Multi-omics data integration to predict gene regulation

Amber Mary Lynn Emmett

Submitted in accordance with the requirements for the degree of Doctorate of Philosophy

The University of Leeds, School of Molecular and Cellular Biology

Submitted for Examination in September 2023

Authorship statement

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 2 of this thesis includes work from the jointly authored publication:

Vijayabaskar, M.S., Goode, D.K., Obier, N., Lichtinger, M., Emmett, A.M.L., Abidin, F.N.Z., Shar, N., Hannah, R., Assi, S.A., Lie-A-Ling, M., Gottgens, B., Lacaud, G., Kouskoff, V., Bonifer, C., Westhead, D.R., 2019. Identification of gene specific cis-regulatory elements during differentiation of mouse embryonic stem cells: An integrative approach using high-throughput datasets. PLoS Comput. Biol. 15, e1007337. https://doi.org/10.1371/journal.pcbi.1007337

The candidate was responsible for data analysis and methodological validation, and contributed to writing and editing the final manuscript. The contributions of other authors are listed in the above reference.

Chapters 3,4 and 5 contain work from the jointly authored publication:

Emmett, A.M.L**.,** Saadi, A., Care M.A., Doody, G.M., Tooze, R.M., Westhead D.R., 2023 Integration of chromatin accessibility and gene expression data with cisREAD reveals a switch from PU.1/SPIB-driven to AP-1-driven gene regulation during B cell activation. https://www.biorxiv.org/content/10.1101/2023.01.09.522862v1.full

The candidate was responsible for data analysis, study design, methodology, validation and writing the manuscript.

Acknowledgements

Firstly, I would like to thank my supervisor Professor David Westhead for all his guidance, support, and encouragement. I have learnt so much from you in this PhD and could not have asked for a better supervisor. I would also like to acknowledge my co-supervisor, Dr Joan Boyes, and collaborators, Professor Reuben Tooze and Dr Matthew Care, for their help and input. For funding I acknowledge the BBSRC White Rose Doctoral Training Program.

I would also like to thank the Westhead group, and wider LIDA community. Particularly thank you to Naz, Vlad, Euan, Michal and Zarnaz for your friendship and support. It has been a pleasure to share laughter, frustrations, and Thai food with you all.

Finally, I would like to thank my wonderful family, especially my Fiancé Joe. You have been a source of unwavering support through the ups and downs of this PhD. Most importantly, I am grateful that you make me cups of tea. None of this would be possible if not for you. I am also grateful to my mum, grandmother, and mother-in-law (to be), for their love, encouragement and help. I would also like to thank my son Elliott for brightening my PhD journey. You have motivated me to achieve this PhD, and I am a stronger person because of you. I love you so much and I dedicate this thesis to you.

Abstract

Each cell in the human body has the exact same genetic sequence but has diverged and specialised to express different proteins and perform different functions. This is because, as a cell differentiates, different combinations of genes are switched on and off. This process is called gene regulation; it is what drives one cell type to become another and allows cells to respond to environmental signals. The study of gene regulation is essential to understanding how healthy gene expression programmes are maintained, and become dysregulated in disease.

Gene regulation is controlled, on the level of transcription, by complex and combinatorial interactions between transcription factor proteins and non-coding DNA sequences, called cis-regulatory elements. These cis-regulatory elements are located in non-coding DNA and can regulate genes across long-genomic distances. Furthermore, they are highly cell-specific, and often only active in certain cellular contexts. For these reasons, characterising gene regulation, and its role in cellular differentiation, is an ongoing challenge.

'Omics sequencing approaches can be used to measure properties of nucleic acids on a genome-wide scale. This thesis explores the integration of multi-omics datasets, using statistics and machine learning methods, in order to predict gene regulation. Over four results chapters this thesis: 1) compares existing methods to predict gene regulation from multi-omics data; 2) develops a new computational method; 3) applies this method to identify how gene regulation drives the key immune process of B cell differentiation; and 4) benchmarks this method against other approaches.

Altogether, the work presented in this thesis contributes novel methodology and knowledge to the fields of bioinformatics, gene regulation and immunology. Specifically, it presents the new cisREAD method which integrates epigenomics and transcriptomics datasets, to prioritise transcription-factor bound, gene-specific cis-regulatory elements important to differentiation. Furthermore, it applies this method to B cell differentiation, and identifies novel mechanisms of transcriptional control. Importantly, it shows that a shift from regulation by PU.1 and SPIB transcription factors, to regulation by AP-1 factor BATF is a key determinant of B cell activation in humans. The new computational method has been accompanied by open-source

software, and the findings of this thesis have been disseminated through publications. Ultimately, this thesis represents a step towards understanding the complex regulatory mechanisms which underpin cellular differentiation and disease.

Table of Contents

List of Tables

List of Equations

List of Figures

List of abbreviations

DNA – deoxyribonucleic acid

RNA – ribonucleic acid

mRNA – messenger ribonucleic acid

RNA pol – RNA polymerase II

CTD – carboxy terminal domain

TSS – transcription start site

PIC – preinitiation complex

TF – transcription factor

CRE – cis-regulatory element

coCRE – community of cis-regulatory elements

bp – base pair

kb – kilobase

Mb – megabase

eRNA – enhancer RNA

CpG – cytosine-phosphate-guanosine

HAT – histone acetyl transferase

HDAC – histone de-acetylase

HMT – histone methyl transferase

TAD – topologically associating domain

EPI – enhancer promoter interaction

LLPS – liquid-liquid phase separation

NGS – next generation sequencing

PCR – polymerase chain reaction

WGS – whole genome sequencing

RNA-seq – ribonucleic acid sequencing

ChIP-seq – chromatin immune precipitation sequencing

ATAC-seq – the assay for transpose accessible chromatin sequencing

DNase-seq – DNase I hypersensitive sites sequencing

DHS – DNase I hypersensitive site

WGBS – whole genome bisulfite sequencing

3C – chromatin conformation capture

Hi-C – high-throughput chromatin conformation capture

PC Hi-C promoter capture high throughput chromatin conformation capture

ChIA-PET – chromatin interaction analysis with paired-end-tag sequencing

PGM – probabilistic graphical model

DBN – dynamic Bayesian network

HMM – hidden Markov model

SVM – support vector machine

ANN – artificial neural network

DNN – deep neural network

LASSO – least absolute shrinkage and selection operator

TP – true positive

FP – false positive

TN – true negative

FN – false negative

PPV – positive predictive value

AUPR – area under the precision recall curve

AUROC – area under the receiver operating curve

MPRA – massively parallel reporter assay

eQTL – expression quantitative trait loci

GWAS – genome wide association study

ESC – embryonic stem cell

HB – haemangioblasts

HE – haematopoietic endothelial cell

HP – haematopoietic progenitor cell

MAC – macrophage

BCR – B cell receptor

TCR – T cell receptor

MHCII – Major histocompatibility complex II

GC – germinal centre

FDC – follicular dendritic cell

SHM – somatic hypermutation

CSR – class switch recombination

AID – activation induced deaminase

ABC – activated B cell

PB – plasmablast

PC – plasma cell

PWM – position weight matrix

DAR – differentially accessible region

DEG – differentially expressed gene

LRT – likelihood ratio test

BH – Benjamini-Hochberg

FDR – false discovery rate

Chapter 1. Introduction

Each cell in the human body contains the exact same instruction manual, encoded in copies of the same Deoxyribonucleic Acid (DNA) molecules stored in the cell's nucleus. In development, a single cell gives rise to the trillions of cells in the adult human body, specialised into 200 distinct cell types. The identity of each of these cells is programmed not through the uniform DNA sequence, but though the distinct combination of genes which are switched on or off. Strict control of gene expression – termed gene regulation – is what drives one cell to differentiate into another cell type, and maintain a healthy cellular identity. Therefore, in order to understand what makes a cell a cell, we need to understand how its genes are regulated.

This thesis focuses on the regulation of transcription, in which the DNA genetic blueprint is transcribed into messenger ribonucleic acid (mRNA), in the first step towards protein production. Specifically, it asks how we can better understand transcriptional regulation through the integration of 'multi-omics' data using methods from statistics and machine learning.

From the completion of the human genome project in 2003, DNA sequencing technologies have progressed at a rapid pace, and have allowed researchers to uncover new insights into the genome on an unprecedented scale. This sequencing revolution has led to the development of various 'omics' fields: genomics (the DNA content of the nucleus), transcriptomics (the RNA content), epigenomics (modifications which occur to DNA), and 3D genomics (the organisation of DNA in the nucleus). Whilst each datatype alone can be hugely informative; integration of multiple omics can provide holistic insight into the biochemical processes in a cell. This is particularly true for transcriptional regulation.

DNA sequencing methods produce huge masses of data, confounded by noise and technical biases, and yielding extremely high-dimensional datasets. Bioinformaticians handling omics data rely on a repertoire of computational tools. These use methodology from statistics to find meaningful patterns in these datasets; helping us understand how cells develop and diseases arise. The development of computational methods to predict transcriptional regulation is an active area of research. There is a

need for widely applicable and interpretable methods, which can be used in real research situations to uncover how gene regulation shapes differentiation.

This thesis presents new work concerning the development and evaluation of computational methods to predict transcriptional regulation from multi-omics data. These methods are applied to 'omics data sets from blood and immune cells, to acquire new knowledge on the molecules, which instruct the differentiation of these systems. This thesis starts with an introduction, which will overview the biomolecular mechanisms controlling transcription; their measurement through high-throughput sequencing; and current computational methods to reconstruct gene regulation from multi-omics data.

1.1 Transcriptional regulation is controlled by cis-regulatory elements, transcription factors and the chromatin environment

Gene regulation in eukaryotes is controlled at multiple levels, firstly at the level of transcription. In order for a protein-coding gene to be expressed in a cell, its DNA template must first be copied into mRNA by the enzyme ribonucleic acid (RNA) polymerase II. To co-ordinate transcription, DNA and protein factors interact to fine-tune the rate of mRNA production. This section will introduce the initiation of eukaryotic transcriptional regulation, and the molecules which orchestrate this process.

1.1.1 Transcription is initiated by general transcription factors at gene promoters

In eukaryotes, transcription of the DNA template requires the loading of RNA polymerase II (RNA pol) onto the transcription start site (TSS) of a gene. RNA polymerase II binds as part of the preinitiation complex (PIC) which, at a minimum, comprises 6 general transcription factors (TFs) alongside RNA polymerase II (RNA pol). The PIC alone can drive basal levels of transcription (Sikorski and Buratowski, 2009). The PIC initiates transcription at the gene promoter, 5' (5 prime) of the TSS. The classic view of transcriptional regulation is that general transcription factors sequentially assemble and recruit RNA polymerase, forming the PIC.

First the TATA-binding protein (TBP), a subunit of the general TFIID transcription factor binds the promoter, creating a sharp bend in its DNA. TFIID then recruits TFIIA and TFIIB, which in turn recruits TFIIF and RNA pol II. TFIIE then joins the complex and recruits TFIIH (Farnung and Vos, 2022). The TFIIH complex has both helicase and kinase activities. It separates the two strands, to help unwind DNA and form the 'transcription bubble'; and it phosphorylates the RNA pol CTD (Carboxyl Terminal Domain), to switch the polymerase to initiate mRNA production (Rimel and Taatjes, 2018).

After transcription has been initiated, elongation begins. Here RNA polymerase II begins to move along the gene, from the 5' end to the 3' (3 prime) end, transcribing DNA to RNA. Following phosphorylation, the polymerase moves onto the DNA and TFIIB dissociates in 'promoter escape', allowing for RNA pol to extend the mRNA molecule (Schier and Taatjes, 2020). In many metazoan genes RNA polymerase II pauses transcription 20-60 nucleotides into elongation. RNA polymerase pausing in early elongation is controlled by negative elongation factors, and is overcome by positive elongation factors (Schier and Taatjes, 2020). Pausing prevents re-initiation and ensures that nascent transcripts are 5' capped (where a methyl 'cap' is added 5' end of the transcript to prevent degradation). The polymerase pause release therefore represents a second layer of transcriptional control (Adelman and Lis, 2012).

1.1.2 Transcription is fine-tuned by cell-specific transcription factors at distal cis-regulatory elements

Whilst the general transcription factors are sufficient to initiate basal levels of transcription, the exact level of transcription in a cell is fine-tuned by cell-specific transcription factors. These transcription factors recognise short, ~6-12 base-pair (bp), recognition sites within larger DNA sequences, hundreds of base-pairs in length, called cis-regulatory elements (CREs) (Spitz and Furlong, 2012). TF binding sites occur within both proximal regulatory elements – gene promoters – and distal regulatory elements, like enhancers and silencers. TFs bound at distal CREs can act across long genomic distances, where they are recruited across hundreds to millions of base pairs by DNA looping, and can transduce their regulatory signals to RNA polymerase II through the mediator complex (Figure 1.1) (Schoenfelder and Fraser, 2019). The mediator comprises 30 subunits in humans and acts as a functional bridge between TFs and

transcriptional machinery. The mediator additionally promotes transcription by recruiting and stabilising the PIC, as well as promoting pol II phosphorylation by TFIIH (Soutourina, 2018).

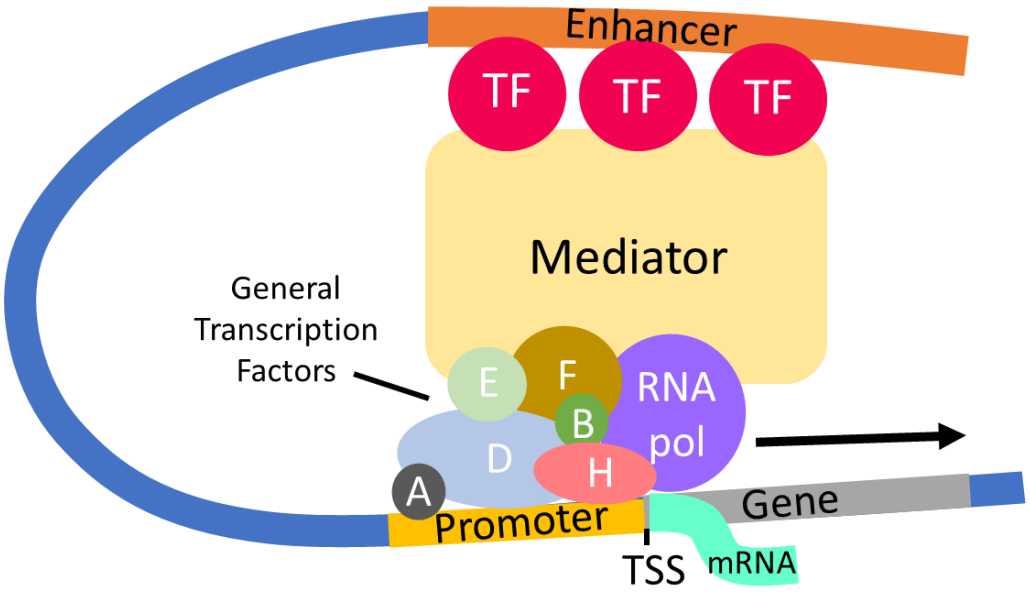


Figure 1.1 Initiation and control of Transcription. Transcription is initiated by formation of the preinitiation complex comprising general transcription factors (including TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH) and RNA polymerase II at the gene promoter. Transcription factors bound at distal enhancers (which loop around to the site of transcription) transduce signals to RNA polymerase through the mediator complex to elevate transcription beyond basal levels. RNA polymerase initiates transcription then clears the promoter and moves along the gene, transcribing DNA to mRNA in the elongation phase.

Transcription factor binding is both cooperative and combinatorial, allowing for complex spatiotemporal control of expression. TF co-binding can either be direct, facilitated by protein-protein interactions at adjacent binding sites, or indirect, where TFs bound to DNA form scaffolds for other TFs (Spitz and Furlong, 2012). Most transcription factors modulate expression through recruiting cofactors. Cofactors include the mediator complex, chromatin re-modellers and histone modifiers (discussed in section 1.1.3). The effects of coactivators can either be activating or repressive. Many TFs can recruit both activating and repressive cofactors, their function can depend on the sequence environment and cofactor availability (Lambert et al., 2018)

Distal cis-regulatory elements can be classed as enhancers or silencers based on their propensity to activate or repress transcription. These cis-regulatory elements are the effectors of signalling cascades, which enable the cell to respond to changes in the environment, or to carry out their innate differentiation programme. Distal cis-

regulatory elements called insulators act to ensure the right enhancers and silencers regulate the right genes, through blocking unwanted cis-regulatory interactions (Shlyueva et al., 2014).

In the four decades since the discovery of the first enhancer in viral genomes (Banerji et al., 1981), enhancers have been extensively studied in a multitude of species and cell types. Enhancers possess several well-characterised properties such as evolutionary conservation and the presence of TF binding sites. When active, enhancers gain chromatin accessibility and undergo histone modifications (section 1.1.3). Enhancers can also be transcribed by RNA polymerase II to produce short-lived non-coding enhancer RNAs (eRNAs) (Ray-Jones and Spivakov, 2021).

The discovery of eRNAs by two landmark studies revealed widespread bidirectional transcription of enhancers (de Santa et al., 2010; Kim et al., 2010). Research in single cells later discovered eRNA transcription is unidirectional, and only appears bidirectional when considering populations of multiple cells (Kouno et al., 2019). The functionality of eRNAs is an area of ongoing research. It has been suggested that eRNA transcripts function to stabilise enhancer-promoter looping interactions (Li et al., 2013), and promote the RNA polymerase pause release (Gorbovytska et al., 2022). In addition, the act of eRNA transcription may mutually stimulate transcription of the promoter (Panigrahi et al., 2018), and some eRNAs can regulate gene expression in trans (Tsai et al., 2018).

Clusters of enhancers, with elevated binding of transcription factors, BRD4 and mediator proteins have been coined super-enhancers. Since 2013, super-enhancers have been reported to regulate cell identity genes in humans and mice. It has been speculated that super-enhancers evolved to enable gene regulation in response to a wide array of cues (Hnisz et al., 2015; Lovén et al., 2013; Whyte et al., 2013).

In contrast to enhancers, silencers have been comparatively understudied, despite being first reported in 1985 (Brand et al., 1985). In efforts to close the gap, a wave of recent studies has elucidated silencing mechanisms and catalogued silencers in a range of species and cells (Cai et al., 2021; Doni Jayavelu et al., 2020; Huang et al., 2019; Ngan et al., 2020; Pang and Snyder, 2020). These have found sequential similarities with enhancers, such as evolutionary conservation and the presence of TF binding sites

(Doni Jayavelu et al., 2020; Huang et al., 2019). However, they have also found crucial differences: there is no characteristic silencer chromatin signature, and silencers may operate through diverse direct and indirect mechanisms. Known silencing mechanisms include anti-looping (in which silencers block enhancer-promoter interactions), competitive TF binding (in which repressive TFs block occupancy of activating TFs at bifunctional CREs), transcriptional interference (in which transcription of intragenic eRNAs interferes with transcription of the host gene), and polycomb-mediated repression (section 1.1.3) (Segert et al., 2021). The concept of super-silencers, long stretches of PRC2-bound silencers analogous to super-enhancers, has recently been suggested (Pang et al., 2023). Importantly, these studies have found widespread bifunctionality of silencers, and have shown that many can function as enhancers in different cellular contacts or chromatin environments (Gisselbrecht et al., 2020; Huang and Ovcharenko, 2022).

Whilst cis-regulatory elements are typically divided into four distinct classes (promoters, enhancers, silencers, and insulators) there is a growing body of evidence which suggests regulatory function is shared between classes, and often fluid. This is exemplified by the bifunctionality of silencers (Gisselbrecht et al., 2020; Huang and Ovcharenko, 2022), alongside the shared sequence, chromatin and architectural features which drive transcription at both enhancers and promoters (Kim and Shiekhattar, 2015). This means that many promoters can enhance activity of other genes (Dao and Spicuglia, 2018). The above findings have led researchers to speculate that enhancers and promoters exist on a continuum (Mikhaylichenko et al., 2018), or are context dependent (Andersson, 2015). Despite similarities, the GC (Guanosine and Cytosine) content of a cis-regulatory sequence differs between enhancers and promoters. Promoters contain more CpG (Cytosine-phosphate-Guanosine) dinucleotides than enhancers, and recruit highly active TFs which bind CpG sites (Andersson and Sandelin, 2020).

1.1.3 Transcriptional regulation is shaped by chromatin structure

Regardless of classification, the chromatin structure of regulatory elements is key to their activity. DNA is condensed around a core of eight histone proteins called nucleosomes, forming a structure termed chromatin. Precisely, 145-147 base pairs (bp)

of DNA is wound around the nucleosome octamer comprised of two of each core histone (H2A, H2B, H3 and H4). The linker histone H1 is also attached outside the core. The structure of compacted chromatin is described as 'beads on a string', with linker DNA connecting one nucleosome to the next (Cutter and Hayes, 2015).

Nucleosomal compaction has a repressive effect on transcription, as RNA polymerase is unable to access condensed chromatin. Cis-regulatory activity, and transcriptional regulation, instead require an accessible chromatin structure. Alterations to the chromatin structure can be achieved through post-translational modifications, which usually occur at the exposed N-terminal (amino-terminal) tails of histones. These modifications can regulate chromatin structure through nucleosome repositioning, both directly and through recruitment of chromatin remodelling enzymes (Bannister and Kouzarides, 2011). Some histone modifications with relevance to gene regulation are listed in Table 1.1

Table 1.1 Histone modifications and their associated chromatin context

Histone modification	Chromatin context
H3K27Ac	Active chromatin
H3K27me3	Repressed chromatin
H3K4me1	Enhancer
H3K4me3	Promoter

Acetylation of lysine (K) residues is dynamically regulated by histone acetyl transferase (HAT) and histone deacetylase (HDAC) enzymes. Acetylation neutralises lysine's positive charge and thereby weakens interactions between histones and DNA (Bannister and Kouzarides, 2011). Acetylated lysine residues are recognised by bromodomain protein domains in histone re-modellers, TFIID subunits, and BRD4 (which is involved in the transcriptional pause release, and marks distal enhancers) (Filippakopoulos and Knapp, 2012). Deposition of an acetyl group of histone 3 lysine 27 (H3K27Ac) occurs at active promoters and enhancers, through the HATs p300 and CBP (Calo and Wysocka, 2013). eRNAs have been found to stimulate deposition of H3K27Ac by CBP (Bose et al., 2017).

H3K27Ac antagonises the repressive histone modification H3K27Me3 (Pasini et al., 2010). Trimethylation of H3K27 is performed sequentially by the histone methyl transferase (HMT) enzyme EZH2. EZH2 is part of the PRC2 polycomb repressive

complex, which serves to compact chromatin through maintenance and further deposition of H3K27me3. These H3K27Me3 marks tightly compact chromatin into 'heterochromatin', which silences gene expression by blocking access of transcription machinery (Hyun et al., 2017). H3K27Me3 enrichment has been observed at a subclass of distal silencers, which repress transcription through chromatin looping (Cai et al., 2021).

Lysine methylation is not always repressive, methylation of H3 lysine 4 is associated with promoter and enhancer activity, respectively. H3K4me3 is preferentially deposited at accessible promoters, and H3K4me1 at accessible enhancers (Heintzman et al., 2007). This is due to the targeting of HMTs SET1A and SET1B (which deposit H3K4me3) to CpG islands. CpG islands are stretches of DNA with a high density of CpG sites which are frequently found in promoters but not in enhancers. Instead H3K4 mono-methylation is performed by other HMTs which do not show preference for CpG sites (Calo and Wysocka, 2013). Enhancers can be switched off during cell-state transitions by removal of H3K4me1 by the histone demethylase LSD1 (Whyte et al., 2012).

Lysine 4 methylation recruits transcriptional co-factors and chromatin modellers with specific DNA binding domains, dependent on the number of methyl groups deposited. H3K4me3 recruits proteins by their PHD finger domains, resulting in recruitment of HATs, which deposit activating chromatin marks to neighbouring nucleosomes, and the TFIID complex, promoting effective formation of the transcriptional initiation complex (Hyun et al., 2017). Deposition of H3K4me1 has been observed to precede enhancer activity, and H3K27 acetylation (Calo and Wysocka, 2013). In numerous systems, including embryogenesis and haematopoiesis, H3K4me1 alone has been associated with enhancer 'priming' (Bonifer and Cockerill, 2017). It is hypothesised that H3K4me1 marks a window of opportunity for enhancer activation, through the binding of pioneer TFs (Calo and Wysocka, 2013). Pioneer TFs are uniquely capable of binding closed chromatin. The initial weak binding to heterochromatin triggers epigenetic remodelling through the recruitment of cofactors, which begin to open chromatin and stabilise occupancy of the pioneer. Pioneer binding and chromatin remodelling serves

to make way for further transcription factor binding, and ultimately facilitate transcriptional activation (Mayran and Drouin, 2018).

Multiple histone modifications are often found at regulatory elements. It has been proposed that combinations of marks form a 'histone code', which specifies the transcriptional activity of chromatin (Jenuwein and Allis, 2001). Combinations of histone marks used to delineate cis-regulatory function are given in Table 1.2. These include H3K4me3 and H3K27me3, which are associated with 'bivalent' promoters, where contradicting modifications allow for signal-responsive regulation in development (Vastenhouw and Schier, 2012). Similarly 'poised' enhancers for key development genes are marked by H3K4me1 and H3K27me3 in embryonic stem cells (ESCs) (Rada-Iglesias et al., 2011). These already exist in looping conformations with (bivalent) gene promoters, which may be mediated by PRC2 across long distances (Cruz-Molina et al., 2017).

Table 1.2 Combinations of histone modifications associated with cis-regulatory elements in eukaryotes. Enhancers are preferentially marked by H3K4Me1, whereas promoters are preferentially marked by H3K4Me3. Additional histone modifications accompanying H3K4 methylation are associated with different enhancer and promoter activities.

Histone modification combination	Cis-regulatory element
H3K4me1 + H3K27Ac	Active enhancer
H3K4me1 alone	Primed enhancer
H3K4me1 + H3K27me3	Poised enhancer
H3K4me3 + H3K27Ac	Active promoter
H3K4me3 + H3K27me3	Bivalent promoter

Alongside modification of histones, DNA undergoes epigenetic modification by the methylation of cytosine bases. 5mC (5 methyl cytosine) marks are applied *de novo* by DNMTs (DNA methyl transferases) DMNT3A and DMNT3B and are reapplied during cell division by DNMT1. DNA methylation is a repressive epigenetic modification that mostly occurs at CpG sites, although CpG sites within CpG islands are rarely methylated (Moore et al., 2013).

Both promoters and distal regulatory elements are subject to changes in methylation during cellular differentiation. Cell-specific enhancers and promoters undergo *de novo* methylation during embryogenesis to repress tissue-specific programmes of gene expression (Koh and Rao, 2013). 5mC (5-methyl cytosine) marks recruit repressive chromatin modifiers and may inhibit transcription factor binding (Baubec and

Schübeler, 2014). Ubiquitously expressed gene promoters, such as those of house-keeping genes, are spared from methylation by their location within CpG islands (Koh and Rao, 2013).

Cell-specific cis-regulatory elements are demethylated during differentiation by TET enzymes, which catalyse the oxidation of 5mc. Demethylation is mediated during enhancer and promoter activation by pioneer TFs, which recruit TET enzymes (alongside chromatin re-modellers) to repressed chromatin (Cedar et al., 2022).

1.1.4 Chromatin architecture sets the stage for transcriptional regulation

Long-range regulatory elements are widely accepted to act through chromatin looping, and thus the chromatin architecture is essential for controlling the effects of distal cis-regulatory elements on gene expression. Interacting cis-regulatory interactions are encased within topologically associating domains (TADs). These are roughly ~1Mb (megabase) in length in humans, and are delineated by insulators (Dixon et al., 2012). Evidence supports a model where chromatin loops form through the process of loop extrusion. Here DNA is pushed through the ring-like structure of the cohesin protein complex until it is anchored at convergent binding sites for the CTCF transcription factor (Davidson and Peters, 2021; Sanborn et al., 2015). These CTCF-delineated TADs offer enclosed environments where regulatory interactions may take place between the right CREs and the right gene promoters.

TADs often host multiple looping interactions, involving numerous enhancers and promoters, anchored together by hubs of transcription factors. These “transcriptional hubs” can either represent dynamic and heterogenous pairwise interactions in a cell population, or multiplex contacts on the same chromosome (Di Giammartino et al., 2020; Tsai et al., 2019). Whilst transcriptional hubs are widely accepted, the mechanisms by which they form are currently debated (Figure 1.2).

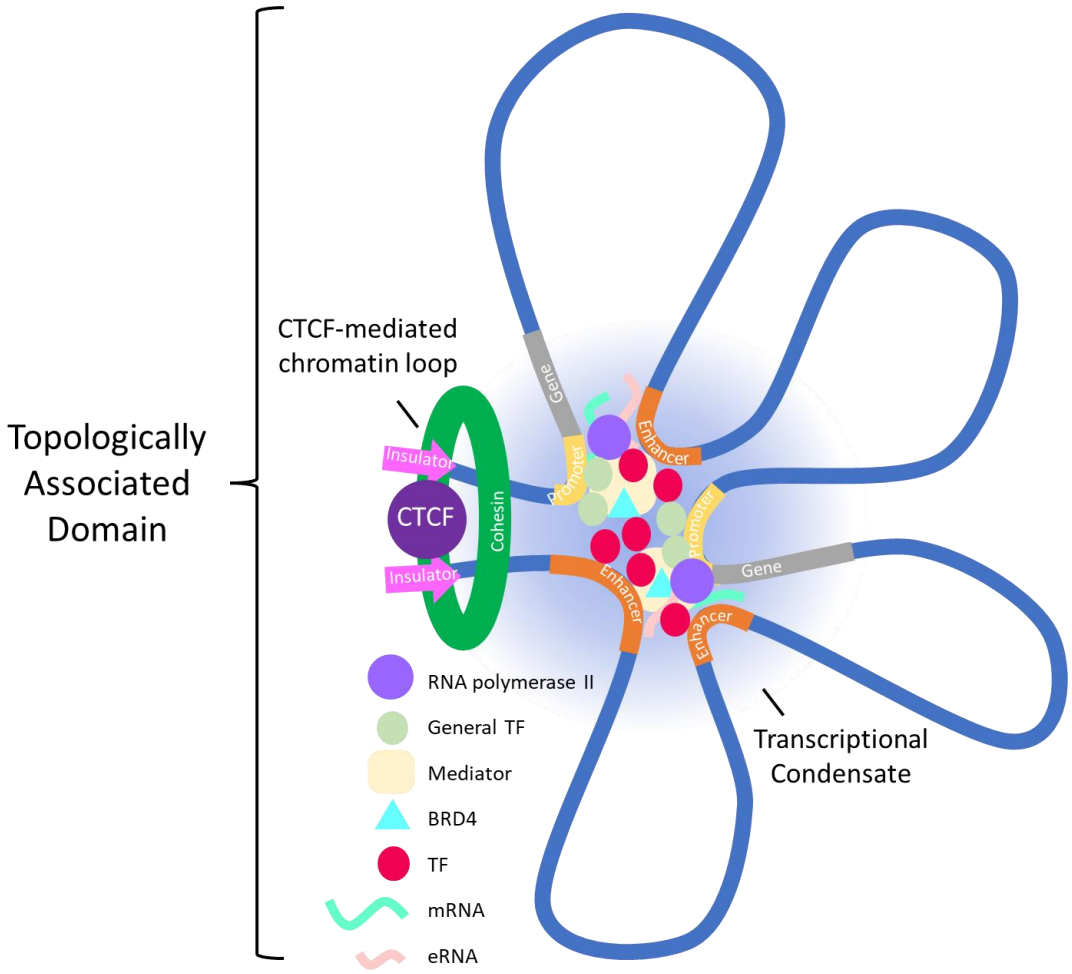


Figure 1.2 Control of transcription by distal cis-regulatory elements. Multi-way cis-regulatory interactions between enhancers and promoters take place within topologically associated domains, which are anchored at CTCF proteins at insulator regions. RNA polymerase, TFs, cofactors and nascent RNAs, associated with interacting enhancers and promoters, form 'transcriptional condensates' through the process of liquid-liquid phase separation. The contributions of CTCF-mediated loop extrusion and LLPS to chromatin architecture and transcription are currently debated.

