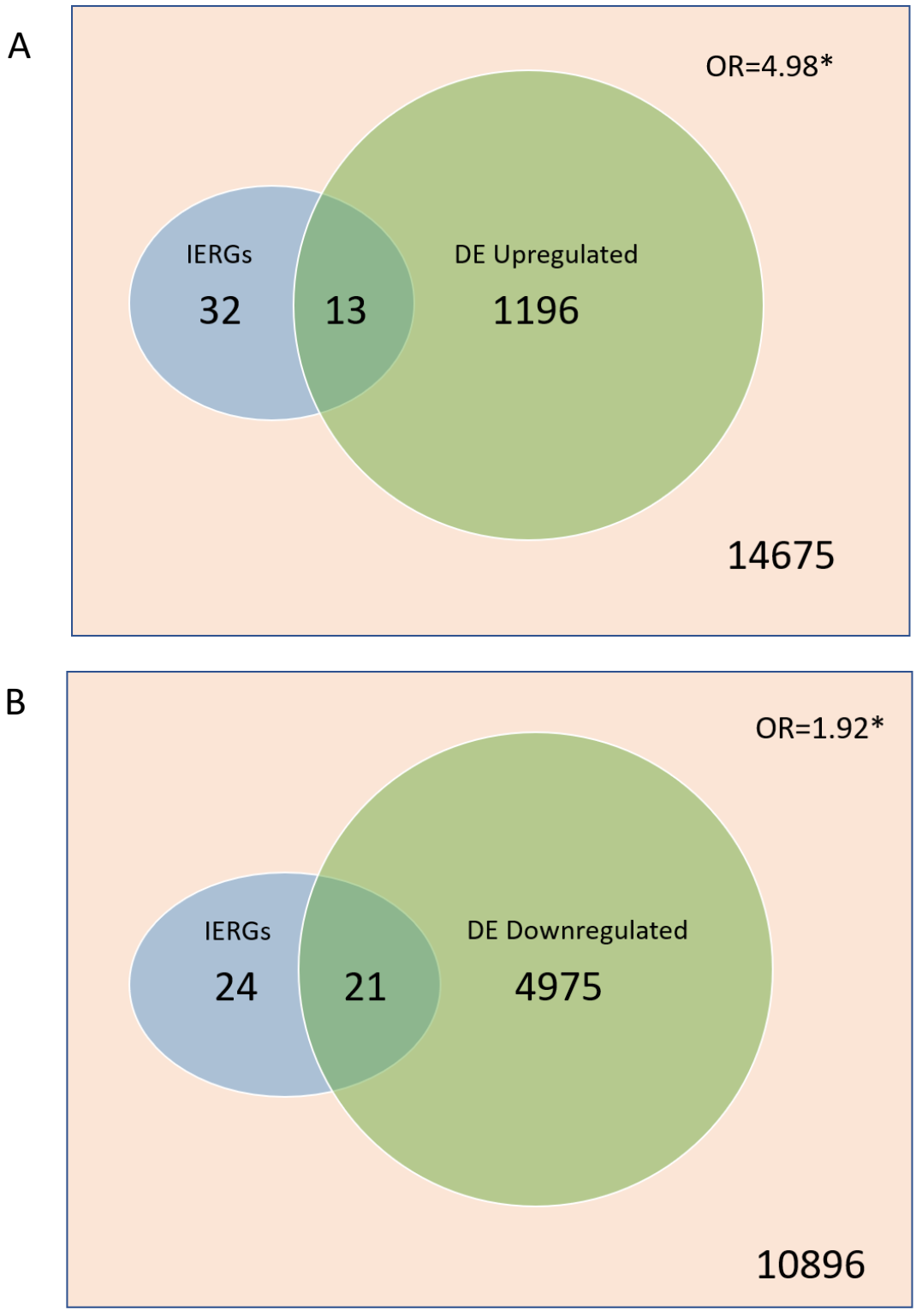


Fig.4.6 Venn diagram showing overlap of immediate early response genes (derived from Tullai et al. 2007) with genes a) differentially upregulated and b) differentially downregulated upon NuMA KD H_2O_2 treatment relative to H_2O_2 treated WT cells. The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. DE = Differentially expressed, IERGs=Immediate early response genes, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

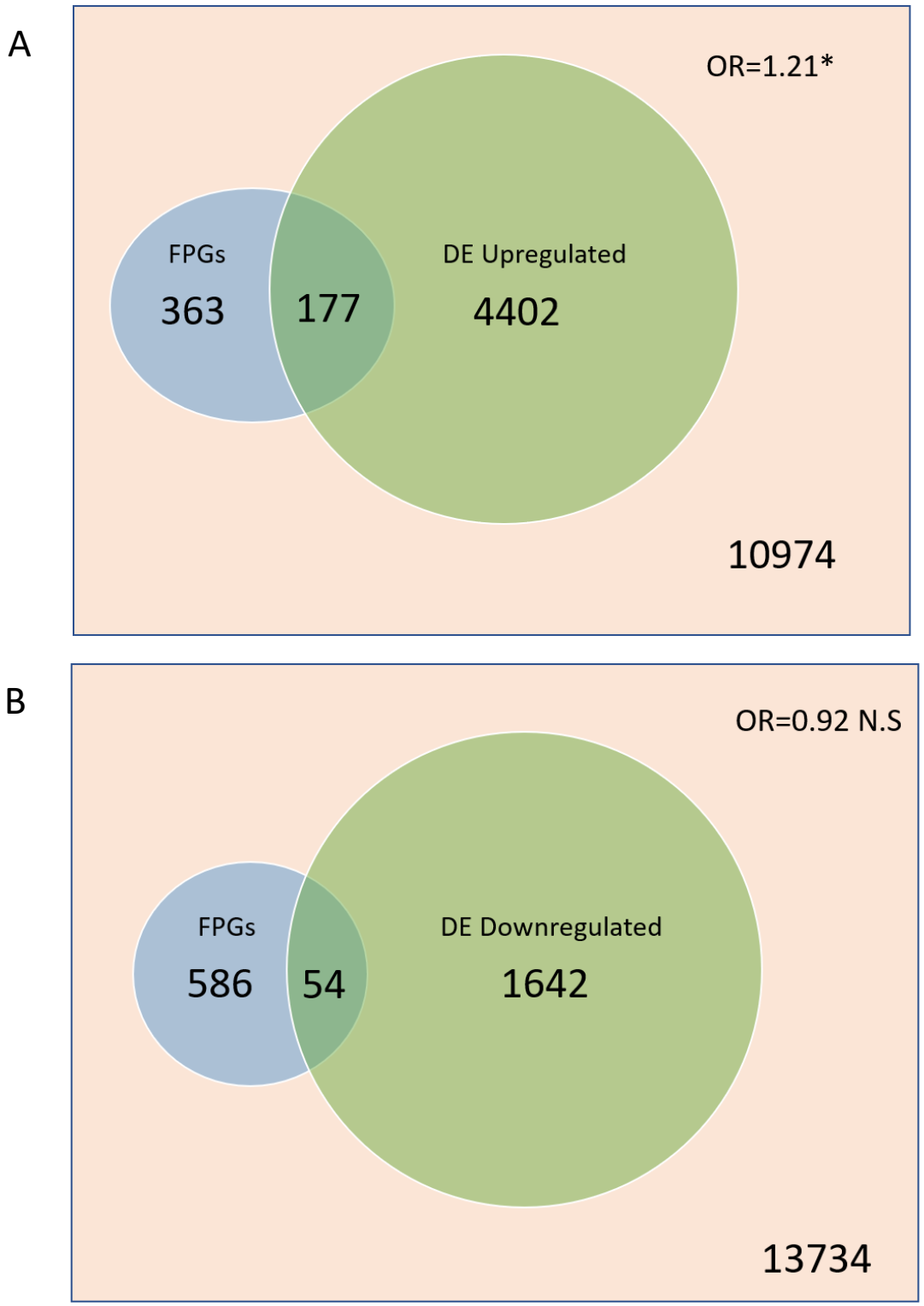


Fig.4.7 Venn diagram showing overlap of genes containing fragile promoters with genes a) differentially downregulated and b) differentially upregulated upon H_2O_2 treatment in WT cells relative to untreated WT cells. The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. DE = Differentially expressed, FPGs = Fragile promoter genes, N.S = Non-significant, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01. * <= 0.001.**

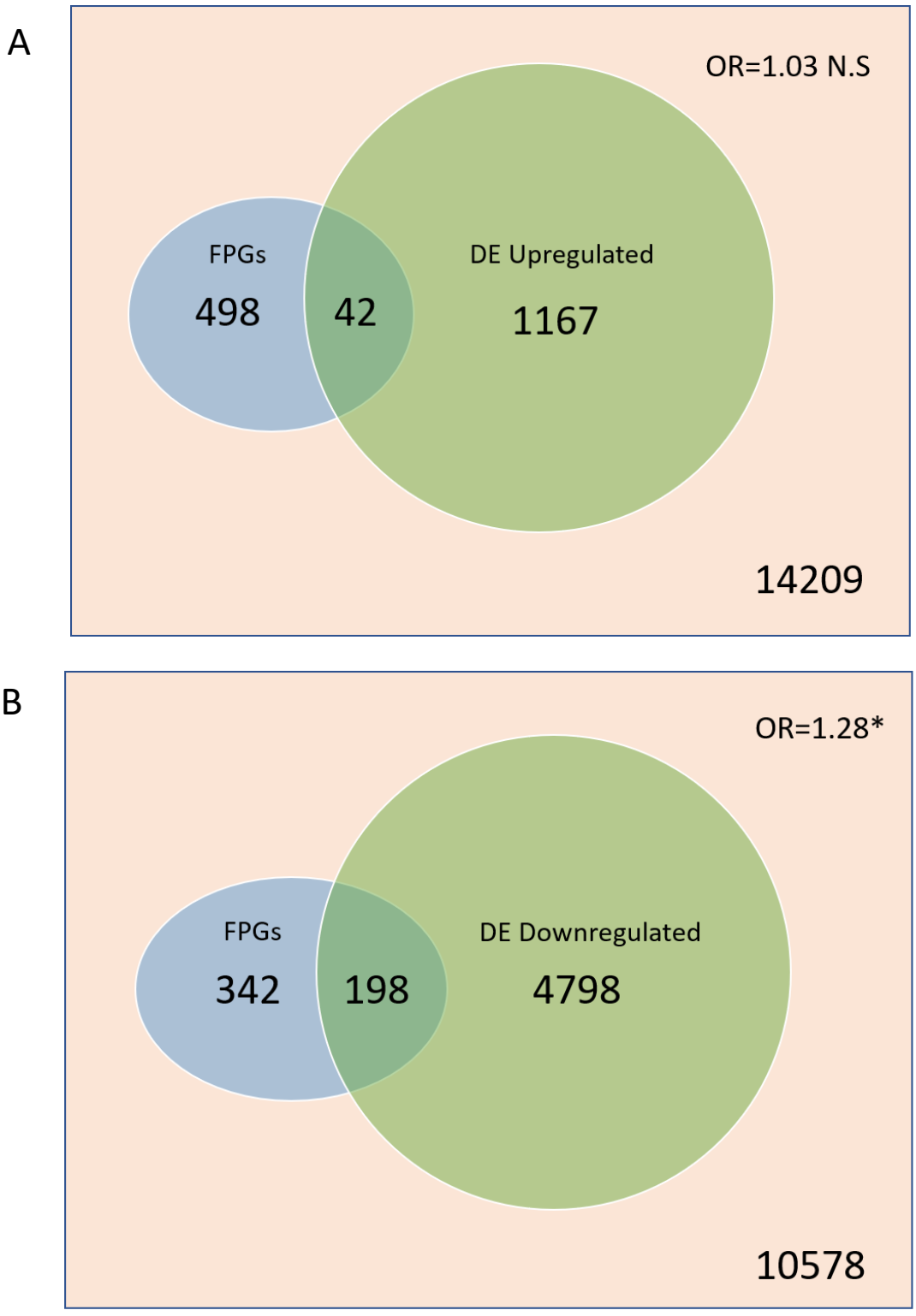
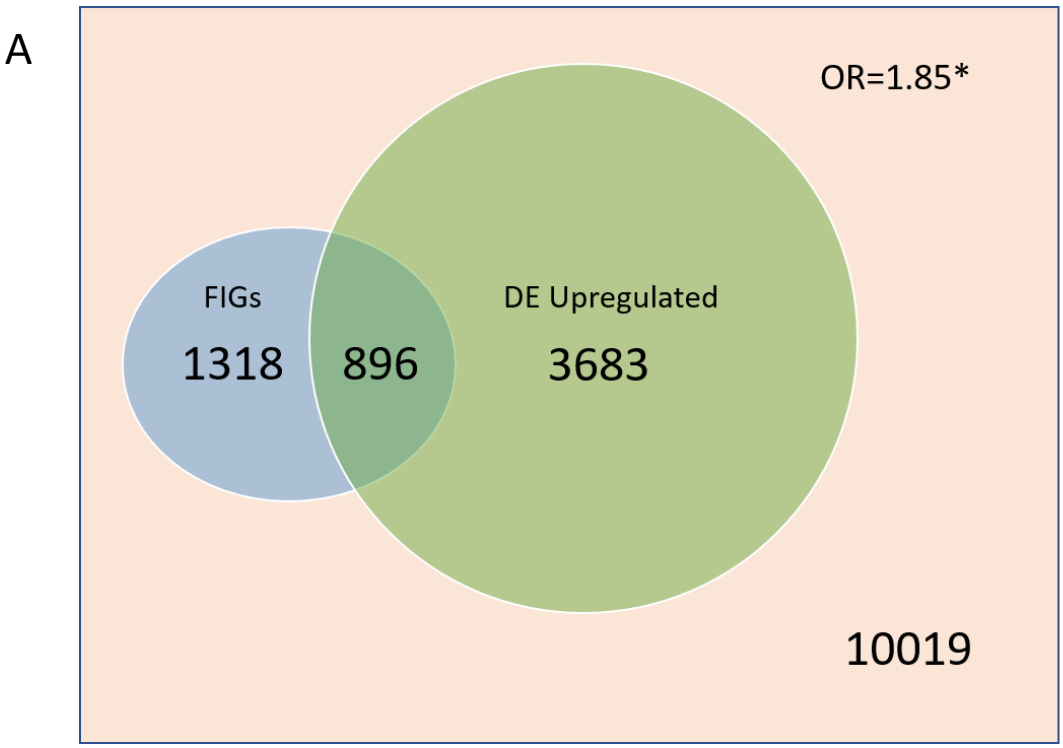


Fig.4.8 Venn diagram showing overlap of paused genes with genes a) differentially upregulated and b) differentially downregulated upon NuMA KD H_2O_2 treatment relative to H_2O_2 treated WT cells. The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. DE = Differentially expressed, FPGs = Fragile promoter genes, N.S = Non-significant, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**



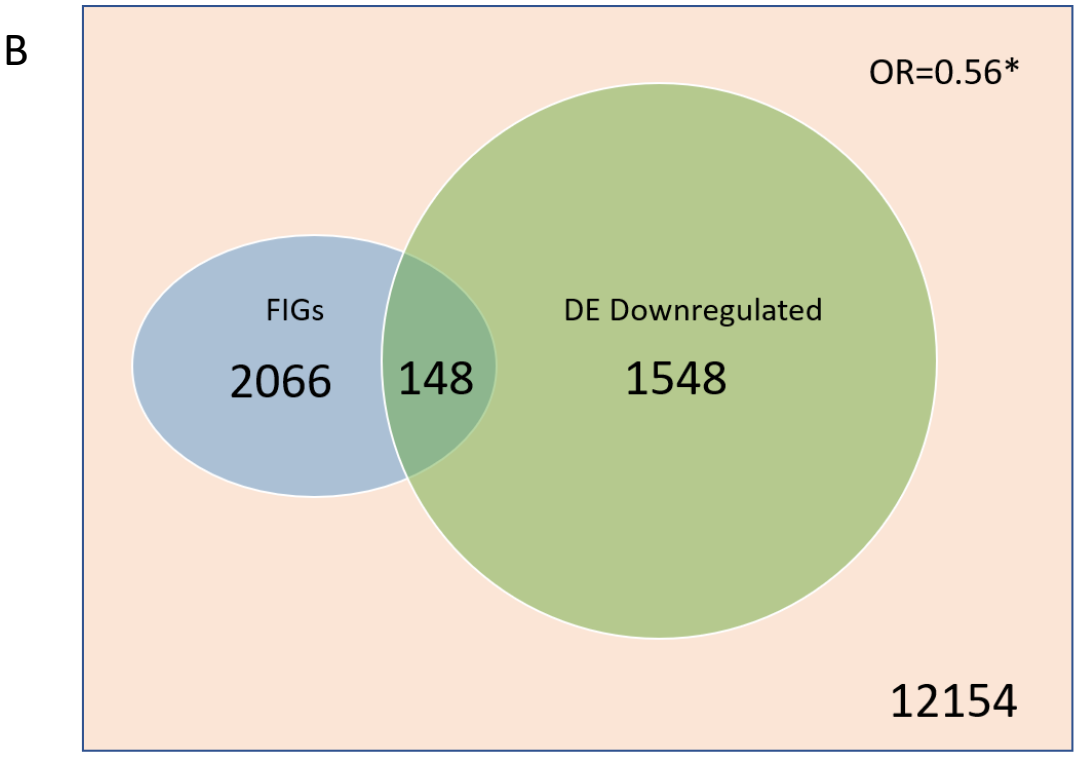


Fig.4.9 Venn diagram showing overlap of genes containing fragile introns with genes a) differentially upregulated and b) differentially downregulated upon H_2O_2 treatment in WT cells relative to untreated WT cells. The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. DE = Differentially expressed, FIGs = Fragile intron genes, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

introns, there was a significant underrepresentation within upregulated genes (OR=0.68) yet an enrichment within downregulated genes (OR=1.83) **(Fig.4.10a,b)**. All in all, the general pattern for these gene categories is a stronger enrichment amongst genes upregulated in H_2O_2 WT cells compared the enrichment for the genes downregulated in untreated WT cells which is non-significant or indeed instead a depletion. The opposite is true for the H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT comparison, where the gene set showed less enrichment in the upregulated than the downregulated genes. The two exceptions to this were the enrichment for paused genes in genes differentially expressed upon H_2O_2 treatment in WT cells and IERGs within genes differentially expressed in H_2O_2 treated NuMA KD cells, although in each case the gene set (upregulated or downregulated) expected to have the greater enrichment contained a greater total number of genes than the opposite set.

4.2.3 Identification of set of genes that switch from upregulated to downregulated in H_2O_2 treated cells upon knockdown of NuMA

Another major question was whether genes that are differentially expressed upon H_2O_2 treatment relative to untreated cells have their expression affected by the absence of NuMA, a topic rendered even more interesting by our existing results. Based on **Fig.4.2**, it was noted that a far higher number of genes are upregulated upon exposure to H_2O_2 than are downregulated. Interestingly, the reverse of this was true when comparing $NuMA_{KD}$ and WT cells both subject to H_2O_2 treatment: many more genes showed downregulation than upregulation. Additionally, the number of genes upregulated upon H_2O_2 treatment was comparable to the number of genes downregulated upon NuMA KD in H_2O_2 treated cells, and vice versa. The patterns of enrichment identified across the different gene categories also matched this observation. If the genes upregulated upon H_2O_2 treatment in WT cells were a similar set to those downregulated in H_2O_2+ $NuMA_{KD}$ cells relative to H_2O_2+ WT cells, this would suggest that their upregulation upon H_2O_2 treatment could be mediated by NuMA. Looking at the overlap between these sets, there was significant enrichment of H_2O_2+ WT vs H_2O_2- WT upregulated genes amongst H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT downregulated genes (OR=3.5). 51% of H_2O_2+ WT vs H_2O_2- WT upregulated genes were present amongst H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT downregulated genes, and 47% were found in the opposite direction, altogether representing a group of 2358 genes **(Fig.4.11)**. This identifies a set of genes whose increased expression upon the induction of oxidative

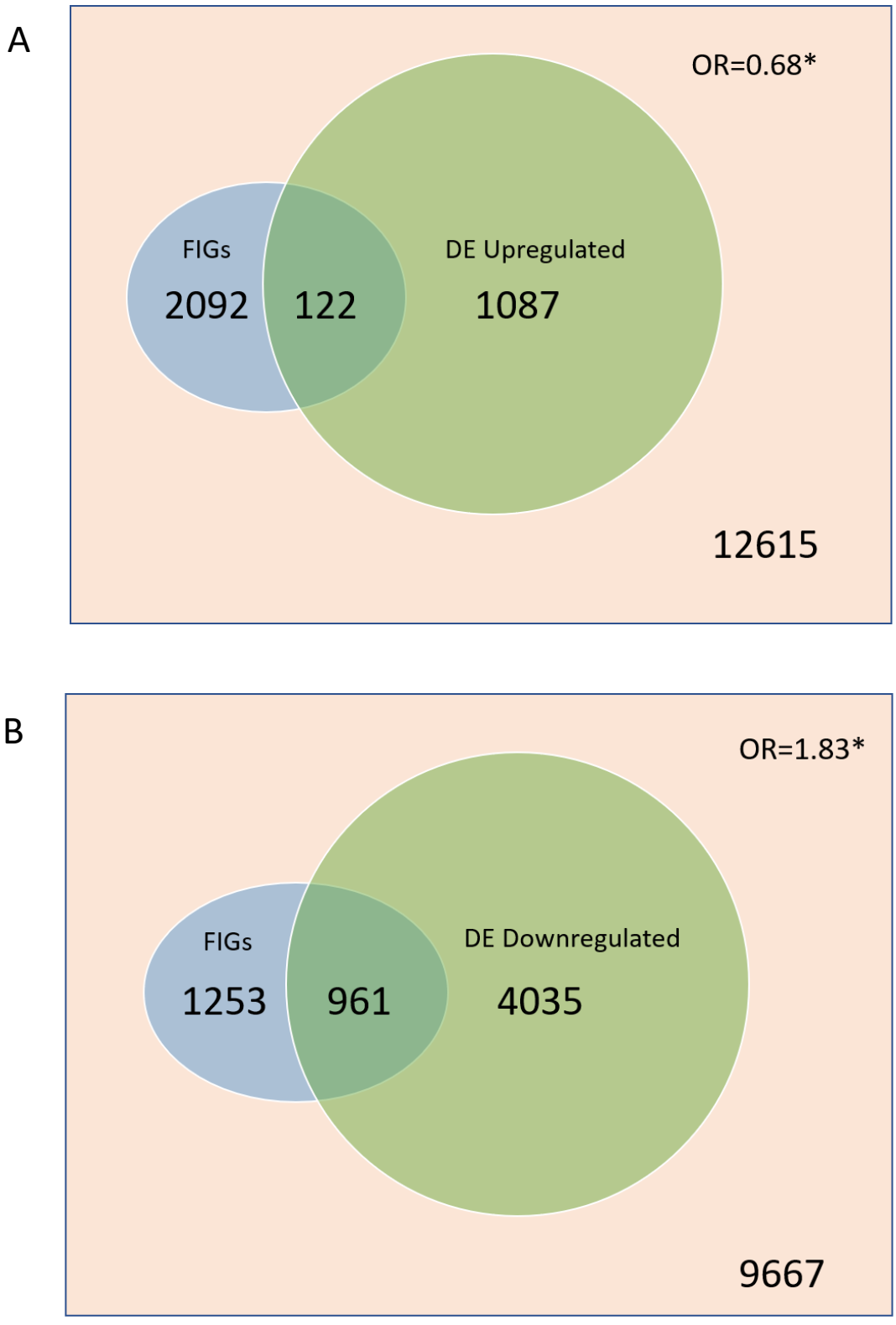


Fig.4.10 Venn diagram showing overlap of genes containing fragile introns with genes a) differentially upregulated and b) differentially downregulated upon NuMA KD H_2O_2 treatment relative to H_2O_2 treated WT cells. The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. DE = Differentially expressed, FPGs = Fragile promoter genes, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

A

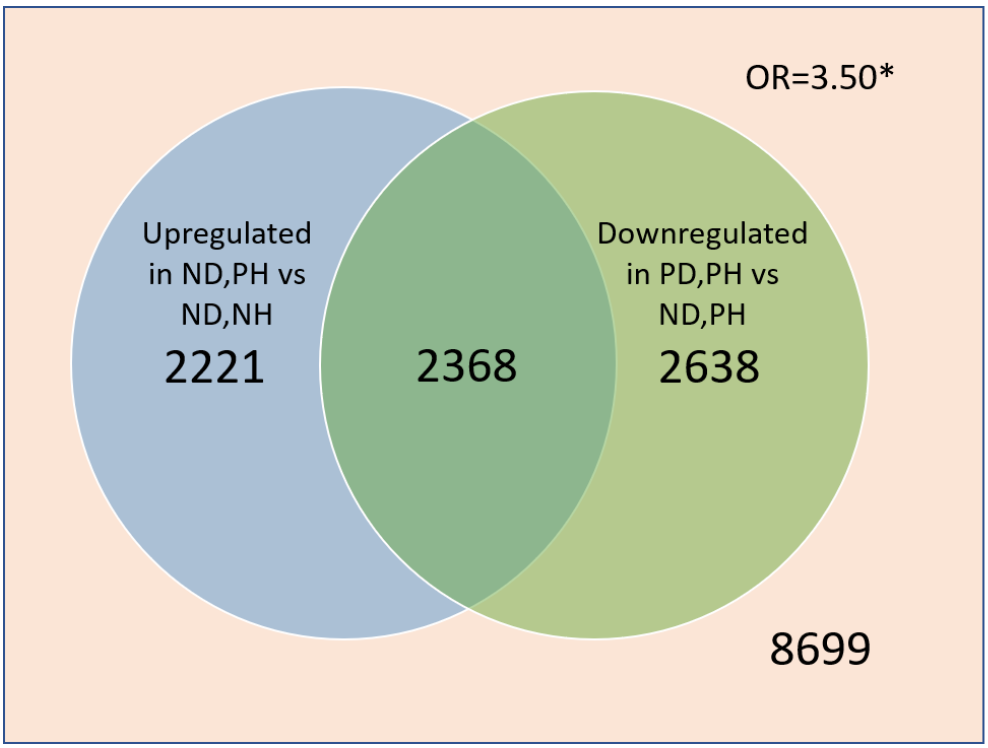


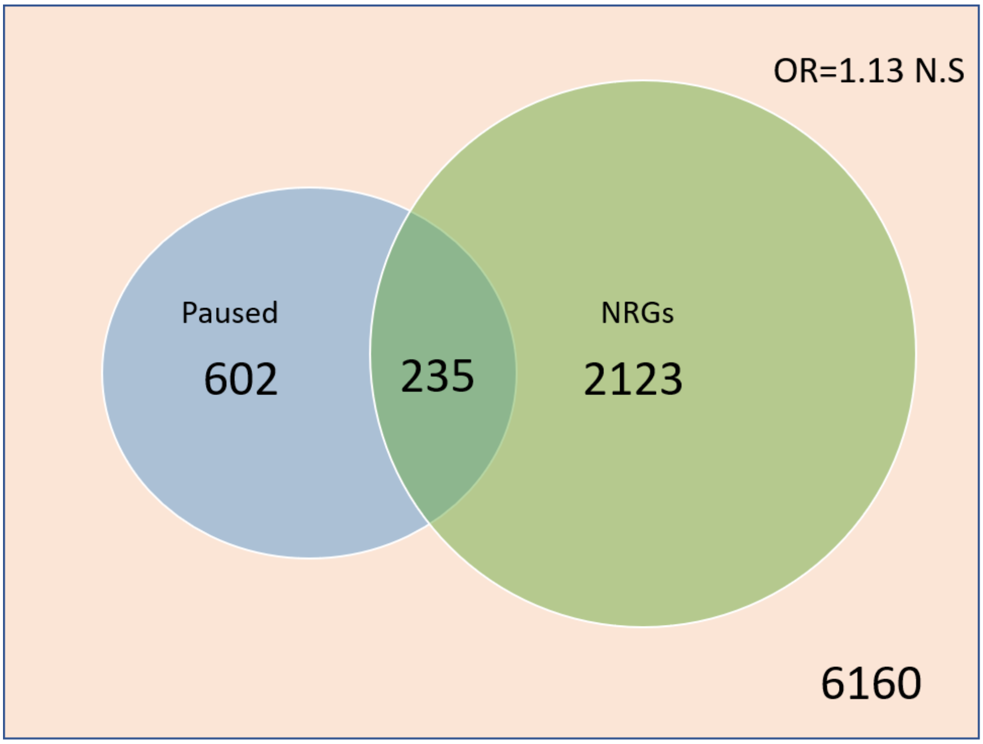
Fig.4.11 a) Venn diagram showing overlap of genes differentially upregulated upon H_2O_2 treatment in WT cells (ND,PH) relative to untreated WT cells (ND,NH) with genes differentially downregulated upon NuMA KD H_2O_2 treatment (PD,PH) relative to H_2O_2 treated WT cells (ND,PH). The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

damage is dependent upon the presence of NuMA, a group henceforth referred to as NuMA regulated genes (NRGs). GO analysis on this set of genes did not reveal any immediately relevant biology.