The well-established chromatin looping model exists alongside the recently introduced model of chromatin organisation through liquid-liquid phase separation (LLPS) (Hyman et al., 2014; Mir et al., 2019; Palacio and Taatjes, 2022). Recent studies have revealed that components of transcription can form discrete 'molecular condensates' through LLPS (Boija et al., 2018; Cho et al., 2018; Sabari et al., 2018).

LLPS occurs spontaneously through weak, multivalent interactions between molecules, and results in the formation of discrete liquid compartments, analogous to 'membrane-less organelles' (Hyman et al., 2014). TFs, cofactors, RNA polymerase, epigenetic modifiers, histones and nascent RNAs are reported to form large biomolecular condensates through LLPS (Boija et al., 2018; Cho et al., 2018; Sabari et al., 2018). According to the LLPS model, these molecular condensates facilitate

interactions between multiple gene promoters and distal CREs, complementary to regulation by chromatin looping (Ray-Jones and Spivakov, 2021).

LLPS can explain recent reports which some researchers argue cannot be explained by the chromatin looping model alone. These include observations from cell imaging studies of 'contactless' enhancers, in which proximity with the promoter is not necessary for transcription (Alexander et al., 2019), or which enhancers move away from the promoter upon activation (Benabdallah et al., 2019).

Whilst the emerging LLPS model can fill in gaps unanswered by chromatin looping, its functionality in chromatin architecture and gene regulation remains controversial due to a lack of conclusive evidence (Palacio and Taatjes, 2022). Researchers have proposed that the alternative models of LLPS-mediated transcriptional condensates and chromatin looping are complementary (Zhu et al., 2021). There is evidence that CTCF-mediated chromatin looping is a pre-requisite for transcriptional condensates (Lee et al., 2022).

1. 2. Next Generation Sequencing enables multi-omics analysis of transcriptional regulation

For decades, gene regulation has been studied on a low-throughput scale using experimental techniques. Since the advent of next generation sequencing (NGS) we have entered the era of high-throughput biology, where 'omics technologies can sequence the entire DNA or RNA content of the nucleus, identify epigenetic changes, and detect chromatin topology on a genome-wide scale.

Since the introduction of commercial platforms in 2005, NGS methods have been used to sequence DNA by synthesis on a massively parallel scale. The introduction of these second-generation NGS methods marked a step-change in DNA sequencing; improving on the slow and costly 'first generation' methods, where DNA fragments were amplified and separated by electrophoresis using Sanger sequencing (Voelkerding et al., 2009).

1.2.1 Next generation sequencing methods perform massively parallel DNA sequencing

Second generation methods, such as Illumina sequencing platforms, follow the same standard workflow of: 1) library preparation, 2) clonal amplification, 3) sequencing by synthesis and 4) bioinformatics analysis.

In library preparation the DNA is fragmented into short sections of double stranded DNA, these are typically 50-300 bp long. These fragments are then ligated to sequencing 'adaptors' to generate sequencing libraries. These libraries may be 'single-end' or 'paired-end', depending on whether the adaptor is ligated to one or both ends of the fragment (Pervez et al., 2022).

Following library preparation, the DNA molecules in the library are attached to a bead or 'flow cell' (a hollow glass slide). In Illumina sequencing, the flow cell contains oligonucleotides which anchor the DNA fragments by their adaptors. The DNA fragments, and their attached adaptors, are then amplified by the polymerase chain reaction (PCR) to make ~1000 identical copies (Voelkerding et al., 2009).

After amplification, the reverse DNA strands are washed off the flow cell, and a primer attaches to the adapter on the forward strand. In Illumina sequencing, a DNA polymerase enzyme then adds fluorescently tagged nucleotides to each DNA strand. The fluorophores tagged onto each nucleotide block polymerase. This causes the enzyme to stop after adding one nucleotide per round. Each of the four nucleotide bases (adenine, thymine, guanosine, and cytosine) is labelled with a different fluorescent dye. This means the sequencing machine can read the emission of the fluorophore to call the base at each position. After calling the base, the fluorophore is washed away so that another labelled nucleotide can be added. Repeated rounds of synthesis enable the sequencer to read the whole DNA fragment (Slatko et al., 2018). In paired end sequencing, this process is then repeated for the reverse strand.

Alongside second-generation sequencing, 'third-generation' technologies have also been emerging. In contrast to 'short read' sequencers, such as Illumina platforms, technologies by PacBio and Oxford Nanopore Technologies employ 'long read' sequencing using far longer DNA fragments (Slatko et al., 2018). Long read sequencing has greater capability to detect large, complex genetic rearrangements and altered

gene splicing, however long read data has not yet been used in computational models of gene regulation. For this reason, the section will describe bioinformatics analysis of short read sequencing data.

1.2.2 Next generation sequencing data undergo bioinformatics analysis

Sequencing machines store the base calls for each fragment in large text files which encode the base calls at each position alongside their quality scores (FASTQ files). Bioinformaticians work with these FASTQ files to check the quality of the sequencing reads (QC), remove the sequencing adaptors and low-quality reads (trimming), and piece them back together by finding their place on a reference genome or transcriptome (mapping). After read mapping, the sequence alignments are stored in Sequence Alignment Map (SAM) files, and due to their size are often compressed to Binary Alignment Map (BAM) files. SAM/BAM files store the aligned reads, alongside details of the alignment. SAM/BAM files can then be filtered to remove unwanted noise, such as reads which map poorly, or map to more than one place (Pereira et al., 2020).

These 'processing' steps are performed using specialised computer programmes and are necessary before any biological analysis of the sequencing data can begin. The exact bioinformatic analysis from this point onwards depends on the aim of the NGS experiment. NGS can be employed with various protocols in order to sequence different aspects of nucleic acids (Figure 1.3). These include:

1) Whole Genome Sequencing (WGS), used to sequence the genome;

2) RiboNucleic Acid sequencing (RNA-seq), used to sequence the transcriptome;

3) DNase I hypersensitive sites sequencing (DNase-seq), the Assay for Transpose Accessible Chromatin with sequencing (ATAC-seq) and Chromatin Immuno-Precipitation Sequencing (ChIP-seq), used to sequence epigenomic changes; and

4) High-throughput Chromosome conformation capture (Hi-C), used to sequence the 3D conformation of chromatin.

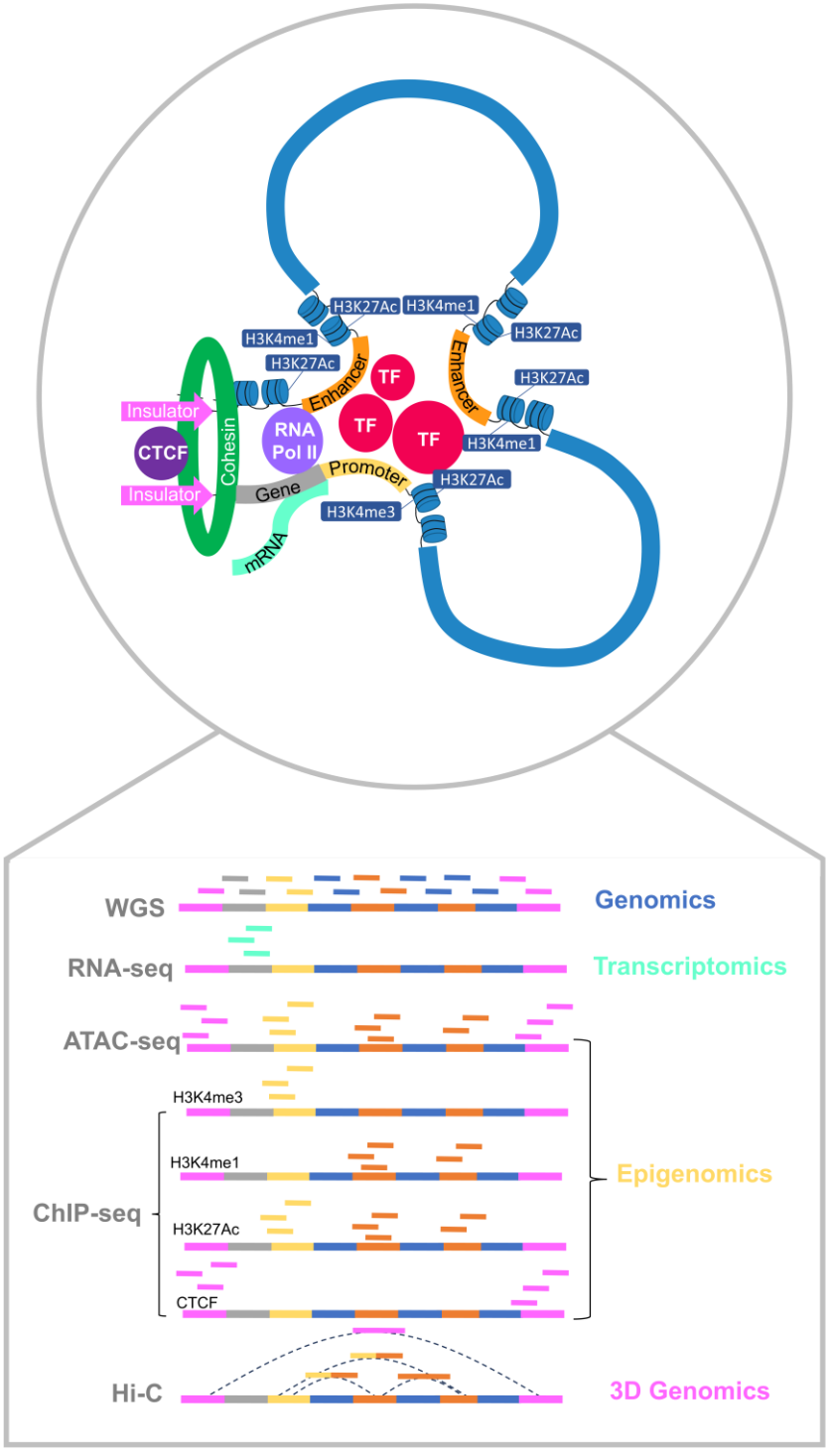


Figure 1.3 Some NGS-based omics technologies relevant to gene regulation. In WGS all DNA in the nucleus is sequenced, reads map to the whole genome. In RNA-seq, the RNA content of the cell is converted to cDNA and sequenced. In mRNA-seq reads map to exons of transcribed genes. In ATAC-seq DNA that is cleared of nucleosomes is sequenced, reads map to accessible regions of the genome. In ChIP-seq protein-DNA interactions are crosslinked, and DNA sequences occupied by a protein of interest are sequenced. When targeting histone modifications, reads map to regions with the chromatin modification. When targeting DNA-binding proteins (*e.g.*, CTCF), reads map to the regions at which they are bound. In Hi-C DNA-DNA interactions are crosslinked, and interacting fragments are sequenced. Read-pairs map to interacting regions.

Each of these technologies, their bioinformatics analysis and their relevance to gene regulation will be introduced in the following sections. Bioinformatics processing steps for each 'omics datatype are shown in Figure 1.4.

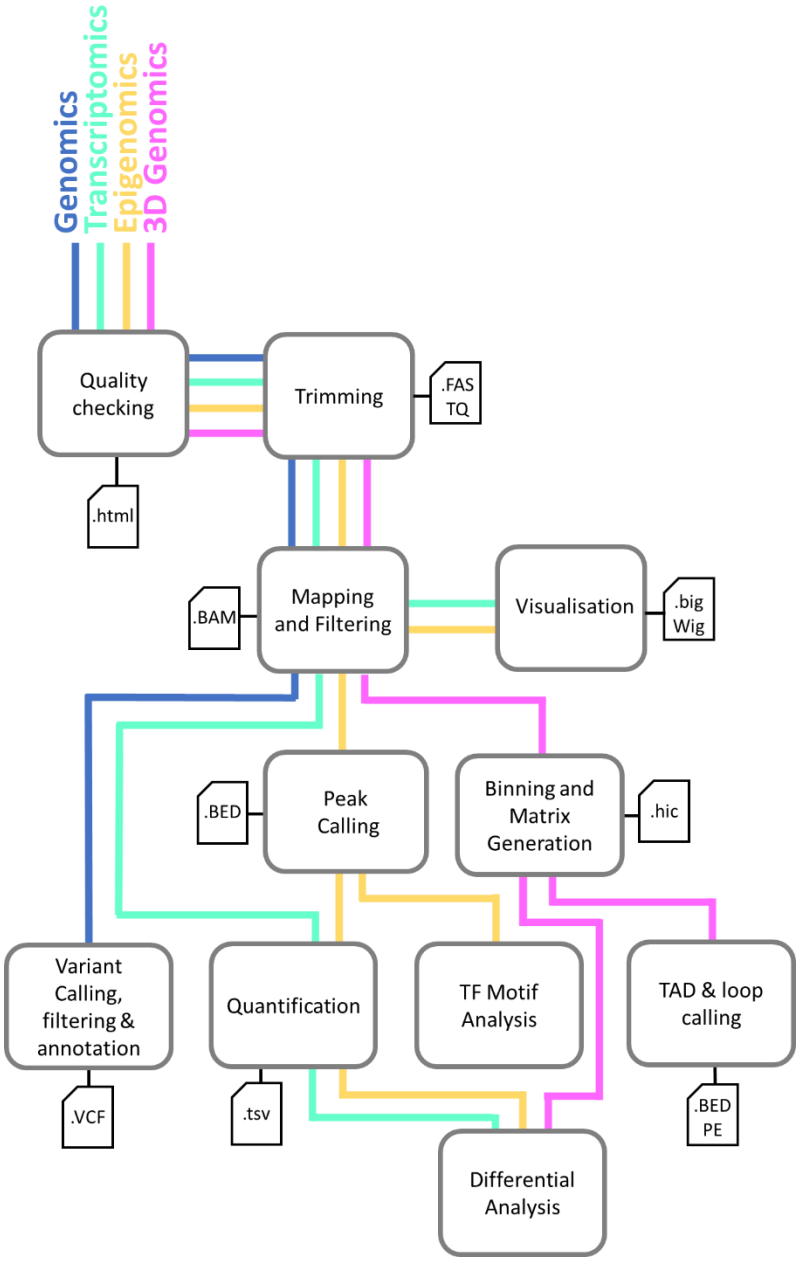


Figure 1.4 Bioinformatics processing and analysis of NGS-based omics data. NGS data from genomics (WGS), transcriptomics (RNA-seq), epigenomics (ATAC-seq, ChIP-seq), and 3D genomics (Hi-C) protocols is supplied as FASTQ files. These undergo common steps of QC, trimming, and mapping and filtering to produce BAM files. After mapping, genomics data can undergo variant calling, producing VCF files which can then be filtered and annotated. Transcriptomics data, in the mapping step, can be aligned to the genome or transcriptome, and the alignment can then be quantified, producing a transcript expression matrix. After mapping, epigenomics data can undergo peak calling to identify accessible (ATAC-seq) or protein-bound (ChIP-seq) chromatin. These peaks are stored in BED files The signal in these reads can then be quantified, Transcription factor binding sites in peaks can be identified by TF motif analysis. Both epigenomics and transcriptomic alignments can be converted to signal tracks in bigWig format, for visualisation. After mapping, 3D genomics data (Hi-C) is binned to reduce noise and a chromatin interaction matrix is generated (.hic is a common format). TADs and loops can then be identified and stored as BEDPE files. Differential analysis of transcriptomic, epigenomic or 3D genomic data from different groups can be performed to identify differentially expressed genes, differentially accessible/bound regions, or differential interactions.

This thesis focuses on the use of the above technologies, with respect to 'bulk' cell populations. However recent years have seen the development of single cell approaches. The introduction of single cell protocols, and methods of bioinformatic analysis and multi-omics integration, is an exciting development in the field of gene regulation. Details on the use of single cell multi-omics for gene regulation are reviewed in Badia-i-Mompel et al., 2023, Hu et al., 2020 and Vandereyken et al., 2023.

1.2.3 Genomics: Whole Genome Sequencing allows for detection of regulatory variants

In whole genome sequencing (WGS), DNA from an individual is sequenced in order to identify genetic variants. WGS is increasingly used in clinical research and healthcare settings to identify genetic changes (such as single nucleotide variants, genetic rearrangements, and copy number variants) associated with disease. Bioinformatics analysis of WGS data involves variant calling, in which specialised software identifies how a sequence varies in respect to a reference. In cancer patients, DNA sequences from a healthy cell can also be provided to identify variants which have occurred somatically in the tumour.

The relevance of WGS to transcriptional regulation comes from the presence of 'regulatory variants', in which mutations in non-coding DNA disrupt gene regulation. Regulatory variants can affect expression through the disruption/creation of transcription factor binding sites or the alteration of chromatin domains. Such regulatory variants can have drastic consequences for human health, driving both cancer and underlying heritable disease (Rojano et al., 2019).

For example, mutations in the promoter of the Telomerase Reverse Transcriptase (TERT) gene are widespread in over 50 cancers. These mutations introduce a binding site for ETS transcription factors and lead to overexpression of TERT, which helps the cancer cell achieve immortality through aberrant telomerase activity (Rachakonda et al., 2021). In another example, inherited deletions of CTCF TAD boundaries in the *EPHA4* locus rewire its enhancers to regulate neighbouring genes (*WNT6*, *PAX3*, *IHH*) and cause limb malformations through altered expression of these developmental genes (Lupiáñez et al., 2015).

Genome Wide Association Studies (GWAS) are able to associate population-level variants with disease incidence. Since 90% of GWAS-identified variants are in non-coding regions, elucidating the functions of regulatory variants is imperative to understand disease pathogenesis and treatment (Cano-Gamez and Trynka, 2020).

1.2.4 Transcriptomics: RNA-seq measures gene expression

RNA-seq measures the RNA content in a sample of cells. The first step in the RNA-seq protocol is isolation of RNA, followed by enrichment or depletion for RNAs of interest. To preferentially obtain mRNA content, RNAs can be selected for presence of a poly A tail (to obtain mature, processed coding RNAs), or ribosomal RNA can be depleted (allowing for inclusion of non-coding RNAs). RNA is then converted to complementary DNA (cDNA) using a reverse transcriptase, which can then be sequenced using methods described above (Stark et al., 2019). Alignment of RNA-seq data to the genome is best performed using a splice-aware aligner to map reads to exons. The alignment step is then followed by quantification, in which the number of RNA-seq reads overlapping a transcript are counted. Recently, 'pseudoalignment' algorithms offer a faster, light-weight alternative to the alignment and quantification process. These work by directly comparing sequencing read to transcripts in order to estimate transcript abundance (Srivastava et al., 2020).

Researchers often employ RNA-seq to measure changes in RNA levels between different experimental groups or time-points. When comparing gene expression, RNA-seq counts must be normalised, to account for differences in library composition and sequencing depth. Normalisation is performed within 'differential expression' analysis, where genes are tested for differences in expression, accountable to biological factors (Van Den Berge et al., 2019). Differentially expressed genes can then be tested for enrichment of pathways and gene signatures in order to assign biological function to differential expression (Maleki et al., 2020). Gene co-expression across conditions can also be analysed through the construction of gene correlation networks. These aim of these analyses is to identify co-regulated genes across a biological process, which form modules enriched in discrete functional pathways (van Dam et al., 2018).

Whilst RNA-seq can be used to measure mRNA content, eRNAs are instead measured by alternative transcriptomic methods including CAGE (5' Cap Analysis of Gene Expression) and GRO-seq (Global Run-on sequencing) (Andersson et al., 2014; Core et al., 2008). These sequence 5' capped ends of RNA molecules and nascent transcription, respectively. CAGE and GRO-seq capture both mRNA and eRNA transcription.

1.2.5. Epigenomics: ATAC-seq and DNase-seq measure chromosome accessibility, ChIP-seq can measure histone modification and transcription factor binding

1.2.5.1 ATAC-seq and DNase-seq

Active cis-regulatory elements are defined by their accessibility; they must be cleared of nucleosomes to make way for transcription factor binding. One of the most useful tools in cis-regulome reconstruction is therefore high-throughput sequencing of chromatin accessibility. Techniques including DNase-seq and ATAC-seq can be used to assay genome wide changes in chromatin accessibility and nucleosome positioning.

DNA digestion with nuclease enzymes has long been used to determine the chromatin structure of DNA. DNase I is an endonuclease which preferentially cleaves regions of open, accessible chromatin. For over 40 years, molecular biologists have used DNase I to identify nucleosome-depleted 'DNase I Hypersensitive Sites (DHSs) in order to determine the chromatin structure of regulatory sequences (Keene et al., 1981; McGhee et al., 1981; Wu et al., 1979). In 2008, the assay for DNase I hypersensitivity was combined with next generation sequencing to identify open chromatin regions throughout the genome (Boyle et al., 2008). The resulting DNase-seq method emerged to be a powerful tool for high-throughput identification of accessible cis-regulatory elements.

DNase-seq requires a large number of input cells. In 2013 a new technique was developed with lower sample requirements (thousands vs millions), a faster protocol and high sensitivity: the Assay for Transpose Accessible Chromatin sequencing (ATAC-seq) (Buenrostro et al., 2015, 2013). ATAC-seq uses a hyperactive Tn5 transposase, which cleaves accessible DNA and inserts sequencing adaptors ('tagmentation'). The

intervening accessible sequences then undergo amplification and sequencing-by-synthesis.

Whilst ATAC-seq has clear advantages, there are a number of technical aspects which must be addressed in bioinformatic analysis. This include the Tn5 cleavage bias, the 9 bp duplication following repair of the Tn5 cleavage site, and the high coverage of mitochondrial reads. ATAC-seq is also capable of identifying mononucleosome, and poly-nucleosomes alongside nucleosome free regions (Yan et al., 2020). Therefore many ATAC-seq pipelines include post-alignment steps to shift reads (+4 bp on the positive strand and -5 bp on the negative strand), filter out mitochondrial DNA (mtDNA), and select for fragments >100 bp in length (identifying nucleosome-depleted regions) (Yan et al., 2020).

Accessible regions of DNA can be identified from both DNase-seq and ATAC-seq by the process of 'peak calling,' in which algorithms identify regions of the genome with an elevated number of mapped reads (Figure 1.5). Accessible transcription factor binding sites can then be identified within these peaks by performing TF motif analyses, which considers the conserved sequences to which TFs bind. TF motif analyses will be described in detail in the introduction of chapter 3. Researchers interested in finding accessibility changes between samples can also perform quantification and differential accessibility analysis (Yan et al., 2020).

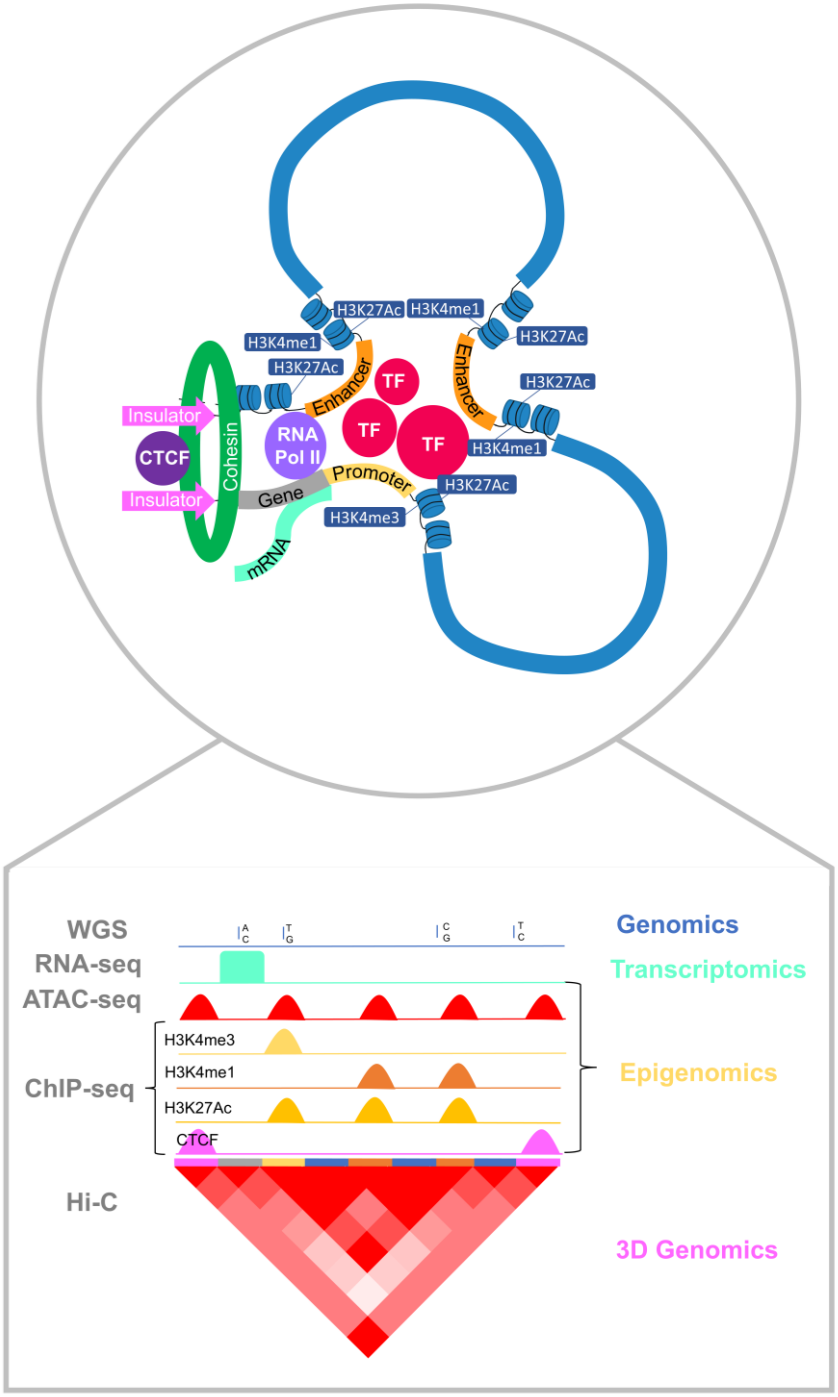


Figure 1.5 Identification of cis-regulatory features from 'omics data. WGS is able to categorise variants in cis-regulatory elements. RNA-seq is able to quantify the expression of genes. ATAC-seq and DNase-seq can identify cis-regulatory elements as 'peaks' and quantify their accessibility. ChIP-seq, targeted to histone modifications, can identify CREs with histone marks that indicate their cis-regulatory function; whilst ChIP-seq targeted to transcription factors can identify CREs bound by a TF of interest. Hi-C can identify chromatin interactions between cis-regulatory elements as bin-pairs in a contract matrix with elevated contact counts above (distance-dependent) background levels.

1.2.5.2 ChIP-seq

Chromatin Immuno-Precipitation sequencing (ChIP-seq) is a versatile tool for identifying DNA regions which are bound by a protein. ChIP-seq emerged in the late 2000s, coupling the popular molecular biology technique of chromatin immunoprecipitation with NGS. The protocol involves chemically crosslinking DNA-protein interactions, shearing the chromatin and using antibodies to select DNA bound by proteins of interest by 'immunoprecipitation' (Johnson et al., 2007). The extracted DNA is then reverse cross-linked, purified, amplified, and sequenced. ChIP-seq experiments can employ antibodies to target TFs, co-activators, histone modifications, RNA polymerase and CTCF to identify cis-regulatory elements.

Like ATAC-seq and DNase-seq, ChIP-seq data can undergo peak calling to identify regions occupied by the protein or marked by the histone modification. Similarly, peak calling can be followed by motif analyses or differential binding analyses (Nakato and Sakata, 2021).

1.2.5.3 Epigenomics analysis can also measure DNA methylation

Alongside changes to chromatin, epigenomics technologies can also probe methylation of DNA. Methylated cytosines can be distinguished from unmethylated cytosines by treatment with sodium bisulfite, which deaminates cytosine when unmethylated. This approach is employed in Whole Genome Bisulfite Sequencing (WGBS), in which DNA fragments undergo bisulfite conversion prior to amplification and NGS (Cokus et al., 2008). Alignment and methylation calling of bisulfite sequencing data requires bespoke software designed for use with this particular datatype (Rauluseviciute et al., 2019).

1.2.6 3D Genomics: Chromosome Conformation Capture technologies measure chromatin topology

Chromosome Conformation Capture (3C) techniques enable researchers to measure the chromatin topology of DNA. High resolution chromatin capture methods have quickly become the gold standard approach to identify chromatin interactions between cis-regulatory elements. 3C and its variants are based on the crosslinking of DNA with formaldehyde to stabilise looping interactions between genomic regions. In

the original 3C technique, crosslinking was followed by restriction enzyme digestion, ligation and PCR to detect topological interactions (Dekker et al., 2002).

In the following years, the original 3C protocol was improved to detect all interactions for a single specified region (4C – one vs all), or all interactions between restriction fragments within a 1Mb region (5C – many vs many). However, it was not until 2009 that the first all-vs-all approach was developed, employing NGS. The introduction of Hi-C marked the first genome-wide chromatin interaction mapping technology. Here all interacting loci are crosslinked, subject to restriction enzyme digest and ligated before fragments are analysed by paired-end sequencing. Since this is a high-throughput assay, Hi-C is considerably more expensive than its low-throughput predecessors (Lieberman-Aiden et al., 2009; Whitaker et al., 2015).

However, the genome-wide scale of Hi-C comes at the expense of resolution. The highest-resolution protocols (such as in situ Hi-C, capable of detecting interactions between 1kb-long segments) require billions of sequencing reads (Rao et al., 2014). Hi-C data also requires careful filtering at the post-alignment stage to ensure only 'valid read pairs' for chromatin contacts are retained. Following alignment, a contact matrix can be generated, giving the interaction count between two genomic regions (Figure 1.5). Due to the noise of Hi-C data, interactions are placed in bins, ranging from kilobases to megabases in scale (Lajoie et al., 2015).

Bioinformatics analysis of Hi-C data can identify chromatin structures. Various algorithms have been designed to segment Hi-C data into TADs or identify chromatin loops as bin-pairs with contact counts elevated above the background level. Whilst Hi-C data can offer direct support for candidate cis-regulatory interactions, it is important to note that many chromatin loops are not regulatory, or even biologically functional. When multiple Hi-C experiments are performed across different conditions or time-points, differential interaction analysis can be performed in order to identify changes in chromatin architecture (Pal et al., 2019).

Chapter 5 of this thesis will also introduce two other 3D genomics technologies: promoter-capture (PC) Hi-C and ChIA-PET (Chromatin Interaction Analysis by Paired End Tagging). These enable genome-wide, high resolution detection of chromatin

interactions involving gene promoters (PC Hi-C), or interactions mediated by a protein of interest (ChIA-PET) (Fullwood et al., 2009; Mifsud et al., 2015).

1.2.7 Web resources and databases host a treasure-trove of NGS data

Both individual research groups and large global consortia have characterised the genomes, transcriptomes, epigenomes and 3D genomes of diverse cell types, species, and diseases. The NGS explosion has produced a vast reservoir of 'omics data, which can be accessed by researchers through online databases (Table 1.3).

In a push towards open science, researchers are encouraged to archive their data to one of more webservers to facilitate data sharing and reproducibility. These include publicly available repositories, such as the Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA). Controlled resources such as the European Genome-phenome Archive (EGA) enable the sharing of personally-identifiable data between verified researchers (Barrett et al., 2013; Leinonen et al., 2011).

Collaborations between scientists, as part of global consortia have played a pivotal role in sequence data generation and curation. These include projects such as ENCODE, Roadmap, FANTOM5 and 4D Nucleome which focus on collating 'omics data with relevance to gene regulation, epigenomics and chromatin architecture. Each of these projects comes with its own web platform, enabling free access to an abundance of sequencing data from cell lines, primary cells and tissues (Andersson et al., 2014; Dekker et al., 2017; Dunham et al., 2012; Roadmap Epigenomics Consortium et al., 2015).

Specialised databases focused on specific diseases or systems provide a valuable resource of clinical NGS data. This is exemplified by the landmark The Cancer Genome Atlas (TCGA) project, which between 2006-2018 generated and catalogued genomic, transcriptomic and epigenomic profiles for 33 tumour types from over 11,000 cancer samples (Hutter and Zenklusen, 2018). In haematology, the Blueprint project aimed to understand gene regulation in blood cells and haematological tissues. Between 2011-2016 the Blueprint consortium generated hundreds of healthy and diseased epigenomes and transcriptomes to aid the study of haematopoietic differentiation,

haematological malignancies and autoimmune disease (Martens and Stunnenberg, 2013).