4.2.4 NRGs are enriched for IER and fragile intron genes, but not paused or fragile promoter genes

Following the identification of NRGs, they were tested for enrichment of the gene categories previously identified as showing enrichment amongst the sets of differentially expressed genes. Paused genes showed no significant enrichment amongst NRGs relative to a combined set of genes differentially expressed in either comparison **(Fig.4.12a)**. This might be because NuMA only regulates the expression of a set of the most highly paused genes, so tests were repeated with a high pausing index subset of genes, defined as those with a pausing index of over 20, but again there was no enrichment **(Fig.4.12b)**. However, when the total set of expressed genes was used as the background set, the results did show significance **(Fig.4.13a,b)**. This and previous results indicate that being paused is a predictor for being differentially expressed in at least one of the two comparisons, either H_2O_2 treatment in WT cells or NuMA KD in H_2O_2 treated cells and therefore NRGs being enriched for paused genes is likely a function of them being differentially expressed upon H_2O_2 treatment and not because they are regulated by NuMA. On the other hand, although there was a substantial decrease in OR from 3.19 to 1.84 when switching from all genes to differentially expressed genes as the background set, IERGs were still significantly enriched amongst NRGs and thus are overrepresented in excess of the enrichment within differentially expressed genes **(Fig.4.14a,b)**. The pattern for genes containing fragile promoters was similar to that for paused genes, although it should be noted this enrichment was very weak even when using all genes as the background **(Fig.4.15a,b)**. Similarly to IERGs, fragile intron gene enrichment was observed and additionally, the decrease in the OR was less dramatic than for IERGs, 2.06 down to 1.75 **(Fig.4.16a,b)**. Although both these sets of genes are enriched in NRGs, they make up only a small proportion of the total 2358, 532 for fragile intron genes and only 16 for IERGs. However, FIGs constitute a substantially larger proportion and so going forward these genes represent a more informative subset of genes when considering NuMA regulated gene expression. This demonstrates that whilst these two categories of genes may in be regulated by NuMA in some fashion, they do not

A



B

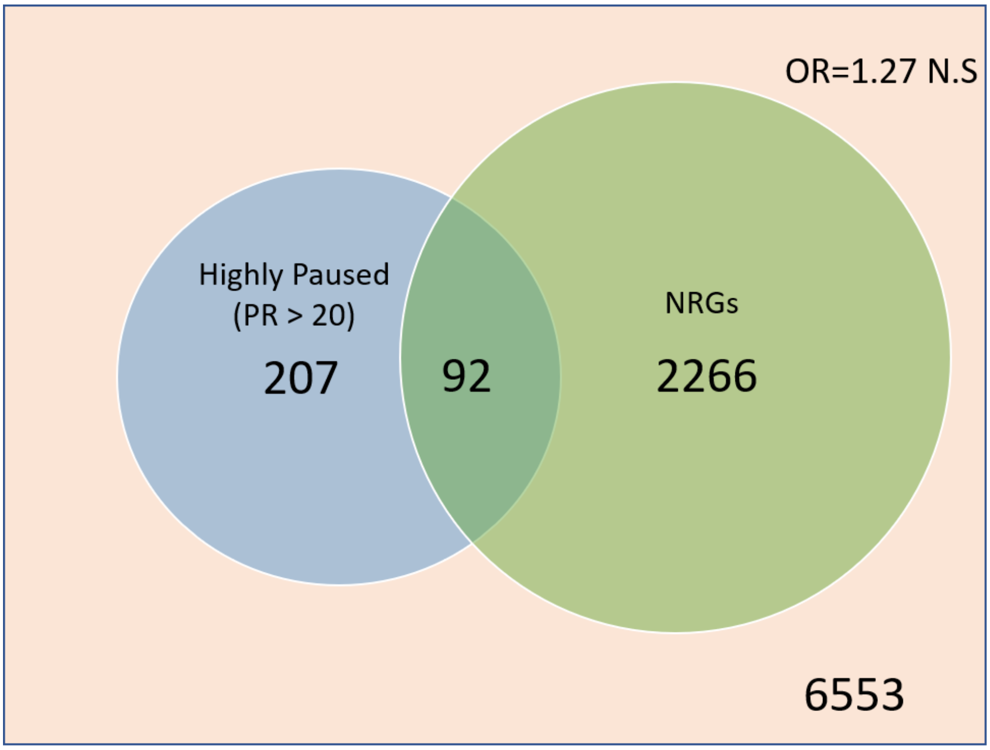


Fig.4.12 Venn diagram showing overlap of a) paused and b) highly paused genes (pausing index/ratio > 20) with NuMA regulated genes. The background gene set is all genes differentially expressed across the H_2O_2+ WT vs H_2O_2- WT and H_2O_2+ NuMA KD vs H_2O_2+ WT comparisons with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. PR = pausing ratio, NRGs = NuMA regulated genes, N.S = Non-significant, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

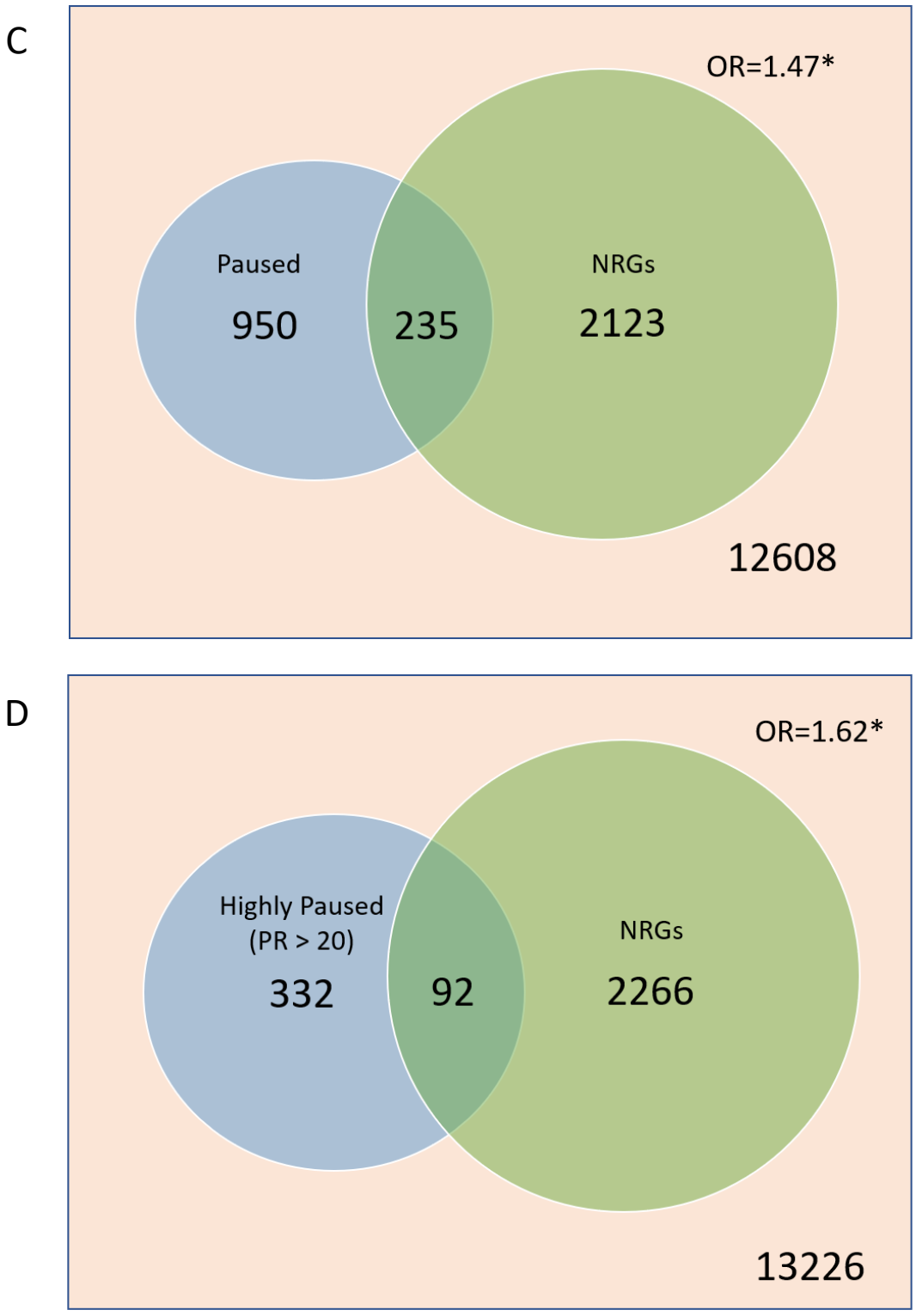


Fig.4.13 Venn diagram showing overlap of a) paused and b) highly paused genes (pausing index > 20) with NuMA regulated genes. The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. PR = pausing ratio, NRGs = NuMA regulated genes, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

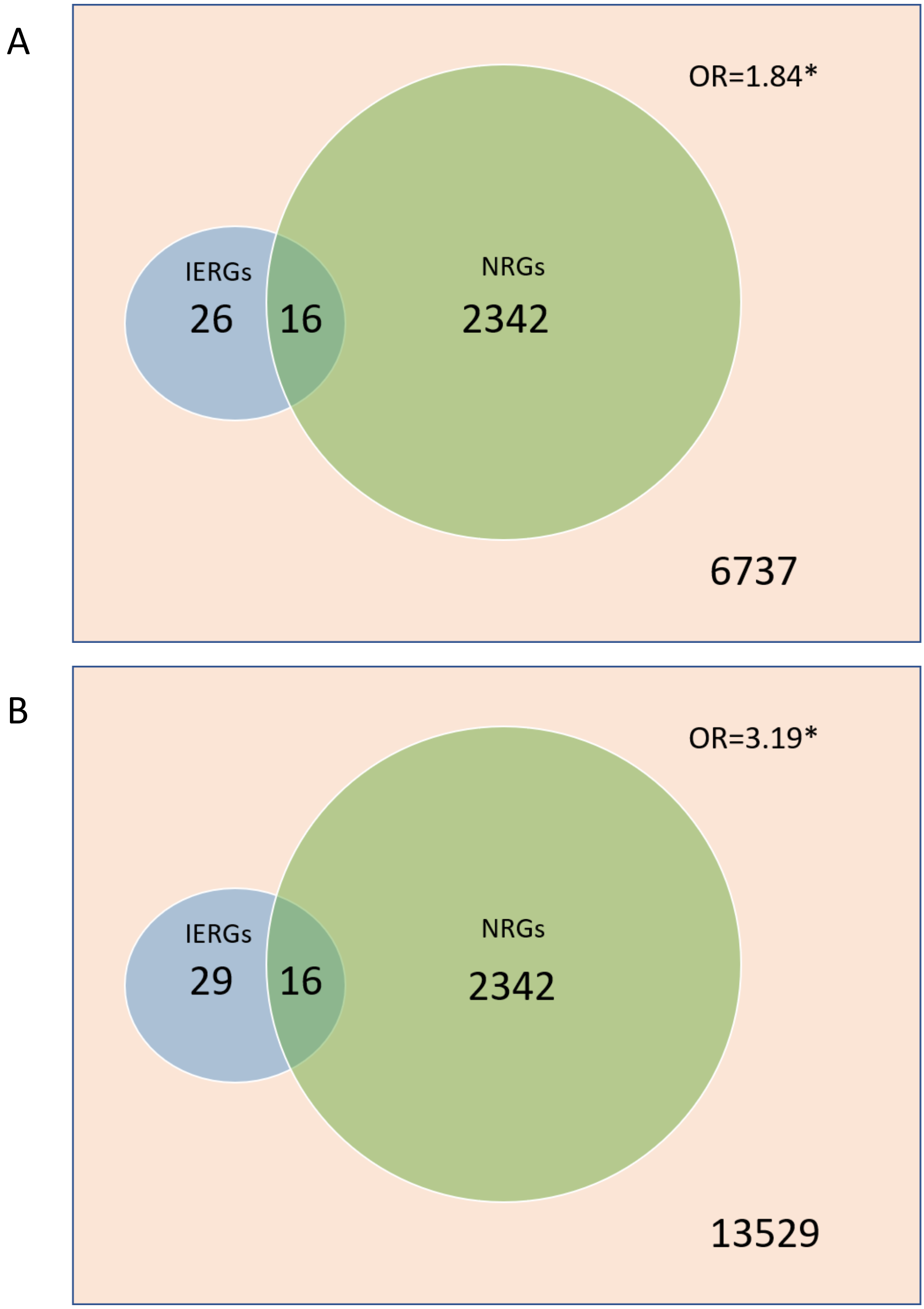


Fig.4.14 Venn diagram showing overlap of immediate early response genes (derived from Tullai et al. 2007) with NuMA regulated genes. a) The background gene set is all genes differentially expressed across the H_2O_2+ WT vs H_2O_2- WT and H_2O_2+ NuMA KD vs H_2O_2+ WT comparisons with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. b) The background gene set is all genes with a TPM value of 1 or greater in both replicates of at least one condition in the 4sU-seq data Significance determined by Fisher's exact test. IERGs = Immediate early response genes, NRGs = NuMA regulated genes, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

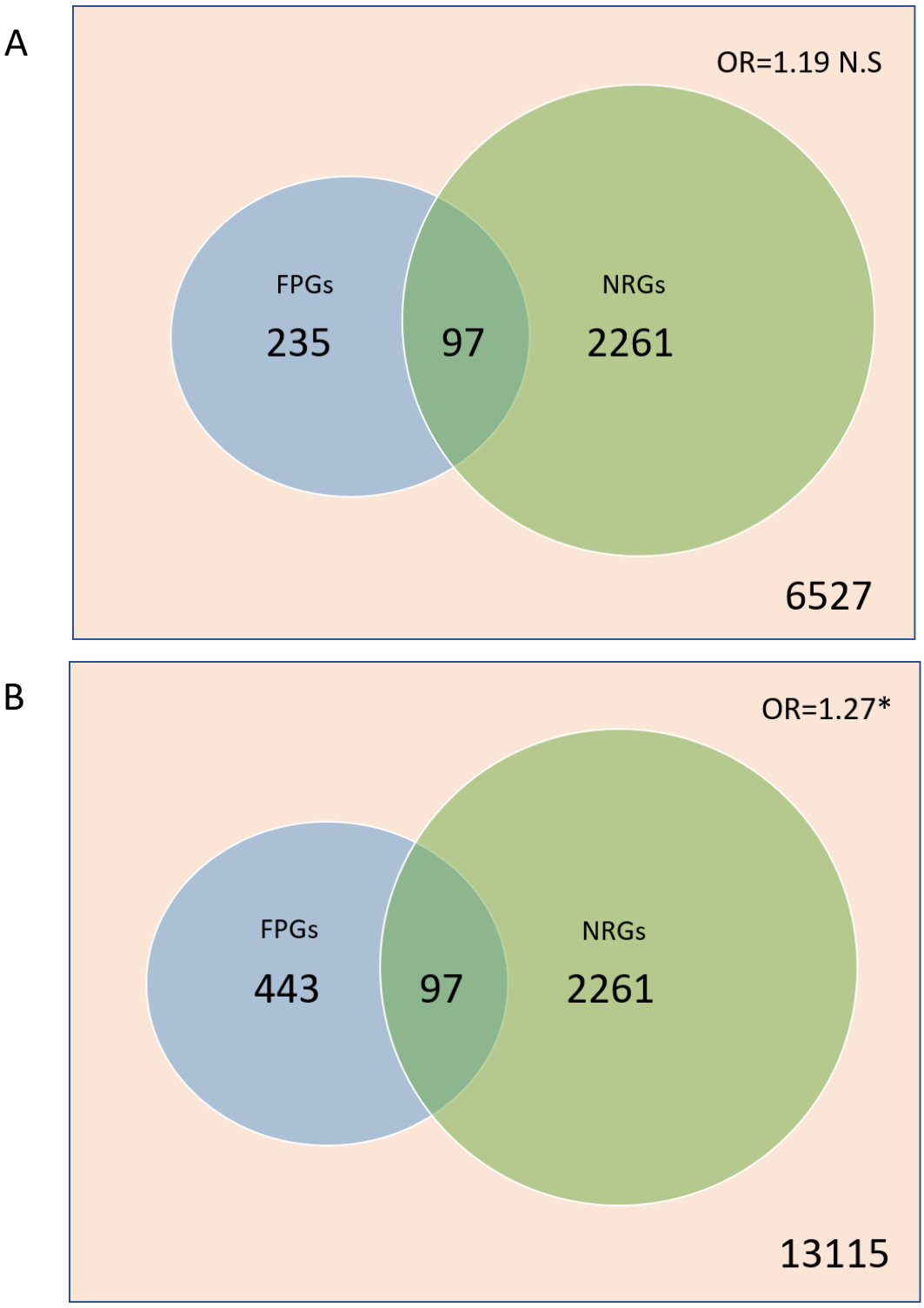


Fig.4.15 Venn diagram showing overlap of genes containing fragile promoters with NUMA regulated genes. a) The background gene set is all genes differentially expressed across the H_2O_2+ WT vs H_2O_2- WT and H_2O_2+ NuMA KD vs H_2O_2+ WT comparisons with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. b) The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data Significance determined by Fisher's exact test. FPGs = Fragile promoter genes, NRGs = NuMA regulated genes, N.S = Non-significant, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

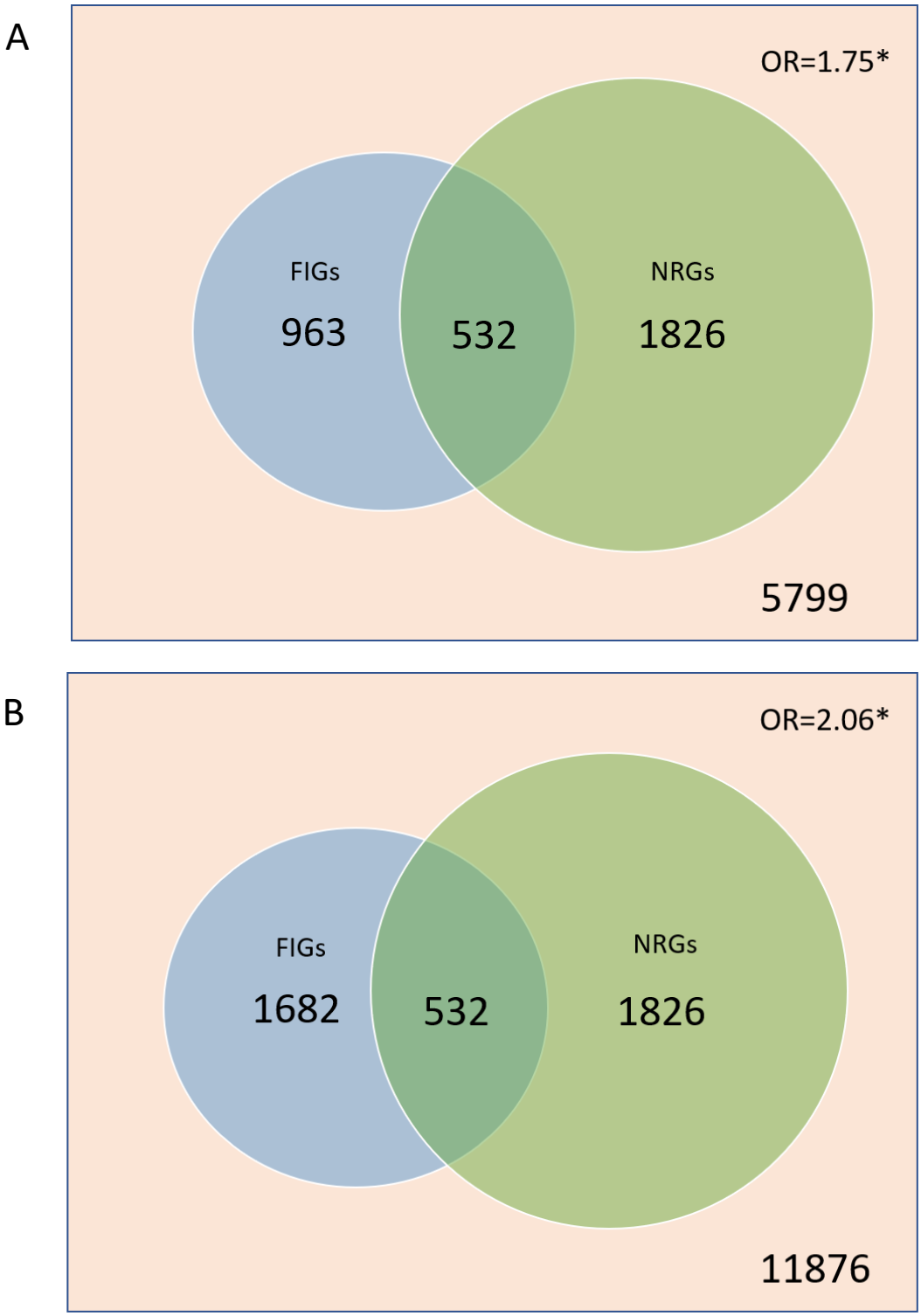


Fig.4.16 Venn diagram showing overlap of genes containing fragile introns with NUMA regulated genes. a) The background gene set is all genes differentially expressed across the H_2O_2+ WT vs H_2O_2- WT and H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT comparisons with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. b) The background gene set is all genes with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. FIGs = Fragile intron genes, NRGs = NuMA regulated genes, OR = Odds ratio. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

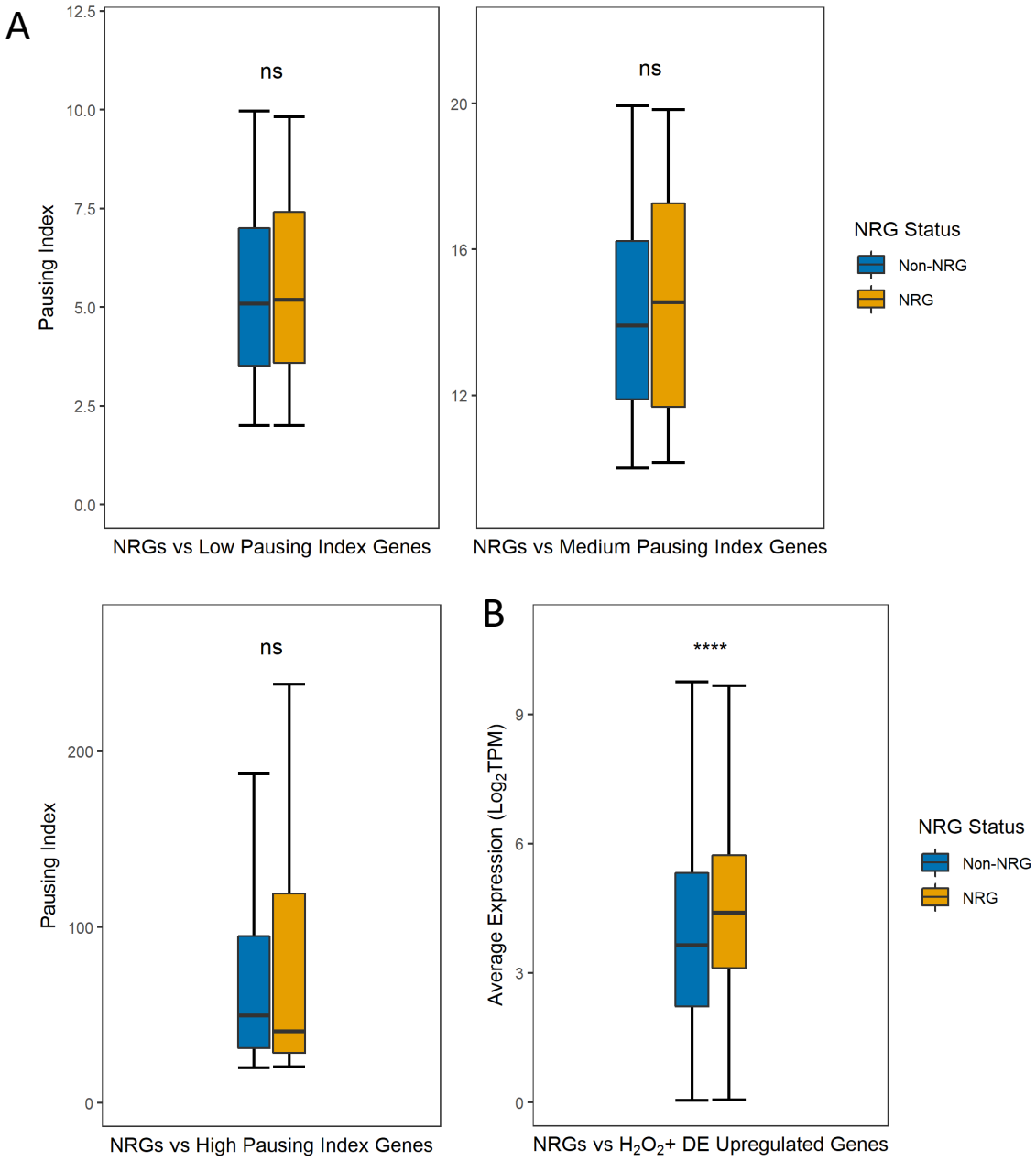
represent the full extent of NuMA mediated gene expression.

4.2.5 NRGs are more highly expressed upon H_2O_2 treatment and enriched for gene length markers relative to Non-NRGs differentially upregulated upon H_2O_2 treatment

Next we assessed whether NRGs differed significantly from genes differentially upregulated upon H_2O_2 treatment across several metrics. The rationale behind this was to investigate whether NuMA might be required to coordinate the expression of genes with certain attributes, for example, longer genes, upon the induction of DNA damage. NRGs did not show a significantly higher pausing index, even when the sets of genes being compared were split apart according to whether they showed a low, medium or high level of pausing **(Fig.4.17a)**. However, they had a significantly higher expression on average **(Fig.4.17b)**. NRGs were also indirectly associated with an increased gene length. **(Fig.4.17c)** shows gene length against log fold-change for the genes differentially upregulated in the H_2O_2+ WT vs H_2O_2- WT comparison, including NRGs, but the logFCs are taken from the H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT analysis. A minor inverse correlation is observed (R=-0.15), meaning genes with low logFCs upon the KD of NuMA in H_2O_2 treated cells tend to have slightly higher gene lengths. We know that NRGs are the only genes within this subset that show significant downregulation in the H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT comparison, and so we can therefore infer that NRGs tend to be slightly longer than their differentially upregulated counterparts. NRGs are also significantly longer than other genes that were called as differentially expressed across the two comparisons, a comparison that accounts for the gene length-differential expression bias **(Fig.4.17d)**. NRGs also have a higher number of introns than other genes doubly called as differentially expressed **(Fig.4.17e)**, and number of introns show a weak inverse correlation (R=-0.14) **(Fig.4.17f)**. However, this could be a function of them being longer or vice versa, as intron number does correlate strongly with gene length **(Fig.4.17g)**.

4.2.6 NuMA is enriched across gene promoters under normal conditions but is lost upon H_2O_2 treatment

Having identified NRGs and demonstrated their enrichment for fragile intron genes and IERGs, we wanted to look at NuMA occupancy across both genic DNA generally and across gene categories already identified as interesting, and how this might change upon the


A
ns
ns
Pausing Index
12.5
10.0
7.5
5.0
2.5
0.0
20
16
12
NRGs vs Low Pausing Index Genes
NRGs vs Medium Pausing Index Genes
NRG Status
Non-NRG
NRG
B
ns
****
200
100
0
9
6
3
0
Average Expression (Log2TPM)
NRGs vs High Pausing Index Genes
NRGs vs H2O2+ DE Upregulated Genes
C
R = −0.15 , p < 2.2e-16
2.5
0.0
-2.5
Log2 FC in H2O2+ NUMA1 KD vs H2O2+ WT
0
1
10
100
1,000
Gene Length (KB) (Log10 Scale)

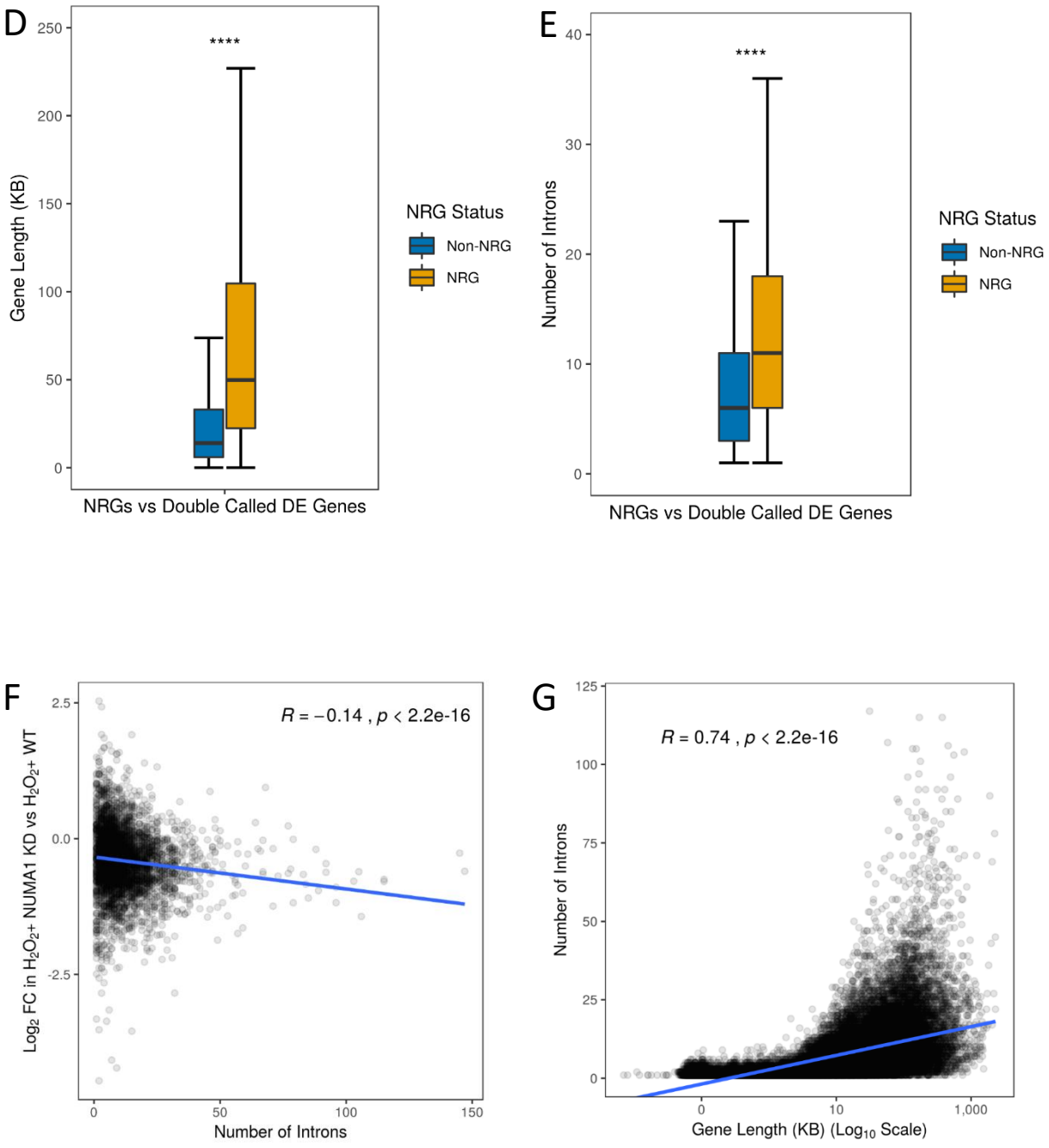

D
Gene Length (KB)
250
200
150
100
50
0
****
NRGs vs Double Called DE Genes
NRG Status
Non-NRG
NRG
E
Number of Introns
40
30
20
10
0
****
NRGs vs Double Called DE Genes
NRG Status
Non-NRG
NRG
F
Log2 FC in H2O2+ NUMA1 KD vs H2O2+ WT
2.5
0.0
-2.5
R = −0.14 , p < 2.2e-16
0
50
100
150
Number of Introns
G
Number of Introns
125
100
75
50
25
0
R = 0.74 , p < 2.2e-16
0
10
1,000
Gene Length (KB) (Log10 Scale)

Fig.4.17 a) Boxplots showing the distribution of pausing indices for NRGs and expressed genes. NRGs and expressed genes are split into those with a low pausing index (top left), medium pausing index (top right) and high pausing index (bottom).

b) Boxplot showing the distribution of average expression levels in Log2 TPM for NRGs and genes differentially upregulated in the H_2O_2+ WT vs H_2O_2- WT comparison.

c) Scatterplot of Log2 fold changes in the H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT comparison for genes differentially upregulated in the H_2O_2+ WT vs H_2O_2- WT comparison against Log10 transformed gene length in kilobases.

d) Boxplot showing spread of gene lengths in kilobases for NRGs and genes also called as differentially expressed in both the H_2O_2+ WT vs H_2O_2- WT and H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT comparisons.

e) Boxplot showing distribution of number of introns for NRGs and genes also called as differentially expressed in both the H_2O_2+ WT vs H_2O_2- WT and H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT comparisons.

f) Scatterplot of Log2 fold changes in the H_2O_2+ $NuMA_{KD}$ vs H_2O_2+ WT comparison for genes differentially upregulated in the H_2O_2+ WT vs H_2O_2- WT comparison against number of introns.

g) Scatterplot of number of introns plotted against Log10 transformed gene length in kilobases for all genes obtained from the R biomaRt hsapiens ensemble 85 mart.

Significance for all boxplots was determined by Wilcoxon rank sum test. The correlation coefficient (R) was determined by Spearman's rank and significance calculated using the asymptotic t approximation for all scatterplots. ns = Non-significant. p-values : * <= 0.05, ** <=0.01, * <= 0.001.**

induction of damage. To this end, NuMA ChIP-seq was carried out in the absence and presence of H_2O_2 treatment by Arwa Abugable of the El-Khamisy lab of the University of Sheffield. Each condition had two replicates and when mapped each sample contained roughly 30 – 40 million mapped reads. It should be noted that this experiment was carried out in RPE cells, whereas the 4sU-seq utilised MRC-5 cells. This data was then used to generate metagene profiles across the transcription start site (TSS), transcription termination site (TTS) and gene body for all genes using both the NuMA ChIP-seq and the IgG controls, with each sample normalised by its number of mapped reads. **Fig.4.18a,b** shows that under normal conditions, NuMA is highly enriched within the promoters of genes, present in moderate but decreasing levels across the length of the gene body and depleted at the TTS. Upon treatment with H_2O_2, NuMA appears to leave both the promoter and gene body, such that the metagene trace for NuMA occupancy upon damage induction is indistinguishable from the trace for the IgG ChIP. However, there were technical problems with the ChIP upon H_2O_2 treatment, and so in subsequent analyses, we mainly focussed on the untreated samples.

4.2.7 NuMA shows increased occupancy within genes upregulated upon H_2O_2 treatment, NRGs, paused genes and genes containing fragile introns

As this was the trace over all genes within our selected annotation, the next step was to see whether this pattern of NuMA occupancy differed between genes generally and specific genes categories of interest. However, an inherent problem with attempting to do this is that reads are being aggregated over different numbers of features, so differences will be inevitable as long as the category of interest does not match the control category in terms of size. Therefore a pipeline was developed into which is passed a list of genes of interest along with a control geneset. After several pre-filtering steps, the control set of genes is randomly sampled to match the target geneset in size, and metagene profiles are generated over these two number matched sets of genes. As a proof of concept, a set of 2500 “target genes” were generated by randomly sampling our pool of 4sU-seq expressed genes and a control set of “non-target genes”, consisting of all the other expressed genes was also created and randomly down-sampled after pre-filtering to match the size of the “target gene” set. If our assumptions regarding this technique for number matched comparative metagenes were correct, we would expect to see very little difference between the

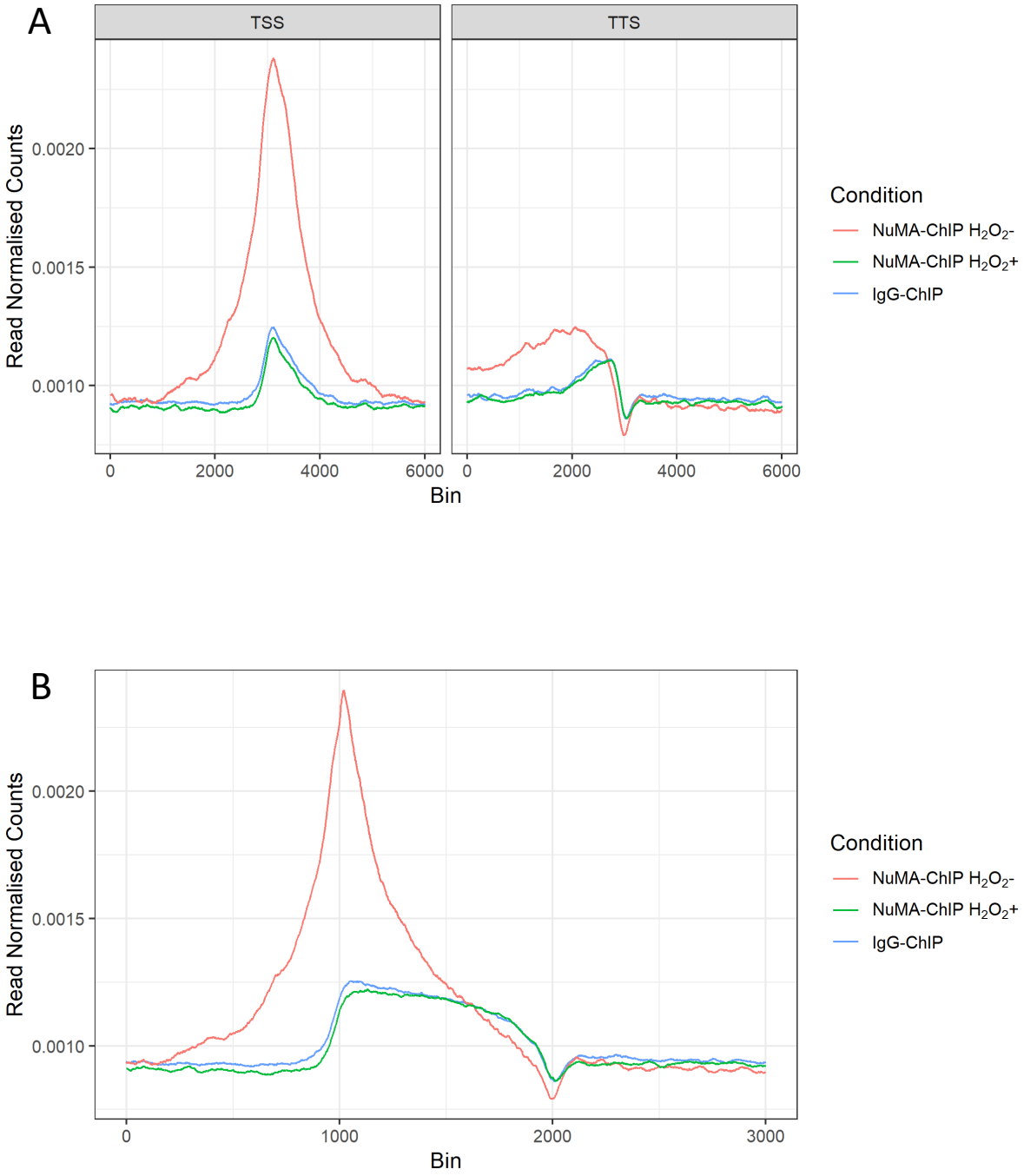
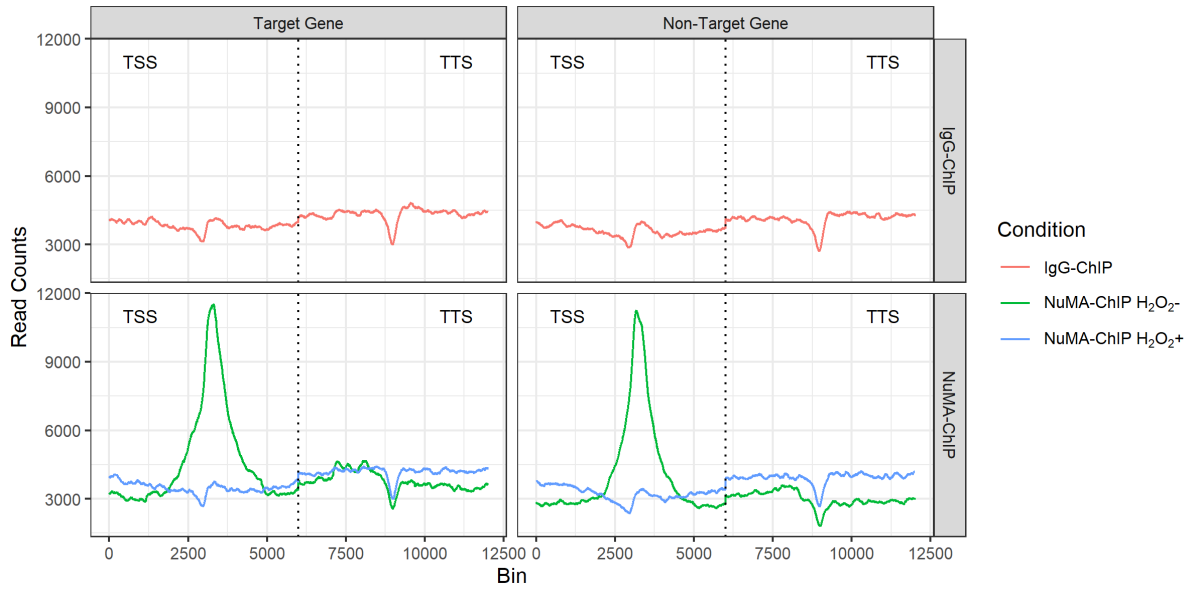


Fig.4.18 Metagene profiles over a) all TSSs, TTSs and b) gene bodies for NuMA H_2O_2-, NuMA H_2O_2+ and IgG ChIP-seq reads. Read counts were sample normalised by the total number of mapped reads within the given sample. Replicates for each condition were averaged within bins. For a) bin 3000 for TSS profile = TSS, bin 3000 for TTS profile = TTS, each bin is equal to one base. For b) bin 1000 =TSS, bin 2000 = TTS. TSS = Transcription start site, TTS = Transcription termination site.

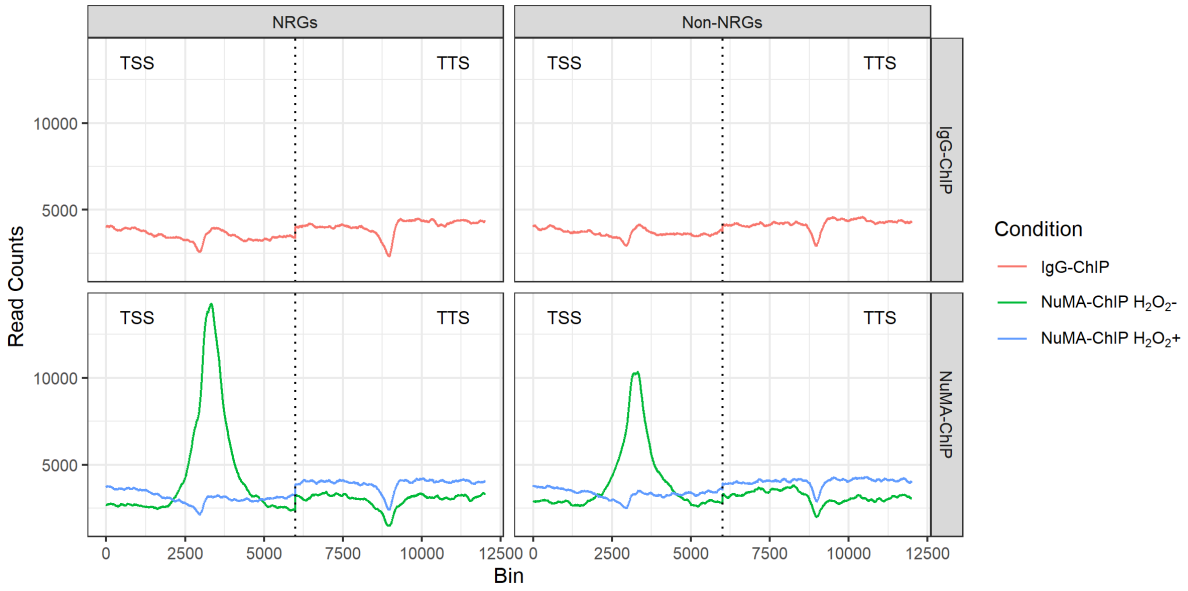
metagenes in the example, as the target set was generated randomly. Indeed, this is what the results in **Fig.4.19a** show, aside from a minor difference across the genes body and the TTS. The peaks across the promoter especially are almost indistinguishable from each other. As the promoter was the area showing least difference between the randomly sampled sets of genes, and was also the region showing the highest NuMA occupancy, it was decided to focus primarily on the promoter and hence the TSS/TTS profiles when analysing these number matched metagenes. Having shown that differences in NuMA promoter occupancy across number matched categories were likely to be non-random, five categories of interest were identified based on our previous work. These were NRGs, paused genes, immediate early response genes, fragile intron genes and finally fragile promoters, and metagenes were calculated over each of there in turn, along with traces over a set of randomly sampled number matched genes not belonging to their category. These graphs showed that NRGs, FIGs and paused genes were enriched for NuMA around their promoters relative to their matched non-category controls (**Fig.4.19b-d**). There was no such enrichment for FPGs (**Fig.4.19e**), and the while results for IERGs do appear to show increased occupancy (**Fig.4.19f**), this is less reliable due to the noisiness of the data, likely due to the small number of genes, 45, over which the metagene was calculated. Therefore, FPGs and IERGs were not taken forward for further metagenes analyses.

As H_2O_2 activated genes, NRGs, FIGs and paused genes are overlapping sets, it is unclear whether each of these factors is an independent determinant of NuMA binding. To test this, sets of genes were generated that showed only one of these properties, removing genes that were in the overlap between sets, along with number matched control sets. **Fig.4.19g** shows that despite the removal of gene categories enriched for promoter NuMA binding, H_2O_2 treatment upregulated genes had more NuMA across their promoters than the number matched controls. Since NRGs were significantly enriched for FPGs and paused genes and likely vice versa, it is possible the promoter enrichment of NuMA in any one of these three categories was being driven by genes from other categories and vice versa, and this relative overrepresentation of NuMA was not a specific feature of the assessed set of genes. Metagenes were generated along with number matched genes, but this time other enriched categories were removed from the FIGs and paused genes: H_2O_2 upregulated and paused genes for FIGs, H_2O_2 upregulated and FIGs for paused genes. Removing H_2O_2

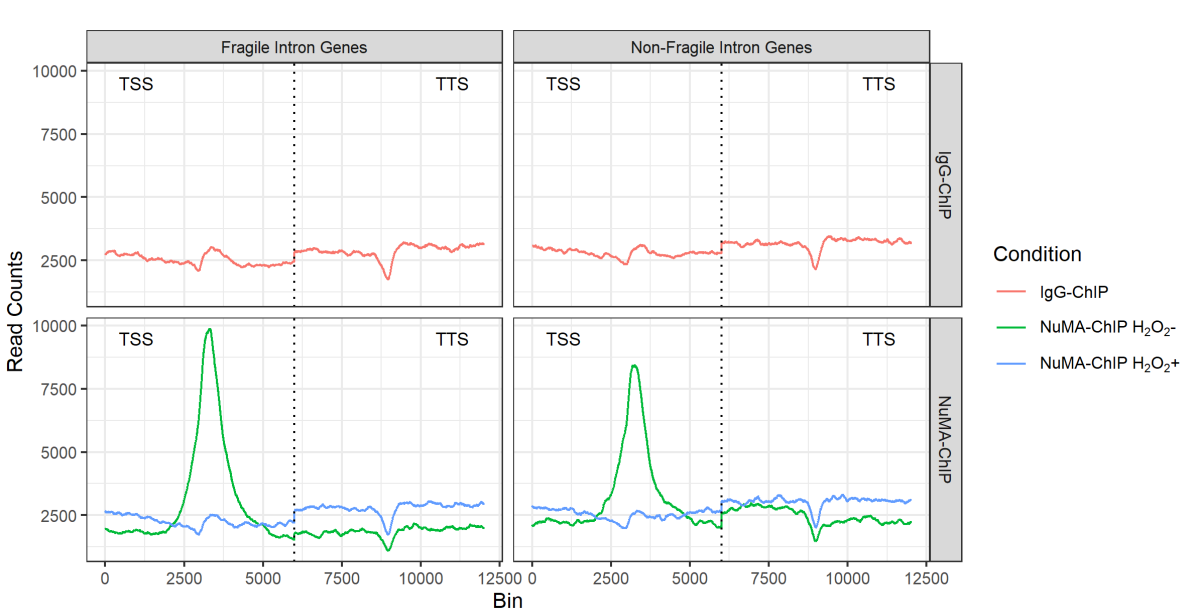
A


Target Gene
Non-Target Gene
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

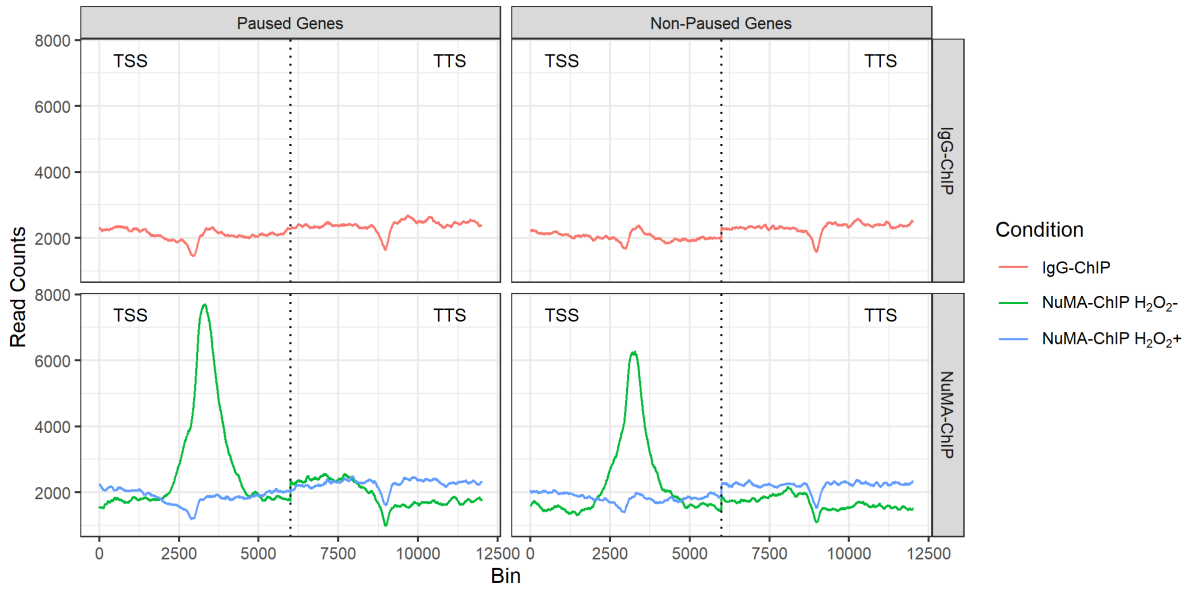
B


NRGs
Non-NRGs
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

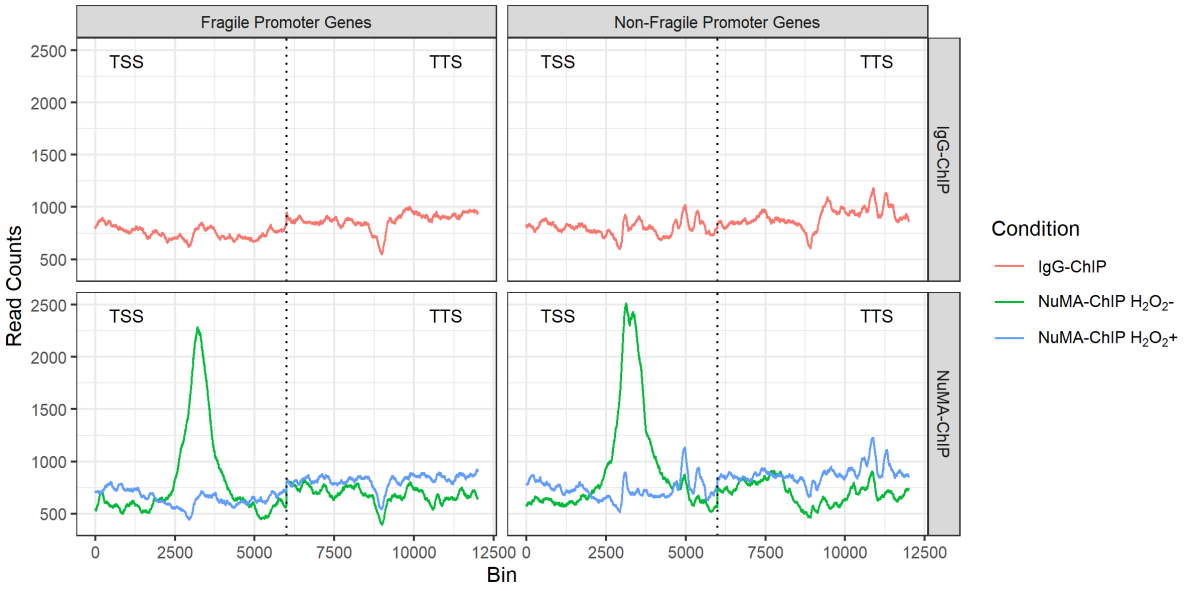
C


Fragile Intron Genes
Non-Fragile Intron Genes
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

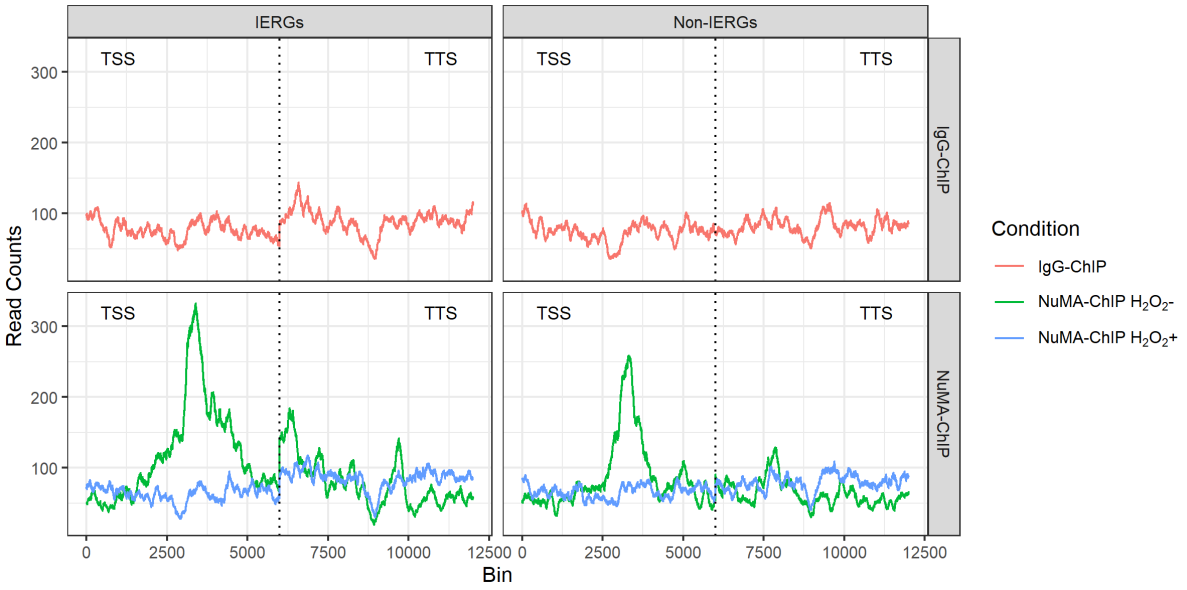
D


Paused Genes
Non-Paused Genes
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