Blueprint, ENCODE, Roadmap, 4D Nucleome and others are united under the banner of the International Human Epigenome Consortium (IHEC). The IHEC started as a global effort to catalogue 1,000 human reference epigenomes (Pratt and Weng, 2018), and the IHEC data portal offers open-access to thousands of epigenomic datasets produced by its member consortia (Stunnenberg et al., 2016).

Table 1.3 Projects and databases hosting NGS data relevant to gene regulation

Project/database	Description	Data Stored
ENCODE *Dunham et al., 2012*	Project aiming to identify functional elements in the genome through mapping transcription, TF binding, histone modification and chromatin structure	Transcriptomics, Epigenomics, 3D Genomics
Roadmap Epigenomics ***Roadmap Epigenomics Consortium et al., 2015***	Project aiming to generate epigenomic maps of human primary cells	Epigenomics, Transcriptomics
FANTOM5 ***Andersson et al., 2014***	Project aiming to catalogue active enhancer and promoter elements through CAGE	Transcriptomics (CAGE)
4D Nucleome *Dekker et al., 2017*	Project aiming to understand nuclear organisation and its role in gene regulation	Epigenomics, Transcriptomics, 3D Genomics
Blueprint *(Martens and Stunnenberg, 2013*	Project aiming to generate epigenomic maps of haematological cells and diseases	Epigenomics, Transcriptomics
The Cancer Genome Atlas (TCGA) *Hutter and Zenklusen, 2018*	Project aiming to characterise genomic changes in cancer	Genomics, Transcriptomics, Epigenomics, 3D Genomics
Gene Expression Omnibus (GEO) *Barrett et al., 2013*	NCBI public repository of raw and processed NGS and microarray data	Genomics, Transcriptomics, Epigenomics, 3D Genomics
Sequence Read Archive (SRA) *Leinonen et al., 2011*	NCBI public repository of raw NGS reads	Genomics, Transcriptomics, Epigenomics, 3D Genomics

1.3 Statistics and machine learning can predict transcriptional regulation from NGS data

This wealth of NGS data has fuelled the development of computational methods, which reconstruct transcriptional regulation using statistics and machine learning. This

section will provide a high-level introduction to machine learning, and overview methods to identify cis-regulatory elements and their target genes.

Machine learning aims to build automated models capable of learning patterns from input data (Li et al., 2015). These models can learn from examples (supervised) or recognise patterns without labels (unsupervised). Supervised machine learning makes use of labelled 'training' data. Here the machine learning algorithm can 'learn' the features associated with different labels to build a model. This model can then be tested, optimised, and used to assign labels to new data (Libbrecht and Noble, 2015). Unsupervised machine learning does not require any training data; instead, the model looks for patterns in the data to identify different classes. If some, but not all data is labelled, machine learning is semi-supervised (Lim et al., 2018).

Alongside supervision, machine learning methods can be defined as either discriminative *or* generative. Discriminative approaches focus on the differences between labelled groups of data to maximise predictive power, these are usually supervised. Conversely, generative approaches build profiles of the features most associated with each group, these are usually unsupervised (Whitaker et al., 2015). All machine learning methods learn their labels from 'features' of the data. These features, also called variables, can either be categorical or numerical.

1.3.1 Machine learning can predict cell-specific cis-regulatory elements from epigenomic and sequence features

Both supervised and unsupervised methods have been applied to the task of cis-regulatory element prediction. These methods aim to annotate active cis-regulatory elements based on their epigenomic and sequence features. This subsection will overview computational methods designed to predict distal cis-regulatory elements from NGS features in mammalian systems, using either unsupervised and supervised machine learning. These methods will be used to introduce concepts in statistics, machine learning and gene regulation which are relevant to work in the thesis.

1.3.1.1 Unsupervised probabilistic graphical models can annotate cis-regulatory elements by genome segmentation

Unsupervised machine learning seeks to find patterns in data from unlabelled inputs. This class of methods includes clustering, where data-points are placed in groups based on measures of similarity, and probabilistic graphical models (PGMs). PGMS capture conditional dependencies of interacting random variables using graphs. These models are typically unsupervised and generative, building full models for each class of data. For this reason, PGMs best lend themselves to the task of genome segmentation, to assign functions to chromosome segments using epigenomic features (Kleftogiannis et al., 2016). Two landmark PGMs are ChromHMM and Segway, launched respectively by the ENCODE and Roadmap consortia to segment and annotate elements of the non-coding genome (Ernst and Kellis, 2012; Hoffman et al., 2012) (Figure 1.6).

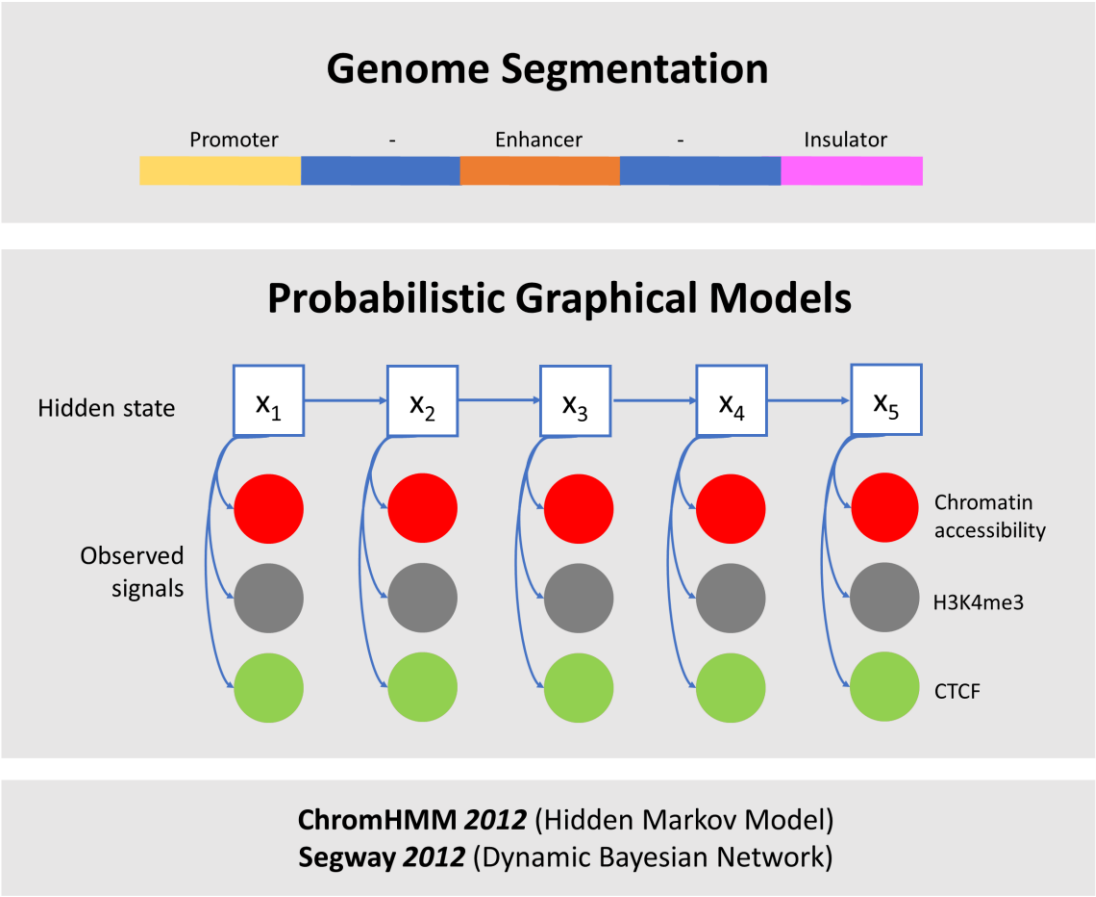


Figure 1.6 Use of probabilistic graphical models for unsupervised genome segmentation. Example shows a Dynamic Bayesian Network predicting labels of 'hidden' chromatin states from observable epigenomic signals. In this example the state of a chromatin segment is a hidden variable, generated by a Markov process where the state of each segment depends on the preceding segment. Here we cannot observe the hidden state of each chromatin segment, but we can observe the signals emitted in NGS data. For example an active enhancer region would emit signals for chromatin accessibility (measured by ATAC-seq/DNase-seq) and H3K4me3 modification (measured by ChIP-seq). Examples of genome segmentation PGMs include ChromHMM and Segway.

ENCODE's ChromHMM model uses multiple histone marks and CTCF binding to annotate genomic regions by 'chromatin state' – with classifications including active TSS, active enhancer, weak enhancer and heterochromatin (Ernst and Kellis, 2012). ChromHMM is based on a multivariate, first-order Hidden Markov Model (HMM) framework. HMMs model 'Markov processes' where the chain of events cannot be observed, yet hidden states can be indirectly measured from related visible states (Li et al., 2015). In an HMM the progression of hidden states (transmission probability) can be inferred from the emission of observed states (emission probability).

HMMs are the simplest form of Dynamic Bayesian Networks (DBNs), where the state in a sequence depends on previous states in the sequence (Li et al., 2015). Segway is a DBN developed by the Roadmap consortium to annotate the genome based on chromatin accessibility, histone modification and TF binding features. Segway performs annotation at every single base (Hoffman et al., 2012), offering a higher-resolution segmentation than ChromHMM which places bases into 200 base-pair bins (Kleftogiannis et al., 2016). Both ChromHMM and Segway have been used to perform joint annotation of cis-regulatory elements in ENCODE cell types (Hoffman et al., 2013).

1.3.1.2 Supervised machine learning models can classify enhancers

Supervised machine learning aims to 'learn' the relationship between input and output variables from labelled 'training' data to make predictions from new, unlabelled data. Most supervised methods are discriminative, and therefore excel at separating classes. Supervised methods can be applied to one of two tasks, regression, or classification, depending on whether the dependent variable is continuous or discrete (Schrider and Kern, 2018). Supervised methods have been widely used for the task of cis-regulatory element classification, and are most frequently used to identify cell-specific active enhancers from epigenomic and sequence features (Figure 1.7).

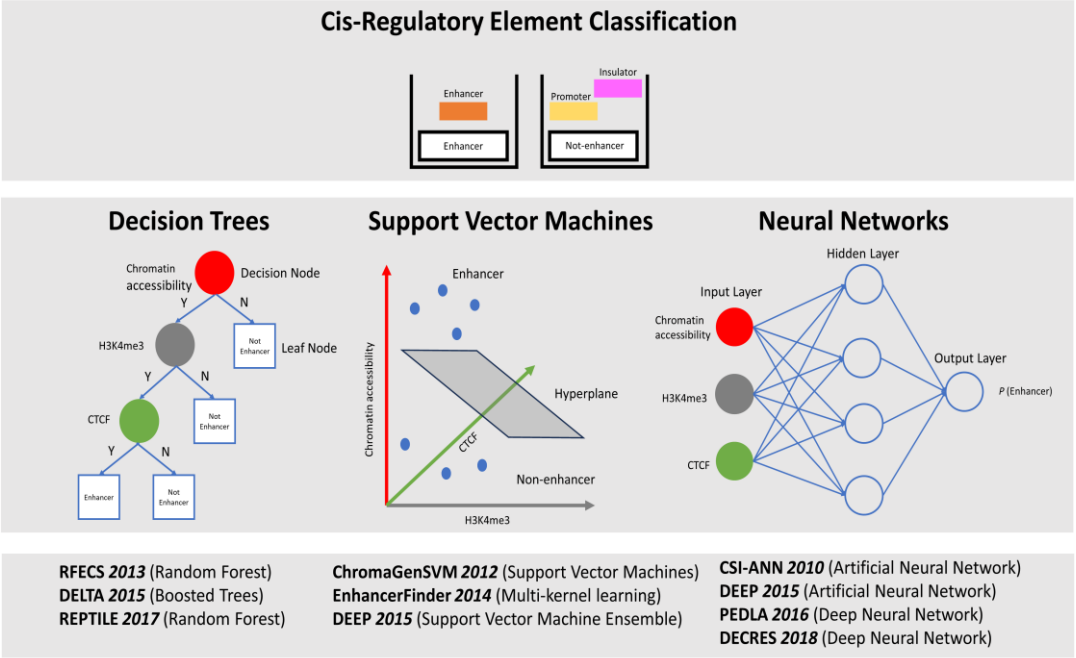


Figure 1.7 Supervised methods to identify cell-specific active enhancers. Supervised machine learning methods to classify cell-specific active enhancers from epigenomic (and sometimes sequence) features. Methods are trained on examples of active enhancers, and employ classifiers to predict unlabelled active enhancers. Popular supervised methods are explained diagrammatically, showing the use of epigenomic features to classify active enhancers and non-enhancers. Decision trees use a flowchart-like model of nodes and branches to split the data based on features. Decision trees can be aggregated by bagging (e.g. Random Forest) or boosting to strengthen performance. Support vector machines use kernels to transform data to higher dimensions, and then separate datapoints by a 'hyperplane'. Artificial neural networks are brain-inspired models which pass data through layers of nodes to make predictions. Neural networks with multiple hidden layers of nodes are deep neural networks.

Decision trees can be used in machine learning to solve classification (or regression) problems. Here each 'leaf' node represents a class label, whilst each 'branch' connection is the set of features leading to that decision (Schrider and Kern, 2018). Decision trees can be aggregated to bolster the strength of the classification, by either bagging (bootstrap aggregation) or boosting.

In bagging, the training data is randomly divided into subsets with replacement, and individual trees perform prediction on each subset. Predictions can then be averaged across this 'ensemble' of trees to produce a more robust decision. Bagging is often performed with the Random Forest method, where each tree is trained using a random combination of features (Breiman, 2001). Random forests are a popular choice for classifying both enhancers and enhancer-gene pairs. The approach forms the core classification mechanism of enhancer-predictors including RFECs and REPTILE (He et al., 2017; Rajagopal et al., 2013).

In boosting, trees are trained consecutively on the full set of data, with the error of one tree passed to the next. By increasing the weight of misclassified datapoints, the next tree works harder to reduce the error in the model. Boosted decision trees are used in the DELTA model, which predicts enhancers from the shape-features of chromatin modification signals (Lu et al., 2015).

Support Vector Machine (SVMs) are discriminative models which seek to separate classes of data by the optimal dividing decision boundary, or 'hyperplane' (see Figure 3f). Here data-points are plotted in high-dimensional space by a 'kernel' function, where a kernel transforms feature-vectors into higher-dimensional space based on measures of similarity (Schrider and Kern, 2018). Models like ChromaGenSVM and DEEP have employed SVMs to classify active enhancers (Fernández and Miranda-Saavedra, 2012; Kleftogiannis et al., 2015). SVMs can also be employed within the context of multiple kernel learning, where a combination of kernels is used. Multiple kernel learning is applied in the EnhancerFinder approach (Erwin et al., 2014).

Artificial Neural Networks take inspiration from neural connections in the brain, applying this architecture to solve machine learning problems. Neural networks (Figure 3e) are composed of multiple layers of nodes – or 'neurons' – which relay inputs from one layer to a series of hidden layers, and finally to the output layer where a classification decision is reached based on a combination of features (Schrider and Kern, 2018). Artificial neural networks have been employed to predict cell-specific active enhancers in methods such as CSI-ANN and DEEP (Firpi et al., 2010; Kleftogiannis et al., 2015). Artificial neural networks with three or more hidden layers are categorised as deep neural networks and perform deep learning. Deep learning is employed by models including PEDLA and DECRES (Y. Li et al., 2018; Liu et al., 2016).

1.3.1.3 Supervised model performance hinges on training data

Supervised machine learning is defined by its use of labelled training data to make predictions from new, unlabelled data. The success of a supervised model therefore hinges on the choice of training data. Failure to choose an appropriate training set can result in overfitting and poor predictive performance. It is important that training data closely represents the full variety and distribution of its data-type (Libbrecht and

Noble, 2015). In the case of enhancer prediction, this means labelled enhancers should be randomly selected from the entire genome and possess a range of sequence and epigenomic properties, representative of their full variation. Curating a set of 'ground truth' enhancers is itself a difficult task. Despite the efforts of databases like VISTA there is no true gold standard of validated, labelled enhancers (Visel et al., 2007).

At the time of writing, the VISTA database has catalogued 1699 human and murine enhancers whose activity has been validated by *in vivo* transgenic mouse experiments. These involved the use of reporter assays, where candidate enhancer sequences were integrated upstream of minimal promoters for a gene with a measurable protein (such as a fluorescent protein). VISTA also reports the tissue where expression was reported in developing mouse embryos (Visel et al., 2007). Enhancers from VISTA have been used to train models including EnhancerFinder. Whilst Erwin *et al.* reported improved performance of EnhancerFinder on previous models, it noted bias towards regions with high conservation in the VISTA training set (Erwin et al., 2014).

The development of Massively Parallel Reporter Assays (MPRAs) such as STARR-seq (Self-Transcribing Active Regulatory Region sequencing), has enabled researchers to screen for cis-regulatory activity on a high-throughput scale (Arnold et al., 2013). Whilst MPRAs have enabled functional validation of cis-regulatory elements on an unprecedented scale, they also face limitations, chiefly the inability to test CREs in their chromatin environment (Inoue and Ahituv, 2015). MPRAs have been used to train predictive methods such as DeepSTARR (de Almeida et al., 2022).

Many supervised methods train their models on enhancers supported by functional 'omics data, not experimental validation. These include methods such as CSI-ANN, ChromaGenSVM, RFECs, DELTA and REPTILE, which train their predictions on distal regions of open chromatin bound by the histone acetyltransferase p300 (Heintzman et al., 2007). Similarly, DEEP (when applied to ENCODE data) was trained on enhancer labels based on joint ChromHMM and Segway annotations (Hoffman et al., 2013), and DECREs was trained on CAGE positive transcribed enhancers from FANTOM5 (Andersson et al., 2014).

1.3.1.4 Model performance can be evaluated by cross-validation or testing

Alongside positive examples of active enhancers, classifiers must be provided with examples of non-enhancers. These negative examples, and the ratio of positives to negatives are highly important to model performance. Negative labels are often random genomic regions or shuffled enhancer sequences, but vary by method (Kleftogiannis et al., 2016). It is important to ensure that unwanted biases are corrected for between positive and negative classes, to prevent model overfitting. Differing numbers of positive and negative labels can lead to the problem of class imbalance and impair performance. Class imbalance is rife within enhancer prediction since there are vastly more non-enhancers than enhancers in the genome. Models trained on class imbalanced datasets are resultantly biased towards non-enhancers (Libbrecht and Noble, 2015). Ensemble classifiers, like DEEP, DELTA and RFECs, are noted to perform better than single classifiers on class-imbalanced training sets. (Kleftogiannis et al., 2015; Lu et al., 2015; Rajagopal et al., 2013).

Once a model has been trained on labelled data its performance can be tested by making predictions on independent datasets with known labels. Many models are also evaluated through cross-validation, in which a fraction of the training set (*n*) is removed and reserved for training. This is performed *n* number of times with the subsection to be held-back rotating each time, cross-validation when *n*=5 is known as 5-fold cross-validation. Whilst models should ideally be tested on independent 'testing' datasets, cross-validation offers a robust method of testing when labelled datasets are limited (Schrider and Kern, 2018).

There are many measures of model performance with no single best metric; these include accuracy, precision and recall. Performance metrics are calculated through evaluating the numbers of true positives (TPs; correctly labelled as enhancer), true negatives (TNs; correctly labelled as non-enhancer), false positives (FPs; incorrectly labelled as enhancer), and false negatives (FNs; incorrectly labelled as non-enhancer). Accuracy is the percentage of correct predictions, precision (also called positive predictive value) is the percentage of true positives of all predicted positives, and recall (also called sensitivity) is the percentage of true positives of all actual positives. A

balanced measure of performance is the F1-score, which is the harmonic mean of precision and recall (Libbrecht and Noble, 2015).

Binary classifiers often predict labels using a threshold, where the model outputs the probability that a prediction belongs to a class label (from 0 to 1) and a decision boundary is chosen. Performance, independent of threshold, can be measured on precision-recall curves, in which precision and recall are plotted for each threshold of the classifier. Threshold-independent performance can then be measured as the Area Under the Precision Recall curve (AUPR). Similarly AUROC scores (Area Under the Receiver Operating Curve) give threshold independent measures of true positive rate (*i.e.* recall) and false positive rate (the percentage of false positives out of all actual negatives) (Libbrecht and Noble, 2015).

1.3.1.5 Active cis-regulatory elements can be predicted from epigenomic, transcriptomic and sequence features

All the methods introduced in this section, regardless of methodology or supervision, learn the epigenomic features associated with cell-specific cis-regulatory elements. Epigenomic features can be derived from NGS applications and include chromatin accessibility, methylation, histone modifications, and occupancy by proteins (including TFs, cofactors, RNA polymerase and CTCF). Epigenomic features can also be accompanied with sequence features including conservation and transcription factor binding sites, as well as GC content and CpG islands. Features considered by NGS-based predictive models, described in this section, are given in Table 1.4.

Table 1.4 Methodological details for models designed to predict cell-specific active cis-regulatory elements from NGS data.

Publication	Software	Inputs	Feature types	Methods	Training Enhancers
Firpi, Ucar & Tan 2010	CSI-ANN	ChIP-seq	Histone modification	Feature Extraction: Fisher Discriminant Analysis Classification: Artificial Neural Network	p300 binding
Fernandez and Saavedra 2012	Chroma-GenSVM	ChIP-seq	Histone modification	Classification: Support Vector Machine	p300 binding
Ernst and Kellis 2012	ChromHMM	ChIP-seq DNase-seq	Histone modification TF binding Chromatin accessibility	Segmentation: Hidden Markov Model	-
Hoffman et al. 2012	Segway	ChIP-seq DNase-seq	Histone modification TF binding Chromatin accessibility	Segmentation: Dynamic Bayesian Network	-
Rajagopal et al. 2013	RFECs	ChIP-seq	Histone modification	Feature selection and classification: Random Forests	p300 binding
Erwin et al. 2014	Enhancer-Finder	DNase-seq, ChIP-seq,	Histone modification TF binding Chromatin accessibility Sequence	Classification: Support Vector Machine	VISTA validated
Kleftogiannis, Kalnis & Bajic 2015	DEEP (DEEP-ENCODE)	ChIP-seq	Histone modifications	Classification: Support Vector Machine Ensembl + Neural Network	ChromHMM and Segway annotations
Lu et al. 2015	DELTA	ChIP-seq	Histone modification	Classification: Boosted decision trees	p300 binding
Liu et al. 2016	PEDLA	ChIP-seq DNase-seq RNA-seq Bisulfite sequencing	Histone modification TF binding Chromatin accessibility DNA methylation Sequence	Classification: Neural Network + Hidden Markov Model	H3K27Ac
He et al. 2017	REPTILE	ChIP-seq Bisulfite sequencing	Histone Modification DNA methylation	Classification: Random Forest	p300 binding
Li, Shi & Wasserman 2018	DECREs	ChIP-seq DNase-seq ChIA-PET	Histone modification TF binding Chromatin accessibility Sequence	Feature Selection & Classification: Deep Neural Networks	eRNA transcription

The methods in Table 1.4 use NGS data to identify enhancers with cell-specific activity. However recent advances in generative artificial intelligence have inspired models which can identify cell-specific enhancers (and their effects on expression) from

sequence features alone. These methods, including Basenji and DeepMind's Enformer model, work by generating tissue-specific epigenomic and transcriptomic signals from DNA sequence (Avsec et al., 2021; Kelley et al., 2018). Recently, machine learning models have been developed to predict silencers from sequence (Doni Jayavelu et al., 2020; Huang and Ovcharenko, 2022; Zeng et al., 2021).

1.3.2 Machine learning models can link cis-regulatory elements to target genes through multi-omics integration

Enhancers and silencers are most likely to regulate proximal genes that are within the same topologically associated domain (Zuin et al., 2022). However cis-regulatory elements do not always regulate their nearest gene, and may control expression over hundreds of thousands of base pairs (Lettice et al., 2003). To accurately identify target genes for cis-regulatory elements, computational methods have been designed to predict cis-regulatory interactions from NGS data. The work presented in this thesis is focused on this task.

Multi-omics, predictive methods work through integrating genomic, epigenomic, transcriptomic and 3D genomics datatypes. This may involve training models on 'validated' enhancer-gene pairs or correlating cis-regulatory features with features of promoters or genes. This section overviews supervised and unsupervised methods which predict gene-specific cis-regulatory elements in mammals through multi-omics integration.

1.3.2.1 Chromatin contacts and eQTLs can link cis-regulatory elements to target genes *in silico*

Interactions between genes and distal cis-regulatory elements can be identified experimentally or detected from 'omics data. 'Omics based approaches to assign cis-regulatory elements to target genes include eQTL (expression Quantitative Trait Loci) analysis and the detection of chromatin interactions from 3D genomics data. This subsection will introduce multi-omics methods to predict cis-regulatory interactions using eQTL and chromatin interaction datasets (Figure 1.8). More detail on eQTLs and

chromatin interactions will be provided in chapter 5, in the context of model evaluation.

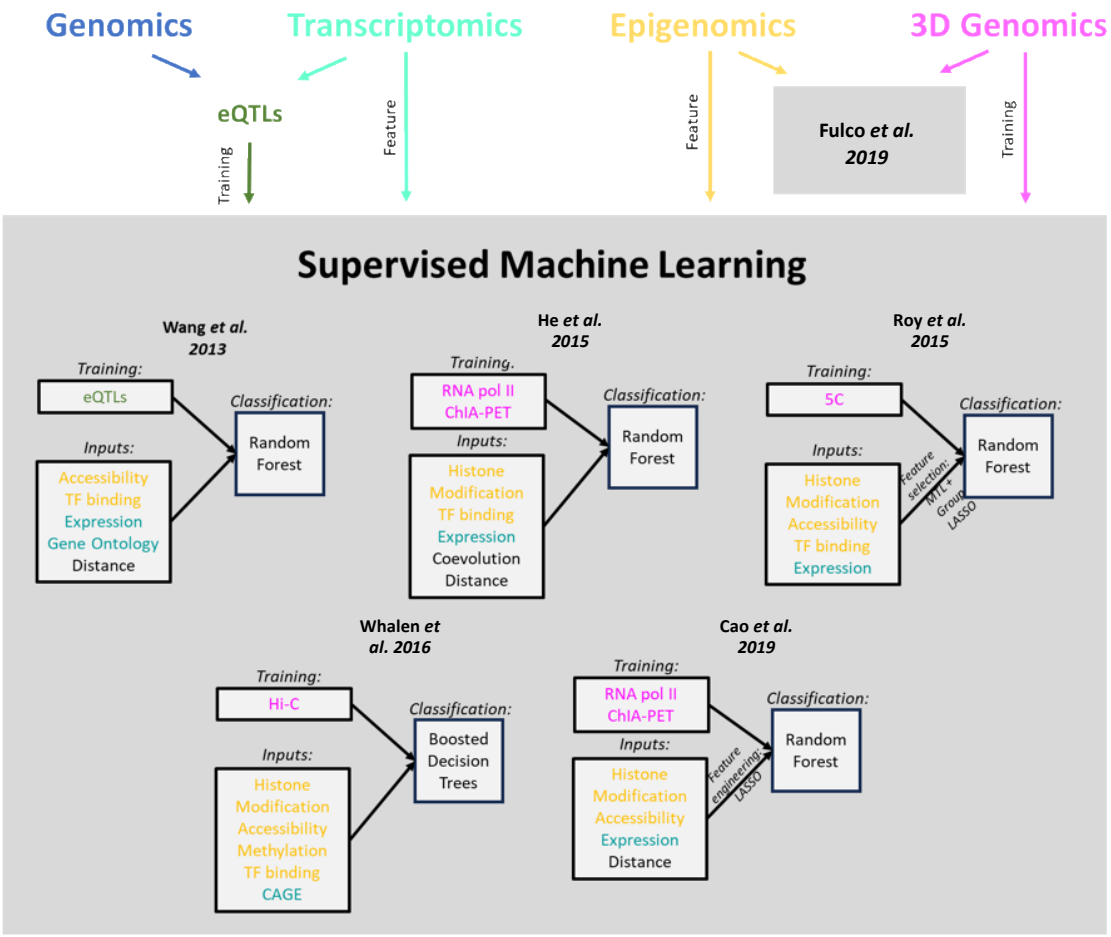


Figure 1.8 Methods to predict cell-specific enhancer-promoter interactions from multi-omics data, using eQTLs or chromatin interactions. Unsupervised EPI predictors (such as the Activity-by-Contact model by Fulco *et al.* 2019) can use chromatin interaction data as a feature, and supervised EPI predictors can use chromatin interactions (or eQTLs) in model training. Diagrams describe workflows of 5 supervised models, with boxes showing training data, input data and classification method. Colours indicate use of omics datatypes (teal – transcriptomics, yellow - epigenomics, pink – 3D genomics) within EPI prediction models.

eQTLs can be identified through integration of population-level genomic and transcriptomic datasets, where genetic variants are tested for association with gene expression (Nica and Dermitzakis, 2013). Since most eQTLs are found outside of promoter regions, this approach can be used to link distal CREs to their target genes. In 2013 Wang et al. used eQTLs to train a random forest classifier to predict target genes for regulatory variants from TF binding, chromatin accessibility, gene expression and distance features. This approach achieved good performance in cross-validation and cross-cell line prediction, and outperformed predictions made by distance alone (Wang et al., 2013).

3C-based datatypes, such as Hi-C and ChIA-PET, can identify cis-regulatory elements involved in chromatin looping interactions. Chromatin interactions, filtered for those involving distal CREs and gene promoters, can therefore be used define training sets for supervised methods. Alternatively, chromatin contacts can be used as a feature for unsupervised prediction. This strategy is used by Fulco et al.'s Activity-by-Contact model which combined epigenomic activity (from DNase/ATAC-seq and H3K27Ac ChIP-seq) with Hi-C-derived chromatin contacts in a rule-based approach. The Activity-by-Contact model was shown to accurately predict experimentally-validated enhancers-gene pairs, and outperformed both distance and supervised machine learning approaches (Fulco et al., 2019)

Despite their demonstrable utility, high-throughput 3D genomics techniques, like Hi-C, are expensive. Large numbers of cells and high sequencing depths are required to achieve resolution capable of detecting interactions between cis-regulatory elements (Rao et al., 2014). This has limited their application to a handful of well-studied cell types. To address this gap researchers have developed supervised machine learning models which aim to learn the features associated with cis-regulatory interactions. These models can then be trained on cell-specific chromatin interaction datasets, and then applied to detect additional cis-regulatory interactions in the same or different cell types.

Supervised models include IM-PET, RIPPLE, TargetFinder and JEME which all aim to identify cell-specific enhancer-promoter interactions (EPIs) using tree-based classifiers (Cao et al., 2017; He et al., 2014; Roy et al., 2015; Whalen et al., 2016). All four models were trained on chromatin interactions from 3C-based datasets (5C, Hi-C or ChiA-PET), which were filtered for interactions involving annotated gene promoters and enhancers. Negative pairs were generated through a variety of methods, such as by assigning active enhancers to random, non-interacting target genes. Epigenomic, sequence and transcriptomic features were then obtained for candidate enhancers, promoters and 'windows' (defined as DNA segments between enhancers and promoters). Prior to classification, RIPPLE and JEME performed an additional step to select (RIPPLE) or weight (JEME) the epigenomic features used in the final classifier. All four methods reported excellent predictive performance (Cao et al., 2017; He et al., 2014; Roy et al., 2015; Whalen et al., 2016).

In recent years, supervised EPI predictors have been criticised for their overfitting to training data. It was shown that random cross-validation schemes, used to test TargetFinder and JEME models, failed to account for shared-features between enhancer-promoter pairs. (Cao and Fullwood, 2019; Xi and Beer, 2018). Xi and Beer retested TargetFinder with chromosomally sorted cross validation folds (ensuring that shared promoter and window features were grouped together) and performance metrics fell drastically. Cao and Fullwood replicated this finding and reported a loss of performance when retesting JEME. They also investigated JEME's training set and found a large distance-bias between positive and negative classes.

1.3.2.2 Correlation and regression techniques can pair cis-regulatory elements to co-active target genes

Neither chromatin interactions nor eQTLs truly represent 'ground truth' cis-regulatory interactions. Whilst useful, both datatypes face several limitations and biases (discussed in chapter 5). These biases may be inherited when training supervised models, or when used as a predictive feature. Avoiding bias introduced by these datatypes, unsupervised methods can link cis-regulatory elements to target genes using correlation or regression approaches. These methods aim to find relationships between CRE activity and gene regulation over datasets comprising multiple cell-types (Figure 1.9). As such their predictions lack cell-specificity, and these methods cannot recognise cis-regulatory relationships which do not involve correlation. These include interactions involving primed enhancers, accessible prior to expression, or bifunctional silencers/enhancers, which remain accessible despite a change in TF-driven activity.

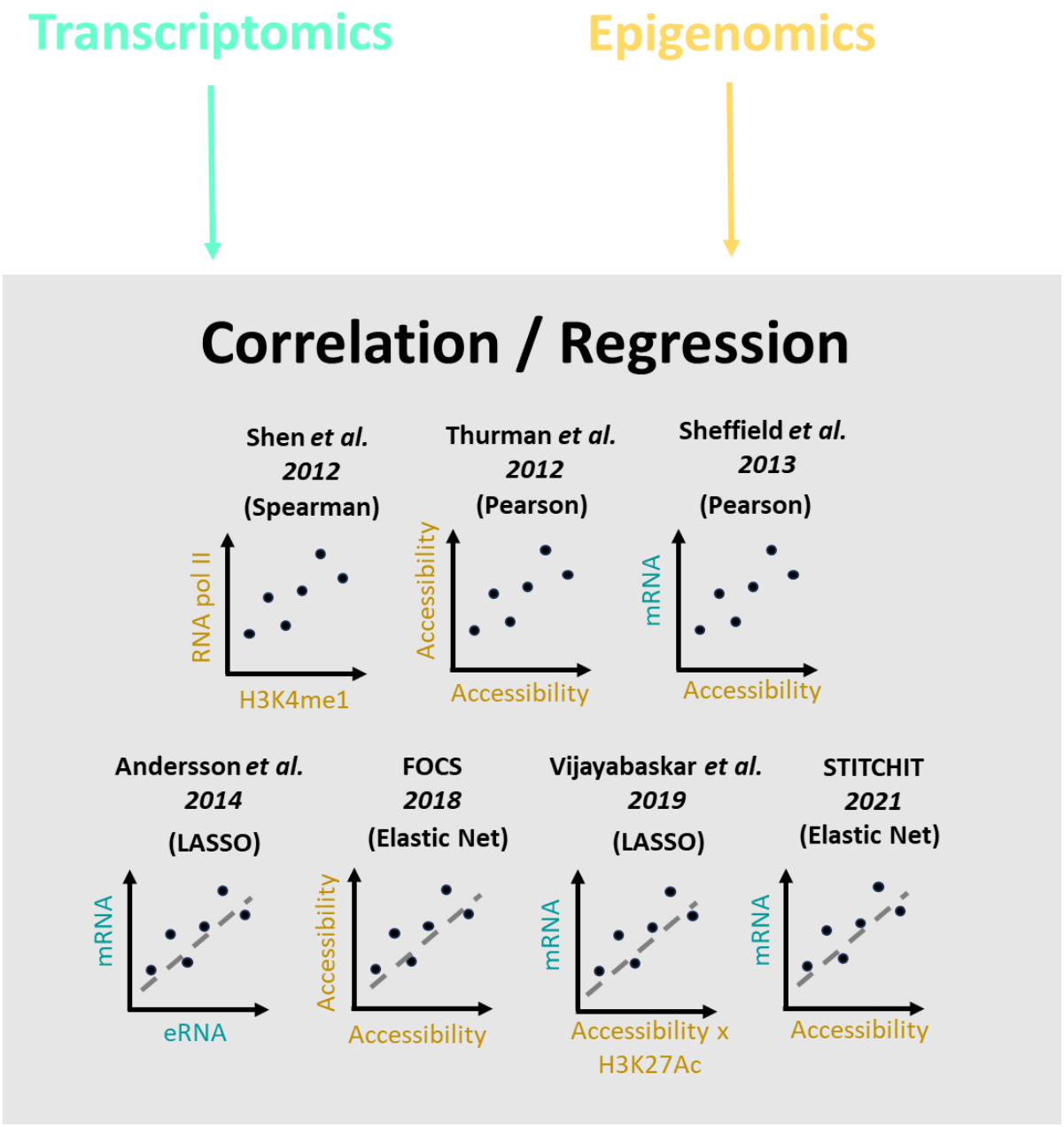


Figure 1.9 Methods to predict gene-specific cis-regulatory elements through correlation or regression of epigenomic and transcriptomic features. Epigenomic features are shown in gold and transcriptomic features are shown in teal. Shen et al., 2012, Thurman et al., 2012 and Sheffield et al., 2013 correlated epigenomic features of enhancers with epigenomic/transcriptomic features of gene promoters using Spearman or Pearson correlation. Andersson et al. 2014, Hait et al., 2018 (FOCS), Vijayabaskar et al. 2019 and Schmidt et al., 2021 (STITCHIT) used penalised regression models (LASSO or Elastic Net) to select enhancers with epigenomic or transcriptomic features that predicted gene promoter activity.

Amongst this class of methods, the simplest approaches look for correlation between epigenomic and/or transcriptomic features at distal CREs and promoters. Examples are found in Shen et al., 2012, which correlated histone marks at candidate enhancers with RNA polymerase binding at gene promoters; Thurman et al., 2012, which correlated chromatin accessibility at candidate enhancers with chromatin accessibility at

promoters; and Sheffield et al., 2013 which correlated chromatin accessibility at candidate enhancers with gene expression at promoters.

Regression analysis can also be used to test the relationship between a dependent variable (*i.e.,* a promoter feature) and one or more independent variables (*i.e.,* distal CRE features). In linear regression**,** the relationship between continuous explanatory and response variables is modelled linearly. Regression analysis estimates coefficients for each independent variable, giving the strength and direction of the relationship. When the dependent variable is binary, logistic regression can be applied by performing regression on a logit (logarithm of the odds) instead of linear scale (Stoltzfus, 2011). Linear regression can be used to predict new values of the dependent variable, and logistic regression can be used to solve binary classification problems.

Regression models can be prone to overfitting, where the model learns noise and outliers in the training data. To avoid overfitting, regularisation techniques can be applied. Regularisation methods apply a penalty to predictor variables, which 'shrinks' the regression coefficients to reduce the complexity and variance of the model. Regularisation techniques are particularly useful when handling high-dimensional datasets, such as those produced by NGS.

Regularisation techniques like ridge regression can reduce overfitting by shrinking regression coefficients using the L2 penalty (where coefficients are penalised by the square of the coefficient) (Hoerl and Kennard, 1970). Alternatively, coefficients can be shrunk through the L1 penalty, equal to the absolute magnitude of the coefficient. The L1 penalty is applied in LASSO (Least Associated Squares Shrinkage Operator) regression (Tibshirani, 1996a). Unlike ridge regression, LASSO regression can shrink less-predictive coefficients down to zero and eliminate them from the model. This is called variable selection. LASSO is often employed for the purpose of variable selection, as well as prediction.

LASSO regression has been applied in gene regulation to perform variable selection on cis-regulatory elements. This approach involves the construction of gene-specific models, where promoter activity, or gene expression, is predicted from the activity of cis-regulatory elements. The FANTOM5 consortium used LASSO regression to select enhancers whose CAGE signal (from eRNA transcription) predicted the CAGE signal at

nearby promoters (from mRNA transcription) (Andersson et al., 2014). LASSO regression was employed by the Westhead group to select candidate CREs whose chromatin accessibility predicted expression of a nearby gene from ENCODE data (Shar et al., 2016).

Whilst LASSO can select gene-specific cis-regulatory elements by constructing sparse models, it is unstable in cases of multicollinearity (in which multiple predictive features are highly correlated). Multicollinearity can cause the model to drop variables at random, whilst selecting others that are highly correlated. In order to mitigate the instability of LASSO regression, the elastic net was proposed. Elastic net regression allows for the L1 LASSO penalty to be mixed with the L2 Ridge penalty (Zou and Hastie, 2005). Elastic net regression was been employed in the FOCS and STITCHIT models to predict gene-specific cis-regulatory elements (Hait et al., 2018; Schmidt et al., 2021). FOCS and STITCHIT were inclusive of diverse cis-regulatory element types, including negatively correlated silencers.

Whilst elastic net regression alleviates the instability of LASSO, it does so at the cost of sparsity. The Westhead group devised an alternative approach to alleviate instability: preceding LASSO regression with a community detection step. This community detection step, introduced in Vijayabaskar et al., 2019, reduced multicollinearity by grouping correlated predictors together, based on co-activity and TF co-binding. The community detection step had the conceptual advantage of grouping together cis-regulatory elements which may coregulate expression together, reflecting regulation by transcriptional hubs.

1.4 There is an outstanding need for applicable, implementable, and interpretable methods to predict gene regulation

Computational prediction of gene regulation is a prolific area of research. A wide range of features and methodologies have been suggested across dozens of predictive models. Whilst these methods have strengthened our understanding of the molecules which orchestrate transcription, there is still no consensus on how best to predict cis-regulatory elements, and to link them to their target genes. Recent advances in machine learning and 3D genomics have inspired methods which boast impressive

performance on well-studied cell types. However, many of the methods overviewed in this chapter are hindered by unachievable input requirements, lack of open-source software and inattention to biological interpretation.

This thesis argues there is an unmet need for methods designed, not just with performance in mind, but for use in real research scenarios. Methods should be designed in line with the needs of researchers: to characterise the transcription factors, cis-regulatory elements and genes which drive systems of differentiation or disease – and to do this easily at low cost.

To address this gap in methods, the overarching aim of the thesis is to develop a method which is 1) applicable, 2) implementable and 3) interpretable. The rationale for these criteria is as follows:

1. **Methods should be applicable**

   Many of the methods overviewed in this section require an abundance of input NGS datatypes for each sample. Many researchers, studying gene regulation on a budget, will be unable to meet the input requirements to use these methods. Amongst the least applicable tools are TargetFinder (which used over 50 NGS datasets per sample) and RIPPLE (which used 23 NGS datasets per sample). It can be argued these methods have little applicability outside of well-studied cell lines. Even the simple Activity-by-Contact model requires cell-specific Hi-C data for optimum performance, which is highly expensive. This thesis argues that methods should be designed with applicability in mind, and therefore should aim to work with minimal data inputs.

2. **Methods should be implementable**

   Scientific research is experiencing a huge push towards reproducibility, and bioinformaticians should be designing tools with reproducibility in mind. Some methods to predict gene regulation are difficult to reproduce due to a lack of usable software. This is particularly true for many of the most ‘applicable’ correlation and regression-based models. Less-implementable approaches include methods like Thurman et al. and Andersson et al. (which rely on users following descriptions in the paper) or models like FOCS (which require users to adapt code made available on a webserver). For methods to be easily

implementable, they should be packaged into software, which is easy to install and use, and available under an open-source licence.

3. **Methods should be interpretable**

   Finally, this thesis argues that methods to predict gene regulation should be biologically interpretable. Interpretable methods give clear explanations of why a cis-regulatory interaction has been predicted and offer a framework for researchers to further explore predicted regulation. Whilst supervised models can offer interpretability by examining feature importance, unsupervised approaches inspired by cis-regulatory mechanisms are most interpretable. Examples of highly interpretable models include Activity by Contact (enhancers regulate genes if they are highly active and in contact with a promoter) and Vijayabaskar et al. (CREs regulate genes together if they are co-bound by TFs and co-active with gene expression).

Over the following chapters, this thesis will work towards developing a method which is applicable, implementable, and interpretable. This will culminate in the application of the method to identify new, biologically important mechanisms of gene regulation in the process of B cell differentiation – in which antibodies are produced during infection. In four results chapters, this thesis will make contributions to the fields of bioinformatics, gene regulation, and immunology as summarised below:

Chapter 2: The method from Vijayabaskar et al. 2019 is compared to the supervised JEME model. The analysis finds the two methods performed comparably and supports the co-authored publication of Vijayabaskar et al. 2019.

Chapter 3: The method from Vijayabaskar et al. 2019 is adapted for streamlined epigenomic and transcriptomic inputs. The resultant method, cisREAD, requires just two input datatypes and is capable of performing bottom-up discovery of key trans-regulators. The method is supported by an open-source R package and is presented in the first-authored publication of Emmett et al. 2023.

Chapter 4: The cisREAD method is applied to ATAC-seq and RNA-seq datasets across an *in vitro* system of B cell differentiation. The results are leveraged on a global scale to

identify a crucial shift in transcription-factor led regulation. cisREAD is also shown to recall known gene-specific cis-regulatory elements, and to generate new hypotheses of transcriptional control for master regulators. This chapter also contributes to Emmett et al. 2023.

Chapter 5: cisREAD is benchmarked against JEME, Activity-by-Contact and Pearson correlation methods using chromatin interaction datasets. cisREAD is found to best identify distal cis-regulatory interactions, which supports Emmett et al. 2023.

These chapters will be followed by a discussion in chapter 6, which brings together each of these contributions, and discusses their impact. This chapter will evaluate whether the development of cisREAD achieved the aim of an interpretable, implementable, applicable method. It will also discuss limitations faced in the thesis and suggest avenues for future work. Importantly it will highlight the potential for this work to uncover mechanisms underpinning gene dysregulation in disease.

Chapter 2. Vijayabaskar et al. and JEME methods comparatively predict gene regulation in murine haematopoiesis

2. 1 Introduction

Chapter 1 introduced supervised and unsupervised methods to predict gene-regulation from multi-omics data. This chapter focuses on comparing the performance of two of these methods: the supervised JEME model from Cao et al., 2017, and Vijayabaskar et al.'s unsupervised regression approach. Work presented in this chapter was published in Vijayabaskar et al., 2019.

In their 2019 paper, Vijayabaskar et al. described a novel approach to identify communities of Cis-Regulatory Elements (coCREs) which co-regulate the expression of differentiation-associated genes. Their method combined RNA-seq, DNase-seq and ChIP-seq data to identify gene-specific cis-regulatory elements in two separate murine lineages, branching from embryonic stem cells to cardiomyocytes, or macrophages.

This chapter describes the application of the JEME model to the macrophage lineage dataset, to compare performance with the Vijayabaskar et al. approach. This was achieved through evaluation of predicted gene-specific cis-regulatory elements against a set of experimentally validated enhancers.

The following introduction will briefly describe blood cell development, with focus on transcriptional control, as well as the Vijayabaskar et al. and JEME methods of enhancer prediction.

2.1.1 Blood cells differentiate from embryonic stem cells by haematopoiesis

Haematopoiesis is the process of blood cell development and differentiation (Figure 2.1). Starting from early embryonic development, haematopoiesis takes place in waves of primitive and definitive haematopoiesis: firstly, to supply the developing embryo, and finally to seed adult blood cell populations. *In vitro* models of haematopoiesis, are able to recapitulate blood cell differentiation; generating blood cell precursors and terminally-differentiated endpoints (e.g. macrophages) from embryonic stem cells (Figure 2.1). These tractable model systems allow researchers to study the regulatory dynamics which control blood cell development in early embryogenesis.

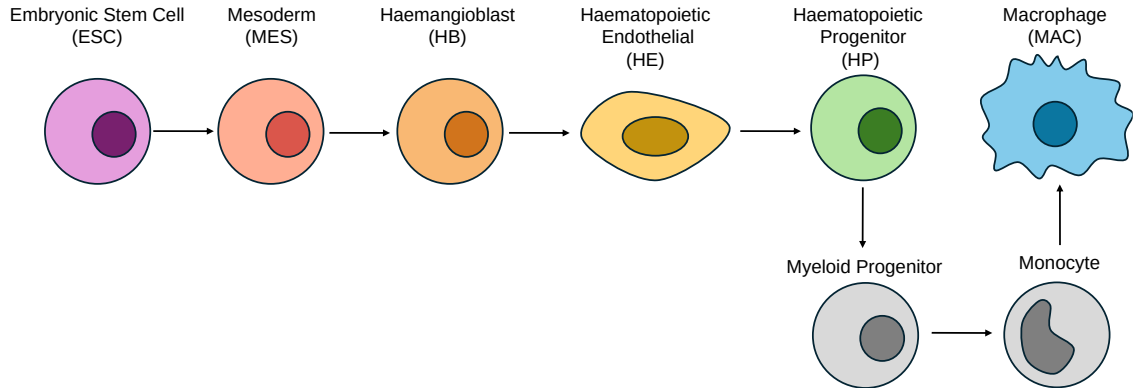


Figure 2.1 Haematopoetic differentiation from embryonic stem cells to macrophages. During gastrulation, embryonic stem cells (ESCs) give rise to the mesoderm (MES). Haemangioblasts (HB), with both endothelial and haematopoietic potential, arise from the mesoderm and differentiate into haematopoietic endothelial cells (HE). Haematopoietic progenitors (HP), also known as haematopoietic stem cells, emerge from this haemogenic endothelium, in an endothelial-to-haematopoietic transition. Haematopoietic progenitors (HP) have potential to produce terminally differentiated blood cell populations. Macrophages (MAC) differentiate from HPs along the myeloid lineage, through myeloid progenitor and monocyte cell states.

The first wave of primitive blood cell development takes place around embryonic day 7.25 (E7.25), shortly after gastrulation, in mouse embryonic development. Here the extra-embryonic mesoderm of the yolk sac brings forth erythroid precursors, macrophages and megakaryocyte progenitors to meet the needs of the developing embryo (Lacaud and Kouskoff, 2017).

This is followed by another wave of haematopoietic differentiation, also in the yolk sac, a day later at E8.25. From this process emerge erythro-myeloid progenitors, capable of giving rise to most erythroid and myeloid lineages. These progenitors expand into the foetal liver where they differentiate into mature blood cells (Mcgrath et al., 2015). The yolk sac also begins to generate lymphoid precursors at around E9.0 (Lacaud and Kouskoff, 2017).

Progenitors capable of giving rise to the full range of adult haematopoietic lineages begin to appear at E10.5, this time in major arteries – notably the dorsal aorta. Haematopoietic progenitors produced here (also referred to as haematopoietic stem cells) seed the foetal liver around E11.5 and expand in population (Medvinsky et al., 2011). This is succeeded by colonisation of the bone marrow by E16.5 (Lacaud and Kouskoff, 2017).

Whilst these processes take place in different structures, and branch to different endpoints, the initial process of transition is broadly similar.

Haemangioblasts are mesoderm-derived clonal precursors with both endothelial and haematopoietic potential (Lacaud and Kouskoff, 2017). Haemangioblasts have been shown to generate populations of haematopoietic endothelial cells in a process regulated by the TAL1 TF (Lancrin et al., 2009).

Definitive haematopoiesis in the yolk sac at E8.25 (erythro-myeloid progenitors) and E9.0 (B and T lymphoid precursors), as well as in the dorsal aorta at E10.5, has been evidenced to arise from the haematopoietic endothelium. Although the first primitive precursors emerge at E7.25 from a cell type with endothelial markers, there is debate over whether this constitutes a haematopoietic endothelium as vasculature has yet to be formed – instead the term 'haemogenic angioblast' has been proposed (Lacaud and Kouskoff, 2017).

Regardless of terminology, all these processes are unified by a common endothelial to haematopoietic transition, marked by changes in cell morphology and motility. The RUNX1 TF drives this transition in all waves of definitive haematopoiesis, but not the primitive E7.25 wave (Chen et al., 2009; Lancrin et al., 2009; Long and Joanna, 2018).

Macrophages are produced during all waves of embryonic haematopoiesis and, like other blood cells, are continuously replenished by HSCs throughout adulthood. This process consists of the progressive differentiation of haematopoietic progenitors to bipotential granulocyte-macrophage progenitors, to monocyte-progenitors, to mature monocytes and finally to terminally differentiated macrophages (Mcgrath et al., 2015). The PU.1 TF plays an important role in macrophage differentiation (Goode et al., 2016).

Regulation of gene expression is essential for establishing and maintaining cell fate specification and cell identity. Networks of core TFs, and their regulatory elements, control these complex transcriptional programs. Experimental studies in blood cells have helped elucidate key players in haematopoietic regulatory circuits (Goode et al., 2016; Schütte et al., 2016).

In 2016, Goode et al. isolated 6 cellular stages on the journey from embryonic stem cell to macrophage. RNA-seq, DNase-seq and ChIP-seq experiments were performed on embryonic stem cells (ESCs), mesodermal cells (MES), haemangioblasts (HB), haematopoietic endothelial cells (HE), haematopoietic progenitors (HP) and macrophages (MAC) to explore how TF dynamics shape blood cell specification. This dataset was used by Vijayabaskar et al. to identify CREs important to haematopoietic differentiation (Vijayabaskar et al., 2019).

2.1.2 The Vijayabaskar et al. method predicts gene-specific cis-regulatory elements by community detection and LASSO regression

Vijayabaskar et al.'s penalised regression approach predicted regulatory elements in a gene-specific manner, based on chromatin features and TF occupancy. The approach was unique in that it considered groups of enhancers, with correlated epigenetic and TF-binding patterns, to co-regulate expression as communities of cis-regulatory elements (coCREs) (Vijayabaskar et al., 2019).

The method first identified overlapping DHS/H3K27Ac peaks within 100kb of a TSS as 'candidate CRE's. Chromatin activity profiles were calculated from DNase-seq and H3K27Ac ChIP-seq signals for each candidate CRE across the dataset. Binary TF binding profiles, where 1 is a binding event and 0 is an absence, were obtained by overlapping CREs with TF ChIP-Seq peaks in cells with TF data. Candidate CREs with correlated with chromatin activity profiles and transcription factor binding profiles were grouped together to form coCREs in a community detection step, using the fastgreedy algorithm (Clauset et al., 2004). Candidates not assigned to a community were termed 'singleton CREs', only singleton CREs within 20kb of a TSS were considered further.

Gene-specific LASSO models were then constructed to identify singleton and coCREs whose chromatin activity profiles (for coCREs the average of individual member

profiles) were most predictive of gene expression (Tibshirani, 1996b). The significance of selected predictors, as they enter the LASSO model, was tested using the covariance test (Lockhart et al., 2014).

The Vijayabaskar et al. model selected cis-regulatory elements for 6,715 differentially expressed genes. These selected CREs were found to be enriched for chromatin activity, TF binding and conservation, compared to all candidate regulators. Vijayabaskar et al. also reported significant overlaps with murine super-enhancers, characterised by TF binding and chromatin modification (Wei et al., 2016; Whyte et al., 2013), and experimentally validated enhancers, active in mouse haematopoiesis (Schütte et al., 2016).

The enrichment of enhancer features, and significant overlap with enhancer sets, indicated that the Vijayabaskar et al. method correctly prioritised gene-specific enhancers, important to murine haematopoiesis. However, the method's performance, relative to other models, had not yet been tested.

2.1.3 The JEME method predicts enhancer-promoter interactions using LASSO regression and a random forests classifier

Cao et al.'s Joint Effects of Multiple Enhancers (JEME) method employed a two-step machine learning framework to predict cell type-specific enhancer-TSS interactions. Firstly, LASSO linear regression models were constructed to estimate the ability of DNase, H3K27Ac, H3K27Me3 and H3K4Me1 signals to predict the expression of a TSS within 1Mb of an enhancer (Tibshirani, 1996b). Secondly, LASSO error terms were input along with DNase and histone enhancer, promoter and window (the chromatin segment between a promoters and enhancer) features into a Random Forest classifier trained on chromosome conformation data (Breiman, 2001).

JEME was originally trained on ChIA-PET data, targeting RNA polymerase II, from human chronic myeloid leukaemia cell line K562 and applied to 127 human samples analysed by the integrated Roadmap and ENCODE Epigenomics project (Dunham et al., 2012; Roadmap Epigenomics Consortium et al., 2015). JEME also made predictions for 808 human samples from the FANTOM5 consortium using an altered feature set where eRNA expression was substituted for epigenetic signals (Andersson et al., 2014).

Cao et al. evaluated the performance of their model by cross-validation, alongside validation against chromatin interactions and eQTLs in same-sample and cross-sample tests. They found their model was better able to link enhancers to target genes than alternative methods. These included assigning enhancers to random genes, assigning enhancers to their nearest gene, or assigning enhancers to genes using a random forests model with distance as the only feature.

2.1.4 Rationale for comparison

To evaluate the performance of predictive methods, it is important to understand how a method performs relative to other methods. Whilst both methods were validated against indicators of enhancer activity, or enhancer-promoter interactions, neither had been compared to alternative published models at the time of writing.

JEME was chosen for comparison to the Vijayabaskar et al. method, due to its status as a recent state-of-the-art method and its similar use of LASSO regression. The decision was also influenced by the availability of most input requirements (JEME was applicable) and the provision of open-source software (JEME was implementable).

Despite some similarities, the two methods were designed for different purposes. JEME is a supervised enhancer-promoter interaction (EPI) predictor. It aims to identify target gene promoters for all active enhancers in a sample. Contrastingly, the Vijayabaskar et al. method is an unsupervised approach to identify cis-regulatory elements (including enhancers and promoters) whose co-activity drives expression of genes important to differentiation. Importantly this means it does not predict all cell-specific EPIs (like JEME) but predicts the interactions most important to a system. This crucial difference must be considered when interpreting any comparisons.

2.2 Aims and Objectives

This chapter aimed to evaluate the ability of the Vijayabaskar et al. method to link enhancers to their correct target genes, in comparison with the JEME model. To achieve this aim the following objectives were set out:

1) to apply the JEME model to predict enhancer-promoter interactions in murine haematopoietic cell stages;
2) to evaluate the performance of JEME on the murine haematopoietic system; and
3) to compare the performance of JEME with the performance of the Vijayabaskar et al. method.

2.3 Methods

2.3.1 Datasets

To meet objective 1, JEME was reapplied to the haematopoietic dataset from Goode et al. 2016, which was used in Vijayabaskar et al. 2019. This required unavoidable alterations to the original method. Since JEME was originally applied to predict EPIs in human cells and tissues, the random forest classifier was retrained for application to murine cells. The feature set input to JEME was altered due to the lack of H3K4me1 ChIP-seq in the haematopoietic dataset. This mark was instead substituted with H3K4me3, which was available. Whilst both these features should be useful for prediction, they are not equivalent. H3K4me1 preferentially localises to enhancers and H3K4me3 to promoters (Calo and Wysocka, 2013). The altered training and input datasets used for application of JEME are shown in Figure 2.2.

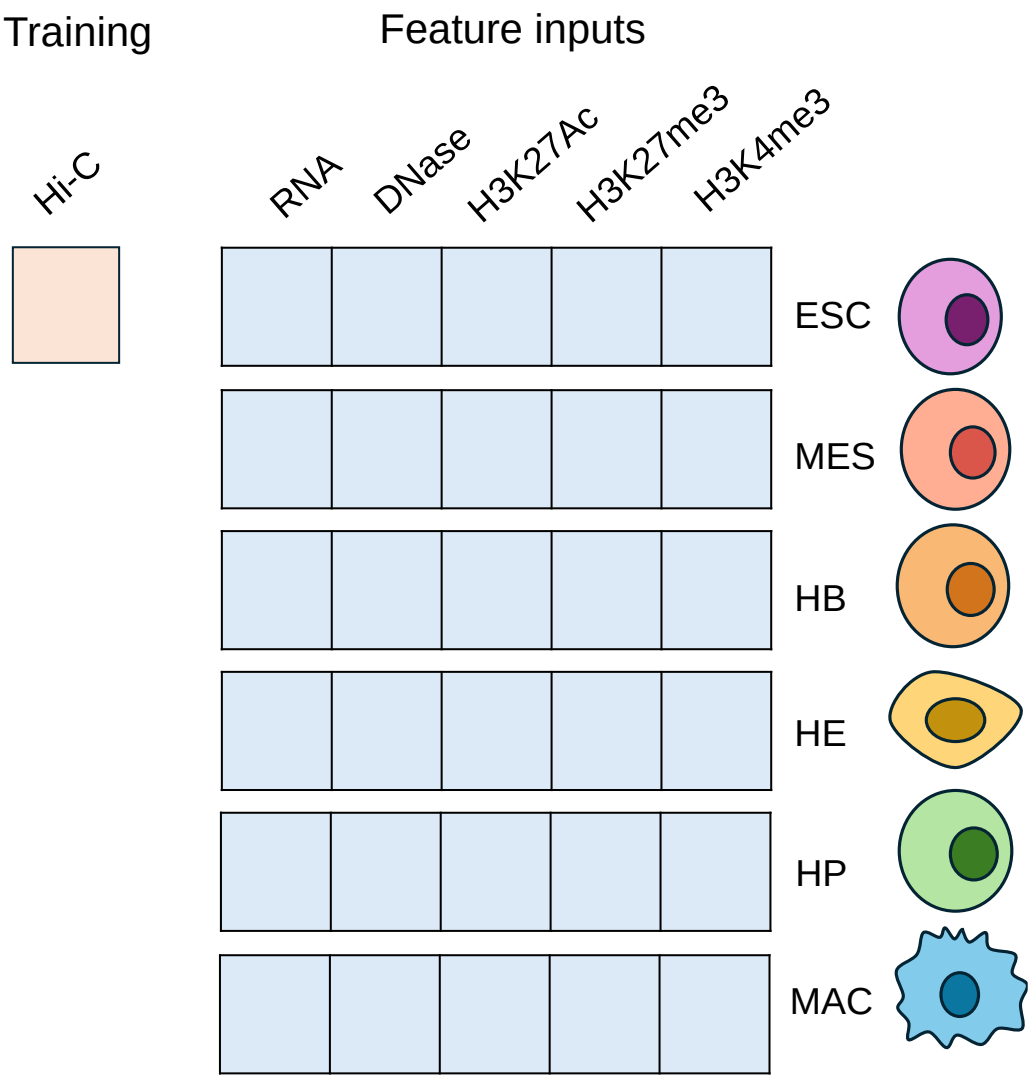


Figure 2.2 Training and input datasets used during reapplication of JEME to the murine haematopoietic system. JEME was retrained on one Hi-C dataset from embryonic stem cells (ESCs). For each cell stage – ESC, mesoderm (MES), haemangioblast (HB), haematopoietic endothelium (HE), haematopoietic progenitor (HP) and macrophage (MAC) – five datasets (RNA-seq, DNase-seq and ChIP-seq for H3K27Ac, H3K27me3 and H3K4me3) were input to JEME.

JEME source code was downloaded from https://github.com/yiplabcuhk/JEME and adapted to the haematopoietic dataset. JEME's Random Forest classifier was also downloaded and retrained on class labels derived from ESC Hi-C data using the WEKA software (Mark et al., 2009). Enhancer-TSS interactions were predicted in MES, HB, HE, HP, and MAC cells using the default threshold of 0.35, used in the original Cao et al., 2017 paper. Input file processing, model training and evaluation of results are described below.

2.3.2 Processing of input datasets

To reapply JEME to the murine haematopoietic system DNase, H3K27Ac, H3K27me3 and H3K4Me3 signals were extracted from .bigWig DNase-seq and ChIP-seq files which were generated by Vijayabaskar et el. 2019. These signals were extracted for 'active enhancer' 'promoter' and 'window' regions as described below.

JEME step 1 involves the construction of pairwise LASSO linear regression models where enhancer features are correlated with the expression of a nearby (<1Mb) gene. To achieve this, JEME requires a set of potential enhancer-TSS pairs alongside DNase and histone enhancer features and TSS expression.

In their original paper, Cao et al. used 15-state ChromHMM predictions to define active enhancers in each sample. ChromHMM made predictions using ChIP-seq signal data imputed for 16 histone marks in the 127 ENCODE + Roadmap cells (Ernst and Kellis, 2015). As equivalent feature data was unavailable, active enhancer predictions were taken from the modified 4-state ChromHMM analysis performed by Goode et al. 2016.

A set of 124,004 enhancers was curated by taking the union of all enhancers predicted active by the custom ChromHMM model in at least one cell. 23,697 protein-coding TSSs were obtained from the mm10 RefSeq Curated annotation using the UCSC table browser (Karolchik, 2003; O'Leary et al., 2016). Enhancers were paired to all protein-coding TSSs within 1Mb, generating a total of 3,658,772 potential pairs.

Enhancers were resized to a uniform length of 2,500bp, centred on the middle co-ordinate, as originally performed by Cao et al., 2017.. DNase and histone ChIP-seq signals were averaged across the 2,500bp regions to generate enhancer features for each cell type. Expression counts in RPKM (Reads Per Kilobase of transcript per Million mapped reads) were calculated for each 'gene' – defined as the 1,000bp window centred on each TSS – and $\log_2$ transformed with a pseudocount of 1.