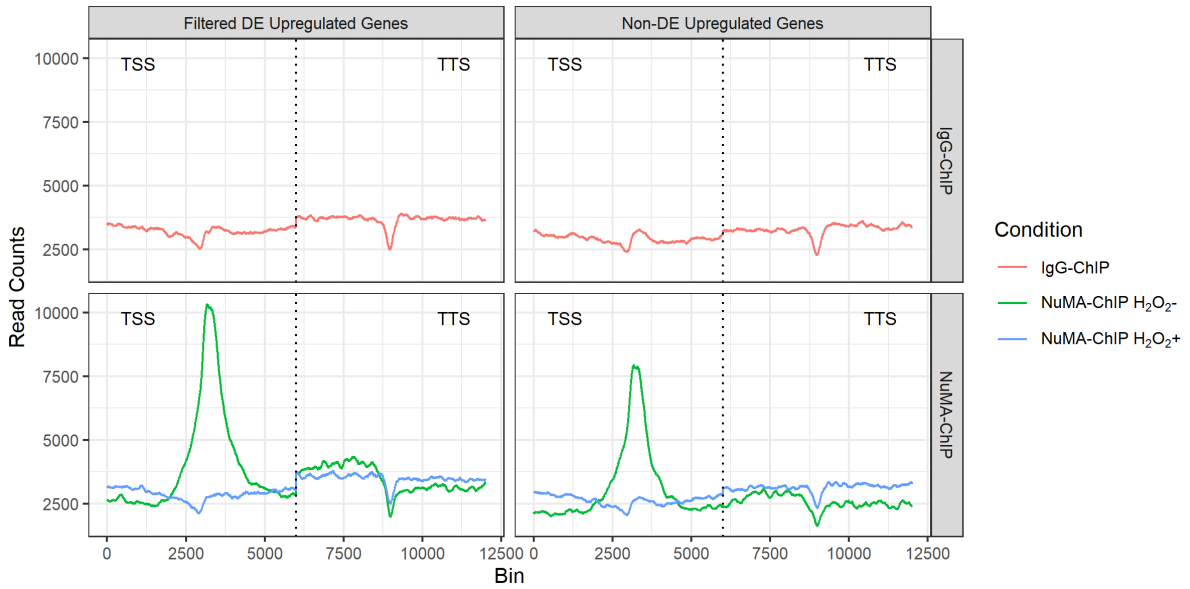
E


Fragile Promoter Genes
Non-Fragile Promoter Genes
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

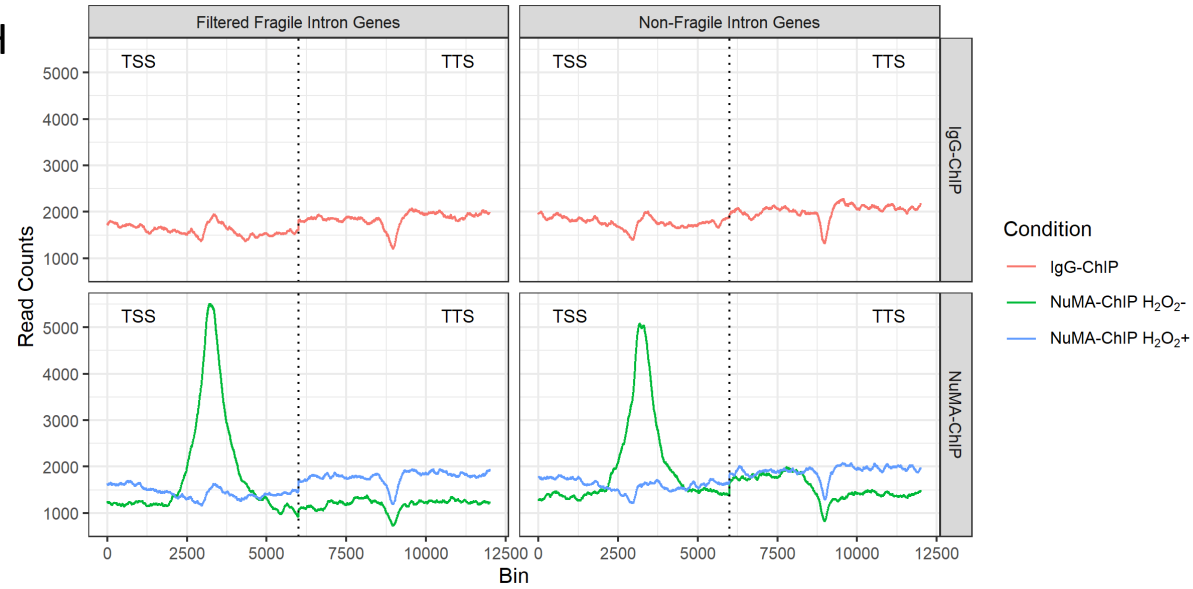
F


IERGs
Non-IERGs
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

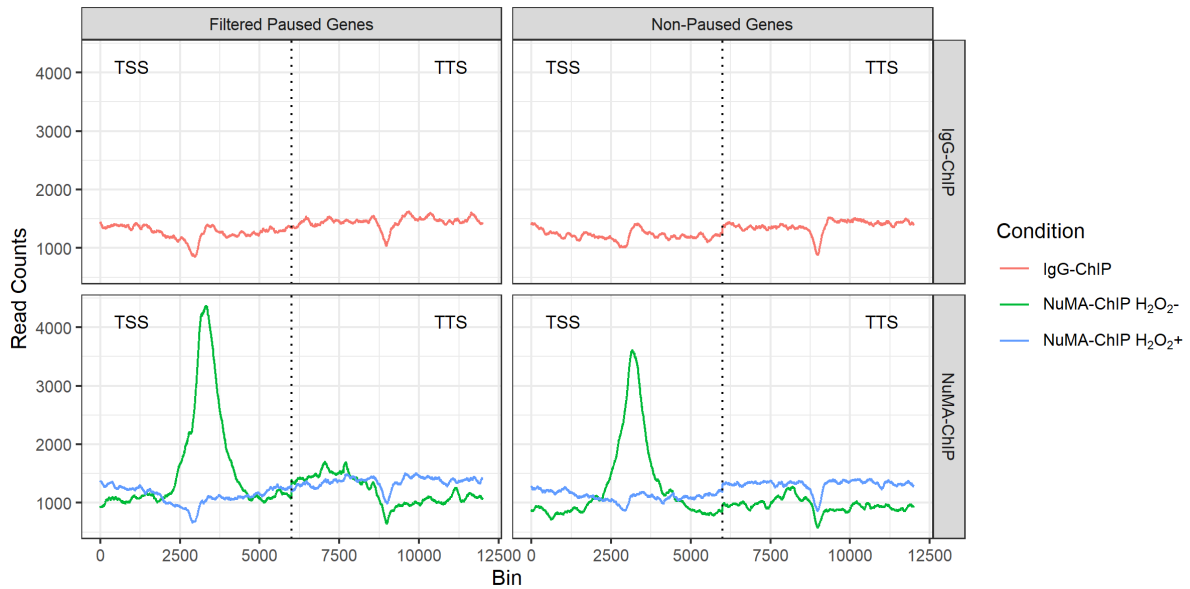
G


Filtered DE Upregulated Genes
Non-DE Upregulated Genes
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

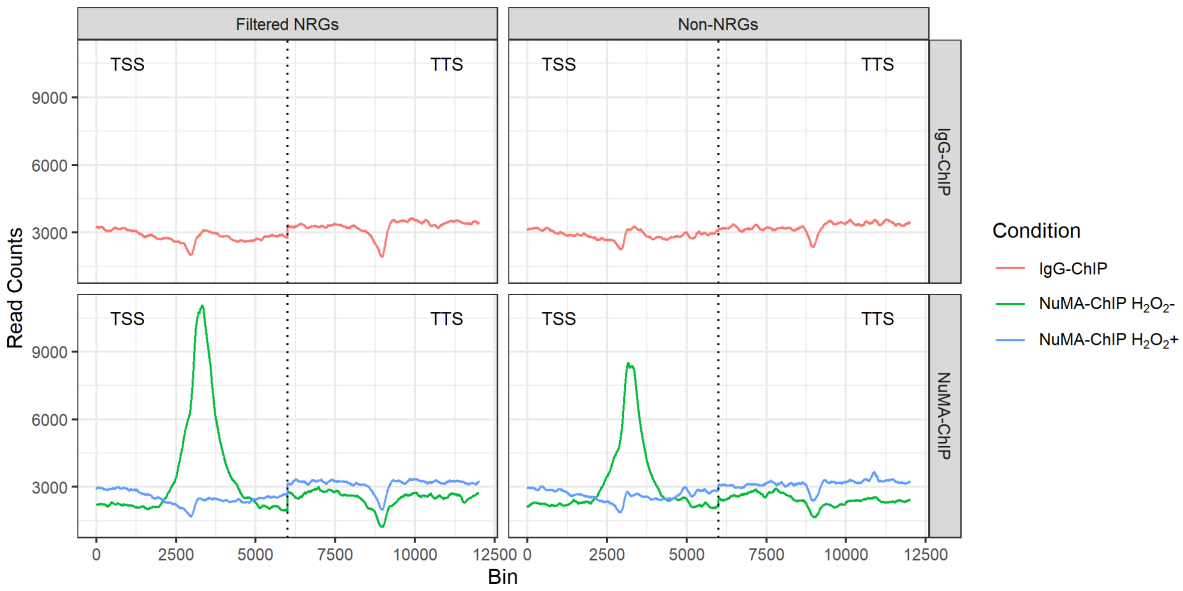
H


Filtered Fragile Intron Genes
Non-Fragile Intron Genes
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

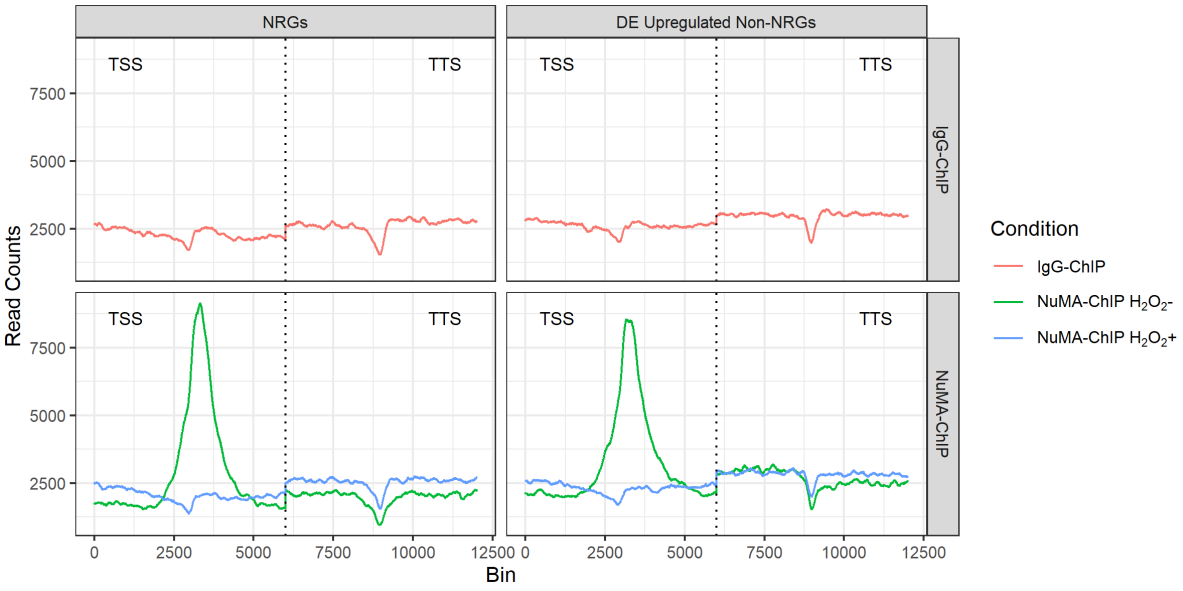
I


Filtered Paused Genes
Non-Paused Genes
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

J


Filtered NRGs
Non-NRGs
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
9000
6000
3000
0
2500
5000
7500
10000
12500
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

K


NRGs
DE Upregulated Non-NRGs
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
7500
5000
2500
0
2500
5000
7500
10000
12500
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

L

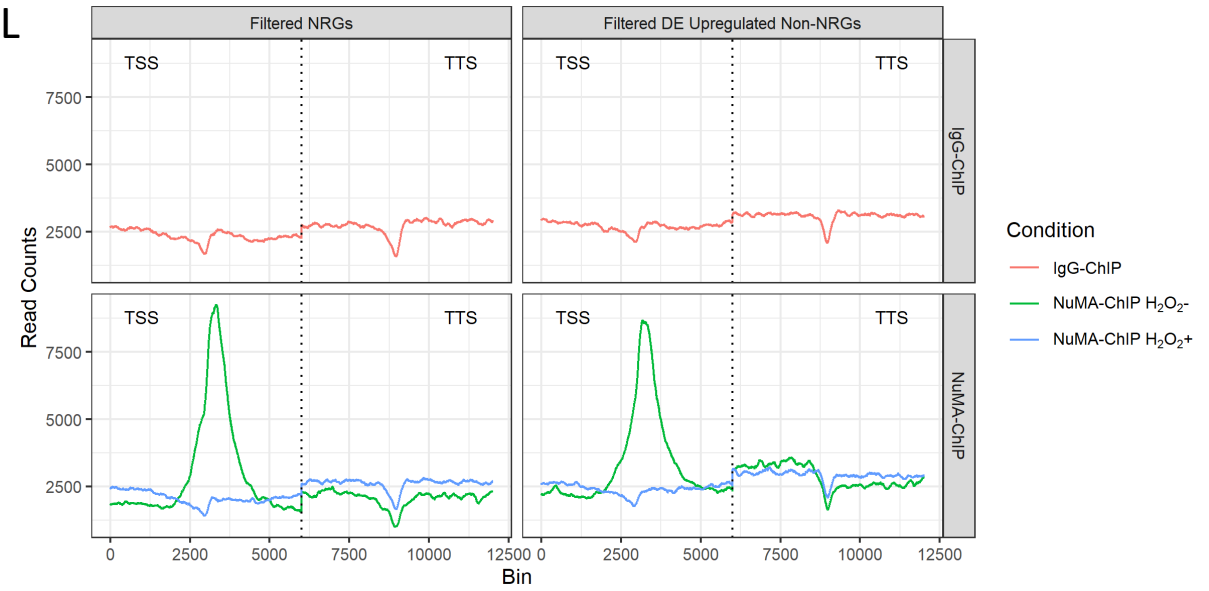

Filtered NRGs
Filtered DE Upregulated Non-NRGs
TSS
TTS
IgG-ChIP
NuMA-ChIP
Read Counts
7500
5000
2500
0
2500
5000
7500
10000
12500
Bin
Condition
IgG-ChIP
NuMA-ChIP H2O2-
NuMA-ChIP H2O2+

Fig.4.19 a-g) ChIP-seq metagenes profiles over the TSS and TTS for a) a randomly selected set of expressed genes (n=2076), b) NRGS (n=2112), c) genes containing fragile introns (n=1596), d) paused genes (n=1164) e) genes containing fragile promoters (n=454) and f) immediate early response genes (n=45) (derived from Tullai et al. 2007) alongside a randomly sampled number-matched set of expressed non-category genes.

g-l) ChIP-seq metagene profiles over the TSS and TTS for g) H_2O_2+ WT vs H_2O_2- WT upregulated genes filtered to remove fragile intron genes (FIGs), NRGs and paused genes (n= 1642), h) FIGs filtered to remove paused genes and H_2O_2+ WT vs H_2O_2- WT upregulated genes (n=1004), i) paused genes filtered to remove FIGs and H_2O_2+ WT vs H_2O_2- WT upregulated genes (n= 694) and j) NRGs filtered to remove FIGs and paused genes (n= 1382) alongside a randomly sampled set of non-category expressed genes. k) Metagenes profiles over the TSS and TTS for a reduced subset of NRGs (n= 1381), so as to number match the control consisting of a randomly sampled number-matched set of H_2O_2+ WT vs H_2O_2- WT upregulated genes. l) Metagenes profiles over the TSS and TTS for a reduced subset of NRGs filtered to remove FIGs introns and paused genes (n= 1382), so as to number match the control consisting of a randomly sampled number-matched set of H_2O_2+ WT vs H_2O_2- WT upregulated genes filtered to remove FIGs and paused genes.

Metagenes were calculated across each NuMA H_2O_2-, NuMA H_2O_2+ and IgG ChIP-seq sample and replicates averaged within bins. Bin 3000 = TSS, bin 9000 = TTS. Each bin is equal to one base. TSS = Transcription start site, TTS = Transcription termination site, NRGs = NuMA regulated genes, DE = Differentially expressed.

upregulated genes also removed NRGs, as all NRGs are also H_2O_2 differentially upregulated genes. Because of this feature of NRGs, it was not possible to remove H_2O_2 upregulated genes from them, and so instead three sets of metagenes were created. One set showed NuMA occupancy over NRGs from which FIGs and paused genes had been removed, the second NuMA occupancy over NRGs with the number matched controls being taken from H_2O_2 upregulated genes and the third the same but both NRGs and the upregulated controls were absent of FIGs and paused genes. Across the FIGs and paused genes metagenes, the target category depleted of other confounding category genes showed a higher level of NuMA across its promoter than the matched control **(Fig.4.19h,i)**. However, for NRGs, the pattern was slightly more ambiguous. **Fig.4.19j** shows NuMA enrichment in NRGs is not driven by FIGs or paused genes. Where NRGs were compared to H_2O_2 treatment upregulated genes, in both cases the total peak size for NRGs was marginally higher than its controls, because it started lower and peaked slightly higher **(Fig.4.19k,l)**. However, it should be observed that the total levels of NuMA occupancy across these regions was very similar if only taking into account the peak summit. The IgG trace also appeared lower for NRGs than for the controls, as did NuMA occupancy across the gene body. Taken together, this demonstrates that H_2O_2 upregulated genes, FIGs and paused genes are each independently enriched for NuMA binding in their promoter regions under normal conditions relative to a random sample of expressed genes. However, for NRGs, the results are less clear and whilst there does appear to be a slight independent enrichment, these results imply that increased NuMA occupancy within NRG promoters is at least in part driven by the same factors that determine promoter NuMA enrichment in H_2O_2 upregulated genes, of which NRGs are a subset.

4.2.8 NRGs are not enriched for fragile first introns or AP-seq signal

Although ChIP-seq had determined that FIGs and NRGs were independently enriched for NuMA at their promoters, it is unclear whether there is something unique about NRGs that contained fragile introns. One possibility following from the observation that NuMA has an affinity for promoter regions, is that fragile intron NRGs could potentially have a fragile first intron. Fragile intron one containing NRGs involved in the cellular response to oxidative damage could be unable to be transcribed because the presence of widespread damage may initiate breakage of their fragile site. NuMA could have a protective effect at these

genes due to its promoter binding bringing it within close proximity of the damaged first intron, preventing or quickly repairing fragile site damage and thereby enabling transcription of these genes in response to DNA damage. To investigate this possibility, we used the publicly available BLISS data from Dellino et al., (2019) to identify genes that had a fragile site overlapping their first intron, and then looked for enrichment of these genes amongst fragile intron NRGs compared to all FIGs. However, no enrichment was found **(Fig.4.20a)**. Continuing with the theme of NuMA acting as a protein performing a protective role to enable transcription under adverse conditions, it was next decided to explore whether NRGs in general were more susceptible to oxidative DNA damage, hence their elevated expression upon DNA damage induction being NuMA-dependent. To this end, sequencing of apurinic sites (AP-seq) across the genome was carried out by Swagat Ray of the El-Khamisy lab of the University of Sheffield. This technique uses biotin-labelled probe that reacts with the aldehydes present at apurinic sites at a neutral pH, so labelling AP-site containing DNA with biotin. After DNA shearing, Biotin labelled DNA fragments can then be isolated using streptavidin and sequenced. For this AP-seq experiment, we used the same four conditions used for the 4sU-seq: untreated WT, H_2O_2 treated WT cells, NuMA KD untreated and NuMA KD H_2O_2 treated. There were 2 replicate samples for each condition and each mapped sample contained roughly 30 – 40 million mapped reads. Number matched metagenes of NRGs and randomly sampled expressed genes were then generated across the samples representing each of these conditions, and samples were normalised for number of mapped reads **(Fig.4.21a)**. Across all of the conditions, there was no difference in the metagene trace for NRGs and expressed genes. The overall trace across the conditions showed regions upstream of the TSS and downstream of TTS had more damage signal than the intragenic regions. There was also signal depletion at the TSS and TTS and an increase along the gene body, often peaking just below the signal for the intergenic regions. Previous research showed exons were depleted of damage whereas introns had signal close to intergenic levels, but as we did not distinguish between introns and exons across the gene body this overall pattern was in-line with previously published results (Poetsch et al., 2018). Whilst here was no difference between NRGs and the randomly sampled expressed genes, the general trend of AP-seq signal across conditions did show some interesting results. There was very little difference in the signal between WT H_2O_2- and WT H_2O_2+ metagenes. However, both the WT and NuMA KD H_2O_2 treated AP-seq samples failed to recapture the

A

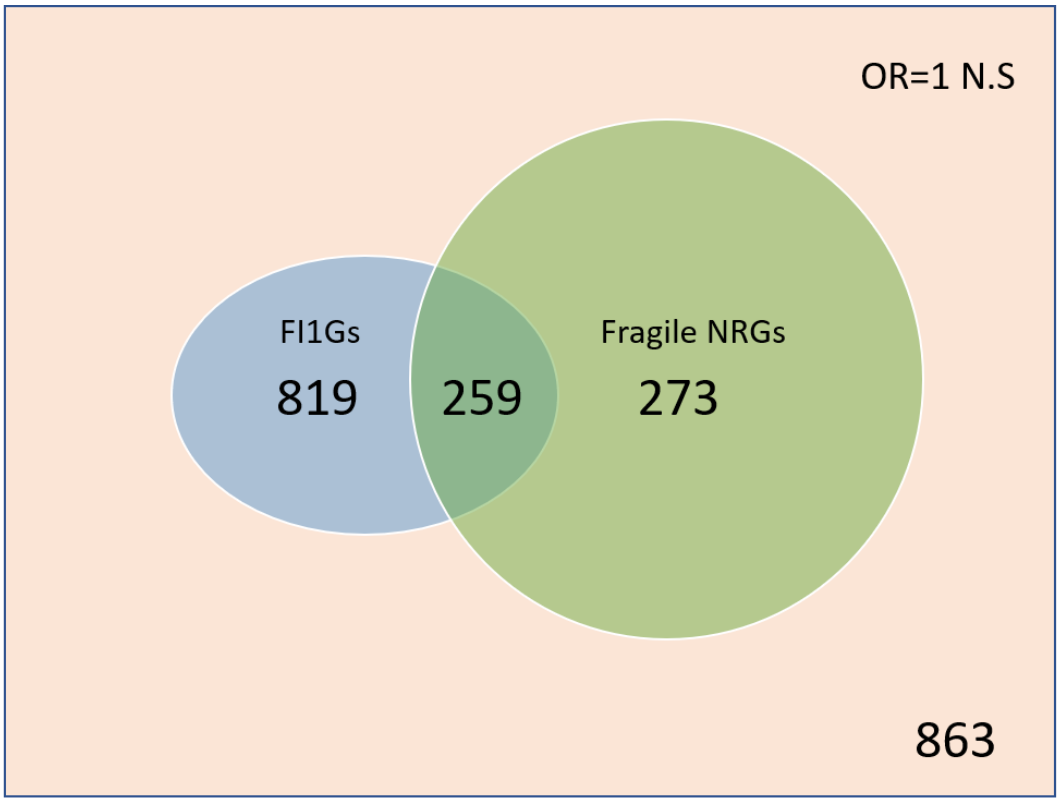


Fig.20a) Venn diagram showing overlap between NRGs containing fragile introns and genes with a fragile site overlapping it's first intron. The background gene set is all genes containing fragile introns with a tpm value of 1 or greater in both replicates of at least one condition in the 4sU-seq data. Significance determined by Fisher's exact test. FI1Gs = Fragile intron one containing genes, NRGs = NuMA regulated genes, N.S = Non-significant, OR = Odds ratio.

A

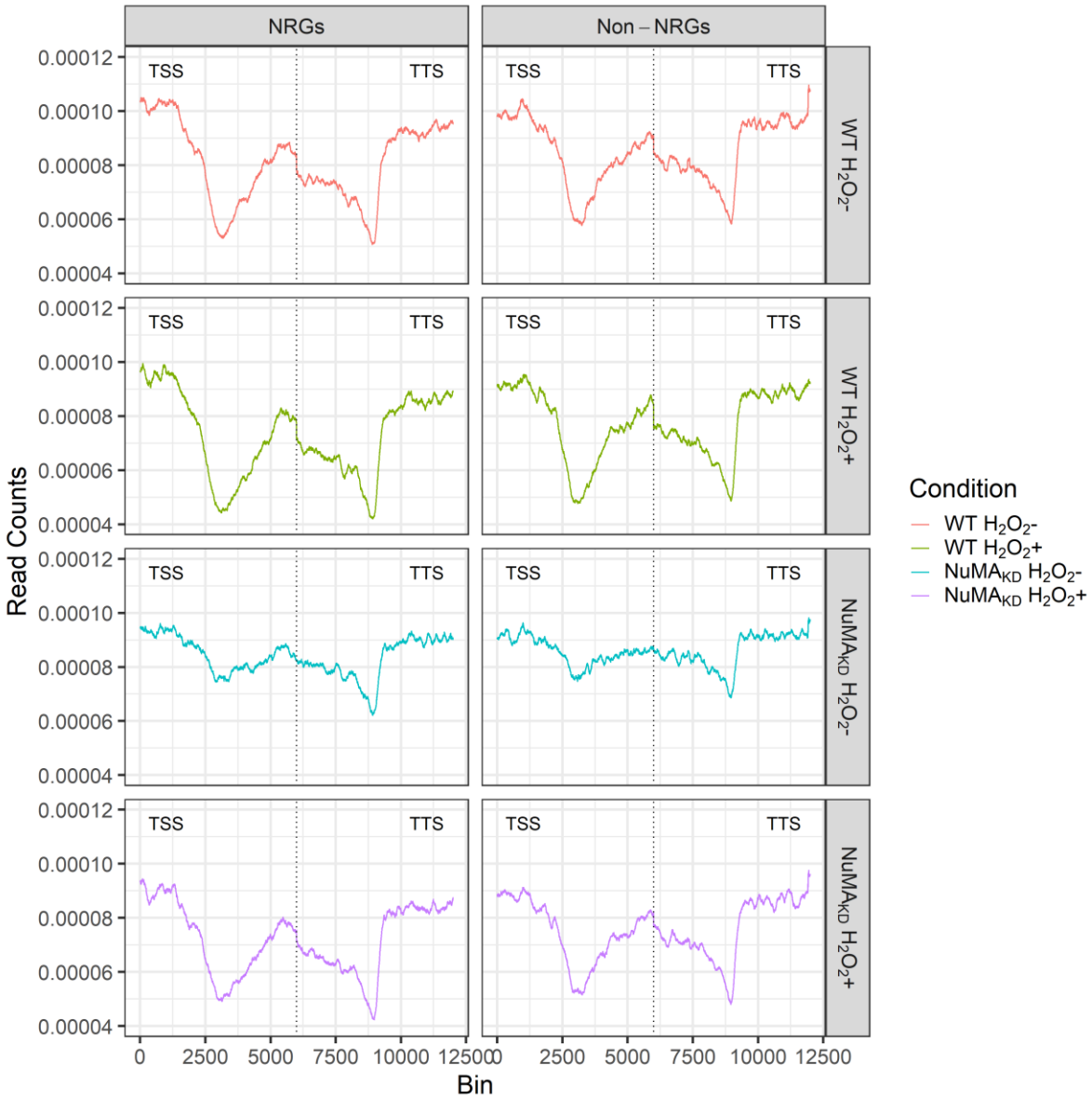


Fig.21a) AP-seq metagene profiles over the TSS and TTS of NRGs (n=2112) and a randomly sampled number-matched set of expressed genes for WT H_2O_2-, WT H_2O_2+, $NuMA_{KD}$ H_2O_2-, $NuMA_{KD}$ H_2O_2+ samples. Read counts were sample normalised by the total number of mapped reads within the given sample. Replicates for each condition were averaged within bins. Bin 3000 = TSS, bin 9000 = TTS. Each bin is equal to one base. NRGs = NuMA regulated genes, TSS = Transcription start site, TTS = Transcription termination site

enrichment of AP-seq signal at specific repetitive regions upon the induction of oxidative damage shown by Poetsch et al., (2018), so it may be the case that the H_2O_2 treated samples are unreliable. The most striking result was that upon knockdown of NuMA the protective effect observed at the TTS and particularly the promoter was severely diminished. Strangely, this increase in AP-seq signal at these regions appeared to be rescued upon H_2O_2 treatment in NuMA KD cells, although as stated, the results from the H_2O_2 treated samples are to be treated with some scepticism.

4.3 Discussion

Through these analyses, we have identified a set of a little more than 2000 genes whose expression upon the induction of oxidative DNA appears to by mediated by NuMA, termed NuMA regulated genes, NRGs. Interestingly, this set of genes is enriched for genes containing fragile introns and immediate early response genes, but not fragile promoter or paused genes, above an already existing enrichment for these gene sets amongst those differentially expressed upon H_2O_2 treatment. The lack of enrichment for paused genes and genes with fragile promoters amongst NRGs compared to the other differentially expressed genes indicates that these categories, as well as fragile intron and early response genes are merely mildy predictive for differential expression upon the induction of DNA damage, which is an interesting finding within itself. Upregulation of IERGs upon a threat to genome stability is expected, as some of these genes are known to be stress responsive, whereas there is no clear reason as to why genes containing fragile sites should be upregulated in this scenario. It might be hypothesised that paused genes are enriched within NRGs because their being paused allows them to respond rapidly to a genomic threat by releasing paused polymerases, thereby generating the necessary protein in a shorter than average time-frame. Whilst there is no definitive answer as to why fragile intron and IER genes are overrepresented amongst NRGs, considering NuMA as a structural protein in the context of its role in the DDR makes some possibilities more plausible than others. As its mechanism of action, NuMA could simply act to facilitate the expression of these genes by modulating the activity of a secondary protein, as it does to p53 (Endo et al., 2013; Ohata et al., 2013). A subset of NRGs could in fact be p53 response genes whose expression is mediated by

NuMA. Alternatively, NuMA could be physically associated with these genes and protect them from damage, an explanation particularly relevant to the expression of genes with fragile introns which may acquire DNA breaks upon coming into contact with damaging agents. These hypotheses are not mutually exclusive and the truth could be either of the two, neither, or different for specific subsets of NRGs. Whilst there are around 500 NRGs that contain fragile introns, making up approximately 1/5th of all NRGs, only 16 NRGs are also IERGs. This indicates that whilst interesting, these gene categories only represent a small subset of total NRGs and that the role of NuMA in the response to DNA damage is a lot broader in scope than merely facilitating expression of these categories and may involve multiple difference mechanisms of action. NRGs also tend to be longer and more highly expressed than other H_2O_2 responsive genes by a statistically significant but modest margin. Their high expression upon H_2O_2 treatment could mean they are particularly necessary for the cellular response to DNA damage, and NuMA has a role similar to its function within the p53 mediated stress response where it facilitates the expression of these damage response genes. The observation that NRGs are longer however, fits more neatly within a protective NuMA framework, as longer genes are more susceptible to acquiring DNA damage by virtue of the fact they contain more DNA, so NuMA could act to offset this and thereby enable effective transcription. A similar mechanism could explain why NRGs have more introns. Introns are known to accrue more DNA damage than exons, and so perhaps NuMA is required to prevent intron damage interfering with transcription when rapid expression of the NRG is required as part of the DDR. However, it is likely that this increased intron number is a function of gene length and further work would be required to disentangle this confounding effect.

The presence of higher levels of NuMA in promoter regions relative to its occupancy elsewhere, along with the observation that loss of NuMA leads to an increase in damage levels in promoters fits with its proposed roles as either a modulator of gene expression or a DNA damage repair/protection protein, although the fact that this occupancy does not appear to be gene specific perhaps favours a damage response explanation. However, the depletion of NuMA from promoters upon DNA damage perhaps requires this hypothesis to be more complex. In contrast to its previously suggested role as a structural scaffold for the recruitment of other proteins at the site of damage, it could act as a sensor of damage and

upon damage detection, leave the promoter to activate DDR proteins. However, it is important not to place too much weight on the results from the H_2O_2 treated ChIP-seq, due to the technical problems encountered with those samples. The observation that NuMA shows enrichment within the promoters of FIGs, NRGs, paused genes and genes differentially upregulated upon H_2O_2 treatment is interesting, not least because one of these categories, paused genes, does not show enrichment in NRGs beyond its overrepresentation within H_2O_2 treatment upregulated genes. This along with the fact that FIGs, paused genes and H_2O_2 upregulated genes all show independent enrichment – that is, not dependent on the presence of genes from any other identifies enriched category, implies that perhaps the role of NuMA at the promoters of these sets of genes is distinct from its function upon the exposure of the cell to oxidative damage. This is further bolstered by the fact that the set of genes whose expression is mediated by NuMA upon H_2O_2 treatment, NRGs, do not appear to be very much enriched for promoter NuMA beyond the occupancy observed for H_2O_2 upregulated genes. The reason behind promoter NuMA enrichment within these genes is not yet known, and there may yet be some common feature or link between them that explains this enrichment. Alternatively, this phenomenon may be linked specifically to the gene category itself and NuMA could carry out diverse roles specific to each of them. It could protect against transcription interfering DNA damage in fragile intron genes and yet have a different function associated with paused genes. It could also be the case that the H_2O_2 activated genes are in fact NuMA regulated but our experiment did not have the statistical power to detect their downregulation upon loss of NuMA, and so this why they are enriched for promoter NuMA.

However, the issue of NuMA playing a DNA damage protection role at the promoters of specific genes is made more complicated by results from the AP-seq data. These results do indeed imply that NuMA protects against oxidative damage within gene bodies, as in the absence of NuMA AP-seq signal is noticeably increased across the promoter and the TTS. However, based on the comparative results between NRGs and their number matched controls, this does not appear to be a category specific effect, and its apparent rescue by H_2O_2 treatment leaves much to be explained. Furthermore, NRGs are not enriched for first intron fragile sites which might have explained their NuMA dependent expression upon damage induction. If a gene contained an early fragile site, this site may be susceptible to

breakage upon exposure to DNA damaging agent which could block transcription of said gene, meaning it required fast and efficient repair in order to be expressed in the presence of DNA damaging agents.

Taken together, these data suggest two potential options. One is that NuMA is acting as an arm of the DNA damage response in all the explored scenarios. It is enriched within the promoters of fragile intron and paused genes because these genes are damage prone, and has some protective role at the promoter detecting from DNA damage generally, hence the increase in promoter oxidative damage upon NuMA loss. It's relationship to NRGs on the other hand is not protective, rather it mediates their expression as part of a wider DDR. Data from the El-Khamisy lab has determined that NuMA interacts with components of DDR, and so this taken along with the data in this thesis strongly suggest NuMA does play a direct role in DNA repair. However, this line of reasoning raises questions. Why is NuMA not enriched amongst genes containing fragile promoters – do fragile sites in promoters have less of an impact on expression than fragile introns? Why are NRGs not enriched for AP-seq signal compared to a control set of genes in any of the conditions? To be coherent, this hypothesis would require the suggestion that relative levels of promoter NuMA between sets of genes have no significance as regards its role in the DDR. A second option is that NuMA occupancy within promoters is indeed not linked to its role in responding to DNA damage, and the DDR role of NuMA with respect to this study is purely restricted to acting as an activator or mediator of NRG expression in response to DNA damage by some mechanism not yet known. NuMA's increased occupancy within certain gene categories relates to a different unknown function or feature of those genes or a collection of disparate functions relating to specific biology surrounding those categories. This again leaves a lot to be explained. It does not account for why NuMA leaves promoter regions upon H_2O_2 treatment if this is not directly related to the DDR. One explanation is that as NuMA is large protein, its presence at the promoter and along the gene body restricts access to the site of DNA damage, and so for effective DNA repair to take place NuMA first needs to leave the DNA. This could in some way parallel the already described phenomenon of opening of closed chromatin upon DNA damage within such regions, in order to allow efficient repair. However, as stated, the results for the H_2O_2 treated ChIP-seq should be taken carry reduced significance compared to other results due to the technical problems with the preparation

of these samples. The interactions between NuMA and DDR proteins observed by the El-Khamisy lab also render this hypothesis unlikely. On balance, the first hypothesis would appear to be a more fitting suggestion, but the significance of differing levels of promoter bound NuMA remains to be determined.

All in all, whilst the discovery of NRGs and the AP-seq results firmly suggest a role for NuMA in the DDR, much more clarifying work needs to be performed. Finally, the results indicating NuMA's increased expression in the cerebellum relative to cerebral tissue is particularly exciting in the context of this thesis. Neural tissues are non-cycling, so any function NuMA has within the brain is highly unlikely to be related to its role in the assembly of mitotic spindles. This leaves open the possibility that NuMA is overexpressed in the cerebellum compared to the cerebrum as part of the DDR. As mentioned previously, it is possible that NuMA dysfunction could lead to neurological disease, following the precedent set my mutations in lamins. This new information suggests that if this were the case, the cerebellum may be particularly affected by loss of NuMA function. Revisiting the concept of somatic mutations in the brain, if NuMA is particularly important and highly expressed in the cerebellum then through the mechanisms of mutation acquisition in the post-mitotic brain, including the "use it and lose it" hypothesis, loss of NuMA in the ageing tissue could have negative consequences for cell viability.

4.4 Methods

4.4.1 Cell culture and treatment

Normal human lung fibroblast MRC-5 cells were cultured in Minimum Essential Medium Eagle (MEM) supplemented with a final concentration of 10% fetal bovine serum, 1% penicillin/streptomycin and 1% L-glutamine at 37°C in a humidified atmosphere containing 5% carbon dioxide (CO_2). The cells were seeded in 15cm plates and grown until 80% confluency, then either left untreated or treated with 10 uM hydrogen peroxide (H_2O_2) in cold PBS.

4.4.2 4sU-seq

RPE1 cells with a Doxycycline inducible NuMA shRNA was grown in DMEM/F-12 medium (Tetracycline-free) supplemented with Puromycin and +/- Doxycycline (for NuMA depletion). Cells were serum starved for 48 hours before addition of H_2O_2 (for the relevant conditions) for 10 mins on ice. Cells were treated with 700uM 4-thiouridine (Sigma) by just adding it to culture media. RNA was isolated after 1-60mins by Trizol, which was added directly to the cells after rinsing with PBS. Approximately 100ug RNA was biotinylated in a volume of 150µl containing 10mM HEPES (pH7.5), 1mM EDTA (pH 8), 5ug MTSEA Biotin-XX (Iris Biotech, dissolved in dimethyl formamide). After incubation in the dark for 60 mins, biotinylated RNA was twice chloroform extracted (which removed excess biotin), phenol chloroform extracted and then ethanol precipitated. uMACS beads from Miltenyi biotech were used for selecting the biotinylated RNA. 50ul beads were blocked with with 1x wash buffer (10mM Tris.Cl pH7.4, 50mM NaCl, 1mM EDTA) and 2ul yeast tRNA (10mg/ml) for 20 minutes at room temp. Columns were washed with nucleic acid equilibration buffer (provided in kit with beads) and three times with 1x wash buffer. The beads were then applied to the column and washed five times with wash buffer. Beads were then eluted from the column by taking it out of the magnet and running 100ul wash buffer through it twice. 200ul bead suspension was combined with the RNA (re-suspended in 1x wash buffer) and incubated at room temp for 20 minutes. The bead RNA suspension was then applied again to the column (back in magnet) and washed three times with wash buffer 1 (10mM Tris.Cl pH 7.4, 6M Urea, 10mM ETDA) warmed to 60 degrees, washed three times with wash buffer 2 (10mM Tris.Cl pH7.4, 1M NaCl, 10mM ETDA) and then warmed to 60 degrees. The RNA was eluted by adding 400ul 0.1M DTT in 4x 100ul aliquots and then ethanol precipitated. The nanodrop 260/280 ratio was used to assess purity before the samples were sent off for sequencing.

4.4.3 4sU-seq Processing and Analyses

Quality of samples was assessed by examining FastQC reports provided by Novogene Plc (Andrews, 2015). The data was mapped to the human reference genome using `STAR` by Novogene Plc (Dobin et al., 2013). Differential expression analyses were performed in R using DESeq2 - the data was normalised using DESeq, false discovery rate calculation after

model fitting was performed using the Benjamini-Hochberg procedure and less than or equal to an adjusted p-value of 0.05 was taken as the cutoff for differential expression, all of which was done by Novogene Plc (Love et al., 2014; R Core team, 2019). Prior to downstream analyses, unexpressed genes were filtered out using transcripts per million (TPM) data for each gene derived from mapping the data using `salmon` (index building parameters: `--type=-fmd kmer=31`, mapping parameters: `quant --libtype A`) (Patro et al., 2017). The resulting quant files were analysed in R studio and imported into R studio using the package tximport (R Studio Team, 2015; Soneson et al., 2015). Any genes that did not have a TPM value of 1 or greater in both replicates of at least one condition were considered to be unexpressed and were discarded. Statistical enrichment was calculated using Fisher's tests through the R function `fisher.test`. Gene ontology enrichment tests were performed using the R package GOseq (Young et al., 2010). Any gene information not included in TPM/count tables, was found using biomaRt in R (Durinck et al., 2005, 2009).

4.4.4 Identification of gene categories

Paused genes analysed in the 4sU-seq analysis were determined using publicly available RNA polymerase II ChIP-seq data from RPE1 cells (Available under the accession GSE60024), whereas the paused genes used in the NuMA ChIP-seq analysis were determined using publicly available RNA polymerase II ChIP-seq data from MRC-5 cells (Available under the accession GSE55171 (Giannakakis et al., 2015)). This was because the 4sU-seq was performed in RPE-1 cells whereas the NuMA ChIP-seq was performed in MRC-5 cells. Paused indices were calculated from from RNA polymerase II ChIP-seq by dividing the average number of reads overlapping 100 bases pairs upstream and 300 base pairs downstream of the transcription start site by the average number of reads overlapping 500 to 2000 base pairs upstream of the transcription start site. This was performed in a jupyter notebook, available at: https://github.com/jdparker101/Pausing_Ratio_Notebook/Pausing_Ratios.ipynb. Genes with a pausing index >= 2 were considered to be paused, and were carried forward to subsequent analyses. Immediate early response genes were collected from Immediate-Early and Delayed Primary Response Genes Are Distinct in Function and Genomic Architecture (Tullai et al., 2007). Gene promoter windows were generated using biomaRt for R to identify

the chromosome and transcription start site (TSS) for each gene. The TSS was then extended by +/- 2500 base pairs to approximate the promoter region and a promoter BED file created. An Intron BED file was downloaded from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables) (options: clade = Mammal, genome = Human, assembly = Dec.2013 (GRCh38/hg38), group = Genes and Gene Predictions, track = NCBI RefSeq, table = knownGene, region = Genome, output format = BED – browser extensible data, Create one BED record per = Introns plus (0)). Overlaps between the promoter and intron BED files and publicly available data detailing endogenous DSB sites identified using BLISS data (available under the accession GSE93038, (Dellino et al., 2019)) was performed using `bedtools intersect` (parameters: `-wa`) in order to identify fragile promoters/introns (Quinlan and Hall, 2010).

4.4.5 Chromatin Immunoprecipitation sequencing (ChIP)

For crosslinking, the cells were first washed with PBS and 10 ml PBS was added to each plate. 270 µl of freshly prepared 37% paraformaldehyde was added to each plate (final concentration = 1%) and incubated at room temperature for 10 minutes at 20 rpm. The crosslinking reaction was quenched by the addition of 1 ml 1.25 M glycine (final concentration = 0.125 M) and incubated at room temperature for 5 minutes at 20 rpm. The plates were then washed twice with 10 ml cold PBS then the cells were scraped in a suitable volume of cold PBS, spun down and the cell pellet was frozen at -80°C or lysed immediately. The cell pellets were then thawed on ice then resuspended in 5 pellet volumes of ChIP Lysis Buffer 1 (50 mM HEPES-KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% Glycerol, 0.5% NP-40, 0.25% Triton X-100) to lyse the cell membrane and incubated for 5 minutes, 4°C, 20 rpm. They were then spun down at 3000 x *g*, 5 minutes, 4°C. The pellet (nuclei) was resuspended in 5 pellet volumes of ChIP Lysis Buffer 2 (10 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA), an isotonic buffer to swell the nuclei and incubated for 10 minutes, room temperature at 15 rpm. The lysate was then spun down at 1500 x *g*, 5 minutes, 4°C and the nuclear pellet was then lysed by resuspending in a suitable volume of ChIP Lysis Buffer 3 (10 mM Tris-HCl, pH 7.4, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Sodium deoxycholate, 0.5% Sodium lauroylsarcosine). All ChIP Lysis buffers contained cOmplete EDTA-free Protease inhibitor Cocktail and phosphatase inhibitor, PhosSTOP. Sonication was optimized using Bioruptor Pico (Diagenode) to yield DNA fragments of the size 100-300 bp.

The lysates were sonicated for a suitable number of cycles in a volume of 100-300 µl per tube and then spun down at 20000 x *g*, at 4°C for 15 minutes and the supernatants were transferred to a new tube. Lysates containing an equal quantity of protein were incubated with either 0.49 ug NuMA (D49H4) Rabbit mAb (Cell Signalling Technology) or 10 ug Rabbit IgG (Invitrogen) overnight at 4°C at 20rpm. 30 µl Dynabeads Protein A beads (Invitrogen) were washed with ChIP Lysis Buffer 3 twice before being resuspended in the original volume, added to each sample and incubated for 2 hours at 4°C at 20 rpm. The samples were then spun down at 1000 x *g* for 5 minutes at 4°C and then placed on a magnet. The beads were washed 5 times in 500 µl RIPA Wash Buffer (50 mM HEPES-KOH, pH 7.5, 500 mM LiCl, 1 mM EDTA, 1% NP-40, 0.7% Sodium deoxycholate, 0.1% Sodium lauroylsarcosine) and once in 1 ml ChIP Final Wash Buffer (10 mM Tris-HCl, pH 7.4, 1 mM EDTA, 50 mM NaCl). For qPCR and sequencing experiments, elution was conducted in 200 µl ChIP Elution Buffer (50 mM Tris-HCl, pH 7.4, 10 mM EDTA, 1% SDS) at 65°C for 30 minutes at 600rpm. For reverse crosslinking, 200µl of the eluate, 8 µl 5 M NaCl was added (final concentration = 0.2 M) and incubated for 16 hours at 65°C. 200 µl 1xTE Buffer with 4 mM calcium chloride and 8 µl 10 mg/ml RNaseA (final concentration = 0.2 mg/ml) was added to the samples and incubated at 37°C for 30 minutes at 600rpm. This was followed by the addition of 8 µl 10 mg/ml Proteinase K (final concentration = 0.2 mg/ml) and incubated at 65°C for 2 hours. To extract the DNA, An equal volume of phenol-chloroform-isoamyl alcohol (Invitrogen) was added to the sample and vortexed thoroughly for 10 seconds. The samples were then centrifuged at 20000 x *g* for 5 minutes at 4°C and the upper aqueous phase was carefully removed and transferred to another tube. An equal volume of chloroform was added for back extraction, vortexed for 10 seconds and centrifuged at 20000 x *g* for 5 minutes at 4°C. The upper aqueous layer was again carefully transferred to another tube. The DNA was then ethanol precipitated by adding 1 µl glycogen, 0.1 times the sample volume 3M sodium acetate pH 5.2, 2.5 times the sample volume 100% ethanol and incubated either at -20°C overnight or -80°C for 1 hour. The samples were centrifuged at 20000 x *g* for 30 minutes at 4°C and the pellet was washed twice with 500 µl 70% ethanol. The pellet was then air-dried and resuspended in a suitable volume (10-20 µl) DEPC-treated Water (Invitrogen) water at 37°C for 10 minutes at 600 rpm. This DNA was then used to check the size of chromatin fragments after shearing by sonication using agarose gel electrophoresis, ChIP-qPCR and ChIP-seq. Using 1 µl of the RNA or DNA sample, the absorbance was measured

spectrophotometrically on a NanoDrop Lite Spectrophotometer (Thermo Scientific) against a blank (elution buffer/water) to determine the concentration and purity of the sample. Absorbance at 260 nm was used to determine the concentration of the sample, while calculating the ratio of the absorbance at 260 nm/ 280 nm (acceptable range: 1.8-2.0) was used to determine the purity of the sample, which was done before the samples were sent off for sequencing.

4.4.6 ChIP-seq processing and analyses

The ChIP-seq data was assessed by examining FastQC files provided by Novogene Plc. and mapping them using `BWA-mem` (parameters: `-M -k 25`) to the hg38 human reference genome as part of the mapping pipeline from CGAT pipelines (Andrews, 2015; Li and Durbin, 2010; Sims et al., 2014). Prior to peak calling and the generation of metagene profiles, duplicates were first removed from the BAM files using `Picard MarkDuplicates` in conjunction with `samtools view` to remove PCR or optical duplicates (parameters: `-F 1024`) (Broad Institute, 2016; Li et al., 2009). Reads with a quality score of less than 30 were also removed using `samtools view` (parameters: `-q 30`). Peak calling was performed using `MACS2` (Zhang et al., 2008). Both narrow and broad peaks were called (parameters: default aside from `-g hs`). For narrow peak calling, fragment pileup and control lambda were output to BEDgraph files using the `MACS2` peak calling option `-B -SPMR`. The sample and control lambda (derived from inputs) BEDgraph signal tracks were then compared using `MACS2 bdgcomp` with the option `-m FE` in order to generate sample BEDgraph files showing fold change relative to input. This was performed as part of a custom pipeline, `pipeline_peaksandprofiles`. Metagene profiles were generated using bam2geneprofile from CGAT scripts using gene annotations derived from Ensembl 85, with reporter set to genes and no normalisation (Zerbino et al., 2018). Both gene profiles, transcription start site, and transcription termination site profiles were generated. Basic metagene profiles were created using `pipeline_peaksandprofiles`.

Number matched metagene profiles in which the genes were split apart by some feature e.g. whether or not they were subject to promoter proximal pausing, and separate metagene traces calculated for an equal number category and control genes, were generated using `pipeline_splitprofilesNumMatched`. Metagenes derived from

`pipeline_splitprofilesNumMatched` only show data from genes determined to be expressed in the 4sU-seq, which is all genes that had at a transcripts per million (TPM) value of 1 or greater in both replicates of at least one condition. In `pipeline_splitprofilesNumMatched`, the provided GTF file was divided by gene category according to a tab-separated file containing the category of each expressed gene using the `GTF iterator` from CGAT.

Both metagene profile pipelines excluded overlapping genes. This was done by converting the GTF(s) to BED files using `gff2bed` (parameters: `--is-gtf`) from CGAT scripts, extending resulting regions by 2500 base pairs upstream and downstream using `bedtools slop` (parameters: `-s`), counting the resulting gene overlaps using `bedtools merge` (parameters: `-c -o count`), identifying overlapping genes as having an overlap count of at least one and removing these from the original GTF(s) using an exclusionary `bedtools intersect` (parameters: `-v`). The genome file passed to `bedtools slop` was a pseudo-contigs files generated from the provided GTF. In `pipeline_splitprofilesNumMatched`, after filtering the category divided GTFs were merged based on transcript and the transcript ID set to the gene ID using `gtf2gtf` from CGAT scripts (parameters: `--method=merge-transcripts/set-transcript-to-gene`. After merging, the GTF containing the control set of genes was randomly sampled using the `pandas` python library to match the target set of genes in number (McKinney, 2010). Across both pipelines, the metagene matrices were generated using `bam2geneprofile` from CGAT scripts with the `--reporter` option set to gene. The `-m` option was set to `geneprofile` and `tssprofile` for gene profile metagenes and TSS/TTS metagenes respectively. In all metagene plots, the trace was averaged across replicates. Where stated, metagenes were normalised by the number of mapped reads, determined using `samtools` (parameters: `view -c -F 4`). Read normalisation was not performed on all metagenes unless comparing traces between conditions was the primary function of metagenes (e.g. AP-seq metagenes (**Fig.4.21)**, total gene metagenes for NuMA ChIP-seq (**Fig.4.18**)). Both of these pipelines were generated using the CGAT Pipelines framework which utilises Ruffus and are available at https://github.com/jdparker101/pipeline_peaksandprofiles and

https://github.com/jdparker101/pipeline_splitprofilesNumMatched (Goodstadt, 2010; Sims et al., 2014).

4.4.7 AP-seq

RPE1 cells with a Doxycycline inducible NuMA shRNA was grown in DMEM/F-12 medium (Tetracycline-free) supplemented with Puromycin and +/- Doxycycline (for NuMA depletion). Cells were serum starved for 48 hours before addition of H_2O_2 (for the relevant conditions) for 10 mins on ice. Genomic DNA was extracted from cells (10cm plate) using the Qiagen Blood and Tissue Kit (roughly 7-10ug) and eluted in 100uL mQ H2O. The DNA was then digested with FpG: 85uL DNA + 5uL FpG + 10uL NEB Buffer 1 +1uL 100x BSA (all reagents in FpG enzyme pouch), precipitated using cold 100% Ethanol and reconstituted in 90uL PBS. 90uL of DNA was labelled with 5mM ARP (in DMF) by incubation with 10uL ARP stock for 2hr at 37C in 2ml tubes to allow for efficient mixing. The DNA was transferred to 1.5ml tube, to which 10uL 3M NaoAc (5M) and 200uL ice-cold Ethanol (100%) was added to precipitate the DNA. After precipitation, the DNA was washed with 70% ethanol, and the pellets reconstituted in 130uL TE buffer (pH8). DNA was sheared to an average peak size of 300bp (2 min cycle 30s on-30s off). Separate 30uL sheared DNA volumes were kept aside at this point for inputs. To prepare the beads, 100uL MyOne Dynabeads/pulldown were taken, washed twice with 1M NaCl in TE buffer, and reconstituted in 100uL 2M NaCl in TE buffer. Each 100uL of reconstituted beads were added to 100uL of DNA and samples were rotated at room temperature for 10 hours. DNA was eluted from beads using 95% formamide and 10mM EDTA for 10 min at 65C (a double elution in 50uL volume each= total 100uL). The Qiagen MinElute Clean Up kit was used for DNA extraction and the DNA eluted in 30uL TE (3x 10uL elution) The DNA was then repaired using the PreCR Repair Mix as per the manufacturer's protocol. At room temperature, 1X ThermoPol Buffer, 100 µM dNTPs, 1X NAD+, 30uL DNA sample were combined and made up to 49 µl with H_2O. 1 µl of the PreCR Repair Mix was added the and solution mixed gently. The repair reaction was incubated for 15-20 minutes at 37°C and then the reactions placed on ice. The DNA was then extracted using MinElute Clean Up Kit, eluted in 13uL mQ H_2O. The nanodrop 260/280 ratio was used to assess purity before the samples were sent off for sequencing.

4.4.8 AP-seq processing and analysis

The AP-seq data was mapped using `BWA-mem` (parameters: `-M -k 25`) to the hg38 human reference as part of the mapping pipeline from CGAT pipelines (Li and Durbin, 2010; Sims et al., 2014). Metagenes for AP-seq across NRGs and a matched number of randomly sampled 4sU-seq expressed genes were generated using `pipeline_splitprofilesNumMatched`, described in detail under **ChIP-seq analyses and peak calling** and available at https://github.com/jdparker101/pipeline_splitprofilesNumMatched.

4.4.9 Data manipulation in R and plots

Within R, data manipulation was performed using the dplyr and tidyr packages (Wickham and Henry, 2019; Wickham et al., 2019). Plots were generated using base R and the ggplot2 package (R Core team, 2019; Wickham, 2016).