Code was run for JEME step 1 and LASSO error terms were calculated for potential enhancer-TSS pairs in each cell type. Enhancer-TSS pairs were modelled according to Equation 2.1.

Equation 2.1 Linear models constructed in step 1 of JEME

$$y = a_{i0} + \sum_{j} a_{ij} x_{ij},$$

Here y represents the expression of a TSS, j is a nearby enhancer and i each predictive feature – summated over all enhancers within 1Mb of the TSS. The term x_{ij} denotes the value of feature i in enhancer j and a_{ij} represents the coefficients learned by LASSO.

Error terms e_{ijk} were computed according to the Equation 2.2 to describe how well TSS expression y can be explained by each feature i of enhancer j independently in sample k.

Equation 2.2 Calculation of LASSO error terms for predictive features in step 1 of JEME

$$e_{ijk} = \left| y_k - \left(a_{i0} + a_{ij} x_{ijk} \right) \right|'$$

During step 2, promoter, window, and active enhancer features are input into the final Random Forest classifier alongside their LASSO error terms and genomic distance. For each sample, only enhancers predicted to be active in that cell type were considered.

DNase-seq and ChIP-seq signals were averaged across uniform 2,000bp wide promoter regions, taken from 1,500bp upstream of a TSS to 500bp downstream (as performed by Cao et al., 2017). Epigenetic signals were also averaged across 'window' regions – the segment of DNA between an active enhancer and its paired TSS. Genomic distances were calculated for potential enhancer-TSS pairs.

Altogether this yielded a total of 17 features for consideration by the final model: 4 epigenetic features (DNase, H3K27Ac, H3K27Me3 and H3K4me3) were considered for active enhancers, promoters and windows, accounting for a total of 12 features. 4 LASSO enhancer-TSS pair error terms (one per epigenetic feature) and genomic distance made up the final five features.

2.3.3 Retraining the JEME Model

Hi-C contacts from murine ESCs, published in Krijger et al. 2016 (uploaded to GEO under the accession GSM2026260) was used to assemble the new set of training pairs. An enhancer-TSS pair was placed in the positive set if a co-ordinate within an enhancer region interacted with a co-ordinate within a gene promoter. The ratio I/N of

interacting pairs I (positive class) to non-interacting pairs N (negative class) was determined by Equation 2.3.

Equation 2.3 Calculation of ratio of positive to negative class labels in the retraining of JEME's random forest model

$$\frac{\mathrm{I}}{\mathrm{N}} = \frac{\mathrm{ER}}{\mathrm{P} - \mathrm{ER}}$$

Here E is the number of active enhancers, P is the number of enhancer-TSS pairs within 1Mb and R is the average number of TSSs per active enhancer (Cao et al., 2017). Aiming for a ratio of 0.15 (the value selected by Cao et al. for their training set) a total of 18,353 positive and 122,353 background pairs were calculated (Table 2.1). Background, non-interacting, pairs were assembled using the 'random targets' method described by Cao et al., in which enhancers from the positive pair set were randomly assigned to non-interacting TSSs within 1Mb (Cao et al., 2017).

Table 2.1 Positive and negative enhancer-TSS pairs in ESC cells used to retrain JEME

	Number
Active Enhancers E	29,743
Enhancer-TSS Pairs P	901,479
Average Number TSSs per Active Enhancer R	3.95
Ratio of Positive to Negative Pairs x	0.15
Interacting Pairs I	18,353
Non-Interacting Pairs N	122,353

2.3.4 Evaluation of JEME through cross-validation

To meet objective 2, the performance of the retrained JEME model was evaluated through cross-validation and validation with external data.

Five-fold cross-validation was performed in WEKA to evaluate the performance of the retrained JEME model (Mark et al., 2009). Cross validation was performed with both random and chromosomally-sorted folds, as recommended in Xi and Beer, 2018 and Cao and Fullwood, 2019. Performance metrics were calculated using definitions defined in section 1.3.1.4 of the introduction.

2.3.5 Evaluation of JEME performance with validated enhancers

The performance of JEME was also compared to haematopoietic enhancers for 9 transcription factor genes, which were validated by reporter assays by (Schütte et al., 2016). To match the HPC-7 cell type used in the study, enhancers predicted for these genes in HP cells were selected for comparison. JEME-predicted enhancers were considered validated as true positives (TPs) if they intersected a Schutte et al. enhancer by at least one base pair, and false positives (FPs) if they overlapped a region found to lack enhancer activity by Schutte et al. True negatives (TNs) were predicted non-interacting enhancers (for the TF genes) which overlapped an inactive region from Schutte et al. Conversely false negatives (FNs) were predicted non-interacting enhancers which overlapped a Schutte et al. enhancer. Precision, recall and F1 scores were calculated from these definitions. WEKA was also used to obtain threshold independent AUPR and AUROC scores. Predictions by Vijayabaskar et al. were also intersected with JEME-predicted and Schutte et al. validated enhancers.

2.3.6 TF binding, DNase I hypersensitivity and H3K27Ac Analysis

Due to the small size of the Schutte et al. dataset, and the large number of predictions made by JEME, additional criteria were defined to suggest the haematopoietic activity of a predicted enhancer. These were DNase I hypersensitivity or H3K27 acetylation, and occupancy by two or more TFs. These regions were identified by intersecting DNase-seq and ChIP-seq peaks in HP (peak calling and cut-off p-values described in Vijayabaskar et al. 2019) with co-ordinates for HP JEME predictions to select novel enhancer candidates for the nine loci from Schütte et al. 2016.

To visualise predicted and validated gene-specific enhancers, .bigWig tracks for 10 TFs (CEBPβ, FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1 and TAL1), DNase hypersensitivity and H3K27Ac in HP cells were uploaded to the Integrative Genomics Viewer along with co-ordinates for validated and predicted enhancers (Robinson, 2012). For Runx1, enhancers from Marsman et al., 2017 were also compared.

2.3.7 Comparison with Vijayabaskar et al. predictions

To meet objective 3, predictions made by JEME were compared to predictions made by Vijayabaskar et al. 2019. Genomic coordinates were overlapped for JEME predicted enhancer-TSS pairs (from any cell type) and selected CREs-gene pairs from Vijayabaskar et al. 2019. JEME predictions were limited to active enhancers paired to genes within 100kb of a TSS considered by Vijayabaskar et al., predicted in one or more samples using bedtools (Quinlan and Hall, 2010). A hypergeometric test was performed in R to calculate the significance of overlap for JEME and Vijayabaskar et al. predictions.

2.3.8 Evaluation of JEME and Vijayabaskar et al. predictions using experimentally validated enhancers

Finally, the set of experimentally validated enhancers from Schutte et al. was used to compare the performance of JEME and Vijayabaskar et al. methods. The two sets of predictions were compared to the validated enhancers, and inactive regions, to true and false positive predictions and calculate performance metrics.

2.4 Results

2.4.1 JEME predicted enhancer-promoter interactions in haematopoietic cell stages

JEME predicted interactions between 106,538 enhancers and 18,056 TSSs, constituting 551,456 unique enhancer-TSS pairs, in at least one of the five cell types (MES, HB, HE, HP and MAC). Predicted EPIs in individual samples ranged from 86,494 in MES to 250,293 in HB cells (Table 2.2). The retrained JEME model predicted more interactions per sample than other predictive methods (FANTOM5 JEME: average 2,392, IM-PET: 17,483-71,536, RIPPLE: 11,696-32,308) (Cao et al., 2017; He et al., 2014; Roy et al., 2015).

The median distance between an enhancer and its paired TSS was under 100kb in all cell lines (Table 2.2), yet JEME predicted pairs were separated by a maximum distance approaching the upper limit of 1Mb. Figure 2.3 shows that distances were similarly distributed across all prediction samples and the ESC training set. Median distances

reported here were similar to those reported for FANTOM5 JEME (33-77kb) and IM-PET (58-123kb) (Cao et al., 2017; He et al., 2014).

Each enhancer was predicted to interact with a mean of 5–14.2 TSSs depending on cell type (Table 2.2, Figure 2.4A). These values were higher than the average number of enhancers per gene in the ESC training set (mean = 2), and the range of means predicted by FANTOM5 JEME (range = 1.3-2) (Cao et al., 2017). Current estimates place the average number of enhancers per gene slightly higher, between 4 and 5, in human cell lines; however the re-applied JEME model still exceeded these values in most samples (Jin et al., 2013; Lam et al., 2015; Mora et al., 2015).

The mean number of TSSs per enhancer was higher in all JEME predicted cells than in the ESC training set (Table 2..2, Figure 2.4B). However FANTOM5 JEME predicted similar numbers of TSSs per enhancer (mean 2.3 to 5.5) (Cao et al., 2017).

The majority (64.6%) of enhancers intersected with an H3K27Ac peak in one or more of the samples in which it was predicted active. A slightly lower number of enhancers (40%) intersected with a DHS, and a total of 35.2% of predictions were both H3K27Ac enriched and DNase I hypersensitive (Table 2.2). Whilst these values varied between cell-lines, H3K27Ac enrichment and DNase I hypersensitivity levels in the ESC training enhancers fell within this range (Figure 2.5).

Table 2.2 Details of genome-wide JEME predictions (threshold=0.35) and positive ESC training pairs

	JEME Predictions						ESC (Train-ing)
	MES	**HB**	**HE**	**HP**	**MAC**	**Any**	
Interacting Enhancers	22,283	50,144	45,802	38,511	35,058	106,538	11,349
Enhancer-TSS Interactions	86,494	250,293	197,664	177,283	143,807	551,456	18,353
Interacting TSSs	17,262	17,598	17,929	17,617	17,651	18,056	9,216
Median Distance to TSS (bp)	54,094	85,047	80,308	69,397	64,482	92,943	62,341
Mean Enhancers per TSS	5.01	14.22	11.02	10.06	8.15	30.54	2.0
Mean TSSs per Enhancer	3.88	4.99	4.32	4.6	4.1	5.2	1.6
H3K27Ac Marked Interacting Enhancers	68.5%	61.4%	76.3%	60.6%	71.5%	64.6%	62.3%
DNase I Hypersensitive Interacting Enhancers	49.9%	31.1%	56.5%	30.9%	60.6%	40.0%	58.8%
H3K27Ac Marked & DNase I Hypersensitive Interacting Enhancers	44.5%	26.5%	50.6%	26.9%	51.5%	35.2%	41.5%
Candidate Enhancer-TSS Interactions	809,147	1,742,019	1,437,406	1,288,875	1,134,995	3,658,772	1,587,423
Candidate Active Enhancers (ChromHMM Active Predictions)	25,826	58,492	53,379	42,268	39,477	124,004	32,389

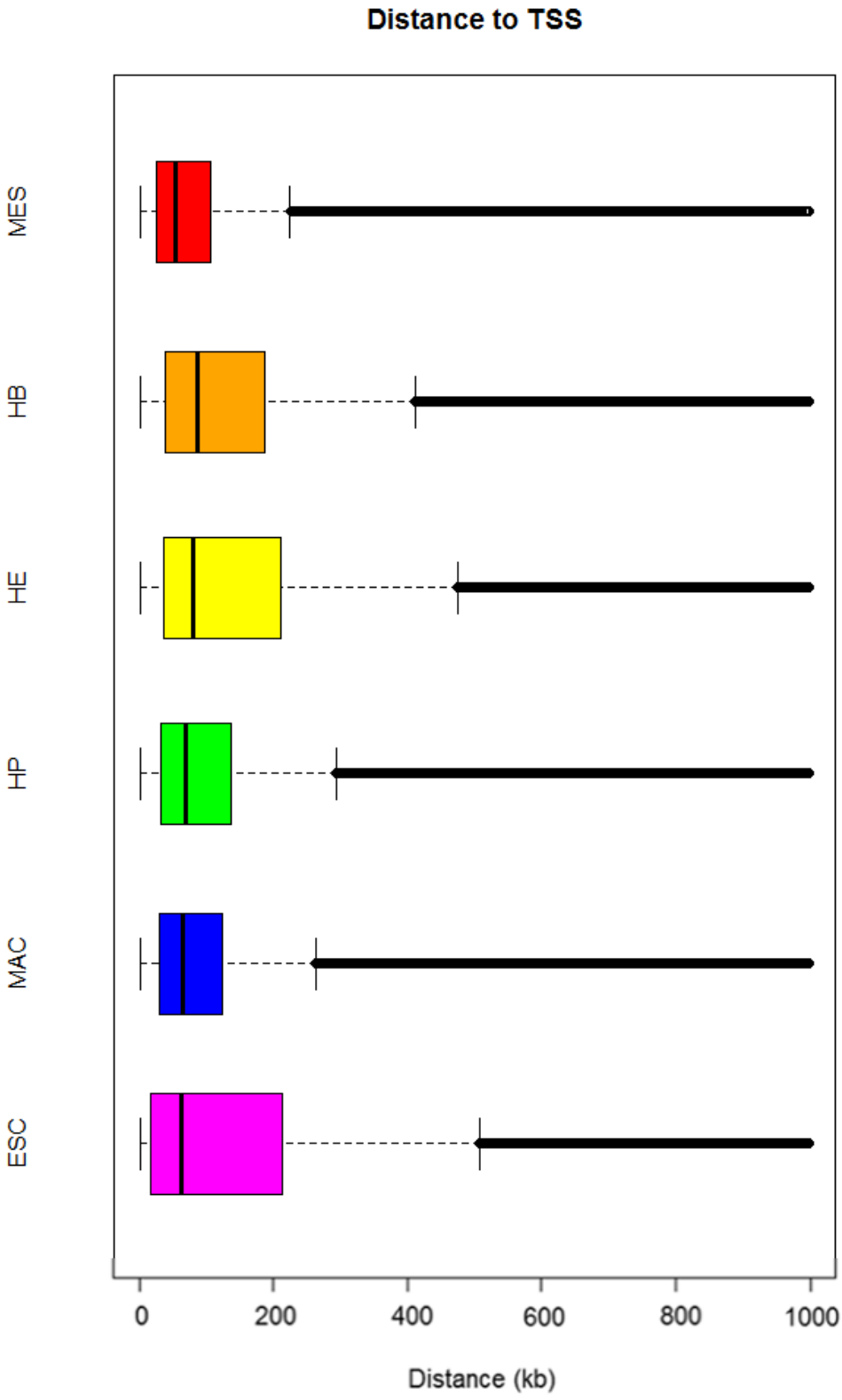


Figure 2.3 Enhancer-TSS distance distributions for JEME prediction and training datasets. Boxplots (coloured by cell-stage) show distributions of distances between TSS and midpoint of enhancer region for MES, HB, HE, HP and MAC predictions and positive ESC training enhancer-TSS pairs.

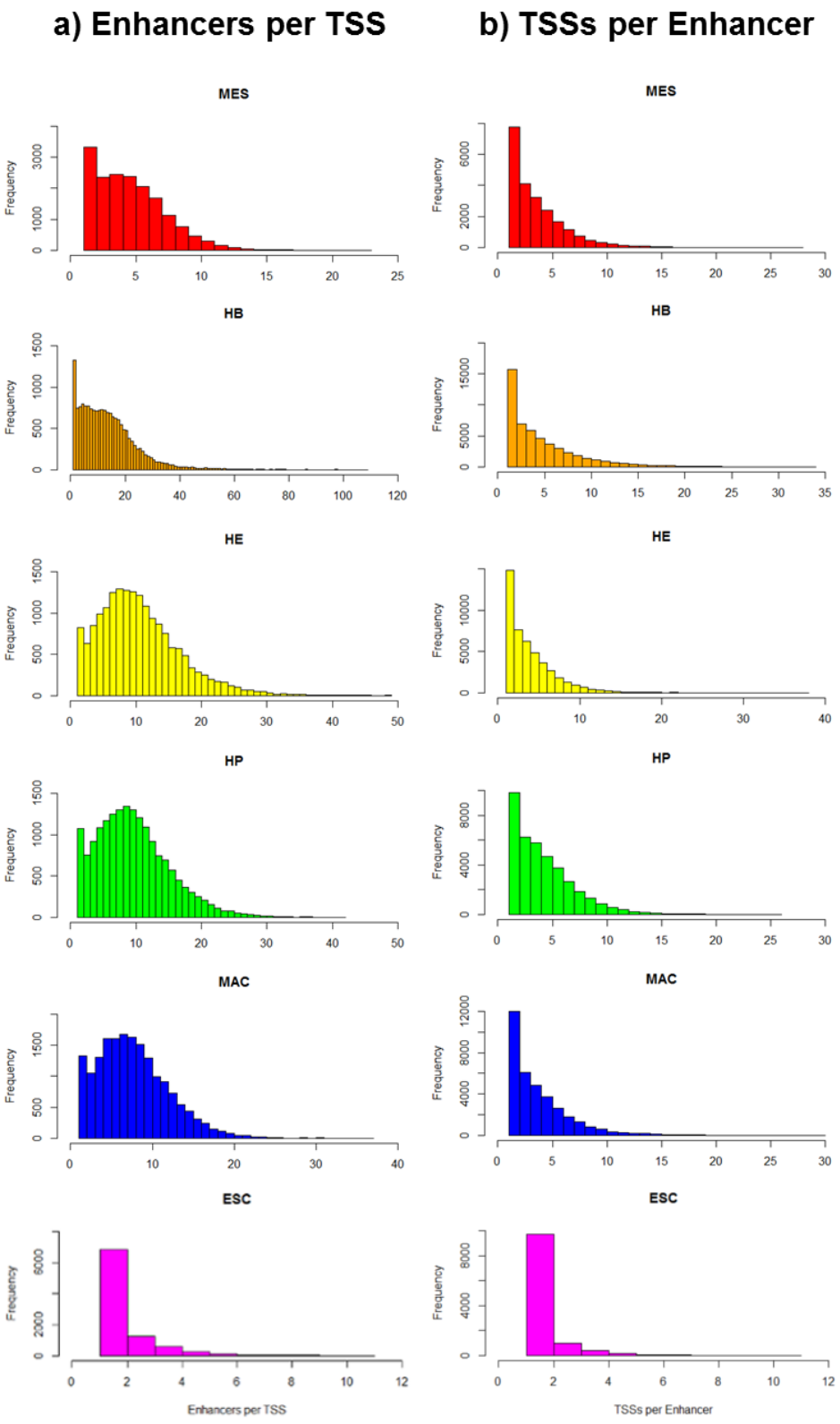


Figure 2.4 Enhancers per TSS, and TSSs per enhancer for JEME prediction and training datasets. A) Enhancers per TSS for MES, HB, HE, HP and MAC predictions and ESC training pairs B) TSSs per enhancer for MES, HB, HE, HP and MAC predictions and positive ESC training pairs. Histograms (bin size = 1, coloured by cell stage) show that JEME predicted more enhancers per TSS, and more TSSs per enhancer, than were assigned by Hi-C in model training on ESCs.

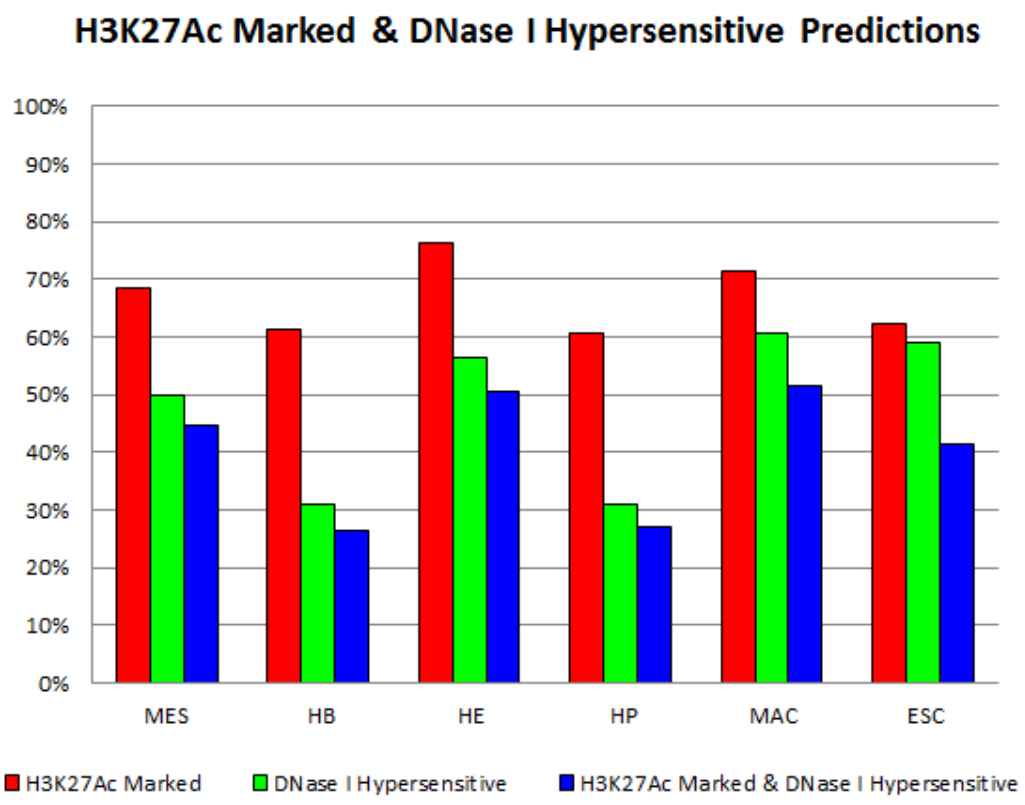


Figure 2.5 Percentage of H3K27Ac enriched and DNase I Hypersensitive enhancers for JEME prediction and training datasets. Bar charts show percentage of interacting enhancers (predicted by JEME in MES, HB, HE, HP and MAC, or identified during training on ESCs) that intersected with H3K27Ac ChIP-seq and/or DNase-seq peaks, which indicate chromatin activity and accessibility respectively. Similar proportions of H3K27Ac-marked and/or DNAse I hypersensitive predictions are observed in training (ESC) and prediction enhancer sets.

2.4.2 JEME showed underwhelming performance in cross-validation

Five-fold cross-validation was performed on ESC training pairs with both random and chromosomally sorted folds. Random five-fold cross-validation reported an F1-score of 0.51, AUROC of 0.834 and AUPR of 0.53, whilst chromosomally sorted cross-validation yielded slightly lower performance metrics (Table 2.3). This is in line with the performance drop observed upon chromosomal sorting by Cao and Fullwood, 2019. The reported F1 and AUPR values combine performance measured by precision and recall, which account for false positives and false negatives, respectively. The metrics reported for the untrained JEME model were lower than those reported in the original JEME paper when using 'random targets' background pairs during random five-fold cross-validation (F1: ~0.66, AUROC: ~0.92, AUPR: ~0.67).

Table 2.3 JEME performance, measured during five-fold cross-validation on ESC training data, using both random and chromosomally sorted cross-validation folds. For each cross-validation strategy, a confusion matrix gives the number of true positives (TP), false positives (FN), true negatives (TN) and false negatives (FN). Performance metrics include the AUROC (area under the receiver operating curve) and AUPR (area under precision recall) scores, which are independent of threshold. Other metrics give performance at the default threshold of 0.35.

Enhancer-TSS Predictions during Cross-Validation		**Random Folds**		**Chromosomally-Sorted Folds**	
		EPI (n=14,280)	Non-EPI (n=125,886)	EPI (n=14,245)	Non-EPI (n=126,461)
ESC Training Pairs	EPI (n=18,353)	8,476 TP	9,877 FN	7,828 TP	10,525 FN
	Non-EPI (n=122,353)	6,344 FP	116,009 TN	6,417 FP	115,936 TN
Accuracy		0.89		0.88	
Precision		0.57		0.55	
Recall		0.46		0.43	
F1-Score		0.51		0.48	
AUROC		0.84		0.82	
AUPR		0.53		0.5	

2.4.3 JEME accurately predicted experimentally validated enhancers, but most predictions were untested

In 2016, Schütte et al. identified enhancers which regulate a network of key transcription factors involved in haematopoietic development. Candidate enhancers were validated by transgenic mouse experiments, reporter assays and ChIP-seq analysis (Schütte et al., 2016).

Here candidate regulatory regions were cloned downstream of the *LacZ* gene, and the reporter vector was delivered to mouse embryos, which were evaluated for haematopoietic *LacZ* staining. Tested enhancers were defined as active if the enhancer drove reporter gene expression in the dorsal aorta or foetal liver of E10-11.5 *LacZ*-reporter transgenic mice. TF binding to each element was studied using ChIP-seq data from haematopoietic progenitor line HPC-7, from Wilson et al., 2010, and 416B myeloid progenitors. Lastly, the effects of TF binding site mutagenesis on enhancer activity were quantified using luciferase reporter assays in 416B cells.

Schütte et al. confirmed a total of 23 enhancers across nine TF genes: E*rg, Fli1, Gata2, Gfi1b, Lyl1, Meis1, Spi1, Runx1* and *Tal1* (Table 2.4).

Table 2.4 Experimentally validated enhancers from Schütte et al. alongside HP-specific JEME predictions, Vijayabaskar et al. predictions and TF binding events in HPC-7 and HP cells

Gene	Enhancer	Predicted by JEME	Predicted by Vijayabaskar et al.	HPC-7 TF binding	HP TF binding
Erg	+65	No	Yes	ERG, FLI1, TAL1	CEBPβ, FLI1
Erg	+75	Yes	Yes	FLI1, PU.1	None
Erg	+85	No	Yes	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA2, GFI1, GFI1B, LMO2
Fli1	-15	Yes	No	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA2, GFI1
Fli1	+12	Yes	No	FLI1	GATA1, PU.1
Gata2	-93	Yes	No	FLI1, RUNX1	FLI1, LMO2, RUNX1
Gata2	-92	Yes	No	FLI1, PU.1	PU.1
Gata2	-3	No	Yes	None	None
Gata2	+3	No	Yes	FLI1, TAL1	FLI1, GATA2, LMO2, TAL1
Gfi1b	+13	Yes	No	GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, TAL1
Gfi1b	+16	Yes	No	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, RUNX1, TAL1
Gfi1b	+17	Yes	No	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
Lyl1	Promoter	Yes	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
Meis1	+48	Yes	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
Spi1	-14	No	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	CEBPβ, FLI1, GFI1, GFI1B, LMO2, PU.1, RUNX1
Runx1	-59	Yes	No	FLI1, GATA2, GFI1B, RUNX1, TAL1	CEBPβ, FLI1, GATA1, GFI1, LMO2, TAL1
Runx1	+3	Yes	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
Runx1	+23	Yes	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	CEBPβ, FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
Runx1	+110	Yes	Yes	FLI1, GATA2, GFI1B, RUNX1, TAL1	CEBPβ,FLI1, GATA1, GFI1, GFI1B, LMO2, PU.1, TAL1
Runx1	+204	No	No (exceeds distance threshold)	PU.1,RUNX1	None
Tal1	-4	No	Yes	None	FLI1

Tal1	+19	Yes	Yes	FLI1, GATA2, PU.1	CEBPβ, FLI1, LMO2, PU.1
Tal1	+40	Yes	Yes	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1

A further 24 tested regions were found to have no haematopoietic-specific regulatory activity (Table 2.5). Using these elements as examples of validated positive and validated negative enhancers, the Schütte et al. dataset was compared with JEME predictions to indicate performance. Predictions for the nine loci in HP cells were selected for comparison. The HP sample was selected due to its developmental equivalence with HPC-7 and the presence of HP populations in E10-11.5 dorsal aorta and foetal liver – the sites used for *in vivo* validation.

Table 2.5 Schütte et al. tested regions which failed to show enhancer activity in HPC-7 alongside HP-specific JEME predictions and Vijayabaskar et al. predictions.

Gene	**Enhancer**	**Predicted by JEME**	**Predicted by Vijayabaskar et al.**	**HPC-7 TF binding**	**HP TF binding**
Erg	+65	No	Yes	ERG, FLI1, TAL1	CEBPβ, FLI1
Erg	+75	Yes	Yes	FLI1, PU.1	None
Erg	+85	No	Yes	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA2, GFI1, GFI1B, LMO2
Fli1	-15	Yes	No	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA2, GFI1
Fli1	+12	Yes	No	FLI1	GATA1, PU.1
Gata2	-93	Yes	No	FLI1, RUNX1	FLI1, LMO2, RUNX1
Gata2	-92	Yes	No	FLI1, PU.1	PU.1
Gata2	-3	No	Yes	None	None
Gata2	+3	No	Yes	FLI1, TAL1	FLI1, GATA2, LMO2, TAL1
Gfi1b	+13	Yes	No	GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, TAL1
Gfi1b	+16	Yes	No	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, RUNX1, TAL1
Gfi1b	+17	Yes	No	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
Lyl1	Promoter	Yes	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
Meis1	+48	Yes	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1

Spi1	-14	No	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	CEBPβ, FLI1, GFI1, GFI1B, LMO2, PU.1, RUNX1
Runx1	-59	Yes	No	FLI1, GATA2, GFI1B, RUNX1, TAL1	CEBPβ, FLI1, GATA1, GFI1, LMO2, TAL1
Runx1	+3	Yes	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
Runx1	+23	Yes	Yes	FLI1, GATA2, GFI1B, PU.1, RUNX1, TAL1	CEBPβ, FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1
Runx1	+110	Yes	Yes	FLI1, GATA2, GFI1B, RUNX1, TAL1	CEBPβ, FLI1, GATA1, GFI1, GFI1B, LMO2, PU.1, TAL1
Runx1	+204	No	Not Considered	PU.1, RUNX1	None
Tal1	-4	No	Yes	None	FLI1
Tal1	+19	Yes	Yes	FLI1, GATA2, PU.1	CEBPβ, FLI1, LMO2, PU.1
Tal1	+40	Yes	Yes	FLI1, GATA2, GFI1B, RUNX1, TAL1	FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1, TAL1

70% of Schütte et al. experimentally verified enhancers and 25% of inactive regions were predicted active by JEME, respectively accounting for 9.9% and 3.7% of JEME predictions across the nine loci. Raising or lowering the 0.35 threshold did not improve overall performance (Table 2.6).