4.4.10 Tabular summary of datasets and analyses

Dataset	Experimental design/Samples utilised	Analyses performed
GTEx V6 RNA-seq	All brain samples	Analysis of NuMA expression across different brain regions (performed using the GTEx portal)
MRC5 polII ChIP-seq (available under the GEO accession GSE55171)	MRC5_RNAPII-untreated	Identification of genes paused in MRC5 cells
RPE1 polII ChIP-seq (available under the GEO accession GSE60024)	RPE-POLII_ChipSeq	Identification of genes paused in RPE1 cells
BLISS-seq DSB table (available under the GEO accession GSE93038)	GSE93038_Dellino_BLISS_Processed_Data_Tier1_DSBs.txt.gz	Identification of fragile promoters and fragile introns
4sU-seq	2 WT H_2O_2- samples 2 WT H_2O_{2+} samples 2 $NuMA_{KD}$ H_2O_2- samples	Differential expression analyses (performed by Novogene)

	2 $NuMA_{KD}$ H_2O_{2+} samples	Category enrichment (Fisher's) tests
NuMA ChIP-seq	1 IgG input sample 1 H_2O_2- input sample 1 H_2O_{2+} input sample 2 IgG ChIP samples 2 H_2O_2- NuMA ChIP samples 2 H_2O_{2+} NuMA ChIP samples	Metagene analyses
AP-seq	2 WT H_2O_2- samples 2 WT H_2O_{2+} samples 2 $NuMA_{KD}$ H_2O_2- samples 2 $NuMA_{KD}$ H_2O_{2+} samples	Metagene analyses

5. Final Discussion

Throughout this thesis I have attempted to explore the intersection between neurological disease, DNA damage and tissue specific disorders through a bioinformatic lens. Primarily, this has focussed on the cerebellum and its differences with the cerebrum in healthy individuals, and how these differences might sensitise the cerebellum to mutations in housekeeping proteins that repair or process DNA, and ultimately lead to the DRDA-ARCAs. A basic point reinforced throughout this body of work is the uniqueness of the cerebellum across a whole host of criteria compared to tissues from the cerebrum. First of all, the cerebellum has a distinct transcriptional profile compared to the cerebrum. This has been shown previously by the fact the cerebellum clusters away from the other brain tissues in multidimensional scaling analyses, but our work builds on this by showing that across various differential gene expression comparisons between the cerebellum and cerebral tissues, many of the same genes are called as differentially expressed, indicating a conserved set of differential genes (Melé et al., 2015). The fact that many of these genes are called as differentially expressed across analysis of different datasets is testament to this difference. The magnitude of the difference between the tissues can be observed when looking at the total number of genes differentially expressed between the cerebellum and the pooled cerebral tissue in the GTEx RNA-seq data, arguably the most robust set of data analysed in chapter 2. Around 16000 genes are called as differentially expressed in this comparison, more than half of the GTEx genes contained within the GTEx dataset after filtering out genes showing negligible expression. When looking into differences in patterns of gene expression, more mitochondrially related genes were found to be downregulated than upregulated in the cerebellum relative to any of the assessed cerebral tissues across both the GTEx and Prudencio healthy control RNA-seq datasets. This pattern remained true even when all the cerebral tissues for the GTEx dataset were pooled together and compared to the cerebellum, showing the robustness of this trend. Many key single strand break repair proteins are also downregulated in the cerebellum compared to the cerebrum, most strikingly DNA polymerase β, an enzyme important for gap-filling at break sites in base excision repair (BER) (Ray et al., 2013). This work was built on in chapter 3 in which the differences between the cerebellum and the cerebrum were further expanded upon. Analysis across multiple versions of a pipeline built to measure mismatches when aligning to

the reference genome revealed that the cerebellum appears to have a higher basal mutation rate compared to the frontal cortex, if not on a genome wide scale then for a larger subset of genes. Consistently more genes were called as having a higher mutation rate in the cerebellum than were called for the frontal cortex. Enrichment analyses also revealed that these genes with a higher rate of mismatch in either tissue were enriched for genes downregulated in that tissue. Based on this work, the cerebellum also appears to have differences in mutational spectra relative to the frontal cortex. C→T mutations which are associated with transcription and are known to be caused by spontaneous cytosine deamination and replication fixed alkylation mutations are increased in the cerebellum relative to the frontal cortex, as are G→T mutations, typically linked to oxidative mutations stabilised by replication (Barnes and Lindahl, 2004; Lodato et al., 2015). Strangely, T→As, not associated with any sort of known chemical attack or type of DNA damage, are enriched in frontal cortex relative to the cerebellum, constituting another difference between these two tissues and highlighting the uniqueness of the cerebellum. The fact that C→Ts and G→Ts are also enriched in the cerebellum relative to the cerebellar hemisphere and not in the hemisphere relative to the frontal cortex does however casts serious doubt on these results, the cerebellum and cerebellar hemisphere samples are essentially duplicates that have been preserved differently. A→G and T→C mutations that had a higher rate of occurrence in both the cerebellum and the cerebellar hemisphere relative to the frontal cortex were thought to be RNA editing events, in the case of T→Cs, specifically those that the script failed to reverse complement correct. This leads us onto one of the most reliable findings in terms of cerebellar specific features – the cerebellum having a higher rate of RNA editing than the frontal cortex, a pattern that holds true across the full length of genes and when restricting our analysis to exons. Also, many more genes have a higher rate of RNA editing in the cerebellum compared to the frontal cortex, again, a feature true across both the total gene and exon versions of pipeline V.2. Both of these patterns also matched across the two duplicate cerebellar samples, increasing their reliability. It should be noted that the phenomenon of higher levels of RNA editing in the cerebellum relative to the cerebral tissues has already been described (Tan et al., 2017). This relative increase in RNA editing was particularly pronounced for repetitive sites, indicative of hyper-editing in the introns and UTRs (Walkley and Li, 2017). This is likely why the number of genes differentially RNA edited between the two tissues goes down upon analysing the results for pipeline V.2 exons.

However, the fact that the cerebellum/cerebellar hemisphere skew remains even when restricting the analysis to exons and also that thousands more genes have a higher rate of RNA editing in the cerebellum/cerebellar hemisphere relative to the frontal cortex makes these results striking but also some of the most robust discussed in the chapter. Finally, the cerebellum appears to have many more SNVs across gene bodies and in exons than the frontal cortex, as revealed through the use of GATK's HaplotypeCaller. Importantly, this is true when comparing both sets of cerebellar samples to the cortex. The main question is, can these differences potentially tie into the sensitivity of the cerebellum to mutations in DNA repair proteins?

The cerebellum evidently has a distinct transcriptional profile relative to the cerebrum. This could tie into disease states in a number of ways. If certain networks of genes have a lower expression in the cerebellum, for example SSBR proteins, then it is possible that upon the loss of one key protein, the cerebellum is more affected because of lower expression of other proteins in that pathway or compensatory proteins, and so shows increased susceptibility to problems arising from the loss of that protein. Potential evidence for this hypothesis comes from the observation that the majority of the key SSBR proteins are downregulated in the cerebellum compared to the cerebrum. The fact DNA polymerase β is the most highly downregulated out of the assessed SSBR genes implies that the cerebellum has a less active base excision repair pathway. This may be because the cerebellum is less susceptible to the type of damage that requires BER for repair. However, when such damage is increased due to loss of specific repair proteins as in the DRDA-ARCAs, the reduced expression of this set of proteins may not be able to cope with the elevated levels of damage imposed on the tissue resulting in genome instability and subsequent cell death. Another consequence of large differences in gene expression between the cerebellum and the cerebrum is that they will likely acquire mutations in different sets of genes. This is because of the much discussed "use it and lose it" hypothesis, relating to the acquisition of somatic mutations in the developed brain (Lodato et al., 2015). To recapitulate, it because mature neurons do not replicate their DNA due to their not undergoing cell division, transcription is the major form of DNA processing that occurs in the adult brain, and so actively transcribed genes are more likely to acquire mutations. The major piece of evidence for this is that mutations are enriched at sites colocalising with

markers of active transcription, such as DNAse I and H3K4me3 histone marks. This implies that part of the reason actively transcribed genes are DNA damage hotspots is because they are areas of open chromatin. Additionally, the vast majority of the high confidence SNVs called were C→T mutations (Lodato et al., 2015). This in itself provides evidence for the transcriptional argument, as C→T mutations can be indicative of spontaneous cytosine deamination. This commonly happens at sites of single stranded DNA, which is particularly susceptible to this form of attack. A common way in which ssDNA is formed are R-loops, a structure formed when the nascent RNA invades the DNA duplex behind the RNA polymerase, thereby displacing the coding strand of DNA and making it single stranded and so susceptible to chemical attack, particularly deamination. By their very nature, R-loops arise at sites of active transcription, and so the very fact that C→Ts dominate the mutational spectra of the brain supports the idea that most damage in the developed brain is linked to transcription. Therefore, those genes which initially show high expression in a particular tissue are more likely to acquire transcriptionally associated mutations, and so over time lose function or normal levels of expression. Our data shows the transcriptional landscape of the cerebellum and the cerebrum are highly distinct. This means that different subsets of genes will be more subject to this transcriptionally linked damage across the different tissues. If the cerebellum has a specific subset of genes that that are both particularly susceptible to this type of damage because of their higher expression and important for the maintenance of genome stability, then it could be possible that over time these genes actually become downregulated in the cerebellum, even though they are initially more important for cerebellar viability. This is very hard to qualify and complicates interpretation of differential expression results, as we cannot say whether a slight downregulatory effect is a true biologically intended difference between the two tissues or the result of the “use it and lose it” hypothesis in action. However, for the “use it and lose it” effect to be the true interpretation of the downregulation of a set of genes, it would have to be remarkably consistent across many different cells in many different samples to give such a downregulation. If this was indeed the case, then this would be remarkable, but it seems unlikely, and perhaps much more likely to manifest as a relatively high frequency set of somatic mutations within a pool of genes. Nevertheless, there remains the possibility that this effect could push some genes from slight upregulation to slight downregulation, and if consistent across a pathway of genes, the combinatorial effect could have negative

consequences for cerebellar health. The ideal way to assess whether a subset of genes consistently switch from up to downregulation would be perform differential expression analyses by age groups and look for genes that show a stepwise decrease in expression as age increases, and then attempt to look for increased acquisition of mutations within these genes. This relates to the second unique feature of the cerebellum with respect to the cerebrum discovered in this thesis: more mitochondrially related genes appear to be downregulated in the cerebellum compared to the cerebrum. The obvious connection to be made to the DRDA-ARCAs is that mitochondrial dysfunction is a common feature of this group of diseases. There is no one clear way this can be linked to downregulation of mitochondrial genes, but some suggestions were made in the discussion section of chapter 1. It is possible that the cerebellum has less mitochondria/less mtDNA than the cerebrum, as has been shown in mice, so when mitochondrial dysfunction occurs as in the DRDA-ARCAs, there is a lower number of healthy mitochondria to compensate for those affected by the disease (Fuke et al., 2011). Another similar interpretation is that this lower level of expression of mitochondrial genes is not linked to any differences in terms of the numbers of mitochondria, but is just true as is. This again could cause problems in disease states in which mitochondrial function is affected, as perhaps not enough mitochondrial proteins are made to compensate for the dysfunction. However, an important consideration is that of compensatory gene expression. It is entirely possible that to offset these factors, the cell could increase expression of these mitochondrial genes, so this initial difference may not contribute to the pathology of the disease (Barthélémy et al., 2001). The third interpretation neatly links back to the "use it and lose it" hypothesis and towards the next differences between the cerebellum and the cerebrum/frontal cortex. Namely, perhaps mitochondrial genes are initially more highly expressed in the cerebellum but lose expression due to the acquisition of mutations. However, when we looked for enrichment of mitochondrial genes amongst genes with a higher rate of mismatch in the cerebellum compared to the frontal cortex, mitochondrial genes were enriched when they were downregulated but depleted when they were upregulated. This same pattern occurred with differentially expressed genes in general: those differentially downregulated were enriched and those upregulated showed underrepresentation. Therefore, it seems likely the enrichment of downregulated mitochondrial genes is due to their being downregulated and not because they are mitochondrial, so this explanation of why more differentially expressed mitochondrial genes

show downregulation in the cerebellum relative to the cerebrum falls short. Nevertheless, moving to focus on the mismatches data, many more genes were called as having a higher rate of mismatch in the cerebellum compared to the frontal cortex. This was true even between the cerebellar hemisphere and frontal cortex, both of which had been snap frozen for preservation as opposed to the fixing that the cerebellum samples were subject to. This pattern did break down at the level of the exons, where the frontal cortex had more significantly mismatched genes with a higher mismatch rate than the cerebellar hemisphere and the total number of mismatched genes fell dramatically, but the result is interesting nonetheless. The reduced number of genes differentially mismatched when just looking at exons also has a biological precedent, as it has been shown that exons are protected against damage in cancer cells and are also protected against oxidative damage (Frigola et al., 2017; Poetsch et al., 2018). However, to pick up these intronic mutations the GTEx RNA-seq datatset would have to contain many unspliced transcripts, which seems implausible. There is additionally the confounding effect of undocumented RNA-editing events being picked up as mismatches. However, bearing these caveats in mind, we can tentatively suggest that we have identified a set of genes that show differences in mutation rate between the cerebellum and frontal cortex.

How might the mutational spectra relate to the susceptibility of the cerebellum to disease? The two most interesting base changes that are significantly different between the cerebellum and the frontal cortex are C→T and G→T mutations. However, the major problem with the for both the C→T and G→T mutations is that the cerebellar hemisphere samples do not show a significant difference for these base changes relative to the frontal cortex and the cerebellum does show a difference relative to the hemisphere. The cerebellum and cerebellar hemisphere sample are essentially the same samples that have been preserved differently. The fact that the frontal cortex and cerebellar hemisphere samples were both preserved by snap freezing, whereas the cerebellum samples were fixed, means the frontal cortex and cerebellar hemisphere samples are more directly comparable to each other. Therefore, the significance of the C→T and G→T base changes for the cerebellum relative to the cerebellar hemisphere and the frontal cortex is extremely likely to be due to differences in preservation. The other significant base change not thought to be RNA editing was T→As, which had a significantly higher rate in the frontal cortex compared

to both cerebellar tissues. However, this base change cannot be linked to any common form of DNA damage, and so the underlying source of this difference rate in T→As remains unsolved.

In contrast to these uncertain findings, the RNA-editing difference was very robust. However, the link to disease is much more difficult to make because how perturbations in RNA editing may relate to pathology has not been readily explored. As mentioned, genes undergoing high levels of RNA editing in the brain are enriched for genes related to neurological disease (Picardi et al., 2015). Perhaps this difference in RNA editing is not a direct cause of differential disease sensitivity between the cerebellum and cerebrum, but rather RNA editing is secondarily affected under disease conditions and this has knock on effects. One can imagine a scenario whereby because the cerebellum has higher levels of RNA editing, RNA editing is therefore more important for regulating gene expression in the cerebellum. Because of this greater importance placed on the process, disturbances in RNA editing are likely to have a greater effect on the cerebellum than the cerebrum. It is plausible that under disease conditions, RNA editing is affected due to general disruption of cellular homeostasis, and this lack of or dysregulated RNA editing could have further negative consequences for cell viability. The other potential scenario briefly covered in the discussion section of this chapter was under the conditions of increased genome instability present in the DRDA-ARCAs, RNA editing sites are mutated, thereby leading to lack of RNA editing and downstream consequences. It was decided this was unlikely however, due to the fact that the major increase in RNA editing between the cerebellum and cerebrum is in repetitive regions which are known to undergo hyper-editing, and although it is not known, it seems unlikely that the mutation of a single site amongst a hyper-edited tract would have large consequences for this form of RNA editing (Tan et al., 2017; Walkley and Li, 2017).

Finally, we have the observation that the cerebellum appears to have more variants called than the frontal cortex. If and how this relates to cerebellar disease obviously depends upon the reasons behind this observation. Aside from differences arising from technical errors such as large differences in sample quality, the two most obvious explanations are either the cerebellum expresses SNP/SNV containing genes more highly, or that somatic mosaic variants are being called as wells as germline ones, and a difference in the number of mosaic mutations that have arisen during development is what is driving this

difference between the two tissues. The first scenario appears implausible, but not to say impossible. The most interesting interpretation for our work is certainly the somatic mosaic hypothesis. If the cerebellum acquires more mutations during development due to high levels of replication and its extended maturation period, then it could already be in some way relatively genomically unstable relative to the tissues of the cerebrum (Bae et al., 2018; El-Khamisy, 2011). Therefore, when this is further compounded by mutations in DNA repair proteins, the mutational burden becomes too much for the cells to cope with, which in turns leads to dysfunction, cell death and disease. A final point to make about how these unique features of the cerebellum may contribute to its sensitivity to neurological disease is that none of the suggested links are necessarily mutually exclusive. The tissue specificity of a disease can be complex issue and multiple factors may play a role in making the cerebellum susceptible to the loss of function of these DNA repair proteins, which could explain why the phenomenon has eluded easy explanation.

A genomic exploration of NuMA, best known as a structural protein involved in spindle assembly, at first may seem like a strange choice for a thesis focussed primarily on cerebellar specific brain diseases and their link with DNA damage. However, the fact that the mutations in the *LMNA* gene encoding the structural proteins lamin A/C can result in neuropathies and the emerging research linking NuMA to DNA repair means that this protein is an exciting candidate for a potential neurological disease related protein (Lamins reviewed in Gonzalo, 2014). Additionally, the fact that NuMA has far higher expression in the cerebellum compared to the cerebral tissues, a role obviously not related to its role during mitosis due to the non-dividing nature of neurons, implies a differential need for NuMA across the two tissues (Gaglio et al., 1995; Merdes et al., 1996, 2000). This could mean that somatic mosaic loss of NuMA function or germline attenuation of NuMA function may disproportionately affect the cerebellum. This in turn could be related to the little explored role of NuMA in the DNA damage response, reaffirming the sensitivity of the cerebellum to loss of function in DNA repair housekeeping genes. In order to shed light on this topic, our work has aimed to bioinformatically investigate the niche NuMA fills within the DDR.

The first major finding of this project is the discovery of a set of genes whose expression upon treatment with H_2O_2 appears to be dependent upon NuMA, as when NuMA

is knocked down, these genes cease to be overexpressed upon the induction of damage. This immediately draws parallels to the role of NuMA in the P53 mediated stress response, where it coordinates the expression of specific response genes. However, these NuMA regulated genes (NRGs) are also enriched for immediate early response genes (IERGs) and genes containing fragile introns (FIGs). As IERGs make up an exceedingly small proportion of NRGs, the most striking result was the FIG enrichment. Why FIGs are enriched amongst NRGs is not understood, but a reasonable interpretation would be that upon damage, these genes are more susceptible to breaks that would interfere with transcription and NuMA has a protective effect upon these genes. However, this merely pushes the problem back as to why FIGs are enriched amongst genes upregulated upon H_2O_2 treatment in the first place, and perhaps the same mechanism that means the majority of NRGs are downregulated upon NuMA KD and H_2O_2 treatment also governs the expression of the FIGs. It is also important to bear in mind FIGs only represent around a fifth of NRGs, and a common feature governing all NRGs has yet to be identified. However, NRGs do tend to be more highly expressed upon H_2O_2 treatment and also longer than control sets of genes.

The second important contribution of this chapter to NuMA studies is the exploration of where NuMA binds in the genome. Under normal conditions, NuMA is highly enriched at promoters. Upon H_2O_2 treatment, this pool of promoter bound NuMA disappears, but due to complications with these samples, this observation is less reliable and requires confirmation. NuMA is also independently enriched in the promoters of fragile intron genes, paused genes and genes overexpressed upon H_2O_2 treatment. Additionally, although the H_2O_2 treated AP-seq samples are questionable, the untreated ones show that upon knockdown of NuMA, damage in the promoter region and transcription termination site increases, another important new finding. Again, these results pose the question: is NuMA acting as a coordinator of the DNA damage response or is it directly involved? Research carried out in the El-Khamisy lab indicates that NuMA directly interacts with several single strand break repair proteins, and this in conjunction with the ChIP and AP-seq results strongly suggests that NuMA does indeed directly participate in DNA repair, most likely as a scaffold for the stabilisation or recruitment of other repair factors. However, the data from the 4sU-seq cannot be adequately explained by the protective effects of NuMA against DNA damage unless the NRGs are in some way more susceptible to damage than the

other genes overexpressed upon treatment with H_2O_2, which interferes with their expression in the absence of NuMA. However, this seems implausible, given that category exploration shows that most NRGs do not contain fragile sites and the AP-seq demonstrates that NRGs do not appear to be subject to more damage compared to a set of controls whether in the presence or absence of NuMA or H_2O_2 treatment. Therefore, this thesis has proposed a dual function of NuMA. This is a model whereby NuMA directly participates in the DNA damage response in some fashion, particularly preventing damage at promoters, which fits with the results from the NuMA ChIP-seq and AP-seq, but also coordinates the expression of DNA damage response genes, specifically, NRGs, independently of this direct involvement, similar to its role in the p53 mediated stress response (Endo et al., 2013; Ohata et al., 2013). There are still pieces of data without a satisfactory explanation, such as why NuMA is enriched at the promoters of NRGs, fragile introns and paused genes, and whether NuMA performs different functions in the promoters of these gene categories or whether there is some as of yet unidentified common link between them. However, we believe this model best fits the available evidence, is an important contribution to the study of NuMA, and will provide the basis upon which to build future work.

In conclusion, despite the lack of definitive explanations reached throughout this thesis, I believe this body of work has contributed to both the concept of the cerebellum as a unique tissue within the brain and how this intersects with its sensitivity to mutations in DNA repair proteins, and also the emerging link between structural proteins and the DNA damage response. The results described here provide much scope for further research and refinement both in consolidating the existing work and taking it in new directions by attempting to answer the questions it has raised, and will hopefully spur further research in these areas going into the future. An obvious follow up to the work carried out in chapter 1 would be to look at differences in mtDNA content and numbers of mitochondria across the tissues of the human brain in order to try and confirm in humans what has already been shown in mouse, as this would perhaps partly explain our observed mitochondrial downregulation. Another ideal experiment would be single cell DNA sequencing and RNA sequencing of the cerebellum and a selection of cerebral tissues and subsequent differential expression analyses and somatic mutation calling to see whether the cerebellum had higher rates of mutation for mitochondrial genes and whether those genes with a higher rate of

mutation had lower levels of expression. This would also be a suitable experiment to build on chapter 2, but across all genes and not just mitochondrial genes. It would help us to tackle both the question of whether the cerebellum has more somatic mosaic mutations as suggested by the increase in germline variant calls in the cerebellum or cell private mutations in specific genes as determined by the mismatches pipeline. As of now, the only publicly available single cell data across brain region is RNA-seq data, so repetition of our analyses in this dataset might give us confidence our observations hold true before moving on the single cell DNA sequencing for more accurate picture. Before this was done however, it would be important to perform some more basic follow ups to clarify the existing work. The obvious next step would be to repeat the analyses but this time include the fixed frontal cortex samples from GTEx as well as the snap frozen ones, so the fixed cerebellum samples had a direct counterpart to which it could be compared. This might help determine whether some of the differential rates of base changes were due to the different ways in which the tissues were preserved. Including other cerebral tissues as well as the frontal cortex might also help to further solidify the observed cerebellar-cerebral differences. It might also be beneficial to run our data through some publicly available somatic variant callers for RNA-seq data and see if the results from these programs support our findings. Another important thing lacking from this work was a positive control for our mismatches pipeline. Unpublished work from another group looking at somatic mutations across the GTEx dataset used skin samples from European and African individuals to test their somatic mutation calling software. According to their mutation caller, sun exposed skin samples from European individuals accrued more mutations than the equivalent samples from African individuals, as would be expected. This was taken to be a proof of principle that their software was performing to their standards and was picking up biologically relevant differences. A repetition of this approach using our mismatches pipeline would be a way to get confirmation as to whether the pipeline is indeed functioning as we intended and that the results are meaningful. It would also be interesting to look at how the list of genes with differential rates of mismatch between the tissues changed with age, and whether this was linked to variations in the differential expression of these genes, in order to further explore how genes with higher rates of mismatch might end up downregulated, as per the “use it and lose it” hypothesis (Lodato et al., 2015). Ultimately, it would ideal if we could perform single cell DNA and RNA sequencing on the cerebellum and cerebrum of individuals who

were DRDA-ARCA patients, to see whether certain subsets of genes had a higher mutation rate in the cerebellum and how this affected gene expression. This could then be cross-referenced with our previous work to see whether it was a similar set of genes that had a less pronounced but still higher rate of mismatch in WT cerebellum, such that the DRDA-ARCAs might represent an accelerated version of what is happening in the WT cerebellum. This would be highly unlikely however, both because of the rarity of some these diseases and the difficulty of acquiring brain samples. One important closing point to make on this body of work is the importance of clear and accurate metadata for publicly available datasets. The difference between the cerebellum and the cerebellum hemisphere was only discovered during the course of writing this thesis, even though such information on tissue preservation for the GTEx RNA-seq data had been looked for prior to this. Had we known that these two sets of samples were actually differentially preserved replicates, the experimental approach taken and subsequent analyses would have been performed differently.

As for the study of NuMA, the important follow up experiments should try to clarify the hypothesis laid out here: that NuMA acts to upregulate a set of genes upon DNA damage and also is directly involved in the repair of such damage. The major unresolved observation why there is increased binding of NuMA to the promoters of specific categories of genes, and so elucidating whether this increased occupancy is linked to direct repair or upregulation upon DNA damage exposure would be crucial. Repetition of the H_2O_2 treated NuMA ChIP-seq would be ideal due to the technical problems mentioned. This might help to see whether NuMA truly leaves promoters upon exposure to DNA damage or whether it is retained on specific subsets of genes, perhaps those particularly susceptible to damage such as paused genes or genes with fragile introns, in order to facilitate repair. Another necessary step would be further work on the AP-seq to investigate whether the H_2O_2 treated samples are accurately recapturing previously published observations, and if not, why not? AP-seq metagene profiles over the promoters of genes with fragile introns and paused genes may show differences in damage between samples where the NRG AP-seq profiles did not, potentially indicating an increased requirement for direct involvement of NuMA in protection from, or repair of, DNA damage at these promoters. Further investigations into the properties of NRGs would also be part of further work. Perhaps some as of yet

unidentified common feature may yet be found that would explain why NuMA is enriched at their promoters, aside from the obvious explanation that its increased binding facilitates their expression upon DNA damage. This hypothesis is unsatisfactory because leads to a complicated scenario where NuMA is performing different roles at the promoters of different categories of genes. Looking at whether NuMA KD in WT cells affects NRG expression would be interesting, it would indicate whether NuMA is required for the basal expression of these genes or whether it is only required for their upregulation in response to DNA damage. Finally, genes with fragile introns was one of the consistently interesting categories explored, being both enriched amongst NRGs and showing elevated levels of NuMA binding in their promoters. Further work on how NuMA interacts with genes with fragile introns could involve looking at NuMA ChIP-seq metagenes across introns, as our current gene body metagenes do not delineate between introns and exons. If NuMA was found to have increased occupancy in fragile introns, it would be strong evidence that NuMA binding at specific genomic features helps to repair or protect from DNA damage at such regions.

Appendix

Software Versions

Software/Package	Version
R	3.2.2
R studio	1.0.136
samtools	1.3.1
FastQC	0.11.3
hisat2	2.1.0
CGAT	Git commit id: 4c3375a
picard MarkDuplicates	1.135
GATK	3.8
bcftools	1.3.1
pysam	0.10.0
pyvcf	0.6.8
org.Hs.eg.db	3.4.0
GO.db	3.4.0
GOseq	1.26.0
Biostrings	2.42.1
polyester	1.10.1
dplyr	0.8.1
tidyr	0.8.3
gplots	3.0.1.1
ggplot2	3.1.1
SRA-toolkit	2.9.1_1
Limma	3.24.15
edgeR	3.10.4
FeatureCounts	1.5.3
gridExtra	2.3
Formattable	0.2.0.1

Software/Package	Version
STAR	2.4.2a
salmon	0.11.4
tximport	1.2.0
biomaRt	2.30.0
bedtools	2.25.0
BWA	0.7.17
MACS2	2.1.1.20160309
pandas	0.21.1
RSQLite	2.1.1
jupyter	1.0.0
jupyter_client	5.0.1
jupyter_console	5.1.0
jupyter_console	4.3.0
jupyter_core	0.7.2
python	2.7.12
ruffus	2.6.3

Bibliography

Abad, Patricia C.; Lewis, Jason; Mian, I. Saira; Knowles, David W.; Sturgis, Jennifer; Badve, Sunil; Xie, Jun and Lelie`vre, S.A. (2007). NuMA Influences Higher Order Chromatin Organization in Human Mammary Epithelium. Mol. Biol. Cell *18*, 348–361.

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat. Rev. Genet. *13*, 720–731.

Agostino, C.D., Ryu, T., Zapotoczny, G., Delabaere, L., Li, X., Khodaverdian, V.Y., Amaral, N., Lin, E., Rau, A.R., and Chiolo, I. (2018). Nuclear F-actin and myosins drive relocalization of heterochromatic breaks. Nature *559*, 54–60.

Ahel, I., Rass, U., El-Khamisy, S.F., Katyal, S., Clements, P.M., McKinnon, P.J., Caldecott, K.W., and West, S.C. (2006). The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates. Nature *443*, 713–716.

Akbari, M., Sykora, P., and Bohr, V.A. (2015). Slow mitochondrial repair of 5'-AMP renders mtDNA susceptible to damage in APTX deficient cells. Sci. Rep. *5*, 1–8.

Alagoz, M., Chiang, S.C., Sharma, A., and El-Khamisy, S.F. (2013). ATM Deficiency Results in Accumulation of DNA-Topoisomerase I Covalent Intermediates in Neural Cells. PLoS One *8*, 1–9.

Allen Insitute, B. (2010). © 2010 Allen Institute for Brain Science. BrainSpan Atlas of the Developing Human Brain. Available from: https://www.brainspan.org/.

Alter, B.P., Rosenberg, P.S., and Brody, L.C. (2007). Clinical and molecular features associated with biallelic mutations in FANCD1/BRCA2. J. Med. Genet. *44*, 1–9.

Ambrose, M., and Gatti, R. a. (2013). Pathogenesis of ataxia-telangiectasia: the next generation of ATM functions. Blood *121*, 4036–4045.

Ambrose, M., Goldstine, J. V., and Gatti, R.A. (2007). Intrinsic mitochondrial dysfunction in ATM-deficient lymphoblastoid cells. Hum. Mol. Genet. *16*, 2154–2164.

Amente, S., Di Palo, G., Scala, G., Castrignanò, T., Gorini, F., Cocozza, S., Moresano, A., Pucci, P., Ma, B., Stepanov, I., et al. (2019). Genome-wide mapping of 8-oxo-7,8-dihydro-2'-deoxyguanosine reveals accumulation of oxidatively-generated damage at DNA replication

origins within transcribed long genes of mammalian cells. Nucleic Acids Res. *47*, 221–236.

Andrews, S. (2015). FASTQC A Quality Control tool for High Throughput Sequence Data. Babraham Inst. http://www.bioinformatics.babraham.ac.uk/projects/.

Anttinen, A., Koulu, L., Nikoskelainen, E., Portin, R., Kurki, T., Erkinjuntti, M., Jaspers, N.G.J., Raams, A., Green, M.H.L., Lehmann, A.R., et al. (2008). Neurological symptoms and natural course of xeroderma pigmentosum. Brain *131*, 1979–1989.

Aquilina, G., Biondo, R., Dogliotti, E., Meuth, M., and Bignami, M. (1992). Expression of the Endogenous O6-Methylgua-nine-DNA-methyl-transferase Protects Chinese Hamster Ovary Cells from Spontaneous G:C to A:T Transitions. Cancer Res. *52*, 6471–6475.

Auguie, B. (2017). GridExtra: Miscellaneous Functions for "Grid''' Graphics." R Packag. Version 2.3 https://CRAN.R-project.org/package=gridExtra.

Azevedo, F.A.C., Carvalho, L.R.B., Grinberg, L.T., Farfel, J.M., Ferretti, R.E.L., Leite, R.E.P., Filho, W.J., Lent, R., and Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J. Comp. Neurol. *513*, 532–541.

Bae, T., Tomasini, L., Mariani, J., Zhou, B., Roychowdhury, T., Franjic, D., Pletikos, M., Pattni, R., Chen, B.J., Venturini, E., et al. (2018). Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science (80-.). *359*, 550–555.

Barnes, D.E., and Lindahl, T. (2004). Repair and genetic consequences of endogenous DNA base damage in mammalian cells. Annu. Rev. Genet. *38*, 445–476.

Barthélémy, C., De Baulny, H.O., Diaz, J., Cheval, M.A., Frachon, P., Romero, N., Goutieres, F., Fardeau, M., and Lombès, A. (2001). Late-onset mitochondrial DNA depletion: DNA copy number, multiple deletions, and compensation. Ann. Neurol. *49*, 607–617.

Barzilai, A., Schumacher, B., and Shiloh, Y. (2017). Genome instability: Linking ageing and brain degeneration. Mech. Ageing Dev. *161*, 4–18.

Bélanger, M., and Magistretti, P.J. (2009). The role of astroglia in neuroprotection. Dialogues Clin. Neurosci. *11*, 281–296.

Bélanger, M., Allaman, I., and Magistretti, P.J. (2011). Brain energy metabolism: Focus on

Astrocyte-neuron metabolic cooperation. Cell Metab. *14*, 724–738.

Bendixen, C., Thomsen, B., Alsner, J., and Westergaard, O. (1990). Camptothecin-stabilized topoisomerase I-DNA adducts cause premature termination of transcription. Biochemistry *29*, 5613–5619.

Blokzijl, F., De Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. Nature *538*, 260–264.

De Boer, J., and Hoeijmakers, J.H.J. (2000). Nucleotide excision repair and human syndromes. Carcinogenesis *21*, 453–460.

Bootsma, D., and Hoeijmakers, J.H.J. (1991). The genetic basis of xeroderma pigmentosum. Ann. Genet. *34*, 143–150.

Bradley, M.O., and Kohn, K.W. (1979). X-ray induced DNA double strand break production and repair in mammalian cells as measured by neutral filter elution. Nucleic Acids Res. *7*, 793–804.

Bransteitter, R., Pham, P., Scharfft, M.D., and Goodman, M.F. (2003). Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc. Natl. Acad. Sci. U. S. A. *100*, 4102–4107.

Bras, J., Alonso, I., Barbot, C., Costa, M.M., Darwent, L., Orme, T., Sequeiros, J., Hardy, J., Coutinho, P., and Guerreiro, R. (2015). Mutations in PNKP cause recessive ataxia with oculomotor apraxia type 4. Am. J. Hum. Genet. *96*, 474–479.

Broad Institute (2016). Picard Tools. p. http://broadinstitute.github.io/picard/.

Brüning-Richardson, A., Bond, J., Alsiary, R., Richardson, J., Cairns, D.A., McCormac, L., Hutson, R., Burns, P.A., Wilkinson, N., Hall, G.D., et al. (2012). NuMA overexpression in epithelial ovarian cancer. PLoS One *7*, e38945.

Buck, D., Malivert, L., Chasseval, D., Barraud, A., Fondane, M., Hufnagel, M., Sanal, O., Plebani, A., and Ste, J. (2006). Cernunnos , a Novel Nonhomologous End-Joining Factor , Is Mutated in Human Immunodeficiency with Microcephaly. Cell *124*, 287–299.

Caldecott, K.W. (2007). Mammalian single-strand break repair: Mechanisms and links with

chromatin. DNA Repair (Amst). *6*, 443–453.

Caldecott, K.W. (2008). Single strand break repair and genetic disease. Nat. Rev. Genet. *9*, 619–631.

Caldecott, K.W., Aoufouchi, S., Johnson, P., and Shall, S. (1996). XRCC1 polypeptide interacts with DNA polymerase β and possibly poly (ADP-ribose) polymerase, and DNA ligase III is a novel molecular "nick-sensor" in vitro. Nucleic Acids Res. *24*, 4387–4394.

Carithers, L.J., and Moore, H.M. (2015). The Genotype-Tissue Expression (GTEx) Project.

Carlson, M. (2019a). org.Hs.eg.db: Genome wide annotation for Human. R Packag. Version 3.1.2.

Carlson, M. (2019b). GO.db: A set of annotation maps describing the entire Gene Ontology. R package version 3.8.2.

Carney, J.P., Maser, R.S., Olivares, H., Davis, E.M., Le Beau, M., Yates, J.R., Hays, L., Morgan, W.F., and Petrini, J.H.J. (1998). The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: Linkage of double-strand break repair to the cellular DNA damage response. Cell *93*, 477–486.

Carson, C.T., Schwartz, R. a, Stracker, T.H., Lilley, C.E., Lee, D. V, and Weitzman, M.D. (2003). The Mre11 complex is required for ATM activation and the G2/M checkpoint. EMBO J. *22*, 6610–6620.

Chen, S.H., Chan, N.-L., and Hsieh, T. (2013). New mechanistic and functional insights into DNA topoisomerases. Annu. Rev. Biochem. *82*, 139–170.

Cheng, K.C., Cahill, D.S., Kasai, H., Nishimura, S., and Loeb, L.A. (1992). 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G → T and A → C substitutions. J. Biol. Chem. *267*, 166–172.

Chiang, S.C., Meagher, M., Kassouf, N., Hafezparast, M., McKinnon, P.J., Haywood, R., and El-Khamisy, S.F. (2017). Mitochondrial protein-linked DNA breaks perturb mitochondrial gene transcription and trigger free radical–induced DNA damage. Sci. Adv. *3*, 1–15.

Chou, W.-C., Hu, L.-Y., Hsiung, C.-N., and Shen, C.-Y. (2015). Initiation of the ATM-Chk2 DNA damage response through the base excision repari pathway. Carcinogenesis *36*, 832–840.

Chun, H.H., and Gatti, R.A. (2004). Ataxia-telangiectasia, an evolving phenotype. DNA Repair (Amst). *3*, 1187–1196.

Cleveland, D.W. (1995). NuMA: a protein involved in nuclear structure, spindle assembly, and nuclear re-formation. Trends Cell Biol. *5*, 60–64.

Colussi, C., Parlanti, E., Degan, P., Aquilina, G., Barnes, D., Macpherson, P., Karran, P., Crescenzi, M., Dogliotti, E., and Bignami, M. (2002). The Mammalian Mismatch Repair pathway removes DNA 8-oxodGMP incorporated from the oxidized dNTP pool. Curr. Biol. *12*, 912–918.

Compton, D.A., Szilak, I., and Cleveland, D.W. (1992). Primary structure of NuMA, an intranuclear protein that defines a novel pathway for segregation of proteins at mitosis. J. Cell Biol. *116*, 1395–1408.

Coudray, A., Battenhouse, A.M., Bucher, P., and Iyer, V.R. (2018). Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. PeerJ *6*, 1–23.

Das, B.B., Dexheimer, T.S., Maddali, K., and Pommier, Y. (2010). Role of tyrosyl-DNA phosphodiesterase (TDP1) in mitochondria. Proc. Natl. Acad. Sci. U. S. A. *107*, 19790–19795.

Date, H., Onodera, O., Tanaka, H., Iwabuchi, K., Uekawa, K., Igarashi, S., Koike, R., Hiroi, T., Yuasa, T., Awaya, Y., et al. (2001). Early-onset ataxia with ocular motor apraxia and hypoalbuminemia is caused by mutations in a new HIT superfamily gene. Nat. Genet. *29*, 184–188.

Dellino, G.I., Palluzzi, F., Chiariello, A.M., Piccioni, R., Bianco, S., Furia, L., De Conti, G., Bouwman, B.A.M., Melloni, G., Guido, D., et al. (2019). Release of paused RNA polymerase II at specific loci favors DNA double-strand-break formation and promotes cancer translocations. Nat. Genet. *51*, 1011–1023.

Díaz-García, C.M., and Yellen, G. (2019). Neurons rely on glucose rather than astrocytic lactate during stimulation. J. Neurosci. Res. *97*, 883–889.

Digiovanna, J.J., and Kraemer, K.H. (2012). Shining a light on xeroderma pigmentosum. J. Invest. Dermatol. *132*, 785–796.

Digweed, M., and Sperling, K. (2004). Nijmegen breakage syndrome: Clinical manifestation

of defective response to DNA double-strand breaks. DNA Repair (Amst). *3*, 1207–1217.

Dizdaroglu, M. (1991). Chemical Determination of Free Radical Induced Damage to DNA. Free Radic. Biol. Med. *10*, 225–242.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dong, Z., and Tomkinson, A.E. (2006). ATM mediates oxidative stress-induced dephosphorylation of DNA ligase IIIa? Nucleic Acids Res. *34*, 5721–5729.

Dringen, R., Pfeiffer, B., and Hamprecht, B. (1999). Synthesis of the antioxidant glutathione in neurons: Supply by astrocytes of CysGly as precursor for neuronal glutathione. J. Neurosci. *19*, 562–569.

Driscoll, M.O., Cerosaletti, K.M., Girard, P., Dai, Y., Stumm, M., Kysela, B., Hirsch, B., Gennery, A., Palmer, S.E., Gatti, R.A., et al. (2001). DNA Ligase IV Mutations Identified in Patients Exhibiting Developmental Delay and Immunodeficiency. Mol. Cell *8*, 1175–1185.

Dumitrache, L.C., and McKinnon, P.J. (2017). Polynucleotide kinase-phosphatase (PNKP) mutations and neurologic disease. Mech. Ageing Dev. *161*, 121–129.

Duncan, B.K., and Weiss, B. (1982). Specific mutator effects of ung (uracil-DNA glycosylase) mutations in Escherichia coli. J. Bacteriol. *151*, 750–755.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Eisenberg, E., and Levanon, E.Y. (2018). A-to-I RNA editing - Immune protector and transcriptome diversifier. Nat. Rev. Genet. *19*, 473–490.

El-Khamisy, S.F. (2011). To live or to die: a matter of processing damaged DNA termini in neurons. EMBO Mol. Med. *3*, 78–88.

El-Khamisy, S.F., Masutani, M., Suzuki, H., and Caldecott, K.W. (2003). A requirement for PARP-1 for the assembly or stability of XRCC1 nuclear foci at sites of oxidative DNA damage. Nucleic Acids Res. *31*, 5526–5533.

El-Khamisy, S.F., Hartsuiker, E., and Caldecott, K.W. (2007). TDP1 facilitates repair of ionizing radiation-induced DNA single-strand breaks. DNA Repair (Amst). *6*, 1485–1495.

El-Khamisy, S.F., Katyal, S., Patel, P., Ju, L., McKinnon, P.J., and Caldecott, K.W. (2009). Synergistic decrease of DNA single-strand break repair rates in mouse neural cells lacking both Tdp1 and aprataxin. DNA Repair (Amst). *8*, 760–766.

Ellenberger, T., and Tomkinson, A.E. (2008). Eukaryotic DNA ligases: structural and functional insights. Annu. Rev. Biochem. *77*, 313–338.

Endo, A., Moyori, A., Kobayashi, A., and Wong, R.W. (2013). Nuclear mitotic apparatus protein, NuMA, modulates p53-mediated transcription in cancer cells. Cell Death Dis. *4*, e713.

Ensminger, M., Iloff, L., Ebel, C., Nikolova, T., Kaina, B., and Löbrich, M. (2014). DNA breaks and chromosomal aberrations arise when replication meets base excision repair. J. Cell Biol. *206*, 29–43.

Faghri, S., Tamura, D., Kraemer, K.H., and DiGiovanna, J.J. (2008). Trichothiodystrophy: A systematic review of 112 published cases characterises a wide spectrum of clinical manifestations.

Fam, H.K., Chowdhury, M.K., Walton, C., Choi, K., Boerkoel, C.F., and Hendson, G. (2013). Expression profile and mitochondrial colocalization of Tdp1 in peripheral human tissues. J. Mol. Histol. *44*, 481–494.

Fogel, B.L., and Perlman, S. (2007). Clinical features and molecular genetics of autosomal recessive cerebellar ataxias. Lancet Neurol *6*, 245–257.

Frazee, A.C., Jaffe, A.E., Langmead, B., and Leek, J.T. (2015). Polyester: Simulating RNA-seq datasets with differential transcript expression. Bioinformatics *31*, 2778–2784.

Frigola, J., Sabarinathan, R., Mularoni, L., Muinõs, F., Gonzalez-Perez, A., and López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. Nat. Genet. *49*,

1684–1692.

Fu, Y., Ito, F., Zhang, G., Fernandez, B., Yang, H., and Chen, X.S. (2015). DNA cytosine and methylcytosine deamination by APOBEC3B: Enhancing methylcytosine deamination by engineering APOBEC3B. Biochem. J. *471*, 25–35.

Fuke, S., Kubota-Sakashita, M., Kasahara, T., Shigeyoshi, Y., and Kato, T. (2011). Regional variation in mitochondrial DNA copy number in mouse brain. Biochim. Biophys. Acta - Bioenerg. *1807*, 270–274.

Fukuhara, N., Nakajima, T., Sakajiri, K., Matsubara, N., and Fujita, M. (1995). Hereditary motor and sensory neuropathy associated with cerebellar atrophy (HMSNCA): a new disease. J Neurol Sci *133*, 140–151.

Gaglio, T., Saredi, A., and Compton, D.A. (1995). NuMA is required for the organization of microtubules into aster-like mitotic arrays. J. Cell Biol. *131*, 693–708.

Gao, Y., Katyal, S., Lee, Y., Zhao, J., Rehg, J.E., Russell, H.R., and McKinnon, P.J. (2011). DNA ligase III is critical for mtDNA integrity but not Xrcc1-mediated nuclear DNA repair. Nature *471*, 240–244.

Garcia-Diaz, B., Barca, E., Balreira, A., Lopez, L.C., Tadesse, S., Krishna, S., Naini, A., Mariotti, C., Castellotti, B., and Quinzii, C.M. (2015). Lack of aprataxin impairs mitochondrial functions via downregulation of the APE1/NRF1/NRF2 pathway. Hum. Mol. Genet. *24*, 4516–4529.

Giannakakis, A., Zhang, J., Jenjaroenpun, P., Nama, S., Zainolabidin, N., Aau, M.Y., Yarmishyn, A.A., Vaz, C., Ivshina, A. V., Grinchuk, O. V., et al. (2015). Contrasting expression patterns of coding and noncoding parts of the human genome upon oxidative stress. Sci. Rep. *5*.

Gilad, F., Yoav, V., Sima, B., Gilad, S., Amit, I., and Oren1, M. (2014). 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome Biol. *15*, R69.

Gonzalez-suarez, I., Redwood, B., Perkins, S.M., Vermolen, B., Lichtensztejin, D., Grotsky, D.A., Morgado-palacin, L., Gapud, E.J., Sleckman, B.P., Sage, J., et al. (2009). Novel roles for A-type lamins in telomere biology and the DNA damage response pathway. EMBO J. *28*,

2414–2427.

Gonzalo, S. (2014). DNA Damage and Lamins. Adv. Exp. Med. Biol. *773*, 377–399.

Goodstadt, L. (2010). Ruffus: A lightweight python library for computational pipelines. Bioinformatics *26*, 2778–2779.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., and Eussen, B.H. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Guo, Z., Kozlov, S., Lavin, M.F., Person, M.D., and Paull, T.T. (2010). ATM activation by oxidative stress. Science (80-.). *330*, 517–521.

Hakonen, A.H., Heiskanen, S., Juvonen, V., Lappalainen, I., Luoma, P.T., Rantamaki, M., Goethem, G. Van, Lofgren, A., Hackman, P., Paetau, A., et al. (2005). Mitochondrial DNA polymerase W748S mutation: a common cause of autosomal recessive ataxia with ancient European origin. Am. J. Hum. Genet. *77*, 430–441.

Harborth, J. (1999). Self assembly of NuMA: multiarm oligomers as structural units of a nuclear lattice. EMBO J. *18*, 1689–1700.

Harborth, J., Elbashir, S.M., Bechert, K., Tuschl, T., and Weber, K. (2001). Identification of essential genes in cultured mammalian cells using small interfering RNAs. J. Cell Sci. *114*, 4557–4565.

Harris, J.L., Jakob, B., Taucher-Scholz, G., Dianov, G.L., Becherel, O.J., and Lavin, M.F. (2009). Aprataxin, poly-ADP ribose polymerase 1 (PARP-1) and apurinic endonuclease 1 (APE1) function together to protect the genome against oxidative damage. Hum. Mol. Genet. *18*, 4102–4117.

Hegde, M.L., Hazra, T.K., and Mitra, S. (2008). Early steps in the DNA base excision/single-strand interruption repair pathway in mammalian cells. Cell Res. *18*, 27–47.

Herrup, K. (2004). Divide and Die: Cell Cycle Events as Triggers of Nerve Cell Death. J. Neurosci. *24*, 9232–9239.

Herrup, K., and Busser, J.C. (1995). The Induction of Multiple Cell-Cycle Events Precedes Target-Related Neuronal Death. Development *121*, 2385–2395.

Hirano, R., Interthal, H., Huang, C., Nakamura, T., Deguchi, K., Choi, K., Bhattacharjee, M.B., Arimura, K., Umehara, F., Izumo, S., et al. (2007). Spinocerebellar ataxia with axonal neuropathy: consequence of a Tdp1 recessive neomorphic mutation? EMBO J. *26*, 4732–4743.

Hoch, N.C., Hanzlikova, H., Rulten, S.L., Tétreault, M., Komulainen, E., Ju, L., Hornyak, P., Zeng, Z., Gittens, W., Rey, S.A., et al. (2017). XRCC1 mutation is associated with PARP1 hyperactivation and cerebellar ataxia. Nature *541*, 87–91.

Hoeijmakers, J.H.J. (2009). DNA Damage, Aging, and Cancer. N. Engl. J. Med. *361*, 1475–1485.

Hsiang, Y.H., Lihou, M.G., and Liu, L.F. (1989). Arrest of Replication Forks by Drug-stabilized Topoisomerase I-DNA Cleavable Complexes as a Mechanism of Cell Killing by Camptothecin. Cancer Res. *49*, 5077–5082.

Hudson, E.K., Hogue, B. a, Souza-Pinto, N.C., Croteau, D.L., Anson, R.M., Bohr, V. a, and Hansford, R.G. (1998). Age-associated change in mitochondrial DNA damage. Free Radic. Res. *29*, 573–579.

Impellizzeri, K.J., Anderson, B., and Burgers, P.M.J. (1991). The spectrum of spontaneous mutations in a Saccharomyces cerevisiae uracil-DNA-glycosylase mutant limits the function of this enzyme to cytosine deamination repair. J. Bacteriol. *173*, 6807–6810.

Interthal, H., Pouliot, J.J., and Champoux, J.J. (2001). The tyrosyl-DNA phosphodiesterase Tdp1 is a member of the phospholipase D superfamily. Proc. Natl. Acad. Sci. U. S. A. *98*, 12009–12014.

Interthal, H., Chen, H.J., and Champoux, J.J. (2005). Human Tdp1 cleaves a broad spectrum of substrates, including phosphoamide linkages. J. Biol. Chem.

Ito, H., Fujita, K., Tagawa, K., Chen, X., Homma, H., Sasabe, T., Shimizu, J., Shimizu, S., Tamura, T., Muramatsu, S., et al. (2015). HMGB 1 facilitates repair of mitochondrial DNA damage and extends the lifespan of mutant ataxin-1 knock-in mice . EMBO Mol. Med. *7*, 78–101.

Iyama, T., and Wilson, D.M. (2013). DNA repair mechanisms in dividing and non-dividing

cells. DNA Repair (Amst). *12*, 620–636.

Jayaraman, S., Chittiboyina, S., Bai, Y., Abad, P.C., Vidi, P.A., Stauffacher, C. V., and Lelièvre, S.A. (2017). The nuclear mitotic apparatus protein NuMA controls rDNA transcription and mediates the nucleolar stress response in a p53-independent manner. Nucleic Acids Res. *45*, 11725–11742.

Kallajoki, M., Harborth, J., Weber, K., and Osborn, M. (1993). Microinjection of a monoclonal antibody against SPN antigen, now identified by peptide sequences as the NuMA protein, induces micronuclei in PtK2 cells. J. Cell Sci. *104*, 139–150.

Kasiviswanathan, R., and Copeland, W.C. (2011). Ribonucleotide discrimination and reverse transcription by the human mitochondrial DNA polymerase. J. Biol. Chem. *286*, 31490–31500.

Kathe, S.D., Shen, G.P., and Wallace, S.S. (2004). Single-Stranded Breaks in DNA but Not Oxidative DNA Base Damages Block Transcriptional Elongation by RNA Polymerase II in HeLa Cell Nuclear Extracts. J. Biol. Chem. *279*, 18511–18520.

Katyal, S., El-Khamisy, S.F., Russell, H.R., Li, Y., Ju, L., Caldecott, K.W., and McKinnon, P.J. (2007). TDP1 facilitates chromosomal single-strand break repair in neurons and is neuroprotective in vivo. EMBO J. *26*, 4720–4731.

Keegan, L.P., Gallo, A., and O'Connell, M.A. (2001). The many roles of an RNA editor. Nat. Rev. Genet. *2*, 869–878.

Kemp, K.C., Cook, A.J., Redondo, J., Kurian, K.M., Scolding, N.J., and Wilkins, A. (2016). Purkinje cell injury, structural plasticity and fusion in patients with Friedreich's ataxia. Acta Neuropathol. Commun. *4*, 1–15.

Khoronenkova, S. V., and Dianov, G.L. (2015). ATM prevents DSB formation by coordinating SSB repair and cell cycle progression. Proc. Natl. Acad. Sci. U. S. A. *112*, 3997–4002.

Kim, D., Langmead, B., and Salzberg, S.L. (2015a). Hisat2. Nat. Methods.

Kim, D., Langmead, B., and Salzberg, S.L. (2015b). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Korcok, J., Dixon, S.J., Lo, T.C.Y., and Wilson, J.X. (2003). Differential effects of glucose on

dehydroascorbic acid transport and intracellular ascorbate accumulation in astrocytes and skeletal myocytes. Brain Res. *993*, 201–207.

Kotula, E., Faigle, W., Berthault, N., Dingli, F., Loew, D., Sun, J.S., Dutreix, M., and Quanz, M. (2013). DNA-PK target identification reveals novel links between DNA repair signaling and cytoskeletal regulation. PLoS One *8*.

Krishnakumar, R., and Kraus, W.L. (2010). The PARP Side of the Nucleus: Molecular Actions, Physiological Outcomes, and Clinical Targets. Mol. Cell *39*, 8–24.

Kubota, Y., Nash, R.A., Klungland, A., Schär, P., Barnes, D.E., and Lindahl, T. (1996). Reconstitution of DNA base excision-repair with purified human proteins: interaction between DNA polymerase beta and the XRCC1 protein. EMBO J. *15*, 6662–6670.

Kuzminov, A. (2001). Single-strand interruptions in replicating chromosomes cause double-strand breaks. Proc. Natl. Acad. Sci. U. S. A. *98*, 8241–8246.

Lamarche, B.J., Orazio, N.I., and Weitzman, M.D. (2010). The MRN complex in double-strand break repair and telomere maintenance. FEBS Lett. *584*, 3682–3695.

Lange, W. (1975). Cell number and cell density in the cerebellar cortex of man and some other mammals. Cell Tissue Res. *157*, 115–124.

Leeuw, R. De, Gruenbaum, Y., and Medalia, O. (2018). Nuclear Lamins : Thin Filaments with Major Functions. Trends Cell Biol. *28*, 34–45.

Lelievre, S.A., Weaver, V.M., Nickerson, J.A., Larabell, C.A., Bhaumik, A., Petersen, O.W., and Bissell, M.J. (1998). Tissue phenotype depends on reciprocal interactions between the extracellular matrix and the structural organization of the nucleus. Proc. Natl. Acad. Sci. *95*, 14711–14716.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, J., and Gilmour, D.S. (2011). Promoter proximal pausing and the control of gene

expression. Curr. Opin. Genet. Dev.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Liu, B., Wang, J., Chan, K.M., Tjia, W.M., Deng, W., Guan, X., Huang, J.D., Li, K.M., Chau, P.Y., Chen, D.J., et al. (2005). Genomic instability in laminopathy-based premature aging. Nat. Med. *11*, 780–785.

Liu, C., Pouliot, J.J., and Nash, H.A. (2002). Repair of topoisomerase I covalent complexes in the absence of the tyrosyl-DNA phosphodiesterase Tdp1. Proc. Natl. Acad. Sci. U. S. A. *99*, 14970–14975.