Table 2.6 Evaluation of HP JEME predictions, at different thresholds, against Schütte et al. validated enhancers for *Erg*, *Fli1*, *Gfi1b*, *Lyl1*, *Meis1*, *Spi1*, *Runx1* and *Tal1* genes

HP JEME Prediction Threshold		**Schütte et al. 2016**		**Accuracy**	**Precision**	**Recall**	**F1-Score**
		Active (*n*=23)	Inactive (*n*=24)				
0.29	Predicted (n=226)	17 TP	8 FP	0.702	0.680	0.739	0.708
	Not Predicted	6 FN	16 TN				
0.32	Predicted (n=187)	17 TP	8 FP	0.702	0.680	0.739	0.708
	Not Predicted	6 FN	16 TN				
0.35	Predicted (n=167)	16 TP	6 FP	0.723	0.727	0.696	0.711
	Not Predicted	7 FN	18 TN				
0.38	Predicted (n=133)	16 TP	6 FP	0.723	0.727	0.696	0.711
	Not Predicted	7 FN	18 TN				
0.41	Predicted (n=101)	15 TP	5 FP	0.723	0.750	0.652	0.698
	Not Predicted	8 FN	19 TN				

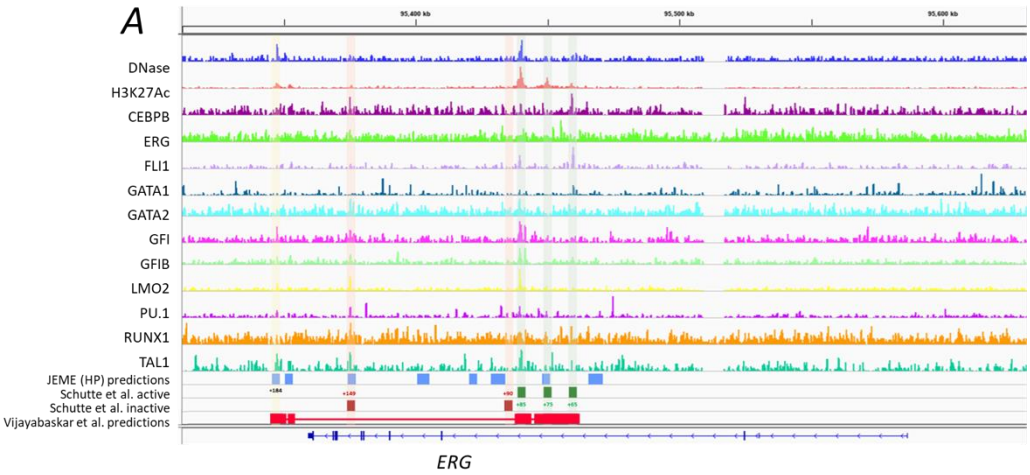
2.4.5 Few untested JEME predictions showed chromatin and TF features of enhancer activity

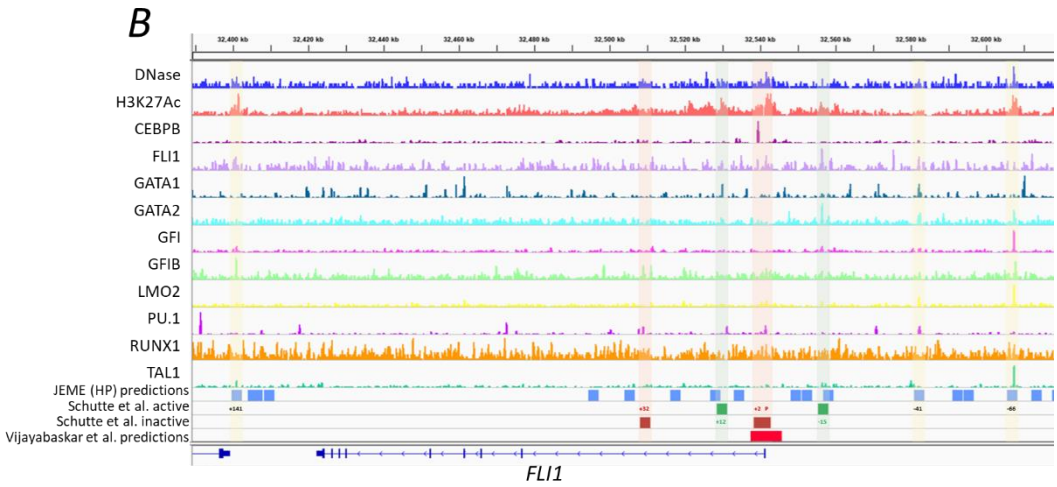
The comparison between JEME predictions and Schutte et al. validated enhancers indicated good performance against regions tested for enhancer activity. However, it also revealed large numbers of untested JEME predictions which may represent true positives or false positives. To explore the potential of these untested enhancers to regulate gene expression in haematopoiesis their chromatin and TF binding features were examined.

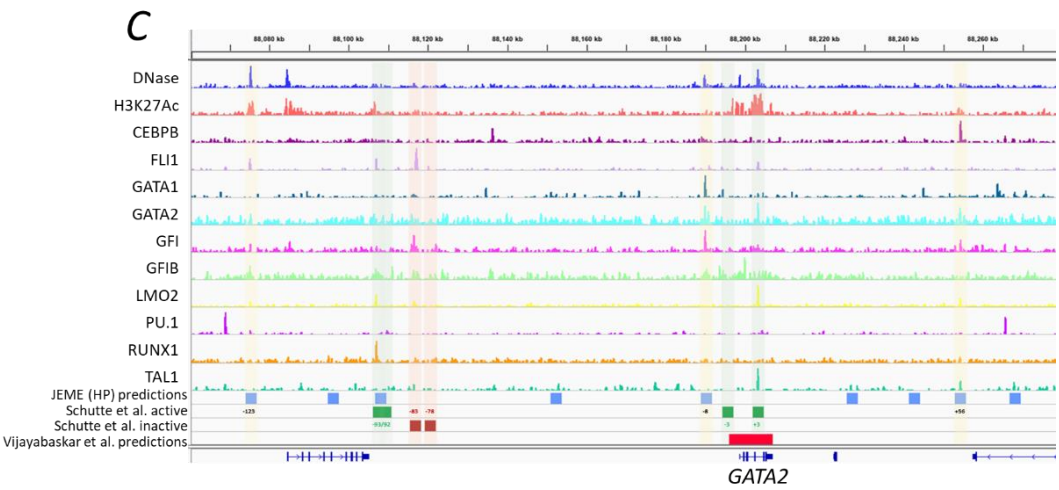
H3K27Ac enrichment, DNase I hypersensitivity and binding profiles for 10 TFs involved in haematopoietic development (CEBPβ, FLI1, GATA1, GATA2, GFI1, GFI1B, LMO2, PU.1, RUNX1 and TAL1) were aligned across these nine loci. JEME predictions were considered alongside Schütte et al. enhancers as well as selected CREs for the same genes from the Vijayabaskar et al. manuscript. Enhancers from Marsman et al. 2017, experimentally validated by transgenic zebrafish experiments and 4C contact data in HPC-7, were also reviewed for the *Runx1* locus (Marsman et al., 2017)

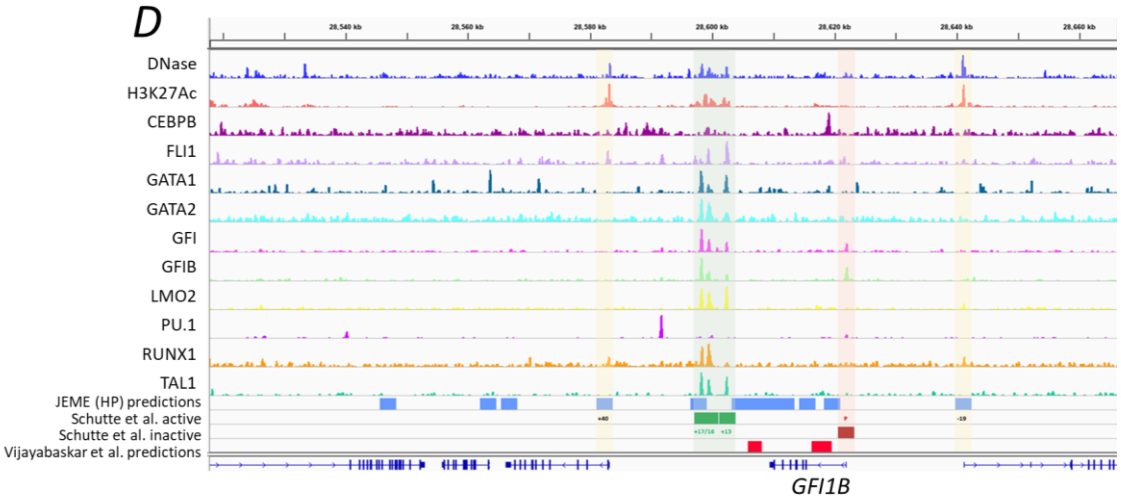
TF binding patterns for Schütte et al. enhancers were reflective of those reported in HPC-7 cells (Table 2.8). TF occupancy by haematopoietic factors, a reliable indicator of haematopoietic activity, was observed in less than half of gene-specific enhancers predicted by JEME for the set of TF genes. Out of the 161 predicted enhancer-TSS interactions, 61 enhancers were bound by at least one of the 10 haematopoietic TFs, 39 by two or more, and 30 by three or more. Furthermore, 39 predicted enhancers were Dnase I hypersensitive and 77 marked by H3K27Ac.

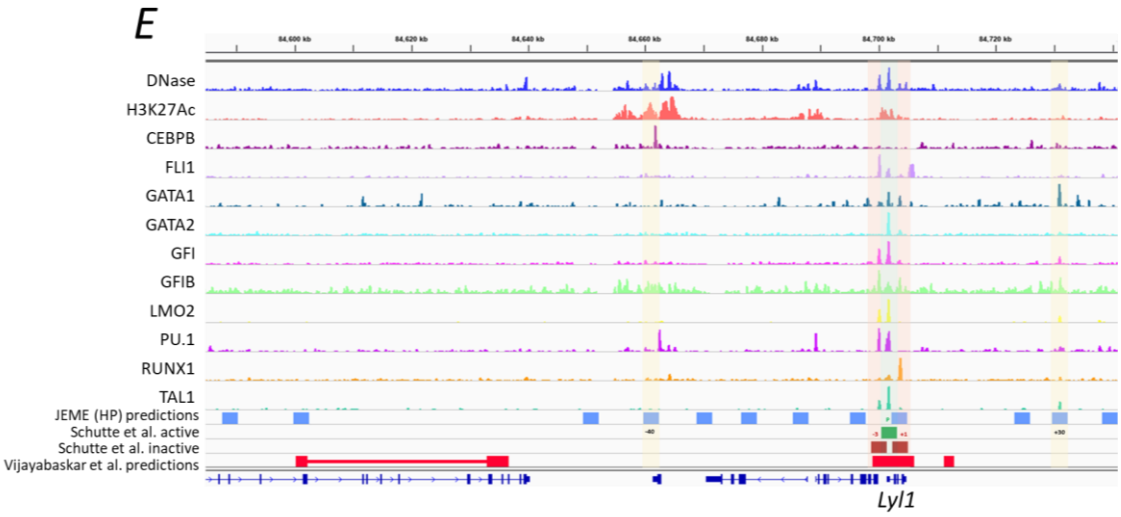
Intersecting Dnase I hypersensitivity, H3K27Ac enrichment and 10 TF binding profiles with JEME-predicted enhancer coordinates resulted in the identification of 19 'potential enhancers' for the haematopoietic TF genes. These regions were defined by reliable indicators of haematopoietic enhancer activity: Dnase I hypersensitivity or H3K27 acetylation, along with the binding of two or more haematopoietic TFs in HP cells. Potential enhancers, meeting these criteria, are listed in Table 2.7, and highlighted in yellow on Figure 2.6. The small proportion of JEME predictions displaying enhancer features indicates that many predictions made by this model may not be regulatory.

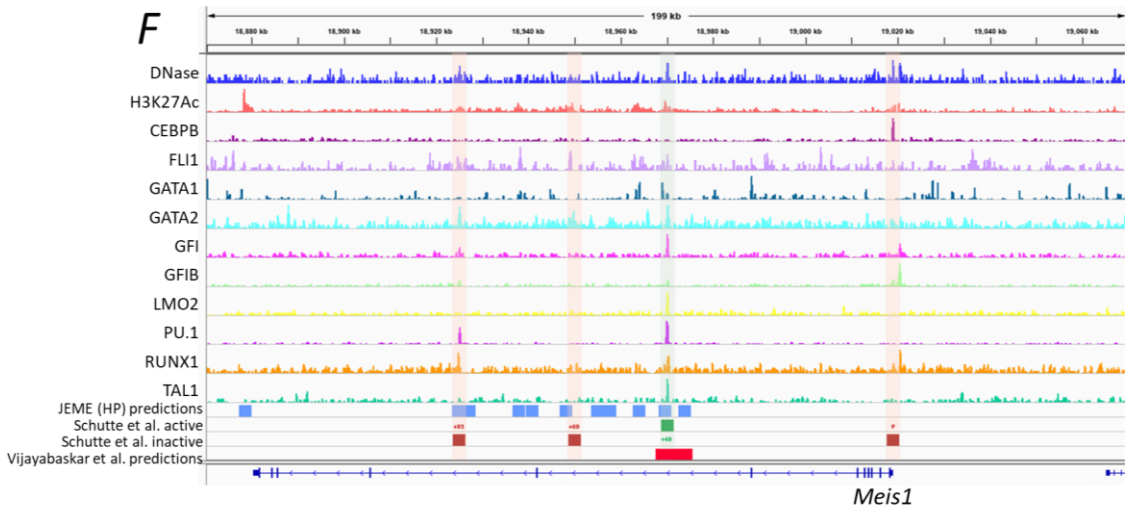

A
DNase
H3K27Ac
CEBPB
ERG
FLI1
GATA1
GATA2
GFI
GFIB
LMO2
PU.1
RUNX1
TAL1
JEME (HP) predictions
Schutte et al. active
Schutte et al. inactive
Vijayabaskar et al. predictions
ERG


B
DNase
H3K27Ac
CEBPB
FLI1
GATA1
GATA2
GFI
GFIB
LMO2
PU.1
RUNX1
TAL1
JEME (HP) predictions
Schutte et al. active
Schutte et al. inactive
Vijayabaskar et al. predictions
FLI1


C
DNase
H3K27Ac
CEBPB
FLI1
GATA1
GATA2
GFI
GFIB
LMO2
PU.1
RUNX1
TAL1
JEME (HP) predictions
Schutte et al. active
Schutte et al. inactive
Vijayabaskar et al. predictions
GATA2


D
DNase
H3K27Ac
CEBPB
FLI1
GATA1
GATA2
GFI
GFIB
LMO2
PU.1
RUNX1
TAL1
JEME (HP) predictions
Schutte et al. active
Schutte et al. inactive
Vijayabaskar et al. predictions
GFI1B


E
DNase
H3K27Ac
CEBPB
FLI1
GATA1
GATA2
GFI
GFIB
LMO2
PU.1
RUNX1
TAL1
JEME (HP) predictions
Schutte et al. active
Schutte et al. inactive
Vijayabaskar et al. predictions
Lyl1


F
199 kb
DNase
H3K27Ac
CEBPB
FLI1
GATA1
GATA2
GFI
GFIB
LMO2
PU.1
RUNX1
TAL1
JEME (HP) predictions
Schutte et al. active
Schutte et al. inactive
Vijayabaskar et al. predictions
Meis1

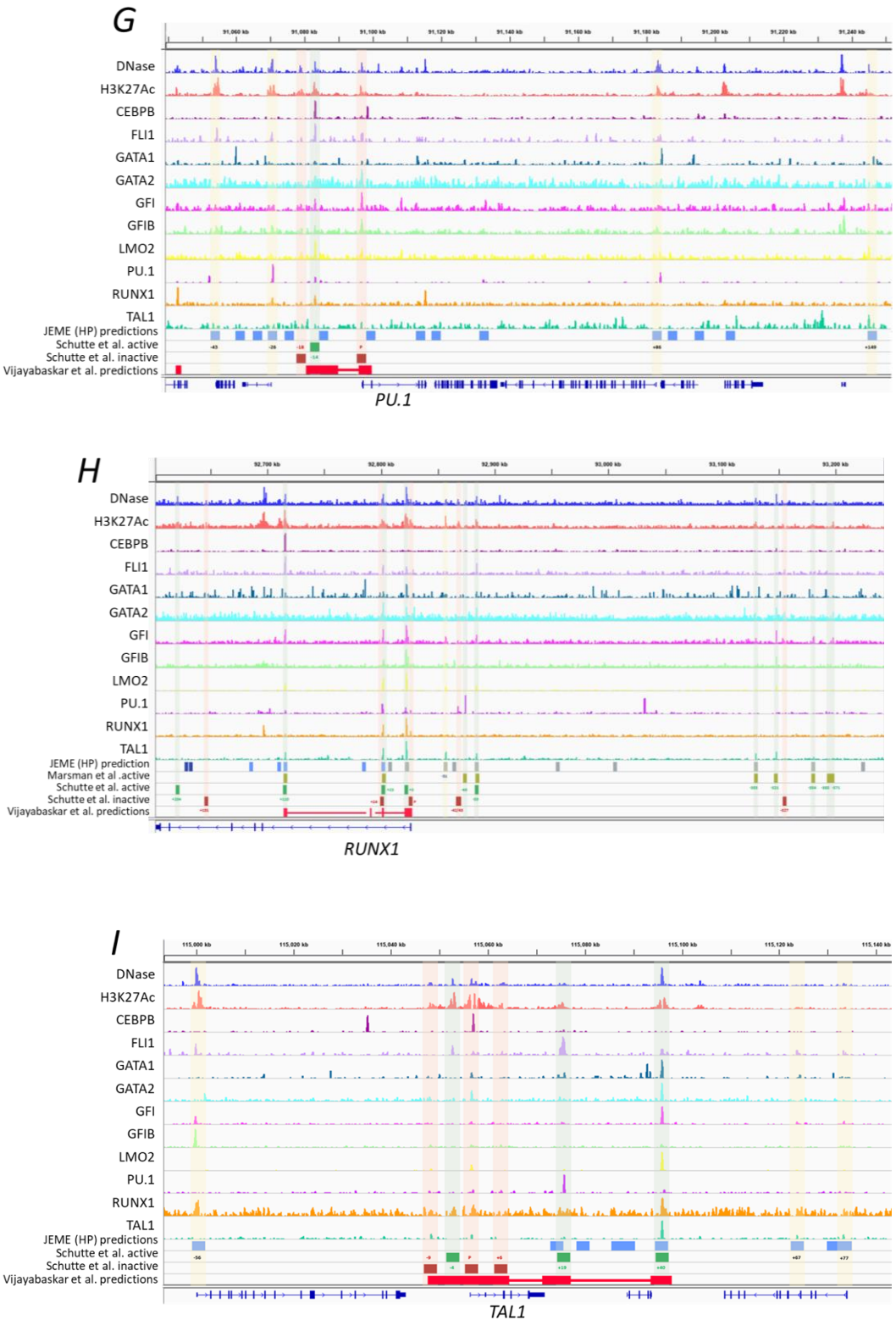


Figure 2.6 JEME and Vijayabaskar predictions compared to validated enhancers and inactive regions for nine haematopoetic TF genes. DNase-seq , H3K27Ac ChIP-seq and haematopoietic TF ChIP-seq signals (for relevant haematopoetic transcription factors) for nine gene loci: *A) Erg (chr16), B) Fli1 (chr9), C) Gata2 (chr6), D) Gfi1b (chr2), E) Lyl1 (chr8), F) Meis1 (chr11), G) Spi1 (chr2), H) Runx1 (chr16), I) Tal1 (chr4)*. Schütte et al. active enhancers/inactive regions and selected coCREs/singleton CREs are shown alongside JEME predictions in HP cells. Schütte et al. active enhancers are highlighted in green, regions absent of enhancer activity are highlighted in red and novel candidates (JEME predictions bound to 2 or more TFs that are Dnase I hypersensitive, H3K27Ac enriched, or both) are highlighted in yellow. In H) JEME predictions coloured grey are for the Runx1 P1 promoter only, dark blue for P2 only, and light blue for both. Marsman et al. active enhancers are also shown for this locus.

Table 2.7 JEME predicted enhancers, untested by Schütte et al., with Dnase I Hypersensitivity (DHS), H3K27Ac and TF binding features (for 2 or more haematopoietic TFs) suggesting enhancer potential.

Gene	Candidate Enhancer	TF binding in HP cells	DHS	H3K27Ac	Notes
Erg	+180	GFI1, GFI1B, LMO2, PU.1	Yes	Yes	coCRE member 80kb downstream of Erg TSS chr16: 95530365
Fli1	+141	GFI1, GFI1B	Yes	Yes	141kb downstream of Fli1 TSS (chr9:32541452)
Fli1	-41	GATA2, GFI1, LMO2, PU.1	Yes	No	41kb upstream of Fli1 TSS (chr9:32541452)
Fli1	-66	GFI1, GFI1B, LMO2, TAL1	Yes	No	66kb upstream of Fli1 TSS (chr9:32541452)
Gata2	-123	GFI1B, LMO2, PU.1	Yes	Yes	123kb upstream of Gata2 TSS (chr6: 88198663)
Gata2	-8	GATA2, GFI1	Yes	No	8kb upstream of Gata2 TSS (chr6: 88198663)
Gata2	+56	CEBPβ, GATA1, LMO2	No	Yes	56kb downstream of Gata2 TSS (chr6: 88198663)
Gfi1b	+40	FLI1, RUNX1	Yes	Yes	40kb downstream of Gfi1b TSS (chr2: 28621982)
Gfi1b	-19	FLI1, LMO2	Yes	Yes	19kb upstream of Gfi1b TSS (chr2: 28621982)
Lyl1	-40	CEBPβ, FLI1, GFI1B, LMO2, PU.1	Yes	Yes	40kb upstream of Lyl1 TSS (chr8:84701468)
Lyl1	+30	GFI1, GFI1B, LMO2, TAL1	Yes	Yes	30kb downstream of Lyl1 TSS (chr8: 84701468)
Spi1	-43	FLI1, GFI1, GFI1B	Yes	Yes	43kb upstream of Spi1 TSS (chr2: 91096677)
Spi1	-26	FLI1, GFI1B, PU.1, RUNX1	Yes	Yes	26kb upstream of Spi1 TSS (chr2: 91096677)
Spi1	+86	GFI1B, PU.1	Yes	Yes	86kb downstream of Spi1 TSS (chr2: 91096677)
Spi1	+149	GFI1, LMO2	No	Yes	149kb downstream of Spi1 TSS (chr2: 91096677)
Runx1	-31	GFI1, LMO2, TAL1	No	Yes	31kb upstream of Runx1 P1 TSS (chr16:92826074)
Tal1	-56	FLI1, GFI1, GFI1B, RUNX1	Yes	Yes	56kb upstream of Tal1 TSS chr4: 115056425
Tal1	-67	FLI1, GFI1, LMO2	No	Yes	67kb upstream of Tal1 TSS chr4: 115056425
Tal1	-77	LMO2, TAL1	Yes	Yes	77kb upstream of Tal1 TSS chr4: 115056425

JEME predicted 14 enhancers interacting with the *Erg* TSS at 95530365 (Figure 2.6A). One of these predictions overlapped the +75 enhancer, and another the +149 inactive element. Of the 12 remaining untested predictions, one region (180kb downstream of the TSS) was identified as a potential candidate enhancer due to Dnase I hypersensitivity, H3K27Ac enrichment and GFI1, GFI1B, LMO2 and PU.1 TF binding. In comparison, Vijayabaskar et al. identified a four-member coCRE overlapping all three Schütte et al. enhancers and no inactive regions. One co-CRE member overlapped the JEME-identified novel candidate at +180.

26 enhancers were paired with the *Fli1* promoter, overlapping both +12 and -15 enhancers and no inactive regions (Figure 2.2B). Three of the 25 untested predictions were identified as having regulatory potential: +141 (DHS, H3K27Ac, GFI1 and GFI1B binding), -41 (DHS, GATA2, GFI1, LMO2 and PU.1 binding) and -66 (DHS, GFI1, GFI1B, LMO2 and TAL1 binding). Vijayabaskar et al.'s singleton CRE overlapped two regions with no enhancer activity (+2 and the gene promoter) but no enhancers.

For *Gata2*, 9 enhancers were predicted (Figure 2.6C). One prediction overlapped two enhancers at +92 and +93, and another three predictions were identified as novel candidates (-123: DHS, H3K27Ac, GFI1B LMO2 and PU.1 binding; -8: DHS, GATA2 and GFI1 binding; +56:H3K27Ac, CEBPβ, GATA1 and LMO2 binding). Two other validated enhancers (-3 and +3) overlapped a singleton CRE. Neither method predicted any Schütte et al. inactive regions.

JEME paired the *Gfi1b* TSS with 18 predicted enhancers, 4 of which overlapped +13, +17 and +18 validated enhancers (Figure 2.6D). One non-enhancer element (the Gfi1b promoter) was predicted, and two untested predictions were selected as candidates (+40: DHS, H3K27Ac, FLI1 and RUNX1 binding and -19: DHS, H3K27Ac, FLI1 and LMO2 binding). coCRE predicted two singleton CREs, both untested by Schütte et al.

16 enhancers were predicted for the *Lyl1* locus. These included the Lyl1 promoter, tested positive by Schütte et al., and the +1 region which tested negative (Figure 2.6E) Two untested candidate enhancers were identified from epigenetic and TF binding profiles: -40 (DHS, H3K27Ac, CEBPβ, FLI1, GFI1B, LMO2 and PU.1 binding) and +30 (DHS, H3K27Ac, GFI1, GFI1B, LMO2 and TAL1 binding). Vijayabaskar et al. predicted

one coCRE and two singleton CREs. One singleton CRE overlapped three Schutte et al. regions: one enhancer (Lyl1 promoter) and two inactive elements (-3 and +1).

JEME made 11 predictions for *Meis1*, including the +48 enhancer and the +93 and +69 inactive regions (Figure 2.6F). No novel candidates were identified from DHS, H3K27Ac and TF binding data. coCRE made one prediction which overlapped the +48 enhancer.

For *Spi1*, encoding the PU.1 TF, JEME predicted 23 enhancers including the -14 validated enhancer and the gene promoter – for which Schütte et al. found no enhancer activity (Figure 2.6G). Of the remaining 21 untested predictions, four were selected as candidates (-43: DHS, H3K27Ac, FLI1, GFI1 and GFI1B binding; -26: DHS, H3K27Ac, FLI1, GFI1B, PU.1 and RUNX1binding; +86: DHS, H3K27Ac, GFI1B and PU.1 binding; +149 (H3K27Ac, GFI1 and LMO2 binding). Vijayabaskar et al. predicted two two-member coCREs; one coCRE overlapped both the gene promoter and -14 enhancer.

The *Runx1* gene has two alternate promoters, P1 and P2, JEME made predictions for both TSS's. JEME predicted 19 unique enhancers for the two TSSs: 11 interacting with P1 only, three interacting with P2 only and five interacting with both (Figure 2.6H). JEME predicted Schütte et al. +110 and +23 enhancers to interact with both promoters, and +3 and -59 enhancers to interact with P1 only. The +24 non-regulatory region was also predicted for both TSSs. Runx1 predictions were also compared with enhancers, identified by genomic alignment to be conserved across mammals, which were validated by Marsman et al. using transgenic zebrafish embryos and 4C in HPC-7 cells. JEME P1 predictions also overlapped with -303 and -354 Marsman et al. enhancers. One untested Runx1 P1 prediction, -31, was selected for regulatory potential due to H3K27Ac enrichment and GFI1, LMO2 and TAL1 binding. Vijayabaskar et al. predicted one singleton CRE and one four-member coCRE. The co-CRE overlapped Schütte et al.'s +110, +23 and +3 enhancers, as well as the +24 non-regulatory region and the P1 promoter (tested negative for enhancer activity).

JEME also considered two TSS's for the *Tal1* locus (Figure 2.6I; ten enhancers were predicted for both. These predictions overlapped +19 and +40 validated enhancers. Three of the 8 untested predictions were chosen as novel candidates: -56 (DHS, H3K27Ac, FLI1, GFI1, GFI1B and RUNX1 binding), -67 (H3K27Ac, FLI1, GFI1 and LMO2

binding) and -77 (LMO2 and TAL1 binding). Vijayabaskar et al.'s three-member coCRE overlapped all Schütte et al. active enhancers (-4, +19 and 40) and regions found absent of enhancer activity (-9, promoter and +6).

2.4.6 64% of Vijayabaskar et al. gene-specific CREs were predicted by JEME

Table 2.4, Table 2.5, and Figure 2.6 show that many gene-specific enhancers were predicted by both Vijayabaskar et al. and JEME in HP cells. However, JEME predicted far more enhancers per gene. To test how well JEME predictions (in any cell type) overlapped Vijayabaskar et al. predictions, the two prediction sets were intersected and a hypergeometric test was performed. Overlaps were performed for unique JEME predictions within 100kb of a TSS considered by the Vijayabaskar et al. method in at least one cell type.

Most selected CRE-gene pairs from Vijayabaskar et al. (64.2%) overlapped a JEME prediction in one or more of the prediction cell types. However due to the higher number of predictions made by JEME, overlaps with Vijayabaskar et al. selected CRE-gene pairs accounted for only 8% of predicted EPIs. Hypergeometric showed the overlap was statistically significant ($p = 1.4x10^{-5}$).

2.4.6 Vijayabaskar et al. identified fewer validated enhancers, but made fewer untested predictions

Finally, to quantitatively evaluate the performance of JEME and Vijayabaskar et al. methods, performance metrics for were calculated using Schutte et al. enhancers and inactive regions. In this analysis only elements within 100kb of one of the nine TF TSSs – or in the gene body itself – were considered. This reduced the number of Schütte et al. active enhancers from 23 to 22 (*Runx1* +204 excluded), inactive regions from 24 to 22 (Erg +149 and *Runx1* +181 excluded) and JEME predictions (in HP cells) from 161 to 106. JEME and Vijayabaskar et al. predictions for these elements were listed in Tables 2.4 and 2.5.

As shown in Table 2.12 and Figure 2.7, JEME predicted more active enhancers (72.7% vs 63.6%) and fewer non-regulatory regions (27.3% vs 45.5%) than the Vijayabaskar et al. method – resulting in greater overall performance (F1: 0.727 vs F1:0.61). However,

it is important to note that these metrics do not account for the larger number of untested JEME predictions. Without further validation data it cannot be certain whether these represent true or false positives.

Table 2.8 Comparative performance of JEME and Vijayabaskar et al. methods across Schütte et al. active enhancers and non-regulatory regions

Schütte et al. 2016		**Vijayabaskar et al.**		**HP JEME Predictions**	
		Predicted (n=27)	Not-Predicted	Predicted (n=106)	Not-Predicted
Enhancer activity	Active (n=23)	14 TP	8 FN	16 TP	6 FN
	Inactive (n=24)	10 FP	12 TN	6 FP	16 TN
Accuracy		0.59		0.73	
Precision		0.58		0.73	
Recall		0.64		0.73	
F1-Score		0.61		0.73	

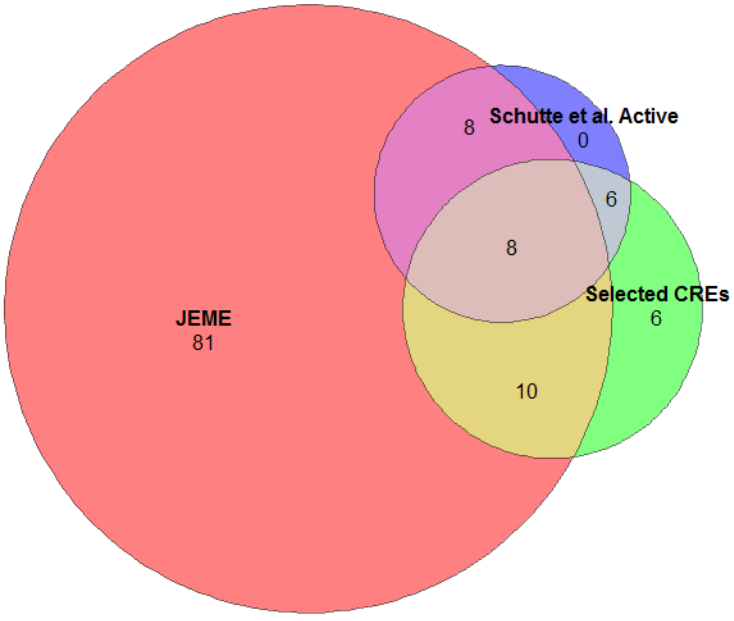


Figure 2.7 Overlaps between predicted and validated gene-specific enhancers. Three-way Euler diagram showing overlap between HP JEME predictions, Vijayabaskar et al. selected CREs and Schütte et al. validated active enhancers.

2.5 Discussion

The aim of this chapter, to compare the predictive ability of JEME and Vijayabaskar et al., proved difficult to achieve. Direct comparison was complicated by the different designs of the two methods and hindered by a lack of gold-standard validation data. It is important to note that any comments on JEME's performance do not necessarily relate to JEME as described by Cao et al., but to JEME as adapted and applied to the haematopoietic dataset.

2.5.1 The Schutte et al. dataset limited validation of model performance

Validation using the Schütte et al. dataset offered a snapshot of comparative performance across a limited number of validated enhancers. At face value, the evaluation of JEME and Vijayabaskar et al. predictions against Schütte et al. active and inactive datasets suggested that JEME performed slightly better; with greater precision and recall (Table 2.8). However, JEME also predicted far more untested regions (Figure 2.7). Without knowing the functionality of these untested regions, model performance could not fairly be evaluated. However, the fraction of JEME predictions which showed both chromatin and TF features of enhancer activity (Table 2.7 and Figure 2.6), and JEME's poorer performance in cross-validation (Table 2.3), suggested that validation on the Schutte et al. dataset (not accounting for untested predictions) overestimated JEME's performance.

Aside from its size, the Schutte et al. dataset was biased by the selection of candidate enhancers for validation, and limited by its use of reporter assays. Firstly, Schütte et al. selected candidate enhancers by their annotation with H3K27Ac, DNase I and TF binding data (Schütte et al., 2016). This means both active enhancers and inactive regions were defined by chromatin features suggestive of enhancer activity.

Secondly, reporter assays involve removing enhancers from their native chromatin environment, which is essential to specifying enhancer activity *in vivo*. This means Schutte et al. 'inactive' regions could still act as enhancers in their correct

chromosomal context, or that validated enhancers could be blocked from gene regulation by chromatin conformation.