Lodato, M.A., Woodworth, M.B., Lee, S., Evrony, G.D., Mehta, B.K., Karger, A., Lee, S., Chittenden, T.W., D'Gama, A.M., Cai, X., et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. Science (80-.). *350*, 94–98.

Lodato, M.A., Rodin, R.E., Bohrson, C.L., Coulter, M.E., Barton, A.R., Kwon, M., Sherman, M.A., Vitzthum, C.M., Luquette, L.J., Yandava, C.N., et al. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. Science (80-.). *359*, 555–559.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lu, T., Pan, Y., Kao, S.Y., Li, C., Kohane, I., Chan, J., and Yankner, B.A. (2004). Gene regulation and DNA damage in the ageing human brain. Nature *429*, 883–891.

Ludérus, M.E., den Blaauwen, J.L., de Smit, O.J., Compton, D.A., and van Driel, R. (1994). Binding of matrix attachment regions to lamin polymers involves single-stranded regions and the minor groove. Mol. Cell. Biol. *14*, 6297–6305.

Lydersen, B.K., and Pettijohn, D.E. (1980). Human-specific nuclear protein that associates with the polar region of the mitotic apparatus: Distribution in a human/hamster hybrid cell.

Cell *22*, 489–499.

Madabhushi, R., Pan, L., and Tsai, L.H. (2014). DNA damage and its links to neurodegeneration. Neuron *83*, 266–282.

Magistretti, P.J., and Allaman, I. (2015). A Cellular Perspective on Brain Energy Metabolism and Functional Imaging. Neuron *86*, 883–901.

Mahen, R., Hattori, H., Lee, M., Sharma, P., Jeyasekharan, A.D., and Venkitaraman, A.R. (2013). A-Type Lamins Maintain the Positional Stability of DNA Damage Repair Foci in Mammalian Nuclei. PLoS One *8*, e61893.

Manju, K., Muralikrishna, B., and Parnaik, V.K. (2006). Expression of disease-causing lamin A mutants impairs the formation of DNA repair foci. J. Cell Sci. *119*, 2704–2714.

Masi, A., Apice, M.R.D., Ricordye, R., Novelli, G., Apice, M.R.D., Ricordye, R., Masi, A., Apice, M.R.D., Ricordy, R., Tanzarella, C., et al. (2008). The R527H mutation in LMNA gene causes an increased sensitivity to ionizing radiation The R527H mutation in LMNA gene causes an increased sensitivity to ionizing radiation. Cell Cycle *7*, 2030–2037.

Matsuura, S., Tauchi, H., Nakamura, A., Kondo, N., Sakamoto, S., Endo, S., Smeets, D., Solder, B., Belohradsky, B.H., Der Kaloustian, V.M., et al. (1998). Positional cloning of the gene for Nijmegen breakage syndrome. Nat. Genet. *19*, 179–181.

Mauldin, S.K., Getts, R.C., Liu, W., and Stamato, T.D. (2002). DNA-PK-dependent binding of DNA ends to plasmids containing nuclear matrix attachment region DNA sequences : evidence for assembly of a repair complex. Nucleic Acids Res. *30*, 4075–4087.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proc. 9th Python Sci. Conf. 51–56.

McKinnon, P.J. (2012). ATM and the Molecular Pathogenesis of Ataxia Telangiectasia. Annu. Rev. Pathol. Mech. Dis. *7*, 303–321.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R.,

Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al. (2015). The human transcriptome across tissues and individuals. Science (80-.). *348*, 660–665.

Merdes, A., and Cleveland, D.W. (1998). The role of NuMA in the interphase nucleus. J. Cell Sci. *111*, 71–79.

Merdes, A., Ramyar, K., Vechio, J.D., and Cleveland, D.W. (1996). A complex of NuMA and cytoplasmic dynein is essential for mitotic spindle assembly. Cell *87*, 447–458.

Merdes, A., Heald, R., Samejima, K., Earnshaw, W.C., and Cleveland, D.W. (2000). Formation of spindle poles by dynein/dynactin-dependent transport of NuMA. J. Cell Biol. *149*, 851–861.

Mimaki, T., Itoh, N., Abe, J., Tagawa, T., Sato, K., Yabuuchi, H., and Takebe, H. (1986). Neurological manifestations in xeroderma pigmentosum. Ann. Neurol. *20*, 70–75.

Moreira, M.-C., Barbot, C., Tachi, N., Kozuka, N., Uchida, E., Gibson, T., Mendonça, P., Costa, M., Barros, J., Yanagisawa, T., et al. (2001). The gene mutated in ataxia-ocular apraxia 1 encodes the new HIT/Zn-finger protein aprataxin. Nat. Genet. *29*, 189–193.

Moreira, M.-C., Klur, S., Watanabe, M., Németh, A.H., Le Ber, I., Moniz, J.-C., Tranchant, C., Aubourg, P., Tazir, M., Schöls, L., et al. (2004). Senataxin, the ortholog of a yeast RNA helicase, is mutant in ataxia-ocular apraxia 2. Nat. Genet. *36*, 225–227.

Müller, K., Wickham, H., James, D.A., and Falcon, S. (2018). RSQLite: "SQLite" Interface for R. R package version 2.1.1. https://CRAN.R-project.org/package=RSQLite.

Nabel, C.S., Manning, S.A., and Kohli, R.M. (2012). The curious chemical biology of cytosine: Deamination, methylation,and oxidation as modulators of genomic potential. ACS Chem. Biol. *7*, 20–30.

Nagele, A. (1995). Poly(ADP-ribosyl)ation as a fail-safe, transcription-independent, suicide mechanism in acutely DNA-damaged cells: a hypothesis. Radiat. Environ. Biophys. *34*, 251–254.

Nance, M.A., and Berry, S.A. (1992). Cockayne Syndrome: Review of 140 cases. Am. J. Med. Genet. *42*, 68–84.

Nikali, K., Suomalainen, A., Saharinen, J., Kuokkanen, M., Spelbrink, J.N., Lönnqvist, T., and

Peltonen, L. (2005). Infantile onset spinocerebellar ataxia is caused by recessive mutations in mitochondrial proteins Twinkle and Twinky. Hum. Mol. Genet. *14*, 2981–2990.

Noordzij, J.G., Verkaik, N.S., Burg, M. Van Der, Veelen, L.R. Van, Bruin-versteeg, S. De, Wiegant, W., Vossen, J.M.J.J., Weemaes, C.M.R., Groot, R. De, Zdzienicka, M.Z., et al. (2003). Radiosensitive SCID patients with Artemis gene mutations show a complete B-cell differentiation arrest at the pre – B-cell receptor checkpoint in bone marrow. Immunobiology *101*, 1446–1452.

O'Connor, E., Vandrovcova, J., Bugiardini, E., Chelban, V., Manole, A., Davagnanam, I., Wiethoff, S., Pittman, A., Lynch, D.S., Efthymiou, S., et al. (2018). Mutations in XRCC1 cause cerebellar ataxia and peripheral neuropathy. J. Neurol. Neurosurg. Psychiatry *89*, 1230–1232.

Ohata, H., Miyazaki, M., Otomo, R., Matsushima-Hibiya, Y., Otsubo, C., Nagase, T., Arakawa, H., Yokota, J., Nakagama, H., Taya, Y., et al. (2013). NuMA Is Required for the Selective Induction of p53 Target Genes. Mol. Cell. Biol. *33*, 2447–2457.

Orii, K.E., Lee, Y., Kondo, N., and Mckinnon, P.J. (2006). Selective utilization of nonhomologous end-joining and homologous recombination DNA repair pathways during nervous system development. Proc. Natl. Acad. Sci. U. S. A. *103*, 10017–10022.

Osterod, M., Larsen, E., Le Page, F., Hengstler, J.G., Van der Horst, G.T.J., Boiteux, S., Klungland, A., and Epe, B. (2002). A global DNA repair mechanism involving the Cockayne syndrome B (CSB) gene product can prevent the in vivo accumulation of endogenous oxidative DNA base damage. Oncogene *21*, 8232–8239.

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2017). Biostrings: Efficient manipulation of biological strings. R package version 2.46.0.

Pannunzio, N.R., Watanabe, G., and Lieber, M.R. (2018). Nonhomologous DNA end-joining for repair of DNA double-strand breaks. J. Biol. Chem. *293*, 10512–10523.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Petersen-Mahrt, S.K., and Neuberger, M.S. (2003). In vitro deamination of cytosine to uracil

in single-stranded DNA by apolipoprotein B editing complex catalytic subunit 1 (APOBEC1). J. Biol. Chem. *278*, 19583–19586.

Picardi, E., Manzari, C., Mastropasqua, F., Aiello, I., Erchia, A.M.D., and Pesole, G. (2015). Profiling RNA editing in human tissues : towards the inosinome Atlas. Sci. Rep. *5*.

Picardi, E., D'Erchia, A.M., Giudice, C. Lo, and Pesole, G. (2017). REDIportal: A comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. *45*, D750–D757.

Poetsch, A.R., Boulton, S.J., and Luscombe, N.M. (2018). Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. Genome Biol. *19*.

Pontel, L.B., Rosado, I. V., Burgos-Barragan, G., Garaycoechea, J.I., Yu, R., Arends, M.J., Chandrasekaran, G., Broecker, V., Wei, W., Liu, L., et al. (2015). Endogenous Formaldehyde Is a Hematopoietic Stem Cell Genotoxin and Metabolic Carcinogen. Mol. Cell *60*, 177–188.

Pouliot, J.J., Yao, K.C., Robertson, C. a, and Nash, H. a (1999). Yeast gene for a Tyr-DNA phosphodiesterase that repairs topoisomerase I complexes. Science (80-.). *286*, 552–555.

Pourquier, P., Ueng, L.M., Kohlhagen, G., Mazumder, A., Gupta, M., Kohn, K.W., and Pommier, Y. (1997). Effects of uracil incorporation, DNA mismatches, and abasic sites on cleavage and religation activities of mammalian topoisomerase I. J. Biol. Chem. *272*, 7792–7796.

Pourquier, P., Ueng, L.M., Fertala, J., Wang, D., Park, H.J., Essigmann, J.M., Bjornsti, M.A., and Pommier, Y. (1999). Induction of reversible complexes between eukaryotic DNA topoisomerase I and DNA-containing oxidative base damages: 7,8-dihydro-8-oxoguanine and 5- hydroxycytosine. J. Biol. Chem. *274*, 8516–8523.

Prudencio, M., Belzil, V. V, Batra, R., Ross, C.A., Gendron, T.F., Pregent, L.J., Murray, M.E., Overstreet, K.K., Piazza-Johnston, A.E., Desaro, P., et al. (2015). Distinct brain transcriptome profiles in C9orf72-associated and sporadic ALS. Nat. Neurosci. *18*, 1175–1182.

Qin, W. Sen, Wu, J., Chen, Y., Cui, F.C., Zhang, F.M., Lyu, G.T., and Zhang, H.M. (2017). The short isoform of nuclear mitotic apparatus protein 1 functions as a putative tumor

suppressor. Chin. Med. J. (Engl). *130*, 1824–1830.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

R Core team (2019). R Core Team. R A Lang. Environ. Stat. Comput. R Found. Stat. Comput. , Vienna, Austria. Http//Www.R-Project.Org/.

R Studio Team (2015). RStudio: Integrated development environment for R. RStudio, Inc. Boston, MA URL http://www.rstudio.com/.

Rantamäki, M., Krahe, R., Paetau, A., Cormand, B., Mononen, I., and Udd, B. (2001). Adult-onset autosomal recessive ataxia with thalamic lesions in a Finnish family. Neurology *57*, 1043–1049.

Rass, U., Ahel, I., and West, S.C. (2007). Actions of aprataxin in multiple DNA repair pathways. J. Biol. Chem. *282*, 9469–9474.

Ray, S., Menezes, M.R., Senejani, A., and Sweasy, J.B. (2013). Cellular roles of DNA polymerase beta. Yale J. Biol. Med. *86*, 463–469.

Redwood, A.B., Perkins, S.M., Vanderwaal, R.P., Biehl, K.J., Gonzalez-suarez, I., Morgado-palacin, L., Sage, J., Roti-roti, J.L., Stewart, C.L., Zhang, J., et al. (2011). A dual role for A-type lamins in DNA double-strand break repair A dual role for A-type lamins in DNA double-strand break repair. Cell Cycle *10*, 2549–2560.

Reijns, M.A.M., Rabe, B., Rigby, R.E., Mill, P., Astell, K.R., Lettice, L.A., Boyle, S., Leitch, A., Keighren, M., Kilanowski, F., et al. (2012). Enzymatic removal of ribonucleotides from DNA is essential for mammalian genome integrity and development. Cell *149*, 1008–1022.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Roberge, M., and Gasser, S.M. (1992). DNA loops : structural and functional properties of scaffold-attached regions. Mol. Microbiol. *6*, 419–423.

Robinson, M.D., and Smyth, G.K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics *9*, 321–332.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rosenberg, B.R., Hamilton, C.E., Mwangi, M.M., Dewell, S., and Papavasiliou, F.N. (2011). Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. Nat. Struct. Mol. Biol. *18*, 230–236.

Rulten, S.L., and Caldecott, K.W. (2013). DNA strand break repair and neurodegeneration. DNA Repair (Amst). *12*, 558–567.

Rydberg, B., and Game, J. (2002). Excision of misincorporated ribonucleotides in DNA by RNase H (type 2) and FEN-1 in cell-free extracts. Proc. Natl. Acad. Sci. U. S. A. *99*, 16654–16659.

Savitsky, K., Bar-Shira, A., Gilad, S., Rotman, G., Ziv, Y., Vanagaite, L., Tagle, D.A., Smith, S., Uziel, T., Sfez, S., et al. (1995). A single ataxia telangiectasia gene with a product similar to PI-3 kinase. Science (80-.). *268*, 1749–1753.

Schrank, B.R., Aparicio, T., Li, Y., Chang, W., Chait, B.T., Gundersen, G.G., Gottesman, M.E., and Gautier, J. (2018). Nuclear ARP2/3 drives DNA break clustering for homology-directed repair. Nature *559*, 61–66.

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M.A., Valcárcel, J., and Eyras, E. (2018). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. *26*, 1426–1426.

Sharma, N.K., Lebedeva, M., Thomas, T., Kovalenko, O. a., Stumpf, J.D., Shadel, G.S., and Santos, J.H. (2014). Intrinsic mitochondrial DNA repair defects in Ataxia Telangiectasia. DNA Repair (Amst). *13*, 22–31.

Shen, J., Gilmore, E.C., Marshall, C.A., Haddadin, M., Reynolds, J.J., Eyaid, W., Bodell, A., Barry, B., Gleason, D., Allen, K., et al. (2010). Mutations in PNKP cause microcephaly, seizures and defects in DNA repair. Nat. Genet. *42*, 245–249.

Sims, D., Ilott, N.E., Sansom, S.N., Sudbery, I.M., Johnson, J.S., Fawcett, K.A., Berlanga-Taylor, A.J., Luna-Valero, S., Ponting, C.P., and Heger, A. (2014). CGAT: Computational genomics analysis toolkit. Bioinformatics *30*, 1290–1291.

Simsek, D., Furda, A., Gao, Y., Artus, J., Brunet, E., Hadjantonakis, A., Van Houten, B., Shuman, S., McKinnon, P.J., and Jasin, M. (2011). Crucial role for DNA ligase III in mitochondria but not in Xrcc1-dependent repair. Nature *471*, 245–248.

Singh, M., Hunt, C.R., Pandita, R.K., Kumar, R., Yang, C.-R., Horikoshi, N., Bachoo, R., Serag, S., Story, M.D., Shay, J.W., et al. (2013). Lamin A/C Depletion Enhances DNA Damage-Induced Stalled Replication Fork Arrest. Mol. Cell. Biol. *33*, 1210–1222.

Siushansian, R., Tao, L., Dixon, S.J., and Wilson, J.X. (1997). Cerebral Astrocytes Transport Ascorbic Acid and Dehydroascorbic Acid Through Distinct Mechanisms Regulated by Cyclic AMP. J. Neurochem. *68*, 2378–2385.

Skourti-Stathaki, K., and Proudfoot, N.J. (2014). A double-edged sword : R loops as threats to genome integrity and powerful regulators of gene expression. Genes Dev. *28*, 1384–1396.

Smyth, G.K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Stat. Appl. Genet. Mol. Biol. *3*, 1–26.

Soneson, C., Love, M.I., and Robinson, M.D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research *4*.

States, J.C., McDuffie, E.R., Myrand, S.P., McDowell, M., and Cleaver, J.E. (1998). Distribution of mutations in the human xeroderma pigmentosum group A gene and their relationships to the functional regions of the DNA damage recognition protein. Hum. Mutat. *12*, 103–113.

Stefanini, M., Lagomarsini, P., Arlett, C.F., Marinoni, S., Borrone, C., Crovato, F., Trevisan, G., Cordone, G., and Nuzzo, F. (1986). Xeroderma pigmentosum (complementation group D) mutation is present in patients affected by trichothiodystrophy with photosensitivity. Hum. Genet. *74*, 107–112.

Stewart, G.S., Maser, R.S., Stankovic, T., Bressan, D.A., Kaplan, M.I., Jaspers, N.G.J., Raams, A., Byrd, P.J., Petrini, J.H.J., and Taylor, A.M.R. (1999). The DNA double-strand break repair gene hMRE11 is mutated in individuals with an ataxia-telangiectasia-like disorder. Cell *99*, 577–587.

Sugawara, M., Wada, C., Okawa, S., Kobayashi, M., Sageshima, M., Imota, T., and Toyoshima, I. (2007). Purkinje cell loss in the cerebellar flocculus in patients with ataxia with ocular motor apraxia type 1/early-onset ataxia with ocular motor apraxia and hypoalbuminemia. Eur. Neurol. *59*, 18–23.

Suzuki, M. (1989). SPXX, a frequent sequence motif in gene regulatory proteins. J. Mol. Biol. *207*, 61–84.

Sykora, P., Croteau, D.L., Bohr, V.A., and Wilson, D.M. (2011). Aprataxin localizes to mitochondria and preserves mitochondrial function. Proc. Natl. Acad. Sci. U. S. A. *108*, 7437–7442.

Sykora, P., Wilson, D.M., and Bohr, V.A. (2012). Repair of persistent strand breaks in the mitochondrial genome. Mech. Ageing Dev. *133*, 169–175.

Tajiri, T., Maki, H., and Sekiguchi, M. (1995). Functional cooperation of MutT, MutM and MutY proteins in preventing mutations caused by spontaneous oxidation of guanine nucleotide in Escherichia coli. Mutat. Res. Repair *336*, 257–267.

Takashima, H., Boerkoel, C., John, J., Saifi, G.M., Salih, M., Armstrong, D., Mao, Y., Quiocho, F., Roa, B., Nakagawa, M., et al. (2002). Mutation of TDP1, encoding a topoisomerase I-dependent DNA damage repair enzyme, in spinocerebellar ataxia with axonal neuropathy. Nat. Genet. *32*, 267–272.

Tan, M.H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A.N., Liu, K.I., Zhang, R., Ramaswami, G., Ariyoshi, K., et al. (2017). Dynamic landscape and regulation of RNA editing in mammals. Nature *550*, 249–254.

Thomas, D., Scot, A.D., Barbey, R., Padula, M., and Boiteux, S. (1997). Inactivation of OGG1 increases the incidence of G · C→T · A transversions in Saccharomyces cerevisiae: Evidence for endogenous oxidative damage to DNA in eukaryotic cells. Mol. Gen. Genet. *254*, 171–178.

Thorslund, T., von Kobbe, C., Harrigan, J.A., Indig, F.E., Christiansen, M., Stevnsner, T., and Bohr, V.A. (2005). Cooperation of the Cockayne Syndrome Group B Protein and Poly(ADP-Ribose) Polymerase 1 in the Response to Oxidative Stress. Mol. Cell. Biol. *25*, 7625–7636.

Tolstonog, G. V, Mothes, E., Shoeman, R.L., and Traub, P. (2001). Isolation of SDS-Stable Complexes of the Intermediate Filament Protein Vimentin with Repetitive , Mobile , Nuclear Matrix Attachment Region , and Mitochondrial DNA Sequence Elements from Cultured Mouse and Human Fibroblasts. DNA Cell Biol. *20*, 531–554.

Tsao, Y.-P., Russo, A., Nyamuswa, G., Silber, R., and Liu, L.F. (1993). Interaction between Replication Forks and Topoisomerase I-DNA Cleavable Complexes: Studies in a Cell-free SV40 DNA Replication System. Cancer Res. *53*, 5908–5914.

Tullai, J.W., Schaffer, M.E., Mullenbrock, S., Sholder, G., Kasif, S., and Cooper, G.M. (2007). Immediate-early and delayed primary response genes are distinct in function and genomic architecture. J. Biol. Chem. *282*, 23981–23995.

Tumbale, P., Williams, J.S., Schellenberg, M.J., Kunkel, T.A., and Williams, R.S. (2014). Aprataxin resolves adenylated RNA-DNA junctions to maintain genome integrity. Nature *506*, 111–115.

Uziel, T., Lerenthal, Y., Moyal, L., Andegeko, Y., Mittelman, L., and Shiloh, Y. (2003). Requirement of the MRN complex for ATM activation by DNA damage. EMBO J. *22*, 5612–5621.

Valentin-Vega, Y. a, Maclean, K.H., Tait-mulder, J., Milasta, S., Dorsey, F.C., Cleveland, J.L., Green, D.R., Kastan, M.B., Dc, W., and Steeves, M. (2012). Mitochondrial dysfunction in ataxia-telangiectasia Mitochondrial dysfunction in ataxia-telangiectasia. Blood *119*, 1490–1500.

Varon, R., Vissinga, C., Platzer, M., Cerosaletti, K.M., Chrzanowska, K.H., Saar, K., Beckmann, G., Seemanová, E., Cooper, P.R., Nowak, N.J., et al. (1998). Nibrin, a novel DNA double-strand break repair protein, is mutated in Nijmegen breakage syndrome. Cell *93*, 467–476.

Verheijen, B.M., Vermulst, M., and van Leeuwen, F.W. (2018). Somatic mutations in neurons during aging and neurodegeneration. Acta Neuropathol. *135*, 811–826.

Vidal, A.E., Boiteux, S., Hickson, I.D., and Radicella, J.P. (2001). XRCC1 coordinates the initial and late stages of DNA abasic site repair through protein-protein interactions. EMBO J. *20*, 6530–6539.

Vidi, P.-A., Chandramouly, G., Gray, M., Wang, L., Liu, E., Kim, J.J., Roukos, V., Bissell, M.J., Moghe, P. V., and Lelievre, S.A. (2012). Interconnected contribution of tissue morphogenesis and the nuclear protein NuMA to the DNA damage response. J. Cell Sci. *125*, 350–361.

Vidi, P.A., Liu, J., Salles, D., Jayaraman, S., Dorfman, G., Gray, M., Abad, P., Moghe, P. V., Irudayaraj, J.M., Wiesmüller, L., et al. (2014). NuMA promotes homologous recombination repair by regulating the accumulation of the ISWI ATPase SNF2h at DNA breaks. Nucleic Acids Res. *42*, 6365–6379.

Voogd, J., and Glickstein, M. (1998). The anatomy of the cerebellum. Trends Neurosci. *21*, 370–375.

Walkley, C.R., and Li, J.B. (2017). Rewriting the transcriptome : adenosine-to- inosine RNA editing by ADARs. Genome Biol. *18*.

Waltes, R., Kalb, R., Gatei, M., Kijas, A.W., Stumm, M., Sobeck, A., Wieland, B., Varon, R., Lerenthal, Y., Lavin, M.F., et al. (2009). Human RAD50 Deficiency in a Nijmegen Breakage Syndrome-like Disorder. Am. J. Hum. Genet. *84*, 605–616.

Whitehouse, C.J., Taylor, R.M., Thistlethwaite, A., Zhang, H., Karimi-Busheri, F., Lasko, D.D., Weinfeld, M., and Caldecott, K.W. (2001). XRCC1 stimulates human polynucleotide kinase activity at damaged DNA termini and accelerates DNA single-strand break repair. Cell *104*, 107–117.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York https://ggplot2.tidyverse.org.

Wickham, H., and Henry, L. (2019). tidyr: Easily Tidy Data with “spread()” and “gather()” Functions. R package version 0.8.3. https://CRAN.R-project.org/package=tidyr.

Wickham, H., Francois, R., Henry, L., and Müller, K. (2019). Package ‘dplyr’. A Grammar of Data Manipulation. R package version 0.8.0.1. https://CRAN.R-project.org/package=dplyr.

Wright, W.D., Shah, S.S., and Heyer, W.D. (2018). Homologous recombination and the repair of DNA double-strand breaks. J. Biol. Chem. *293*, 10524–10535.

Wu, J., and Liu, L.F. (1997). Processing of topoisomerase I cleavable complexes into DNA damage by transcription. Nucleic Acids Res. *25*, 4181–4186.

Wu, J., Xu, Z., He, D., and Lu, G. (2014). Identification and characterization of novel NuMA isoforms. Biochem. Biophys. Res. Commun. *454*, 387–392.

Xia, G., McFarland, K.N., Wang, K., Sarkar, P.S., Yachnis, A.T., and Ashizawa, T. (2013). Purkinje cell loss is the major brain pathology of spinocerebellar ataxia type 10. J. Neurol. Neurosurg. Psychiatry *84*, 1409–1411.

Yan, W.X., Mirzazadeh, R., Garnerone, S., Scott, D., Schneider, M.W., Kallas, T., Custodio, J., Wernersson, E., Li, Y., Gao, L., et al. (2017). BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. Nat. Commun. *8*, 1–9.

Yang, C.H., Lambie, E.J., and Snyder, M. (1992). NuMA: An unusually long coiled-coil related protein in the mammalian nucleus. J. Cell Biol. *116*, 1303–1317.

Yang, M.Y., Bowmaker, M., Reyes, A., Vergani, L., Angeli, P., Gringeri, E., Jacobs, H.T., and Holt, I.J. (2002). Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication. Cell *111*, 495–505.

Yang, S.W., Burgin, A.B., Huizenga, B.N., Robertson, C.A., Yao, K.C., and Nash, H.A. (1996). A eukaryotic enzyme that can disjoin dead-end covalent complexes between DNA and type I topoisomerases. Proc. Natl. Acad. Sci. U. S. A. *93*, 11534–11539.

Yizhak, K., Aguet, F., Kim, J., Hess, J.M., Kübler, K., Grimsby, J., Frazer, R., Zhang, H., Haradhvala, N.J., Rosebrock, D., et al. (2019). RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science (80-.). *364*.

Yoshihara, M., Jiang, L., Akatsuka, S., Suyama, M., and Toyokuni, S. (2014). Genome-wide profiling of 8-oxoguanine reveals its association with spatial positioning in nucleus. DNA Res. *21*, 603–612.

Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. *11*, R14.

Yuce, O., and West, S.C. (2013). Senataxin, Defective in the Neurodegenerative Disorder Ataxia with Oculomotor Apraxia 2, Lies at the Interface of Transcription and the DNA Damage Response. Mol. Cell. Biol. *33*, 406–417.

Zeman, M.K., and Cimprich, K.A. (2014). Causes and consequences of replication stress. Nat.

Cell Biol. *16*, 2–9.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res. *46*, D754–D761.

Zhang, H., and Pommier, Y. (2008). Mitochondrial topoisomerase I sites in the regulatory D-loop region of mitochondrial DNA. Biochemistry *47*, 11196–11203.

Zhang, H., Barceló, J.M., Lee, B., Kohlhagen, G., Zimonjic, D.B., Popescu, N.C., and Pommier, Y. (2001). Human mitochondrial topoisomerase I. Proc. Natl. Acad. Sci. U. S. A. *98*, 10608–10613.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhou, W., and Doetsch, P.W. (1994). Transcription Bypass or Blockage at Single-Strand Breaks on the DNA Template Strand: Effect of Different 3’ and 5’ Flanking Groups on the T7 RNA Polymerase Elongation Complex. Biochemistry *33*, 14926–14934.

Zou, Y., Liu, Q., Chen, B., Zhang, X., Guo, C., Zhou, H., Li, J., Gao, G., Guo, Y., Yan, C., et al. (2007). Mutation in CUL4B, which encodes a member of cullin-RING ubiquitin ligase complex, causes X-linked mental retardation. Am. J. Hum. Genet. *80*, 561–566.