Furthermore, the comparison between JEME and Vijayabaskar et al. predictions was complicated by the fact that JEME makes cell-specific predictions, and Vijayabaskar et al. makes predictions across a dataset. Since Schütte et al.'s validation strategy involved the use of HP cell-line HPC-7, and embryonic sites associated with HP populations, it naturally followed to compare with JEME predictions in HP cells. Unlike JEME, Vijayabaskar et al. combined feature data across all cell types – including the terminally differentiated macrophage stage – to make more general predictions relevant to haematopoietic differentiation. It is possible that the 'false positive' enhancers predicted by Vijayabaskar et al. were active in other stages of the murine system. Specifically, the Vijayabaskar et al. method is capable of predicting cis interactions specific to embryonic stem cells or macrophages, which have distinct regulatory networks from other haematopoietic cell types (Goode et al., 2016).

2.5.2 JEME predicted many enhancer-promoter interactions

The retrained JEME model predicted many more enhancer-promoter interactions than the Vijayabaskar et al. method, and other cell-specific EPI predictors like IM-PET, RIPPLE and JEME applied to FANTOM5 data (Cao et al., 2017; He et al., 2014; Roy et al., 2015). Chromatin and TF data shown in Figure 2.6 suggested that many JEME predictions could be false positives, however this could not be established without additional validation data. The retrained JEME model may have predicted many non-functional enhancer-promoter interactions. This could have been influenced by several factors.

Firstly, JEME could have predicted more EPIs because it considered more candidate enhancers. Cell types with more candidate enhancer-promoter interactions (ChromHMM predictions matched to genes within 1Mb) had more predicted EPIs, more predicted enhancers per TSS and a greater median distance between pairs (Table 2.2). It is possible that JEME predicted more EPIs since it considered more active enhancers, which could be influenced to the coarse-grained 4-state ChromHMM model employed by Goode et al.

Secondly, it is possible that JEME's default threshold of 0.35, optimised to Cao et al.'s data, should be higher. Increasing the threshold would reduce the number of overall predictions yet would also affect JEME's predictive ability. Threshold optimisation would require evaluation against genome-wide data such as a Hi-C dataset from another cell type, or ESC data during cross-validation. However, as shown in Table 2.6, raising the threshold did not improve performance across Schütte et al. enhancers and inactive regions, suggesting that the 0.35 threshold was appropriate.

Finally, the ratio of positive to negative pairs used in model training would have influenced the ratio of positive and negative predictions. Cao et al. determined that the ratio of interacting to non-interacting pairs should be ~0.15, using Equation 2.3. The rationale behind this ratio is not clear; it is neither a class-balanced training set nor reflective of the ratio of interacting to non-interacting pairs detected by chromatin interaction data. The ratio of interacting Hi-C pairs to non-interacting potential pairs in the training dataset was much lower at 0.01 (Table 2.2). Retraining JEME using other class ratios could further investigate the model's performance.

2.5.3 Low resolution Hi-C data may have hindered the performance of the retrained JEME model

The performance metrics measured during cross-validation of the retrained JEME model were lower than those reported for the original JEME model applied to Roadmap data (Cao et al., 2017). This suggests that alterations to the JEME model (necessary for its application to the murine haematopoietic dataset) could have impaired performance. One factor may have been the substitution of H3K4me1 (more predictive of enhancers) with H3K4me3 (more predictive of promoters). In addition, the poorer quality of training data likely hindered performance.

In their paper, Cao et al. trained JEME on RNA pol II ChIA-PET interactions in K562 human cells. ChIA-PET identifies long range interactions, mediated by a protein of interest, on the same scale of ChIP-seq peaks (hundreds of bases). In contrast, this chapter saw JEME trained on low-resolution Hi-C data from murine ESCs, due to a lack of high-resolution datasets appropriate for use with the haematopoietic system. These Hi-C contacts would be far less likely to reflect cis-regulatory interactions than ChIA-

PET loops. Since the Hi-C contacts used in training were not binned (as to identify regulatory interactions on the scale of enhancers and promoters) it is likely that the training dataset contained considerable noise. This would have made it harder for JEME to learn the features associated with enhancer-promoter interactions. Whilst this is a major flaw in the analysis, it highlights the inapplicability of methods like JEME to non-human systems where high-resolution 3D genomics data is unavailable.

2.5 Conclusion

In conclusion, limitations of the validation dataset made it hard to fully compare the performance of Vijayabaskar et al. to JEME. Whilst JEME predicted far more enhancer-promoter interactions, there was still significant overlap with gene-specific cis-regulatory elements from the Vijayabaskar et al. method. The reported performance advantage for JEME was not conclusive, due to limitations of the small validation dataset. Without further validation, using high-throughput datasets such as eQTLs and high-resolution chromatin interactions, it was difficult to fairly assess performance.

In the next chapter of this thesis, a new method, cisREAD, is designed. cisREAD builds on the community detection and LASSO regression mechanism from Vijayabaskar et al. and is applied to human B lymphocytes. Chapter 5 will revisit the comparison with JEME, through benchmarking cisREAD against other predictive methods using high-throughput validation datasets.

Chapter 3. Integrating chromatin accessibility and gene expression with cisREAD identifies transcription factor led gene regulation in B cell differentiation

3.1 Introduction

Chapter 2 described a comparison between the Vijayabaskar et al. method to predict gene-specific cis-regulatory elements, and the JEME method to predict enhancer-promoter interactions. The Vijayabaskar et al. method used community detection and LASSO regression to integrate chromatin accessibility (DNase-seq), histone modification (ChIP-seq), transcription factor binding (ChIP-seq) and gene expression (RNA-seq) datasets across murine haematopoietic differentiation. This approach was designed to prioritise cis-regulatory elements important for transcriptional control of differentiation associated genes, however it required an abundance of input datasets, which would limit its application.

In the work described in chapter 3, we were presented with chromatin accessibility (ATAC-seq) and gene expression (RNA-seq) datasets across an *in vitro* system of human B cell differentiation. This raised the challenge of inferring transcription factor binding in the absence of direct protein-DNA interactions from ChIP-seq. This chapter describes the exploration of computational methods to detect transcription factor binding from chromatin accessibility data, for integration into a workflow to predict gene-specific cis-Regulatory Elements Across Differentiation: cisREAD.

3.1.1 B Cell Differentiation

cisREAD was developed to predict transcription factor-led gene regulation throughout the system of B cell differentiation. The maturation of B cells to plasma cells is a critical process in the adaptive immune response, defending the host against pathogens through the production of antigen-specific antibodies (Figure 3.1). Defects in the differentiation process can lead to immunodeficiency, autoimmune disease, or B cell cancers (Lebien and Tedder, 2008). Therefore, it is important to understand how transcriptional control of gene expression coordinates the differentiation of mature B cells.

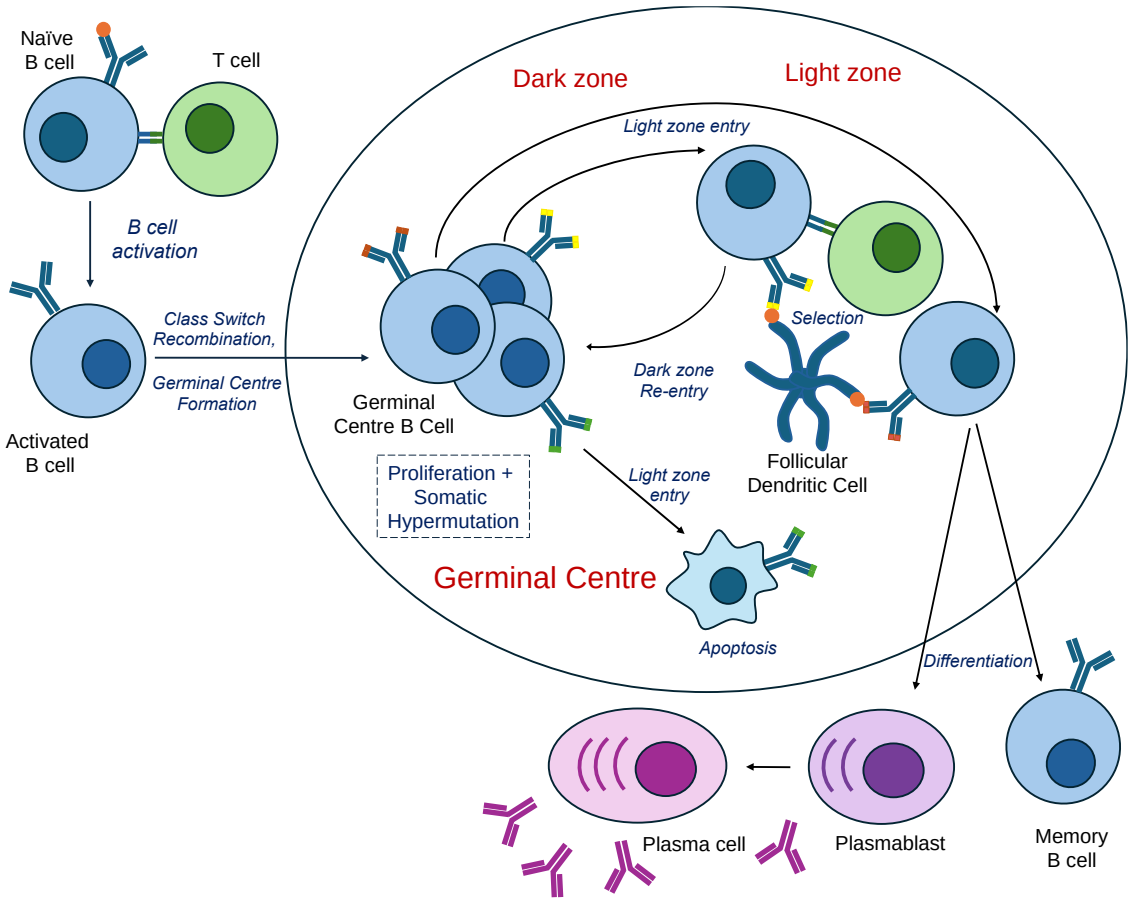


Figure 3.1 B cell activation, germinal centre reaction and plasma cell differentiation. The activation of naïve B cells is initiated when a B cell recognises an antigen on its B cell receptor. Follicular helper T cells can interact with the B cell to further stimulate activation. Activated B cells undergo class switch recombination, to switch their antibody type, and form germinal centres. In the dark zone of germinal centre B cells proliferate and further mutate their antibody genes through somatic hypermutation, before entering the light zone. In the light zone, B cells compete with resources. This functions to select for B cells with high-affinity antibody mutations. In the light zone follicular dendritic cells present antigens to high-affinity B cells, which interact with T cells and receive survival signals. Low-affinity B cells do not receive this T cell 'help' and undergo apoptosis. Surviving B cells either re-enter the dark zone for further mutation, or commit to memory B or plasma cell fate and exit the germinal centre.

3.1.1.1 Activation

The differentiation of mature B cells to plasma cells is initiated by B cell activation, in which an antigen is recognised by the B Cell Receptor (BCR) on the B cell surface (Figure 3.1). Upon BCR engagement by a protein antigen, the B cell receives 'activation' signals and the cognate antigen is internalised, digested, and presented as peptides on the major histocompatibility II (MHC II) complex. T follicular helper cells, stimulated by the same antigen, can recognise these peptides and engage the B cell's MHC II through its T Cell Receptor (TCR). T:B interactions provide further activation signals. Stimulation of the CD40 receptor, bound to the B cell membrane, by the CD40 ligand (CD40L), bound to the T cell membrane, triggers a signalling cascade promoting resting B cells to

enter the cell cycle. Secretion of additional cytokines, including interleukins (ILs) and interferons (IFNs), by the T cell provides the B cell with further stimuli promoting proliferation and differentiation (Cyster and Allen, 2019). This process of T-cell dependent activation can be stimulated *in vitro* through provision of antigens/antibodies, CD40L and cytokines (Cocco et al., 2012).

3.1.1.2. Proliferation, mutation, and germinal centre reaction

Following activation, B cells proliferate and mutate their immunoglobulin (Ig) genes, in preparation to secrete large quantities of antigen-specific antibodies. During the process of Class Switch Recombination (CSR) B cells switch their antibody 'constant' region from one isotype to another (e.g. IgM to IgG), allowing the antibody to interact with different effectors (Roco et al., 2019).

B cells activated in secondary lymphoid organs (lymph node, spleen, or tonsils) then congregate to form transient structures called germinal centres (GCs); which become the primary site of proliferation and mutation (Figure 3.1). Outside of the germinal centre, a small number of B cells proliferate and mutate to produce an initial wave of short-lived antibodies to control the infection (Akkaya et al., 2020).

Germinal centres are divided into a 'dark zone' of rapidly dividing B cells (centroblasts) which undergo 'somatic hypermutation' (SHM) of 'variable' regions of antibody genes, and a light zone of non-dividing centrocytes. Somatic hypermutation is performed by the Activation Induced Deaminase (AID) enzyme, which deaminates cytosine into uracil. Substitutions are then induced during error-prone DNA repair of the uracil lesions. Somatic hypermutation of a proliferating centroblasts population produces a diverse range of high-affinity and low-affinity antibodies, presented on B cells (Pilzecker and Jacobs, 2019). As B cells move to the light zone, they face intense competition for resources. In the light zone follicular dendritic cells (FDCs) present antigens to B cells. Centrocytes with high affinity for the antigen are favoured for 'positive selection' by T cells. Only high affinity B cells are able to conjugate fully with T cells and receive survival signals; low-affinity B cells fail to compete and undergo apoptosis. B cells cycle through germinal centre light and dark zones, undergoing

successive rounds of hypermutation and selection to increase antibody affinity (Young and Brink, 2021).

3.1.1.3 Commitment to plasma cell fate

B cells continue to cycle through the germinal centre, before receiving signals to exit either as memory B cells, capable of reactivation by the same antigen (followed by rapid plasma cell differentiation), or long-lived plasma cells, capable of secreting large quantities of antigen-specific antibodies (Figure 3.1). There is evidence that B cells with higher-affinity BCRs are more likely to undergo plasma cell differentiation, whereas lower-affinity cells are steered towards memory B cell differentiation (Akkaya et al., 2020).

Plasma cell vs memory B identity is programmed through the differential induction of key transcription factors contingent on the strength of CD40 and BCR engagement (in turn determined by antigen-affinity) (Akkaya et al., 2020).

Along the plasma cell trajectory, B cells pass through a transitory plasmablast (PB) stage, characterised by proliferation, antibody production and migration to the bone marrow survival niche, which is necessary to sustain long-lived plasma cells. Long-lived plasma cell generation can be driven *in vitro* through provision of niche signals like APRIL (Stephenson et al., 2022).

3.1.1.4 Transcriptional regulation

This process of B cell differentiation is controlled through a dynamic gene regulatory network, shaped by epigenomic remodelling and transcriptomic reprogramming. These changes are stimulated both intrinsically and extrinsically, through lineage-specifying and signal-inducible transcription factors. As B cells transition through proliferative activated B cell (ABC) and plasmablast states, regulatory control shifts from a transcriptional circuit upholding B cell identity (including BACH2 and PAX5) to a mutually antagonistic network promoting plasma cell fate (including IRF4, PRDM1 and XBP1) (Tellier and Nutt, 2019; Trezise and Nutt, 2021). Changes in gene expression are also driven through induction of transcription factors by external signals, including NF-kB downstream of BCR, CD40 and APRIL, as well as STAT3 downstream of IL-21 (Berglund et al., 2013; Cornelis et al., 2020; Luo et al., 2018). Whilst many B lineage

transcriptional regulators are well characterised, there is an incomplete understanding of how transcription factors reprogram downstream regulatory networks in response to activation and differentiation stimuli.

3.1.1.5 Studying transcriptional regulation of *in vitro* B cell differentiation

Integration of epigenomic and transcriptomic data has allowed researchers to study the genetic regulation of B cell maturation across a number of organisms, primary cells and differentiation systems (Joyner et al., 2022; Moroney et al., 2020; Price et al., 2021; Scharer et al., 2018). This chapter describes the development of a method to integrate ATAC-seq and RNA-seq, and its application to samples taken across human *in vitro* B cell differentiation, where activation is stimulated through CD40L and antibodies, and long-lived plasma cell differentiation is driven through APRIL (Cocco et al., 2012; Stephenson et al., 2022).

A challenge in adapting the method to the *in vitro* B cell dataset is the lack of transcription factor binding data. In the absence of ChIP-seq, transcription factor occupancy can be inferred on a high-throughput scale using methods which identify transcription factor binding sites in accessible chromatin.

3.1.2 Identifying transcription factor binding from ATAC-seq data

3.1.2.1 Transcription factor binding motifs

Cis-regulatory element activity is contingent on both its chromatin environment and the temporal, combinatorial recruitment of sequence-specific transcription factors. Transcription factors bind conserved, degenerate, consensus sequences, specified by structural protein motifs in the TF's DNA binding domain (Spitz and Furlong, 2012). TF sequence specificity is determined by protein motifs which recognise the chemical signature of DNA bases or the sequence-dependent shape of DNA. For example, many mammalian TFs feature tandem arrays of C2H2 zinc finger protein structures in their DNA binding domain. Each zinc finger binds the major groove of DNA at close intervals, making sequence-specific contacts (Lambert et al. 2018).

The short (~6-12bp) DNA sequences to which TFs bind are termed transcription factor

binding motifs. TF motifs are obtained by aligning the experimentally determined binding sites of TFs and summarizing the frequencies at which each nucleotide occurs. These motifs can be represented numerically by position weight matrices (PWMs), where the frequency of each nucleotide is scored at each position. Motifs can also be represented visually by sequence logos, where the size of each nucleotide corresponds with its frequency (Schneider and Stephens, 1990; Stormo, 2000). Multiple adjacent copies of the same or different motifs, can occur in homotypic or heterotypic clusters for robust, combinatorial gene regulation (Erceg et al., 2014).

Transcription factor consensus sequences occur millions of times throughout the genome. Of all the genomic sequences matching a PWM, only a small fraction will be biologically functional in a given cell. This problem has been termed the 'futility theorem', and exists because inaccessible chromatin structures or binding partner requirements often prohibit motif occupancy (Lambert et al., 2018; Wasserman and Sandelin, 2004).

The chromatin state of a motif determines its availability to transcription factor recognition. Most transcription factors are incapable of binding nucleosomal DNA, and require that the motif be cleared of nucleosomes. In exception, a small number of 'pioneer transcription factors' (e.g., FOXA and GATA TFs) have the unique capacity to bind nucleosomal chromatin and open the regulatory element. This allows for recruitment of other transcription factors, co-factors and chromatin remodelling complexes; establishing cell-specific regulatory programmes during differentiation (Zaret, 2020).

In order to identify motifs capable of transcription factor recognition, PWM scans can be limited to sequences with cell-specific accessibility (determined by DNase-seq or ATAC-seq), activating histone modifications (e.g., H3K27Ac ChIP-seq) or transcription factor occupancy (also with ChIP-seq). Given a set of ATAC/DNase/ChIP-seq peaks, potential regulatory sequences can be scanned for PWMs obtained through one or more databases. JASPAR, TRANSFAC and HOCOMOCO all host a wealth of curated PWMs determined experimentally using techniques including protein binding microarrays, ChIP-seq and SELEX (Fornes et al., 2020; Kulakovskiy et al., 2018;

Wingender, 2008). Whilst valuable resources, known motif databases are incomplete and include large numbers of redundant motifs. Motif redundancy results from related transcription factors, with structurally similar DNA-binding domains, which occupy near-identical consensus sequences. This can make it difficult to determine which transcription factor(s) occupy a given motif.

3.1.2.2 *De novo* motif discovery

The technique of *de novo* motif discovery circumvents the caveats of PWM scanning by using pattern discovery methods to identify motifs without prior knowledge of transcription factor binding sites. This class of methods includes programmes which use probabilistic (e.g., MEME), or word-based (e.g., STREME, HOMER) algorithms to find over-represented DNA *k*-mers in a set of primary sequences compared to a background distribution (Bailey and Elkan, 1994; Bailey, 2021; Heinz et al., 2010). *De novo* motif discovery methods typically perform clustering of 'discovered' motifs to avoid redundancy.

Both MEME and STREME have been optimised for use with high-throughput datasets, like ATAC-seq or ChIP-seq peaks, through the MEME-ChIP wrapper (Machanick and Bailey, 2011). Here *de novo* discovery is performed using both MEME and STREME, and resultant *de novo* motifs are matched to known PWMs using the TOMTOM programme. Another popular software choice is HOMER which, similar to STREME, counts the number times a 'word' of length *k* occurs in both primary and background sequences, before calculating its over-representation with a statistical or binomial test. HOMER performs both *de novo* discovery and known PWM matching in one step (Heinz et al., 2010)

The careful selection of appropriate background sequences is imperative to successful *de novo* motif discovery: it must be similar in nucleotide composition and sequence length to the primary set (Simcha et al., 2012). Both MEME-ChIP and HOMER can generate their own set of bias-corrected background sequences, or the user can supply their own sequences, which also undergo bias-correction. This option facilitates differential enrichment analyses, to discover motifs are over-represented in one set of

chromatin regions compared to another. This is useful for identifying cell or condition specific differences in transcription factor binding.

MEME and STREME construct artificial random sequences using an n^{th}-order Markov model. Here DNA sequences are modelled as Markov chains of order *n*, where the probability of a nucleotide occurring in the sequence depends on the preceding *n* nucleotides. Users can select the order of the model to adjust for bias. When *n*=1 each nucleotide is dependent on one preceding nucleotide, which adjusts for dimer biases like CpG sites. Background sequences are constructed by the Markov model either by shuffling the primary sequences (default) or a user-supplied set of control sequences (Bailey and Elkan, 1994). HOMER generates its default background set by randomly selecting DNA sequences from the genome (not included in the primary set) and matching the background GC content distribution to the primary GC-content distribution. HOMER also performs auto-normalisation to remove imbalances in DNA *k*-mers where *k*=1,2 and 3. User-specified background sequences also undergo bias correction and normalisation with HOMER (Heinz et al., 2010)

3.1.2.3 Predicting binding site occupancy

Whilst *de novo* motif discovery can identify enriched transcription factor binding sites in a set of regions, it provides no indication that each motif instance is bound. A number of methods have been designed to identify bound motifs from DNA sequence and chromatin accessibility data. One popular choice of method is computational 'footprinting'; where accessible chromatin regions identified by DNase-seq or ATAC-seq is probed for dips in read coverage where enzymatic cleavage is blocked by TF occupancy. Computational tools like HINT and Wellington have been developed to detect these footprints in DNase-seq peaks, prior to matching the footprint sequence to known TF binding-sites (Gusmao et al., 2014; Piper et al., 2013). The HINT framework was later adapted to account for Tn5 cutting-site bias in ATAC-seq and packaged as the tool HINT-ATAC (Li et al., 2019). HINT-ATAC works by generating cleavage signals from sequencing libraries following fragment size filtering, normalisation, and correction of cleavage bias. Then a hidden Markov model (HMM), trained on transcription factor ChIP-seq in a semi-supervised manner, segments the signal to locate footprints. HINT-ATAC outperforms footprinting methods developed

for DNase-seq and achieves good results with a moderate number of reads per sample (~50 million).

Whilst many studies successfully employ ATAC-seq footprinting to interrogate TF binding dynamics (R. Li et al., 2018; Scharer et al., 2018; Vierstra et al., 2020) it has been noted that many transcription factors do not leave strong footprints, particularly those with short DNA residency times (Baek et al., 2017; D'Oliveira Albanus et al., 2021; Sung et al., 2014). This is particularly a concern for transiently binding TFs such as STAT3 and NF-kB. To address the limitations of footprinting, alternative methods have been developed. One such method is BMO, the "bee model of occupancy" which predicts TF binding through negative binomial models of chromatin accessibility and motif co-occurrence. BMO likens TFs to "Brownian bees", which are more likely to visit flowers (motifs) which are accessible and plentiful (D'Oliveira Albanus et al., 2021).

3.2 Aims and Objectives

In this chapter, we aimed to develop a method capable of identifying gene-specific cis-regulatory elements, from ATAC-seq and RNA-seq data, with the ultimate goal of identifying transcription-factor led regulation throughout B cell differentiation. To meet this aim the following objectives were set out:

1) to expand the LASSO-based method in chapter 2 to predict transcriptional regulation from chromatin accessibility and gene expression;
2) to investigate tools to identify transcription factor binding site enrichment and occupancy from ATAC-seq data; and
3) to apply the resultant method to ATAC-seq and RNA-seq datasets to predict gene regulation during B cell differentiation.

These aims will be evaluated in the context of method performance, prior to biological interpretation of predicted regulation in chapter 4.

3.3 Methods

This section will explain the rationale, development, and application of cisREAD: an integrative 'omics approach to identify gene-specific cis-Regulatory Elements Across Differentiation. It begins with an introduction to the B cell differentiation dataset,

followed by an outline of the cisREAD method, highlighting adaptations made to the LASSO-based method in chapter 2. Finally, the application of cisREAD to B cell differentiation will be described, alongside evaluation of included TF binding analysis tools.

3.3.1 Dataset

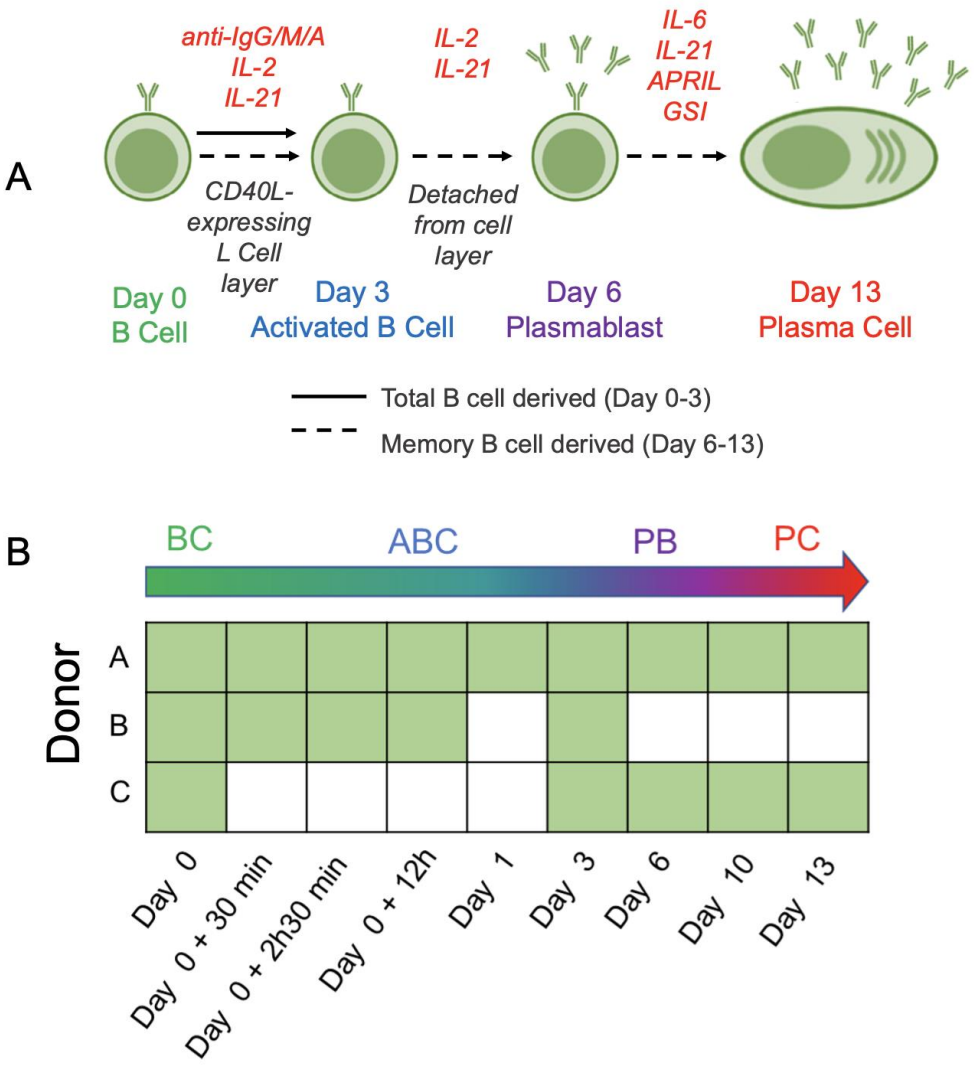


***Figure 3.2 In vitro* system of human B cell activation plasma cell differentiation.** A) B cells are first extracted from the peripheral blood of donors and are cultured and activated in the presence of the CD40 ligand, anti-Ig and IL-2 and IL-21 cytokines. Activated B cells then undergo plasma cell differentiation, driven through removal of CD40L and anti-IgG/M/A and are pushed towards a mature plasma cell phenotype through removal of IL-2 and provision of APRIL, γ-Secretase inhibitor (GSI) and IL-6. B) Donor (A, B, C) matched ATAC-seq and RNA-seq datasets taken at nine time-points across *in vitro* plasma cell differentiation, from 1-3 biological replicates, yielding a total of 19 samples. BC, B cell; ABC, activated B cell; PB, plasmablast, PC; plasma cell. Samples with matched ATAC-seq and RNA-seq data are shown in green. Due to issues with population expansion, day0-3 samples were derived from total B cells, and day6-13 samples were derived from isolated memory B cells.

The *in vitro* B cell differentiation dataset derives from a system of human B cell activation and plasma cell differentiation produced by the Tooze group. The dataset was generated by Dr Amel Saadi, as part of a project led by Professor Reuben Tooze and Dr Gina Doody. B cell populations were isolated from the peripheral blood of 3 donors using MACS (Magnetic activated cell sorting) separation (total B-cells for time-points to day 3, and memory B cell-enriched for time points after day 6 and 13) and long-lived plasma cells were generated *in vitro* following published protocols (Figure 3.2A) as previously described (Cocco et al., 2012; Stephenson et al., 2022). Briefly, B cells were exposed to activating conditions including F(ab')2 anti-IgG/A/M, IL-2, IL-21 and irradiated CD40L L-cells at day 0 to stimulate B cell activation. Cells were sampled by careful removal from the stromal cell layer at indicated time points. Cells were transferred at day 3 to conditions with cytokines IL-2 and IL-21 alone, and plasmablasts at day 6 were driven towards a long-lived plasma cell phenotype through further cytokine signalling (IL-21, IL-6, and APRIL) and the addition of γ-Secretase Inhibitor (GSI). ATAC-seq and RNA-seq experiments, measuring chromatin accessibility and RNA levels, were performed at 9 time-points across the *in vitro* differentiation process, yielding a dataset of 19 samples (Figure 3.2B).

3.3.2 cisREAD methodology

cisREAD takes the core mechanisms of community detection and LASSO regression, which were validated in chapter 2, whilst adapting the method to derive transcription factor binding from accessibility data. This forms a four-step workflow (outlined schematically in Figure 3.3) of differential accessibility/expression, transcription factor analysis, community detection and LASSO regression. cisREAD is designed to work on sample-matched chromatin accessibility and gene expression datasets, taken across a system of cellular differentiation. The first two steps use common bioinformatics tools to extract accessibility, TF binding and expression features from sequencing files. The last two steps, which construct gene-specific models to predict regulation, are implementable in the cisREAD R package at https://www.github.com/AmberEmmett/cisREAD .

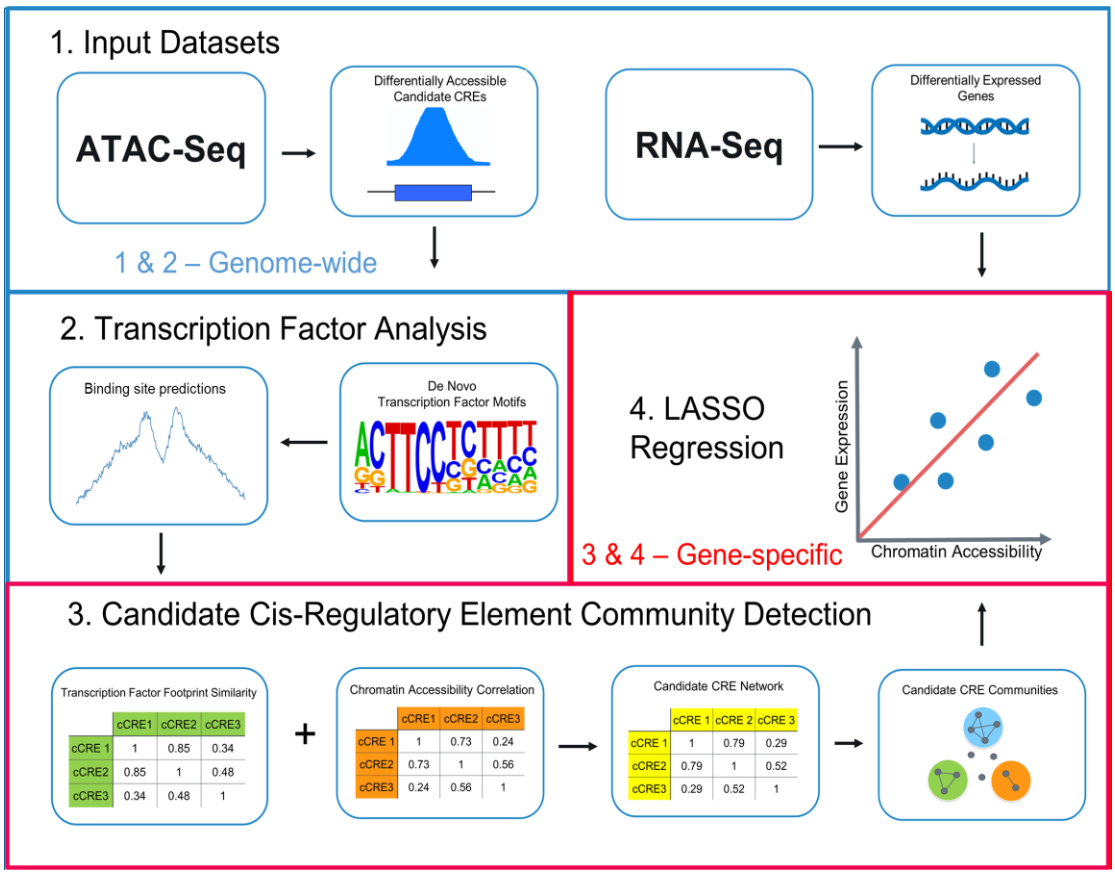


Figure 3.3 Overview of cisREAD methodology. cisREAD is designed to identify gene-specific cis-regulatory element communities across cellular differentiation from ATAC-seq and RNA-seq datasets**.** Step 1: candidate CREs are first identified as differentially accessible ATAC-seq peaks and differentially expressed genes are identified for gene-specific modelling. Step 2: enriched transcription factor motifs are curated through de novo motif discovery across all genome-wide candidate CREs and matched to transcription factor footprints called in the candidate CREs which are accessible in each differentiation-stage. Step 3: The sample-specific chromatin accessibility of each candidate CRE is characterised, alongside its differentiation-stage specific transcription factor occupancy events and all candidate CREs within 100kb of a target genes TSS, with 3 or more occupancy events, are considered gene-specific candidates. The chromatin correlation and transcription factor similarity of each possible candidate CRE pair is calculated and multiplied to produce integrated similarity scores, used to construct a candidate CRE network. Candidate CREs with similar chromatin accessibilities and transcription factor occupancy events are then grouped together through infomap community detection. Step 4: the chromatin accessibility of each candidate CRE (or mean accessibility of each coCRE) is considered to predict gene expression across all samples in gene-specific LASSO models. LASSO models select candidate (co)CREs whose chromatin accessibility best predicts gene expression and rejects any others. LASSO models are constructed for all differentially expressed genes, and significant predictors of gene expression are considered high-confidence gene-specific predicted CREs.

3.3.2.1 Step 1. Differential Accessibility and Expression

To predict gene-specific cis-Regulatory Elements Across Differentiation, we first identify the genes and regulatory elements which differ across the lineage. To do this genes and accessible chromatin region are tested for differential expression and differential accessibility. Likelihood ratio tests (comparing model fit with and without the cell-stage variable) are used to identify differentially accessible regions (DARs) and differentially expressed genes (DEGs) whose activity varies by differentiation-stage.

3.3.2.2 Step 2. Transcription factor binding site analysis

In the absence of differentiation stage-specific transcription factor binding data (e.g., from ChIP-seq) binding sites are derived from chromatin accessibility data. In order to define a set of linage-specific transcription factors to take forward, a *de novo* motif analysis is first performed using tools like HOMER and MEME-ChIP. This approach identifies binding sites enriched in DARs compared to constitutively-accessible chromatin, indicating that the occupying transcription factor exhibits dynamic binding during differentiation. Importantly, *de novo* motif discovery requires no prior knowledge relevant transcription factors, allowing for the data-driven selection of TFs important to the lineage.

Transcription factor binding sites are then identified through scanning DARs with *de novo* PWMs, and predicting motif occupancy from chromatin accessibility data using either footprint-based or footprint-independent methods.

3.3.2.3 Step 3. Community detection

After characterising the chromatin accessibility and transcription factor occupancy of all differentially accessible regions, candidate CREs are linked to target genes. Prior to selecting or rejecting candidate CREs to regulate transcription, through gene-specific LASSO regression models, a community detection step is performed. Community detection groups together gene-specific candidate CREs with similar patterns of chromatin accessibility and transcription factor occupancy. This step serves a dual purpose: to identify TF-bound regions which may co-operatively regulate gene expression, and to alleviate multicollinearity in LASSO regression models (by grouping together correlated predictors for singular input into the model). The community detection step is performed as described below.

All DARs are described by a $\log_2$ normalised ATAC-seq count matrix and a binary transcription factor occupancy matrix – indicating the presence (1) or absence (0) of *de novo* motif occupancy in each differentiation-stage. DARs with fewer than 3 TF occupancy events are discarded, and remaining DARs are assigned to nearby DEGs to identify gene-specific candidate CREs. DARs are matched to DEGs if their transcription start site (TSS) is within 100kb of the DAR's midpoint. This distance is chosen as most

mammalian regulatory elements have been shown to operate within 100kb of their target gene (Fulco et al., 2019). Whilst the distance threshold can be increased in the cisREAD R package, this would also increase the number of predictors (and thus multicollinearity) in LASSO regression.

For all gene-specific candidate CREs, a chromatin accessibility correlation matrix and TF footprint similarity matrix are produced. The chromatin accessibility correlation matrix gives the Pearson correlation coefficient between each pair of gene-specific candidate CREs, and the transcription factor occupancy matrix gives the Dice similarity coefficient between each pair. The dice coefficient is calculated as twice the number of binding events common to both CREs divided by the sum of the number of binding events for each CRE (Dice, 1945). These two matrices are then multiplied to give pairwise integrated similarity scores. These similarity scores range from 0 to 1 and indicate the extent to which candidate CREs are accessible in similar differentiation-stages, and similarly occupied by transcription factors.

To group together similar candidate CREs, a weighted undirected graph is produced from the integrated similarity matrix where nodes are gene-specific candidate CREs and edges are their similarity scores. Integrated similarity scores are used for edge weighting due to a theoretical assumption that co-accessibility and TF co-binding are equally important for co-regulation. Edges were only drawn between nodes if the integrated similarity score exceeded 0.3. The threshold of 0.3 was chosen empirically, by varying the threshold for 10 test genes, as it produced communities with few intercommunity connections. This edge threshold can be adjusted in the R package: smaller values will result in larger, looser communities and larger values in smaller, tighter connections. Communities of highly-connected nodes, representing similarly accessible and occupied candidate CREs, are detected from this network using the infomap algorithm, implemented in the igraph R package. Infomap community detection is an optimal, information theory approach using random walks, which performs well on small networks with few inter-community connections (Rosvall and Bergstrom, 2008; Yang et al., 2016). After community detection, the chromatin accessibility of each coCRE is then given as the mean of its constituent members.

3.3.2.4 Step 4. LASSO Regression

To select which candidate coCREs or individual CREs regulate each DEG, gene-specific LASSO regression models are constructed. Here LASSO models are used for variable selection; to select candidate (co)CREs whose chromatin accessibility is predictive of gene expression. This step identified regulatory elements which are active in the same differentiation-stages that the gene is expressed. LASSO regression and its implementation in cisREAD is explained below:

LASSO (Least Absolute Squares Shrinkage Operator) regression is a form of penalised (L1-regularised) regression which applies a penalty term ($|\beta|$), equal to the absolute magnitude of their coefficient (β), is applied to the regression coefficients (Tibshirani, 1996a). This enables non-predictive variables to be shrunk to zero and eliminated from the model. Within cisREAD, LASSO regression models are fit to predict expression of a DEG from the chromatin accessibility of candidate (co)CREs according to Equation 3.1.

Equation 3.1 Linear model to predict transcription from cis-regulatory element accessibility

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij,}$$

Here y_i (the dependent variable) gave the expression of gene y in sample i and x_{ij} (the independent variable) gave the chromatin accessibility in in sample i, in candidate CRE j. A response vector, giving the expression of a gene in each sample, and predictor matrix, giving the chromatin accessibility in each sample for all predictors (p), were input to gene-specific models. Both chromatin accessibility and gene expression were $\log_2$ transformed (adding a pseudocount of 1) and standardised as z-scores prior to LASSO regression.

Equation 3.2 Coefficient estimation in LASSO regression

$$\min_\beta \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 - \lambda \sum_{j=1}^{p} |\beta j|$$

LASSO assigns coefficients (Equation 3.2) by minimising the error term $y_i - \bar{y}_i$ (the difference between actual and predicted gene expression), whilst applying the penalty ($|\beta|$) to perform regularisation. The extent of regularisation is determined by the tuning parameter λ, where greater values of λ result in sparser models where more

coefficients are shrunk to zero. The optimum λ can be determined by cross-validation, finding the value which minimises mean cross-validated error (λ_{min}) (Tibshirani, 1996).

cisREAD employs the glmnet package in R to perform model selection through random 5-fold cross-validation, before fitting the final regression model at λ_{min}. Candidate (co)CREs with $\beta = 0$ are rejected to regulate the gene, whilst those with $\beta \neq 0$ are selected as gene-specific CREs. Selected variables, with $\beta \neq 0$, are subject to significance testing through the method of 'selective inference', implemented in the selectiveInference R package. Selective inference performs significance testing of predictors whilst accounting for how variables are cherry-picked from a larger pool of candidates (Taylor and Tibshirani, 2018).To account for multiple testing, gene-specific p-values are assigned to each model, equal to that of its most significant predictor, and subject to Benjamini Hochberg (BH) adjustment (Benjamini and Hochberg, 1995). Selected (co)CREs with $p < 0.05$ and gene-specific False Discovery Rate (FDR) < 0.05 are considered statistically significant predictors of gene expression. These significant predictors should be interpreted as high-confidence gene-specific CREs.

The application of cisREAD to predict gene-specific regulation is exemplified in Figure 3.4. Here 23 TF-bound DARs are located within 100kb of the TSS for the BATF gene, community detection identifies two coCREs, bound by similar transcription factors and accessible in similar time points (Figure 3.4A). LASSO regression models are constructed, where gene expression is predicted by 16 coCREs or lone CREs. The optimum LASSO model (at λ_{min}) selected 5 predictors to control BATF (Figure 3.4B), including both coCREs. Expression predicted by the model is highly correlated with BATF expression (Figure 3.4C), and the accessibility of significant predictors mirrors expression of the gene (Figures 3.3D and 3.3E).

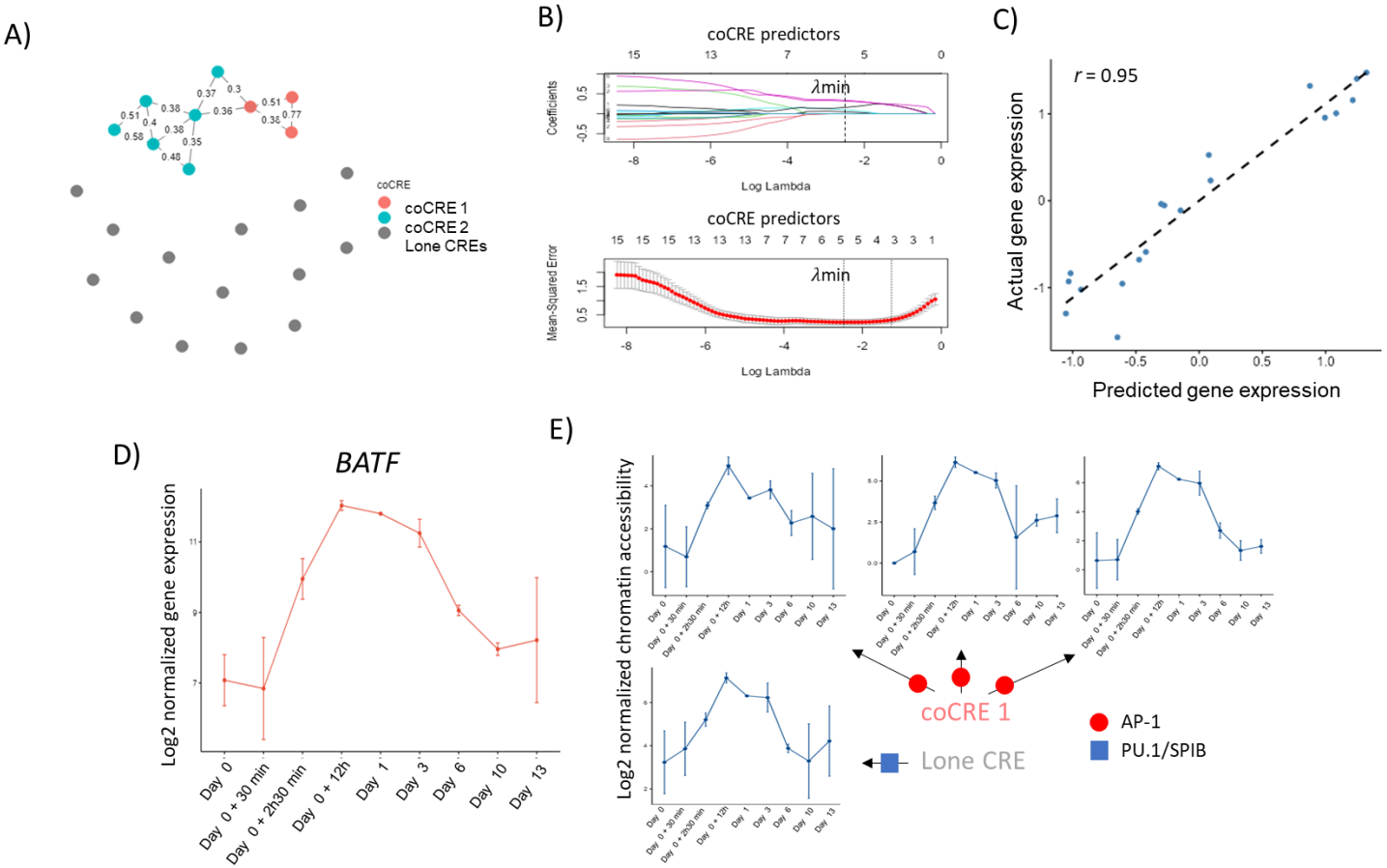


Figure 3.4 Example cisREAD model for *BATF* gene. A) Community detection reduced 23 gene-specific candidate CREs (DARs within 100kb of the gene's TSS) into two co-accessible and TF co-bound communities and 14 lone CREs. B) A LASSO model was constructed where *BATF* expression was predicted from the chromatin accessibility of the 16 (co)CRE predictors. Cross-validation found an optimum λ which minimised cross-validated error, a total of 5 predictors were selected at λ_{min}. C) BATF expression predicted by this LASSO model was highly correlated with observed gene expression (Pearson r = 0.95). D) BATF gene was mirrored by the chromatin accessibility of significant predictors in Figure 3.4E. E) Significant predictors included a lone CRE and coCRE 1 from Figure 3.4A. These were bound by common TFs and were co-accessible when the gene was expressed.

3.3.3 Data processing and cisREAD implementation

In order to predict gene-specific cis-regulatory elements across B cell differentiation, the cisREAD was applied to the *in vitro* B cell differentiation dataset. This involved processing and preliminary analysis of the data, followed by differential analysis (step 1) and an exploration of transcription factor motif and binding site tools for incorporation into step 2. Once *de novo* motifs and binding sites had been derived, the cisREAD R package was used to perform community detection (step 3) and LASSO regression (step 4) to predict gene-specific CREs. Finally, the influence of the TF binding site prediction method on cis-regulatory element prediction was evaluated using experimentally determined binding sites and TF-target genes.

3.3.3.1 RNA-seq processing

RNA-seq FASTQ files were first quality checked with FastQC (Andrews, 2010). Trim Galore was then used with default settings to remove Illumina sequencing adaptors and low-quality reads (Q < 20) (Kreuger, 2012). Trimmed RNA-seq reads were mapped against the GRCh38 decoy-aware human transcriptome and quantified with Salmon, correcting for GC bias and sequence bias (Patro et al., 2017). Transcript counts were aggregated to gene level with the tximport R package and normalised using DESeq2's median of ratios method (Love et al., 2014).

3.3.3.2 ATAC-seq processing

Trimmed, paired ATAC-seq reads were aligned to the human genome (NCBI GRCh38 decoy version) using bowtie2 (--very-sensitive) (Langmead and Salzberg, 2012). Post-alignment, low-quality mappings (MAPQ < 20) were filtered out with Samtools and duplicates were removed with Picard tools (Broad Institute, 2019; Li et al., 2009). Bedtools was also used to filter out ENCODE and mitochondrial blacklists (Amemiya et al., 2019), and select fragments <100bp to isolate nucleosome-free regions (Quinlan and Hall, 2010). Coordinates were shifted with deepTools +4 bp on the positive strand and −5 bp on the negative to centre on the Tn5 cutting site (Ramírez et al., 2016). Narrow peaks were called with Macs2 in paired-end mode with q < 0.05 (Zhang et al., 2008). Consensus peak sets were constructed using the DiffBind R package to retain peaks present in at least half of donors for each time-point (Stark and Brown, 2011). A

count matrix was then produced from the union of all consensus peaks, across all samples, and accessibility signal was normalised using DESeq2s median of ratio's method (Love et al., 2014).

3.3.3.3 Differential Analysis

Raw ATAC-seq and RNA-seq count matrices were VST (Variance Stabilising Transformation) transformed within DESeq2 and subject to preliminary analysis in R. Principal components were computed and a PCA biplot was produced with the ggplot2 package. Euclidean sample distances were hierarchically clustered and visualised with the pheatmap package.

Differential accessibility and expression analyses were performed on raw ATAC-seq and RNA-seq count matrices (cisREAD step 1) using likelihood ratio tests (LRTs) in DESeq2. Each LRT compared a model design of Differentiation-Stage + Donor, to a model with Donor only.

3.3.4 Comparative *de novo* motif discovery and TF occupancy prediction

To identify transcription factor binding sites important to B cell differentiation (cisREAD step 2), *de novo* motifs were discovered in DARs using HOMER and MEME-ChIP, and occupancy was predicted using HINT-ATAC and BMO. The purpose of this comparison was to curate a final set of motifs for use with cisREAD, and to explore the impact of binding site prediction on cisREAD-predicted gene regulation.

De novo motifs were discovered using HOMER and MEME-ChIP in differentially accessible regions, compared to chromatin regions whose accessibility did not differ across the time-course (LRT BH-adjusted $p > 0.01$). Motif discovery was also performed against control sequences generated using the software's default settings. *De novo* motifs were considered enriched based on the criteria of statistical enrichment ($p < 0.05$), frequency (present in >2.5% of DARs) and non-redundancy (enriched motifs should represent distinct TFs).

HOMER findMotifsGenome.pl was used to discover enriched *de novo* motifs and match *k*-mers to known PWMs from JASPAR Vertebrate Core 2020 (Fornes et al., 2020), HOCOMOCO v11 (Kulakovskiy et al., 2018), and HOMER itself (Heinz et al., 2010).

Default parameters were used aside from -size given, to discover motifs in the whole ATAC-seq peak instead of those in the centre of the peak.

MEME-ChIP was used to discover *de novo* motifs (through MEME and STREME programs), calculate enrichment (through CENTRIMO) and match *k*-mers to known PWMs from JASPAR Vertebrate Core 2020 and HOCOMOCO v11 (through TOMTOM) (Machanick and Bailey, 2011). MEME-ChIP was run using '-meme-mod anr' (to assume that each sequence may contain any number of motif repetitions), -meme-searchsize 0 (to discover motifs in all ATAC-seq peaks) and -centrimo-local (to calculate enrichment of motifs in the whole of each peak).

Following selection of the final motif set, TF occupancy was predicted in each differentiation-stage using HINT-ATAC and BMO with default parameters. ATAC-seq footprints were called by HINT-ATAC using the 'rgt-hint footprinting command' (Li et al., 2019). HINT-ATAC was supplied with differentiation-stage specific BAM files – produced by merging BAMs of biological replicates (Samtools) – and differentiation-stage specific peak files – produced by intersecting DARs with stage-specific consensus ATAC-seq (bedTools). The resultant footprints were then scanned with PWMs of enriched *de novo* motifs, using HOMER findMotifsGenome.pl in -find mode, to predict TF occupancy at each stage.

Motif occupancy was also predicted using BMO with default settings (D'Oliveira Albanus et al., 2021). BMO was supplied with genome-wide motif scans for each *de novo* motif (obtained using HOMER scanMotifGenomeWide.pl), differentiation-stage specific BAM files and differentiation-stage specific peak files.

3.3.5 Predicting gene-specific cis-regulatory elements with cisREAD

Following binding site prediction with HINT-ATAC and BMO, differentiation-stage specific binding sites were summarised into TF occupancy matrices using custom python scripts. TF occupancy, differentially accessible peak and differentially expressed gene matrices were input into the cisREAD R package to perform community detection (cisREAD step 3) and LASSO regression (cisREAD step 4). cisREAD was run separately with HINT-ATAC-predicted TF occupancy and then with BMO-predicted TF occupancy.

HINT-ATAC + cisREAD successfully modelled regulation for 8,215 DEGs. Each gene was assigned a mean of 8.4 candidate CREs (range 2-34) and, following community detection, each model had a mean of 6.9 predictors (range 2-26). Of these a mean of 4.7 were selected by the LASSO model (range 1-22), and 2.1 were statistically significant (range 1-16). Application of cisREAD resulted in the prediction of 38,554 CRE-gene relationships in total, of which 9,440 were statistically significant. 24,186 relationships were characterised by upregulation at enhancers or promoters (of which 6,609 were significant) and 14,368 were characterised by downregulation (3,626 significant). Similar, slightly higher, numbers of predictions were reported for BMO + cisREAD (Table 3.1).

Table 3.1 Summary of gene-specific regulatory elements predicted by cisREAD, using either HINT-ATAC or BMO to predict transcription factor occupancy. Values in last four rows give mean and range.

		cisREAD + HINT-ATAC			cisREAD + BMO		
		Modelled	**Selected**	**Significant**	**Modelled**	**Selected**	**Significant**
Genes		8,215	8,215	4,959	8,539	8,539	5,025
CREs		35,064	24,648	8,600	40,506	29,064	10,702
CRE-gene relationships	**All**	69,643	38,554	10,235	82,557	46,594	12,758
	Positive	-	24,186	6,609	-	29,339	8,542
	Negative	-	14,368	3,626	-	17,255	4,213
CREs per gene		8.4 (2-34)	4.7 (1-22)	2.1 (1-16)	9.7 (2-46)	5.7 (1-37)	2.5 (1-22)
predictors/(co)CREs per gene		6.9 (1-26)	3.9 (1-16)	1.7 (1-14)	6.8 (1-23)	3.9 (1-17)	1.7 (1-11)
coCRE membership		2.7 (2-12)	2.7 (2-12)	2.7 (2-11)	3.1 (2-17)	3.1 (2-16)	3.1 (2-16)
Genes per CRE		2.0 (1-16)	1.6 (1-12)	1.2 (1-8)	2.0 (1-16)	1.6 (1-13)	1.2 (1-7)

3.3.6 Evaluation of predicted TF binding and gene regulation

To evaluate motif occupancy predicted by HINT-ATAC and HOMER, binding sites predicted by HINT-ATAC and BMO in day 3 activated B cells were compared to experimentally determined ChIP-seq binding sites in GM12878 lymphoblastoid cell lines – which resemble the day 3 differentiation state. ChIP-seq peaks for transcription factors recognised to bind each motif were downloaded from ENCODE (IDR threshold peaks aligned to hg38) and intersected with HINT-ATAC and BMO-determined binding site predictions using bedtools. ChIP-seq-determined binding site enrichment was

calculated for motifs predicted occupied, compared to motifs predicted unoccupied using a two-sided Fisher test.

Whilst the above exercise indicated whether predicted binding reflects measured binding, it did not reflect whether predicted transcription factor binding sites could be linked to gene expression. To test this HINT-ATAC-derived and BMO-derived cisREAD predictions were compared to known transcription factor target genes, curated from MSigDB and Lymphochip (Staudt lab) datasets (Liberzon et al., 2015; Shaffer et al., 2006). The dataset used for this analysis was curated by filtering the Lymphochip gene signatures for the 'Transcription factor target' category, and then filtering again by subcategory for the TF of interest (NF-kB or IRF4). The 'HALLMARK_TNFA_SIGNALING_VIA_NFKB' signature from MSigDB was also included for NF-kB.

Gene set enrichment was calculated for genes whose selected CREs were predicted occupied by the TF, compared to genes whose selected CREs were predicted occupied by the TF using a two-sided fisher test.

3.4 Results and Discussion

3.4.1 Chromatin accessibility and gene expression are reprogrammed during *in vitro* B cell differentiation

To assess the suitability of the B cell differentiation dataset for the planned analysis, unsupervised exploratory analyses were performed. Hierarchical clustering (Figure 3.5A) and principal components analysis (Figure 3.5B) of ATAC-seq and RNA-seq samples showed that B cell epigenomes and transcriptomes undergo global reprogramming upon B cell activation and plasma cell differentiation. Samples from each differentiation stage were observed to cluster together, showing no evidence of batch effects or outliers. Therefore all 19 ATAC and RNA-seq sample pairs were retained for further analysis.

To identify regulatory elements and genes whose activity changed across the differentiation time-course, differential accessibility and expression analyses were performed. 97,707 ATAC-seq peaks (LRT FDR < 0.01) and 9,082 protein coding genes

(LRT FDR < 0.001) were defined as significantly differentially accessible or expressed.

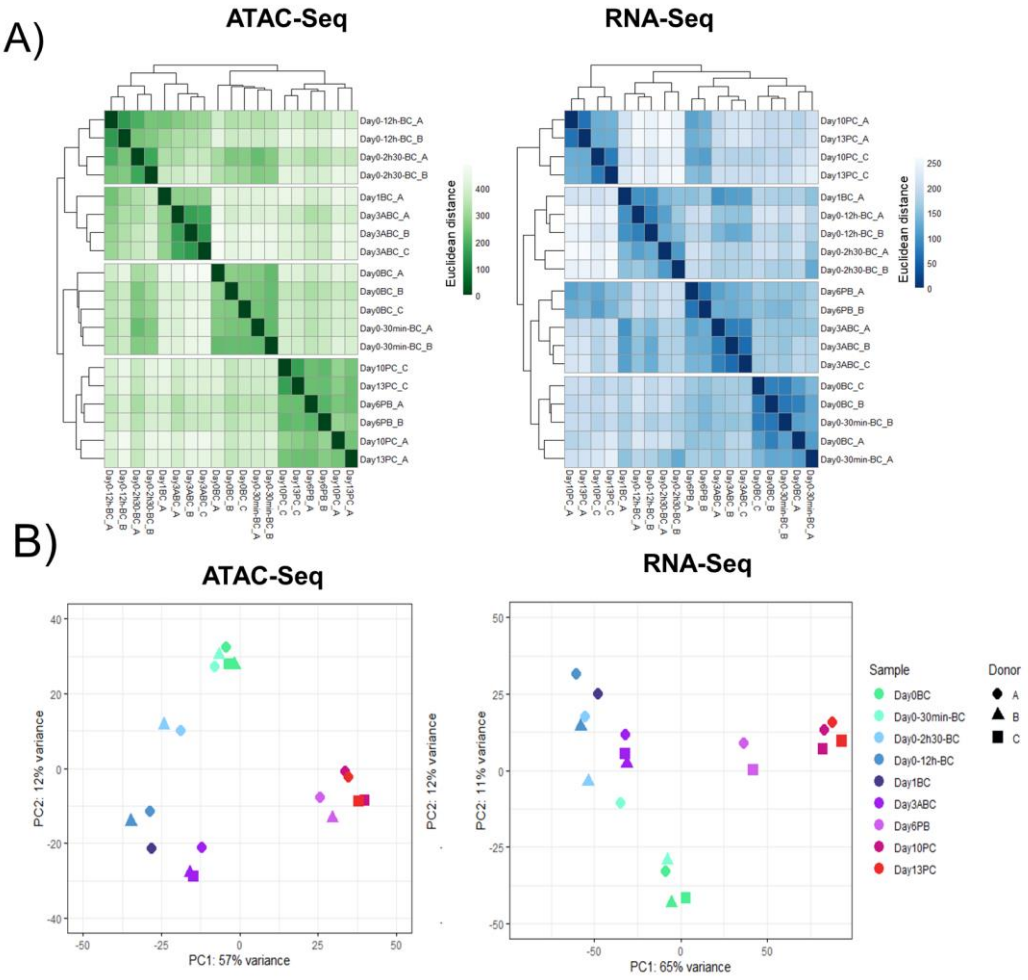


Figure 3.5 Hierarchical clustering of ATAC-seq and RNA-seq datasets. Hierarchical clustering was performed on Euclidean sample distances calculated from VST normalised counts for all consensus ATAC-seq peaks and genes. B) Principal components analysis biplot for ATAC-seq and RNA-seq datasets, showing PC1 and PC2 calculated from VST normalised counts. BC, B cell; ABC, activated B cell; PB, plasmablast, PC; plasma cell.

3.4.2 Key transcription factor motifs are enriched in differentially accessible regions

Following the identification of differentially accessible regions, *de novo* motif discovery was performed to identify enriched, frequent, and non-redundant transcription factor binding sites. Analysis with both HOMER and MEME-ChIP found enrichment of *de novo* motifs matching known PWMs for AP-1, PU.1/SPIB, IRF4, RUNX, OCT2, NF-kB, SP/KLF, E-Box, MADS-Box and CTCF factors in DARs (Figure 3.6). These motifs were repeatedly discovered when using different discovery algorithms and control sequences, and their

involvement in mature B cell differentiation is supported by the literature (Brescia et al., 2018; Cao et al., 2010; Gerondakis and Siebenlist, 2010; Klein et al., 2006; Pérez-García et al., 2017; Watanabe et al., 2010; Willis et al., 2017; Wöhner et al., 2016).

In addition to the above motifs, HOMER reported enrichment of PAX5 and CREB/ATF motifs when using both default (Figure 3.6A) and non-differentially accessible (non-DA) background sequences (Figure 3.6B). Motif discovery and clustering identified separate motifs for and CREB/ATF factors when using default background sequences. However, a single broad *k*-mer which weakly matched both PAX5 and CREB/ATF PWMs, was discovered when using a non-differentially accessible background set.

HOMER + nonDA also discovered motifs matching PWMs for STAT3 and ZBTB33. STAT3 is induced downstream of IL-21 (Berglund et al., 2013), and therefore is induced at every time point and ZBTB33 has previously been implicated in germinal centre formation (Koh et al., 2013). The STAT3 motif however includes a GA dimer preceding the CTCCGGAA consensus and so would only match a subset of STAT3 sites.

MEME-ChIP + nonDA showed the lowest sensitivity of all analyses, identifying the fewest number of enriched motifs, including 3 low-complexity motifs which poorly matched known PWMs (Figure 3.6D). This may indicate that HOMER better corrected biases between primary and user-specified control sequences. A greater number of higher-quality motifs were identified when using default control sequences (Figure 3.6C). MEME-ChIP + default identified 10 motifs also discovered by HOMER and an additional motif matching the PWM for IKAROS, known to be important for B cell development and activation (Sellars et al., 2011).

The 13 *de novo* motifs from HOMER + nonDA were taken forward for use within cisREAD (Figure 3.6B). These were selected since this analysis identified the most *de novo* motifs, all of which had relevance to the system. Whilst the quality of PAX5/CREB/ATF (too broad) and STAT3 motifs (too specific) may hinder downstream application, these additional motifs were still included due to the importance of the factors in B cell differentiation.

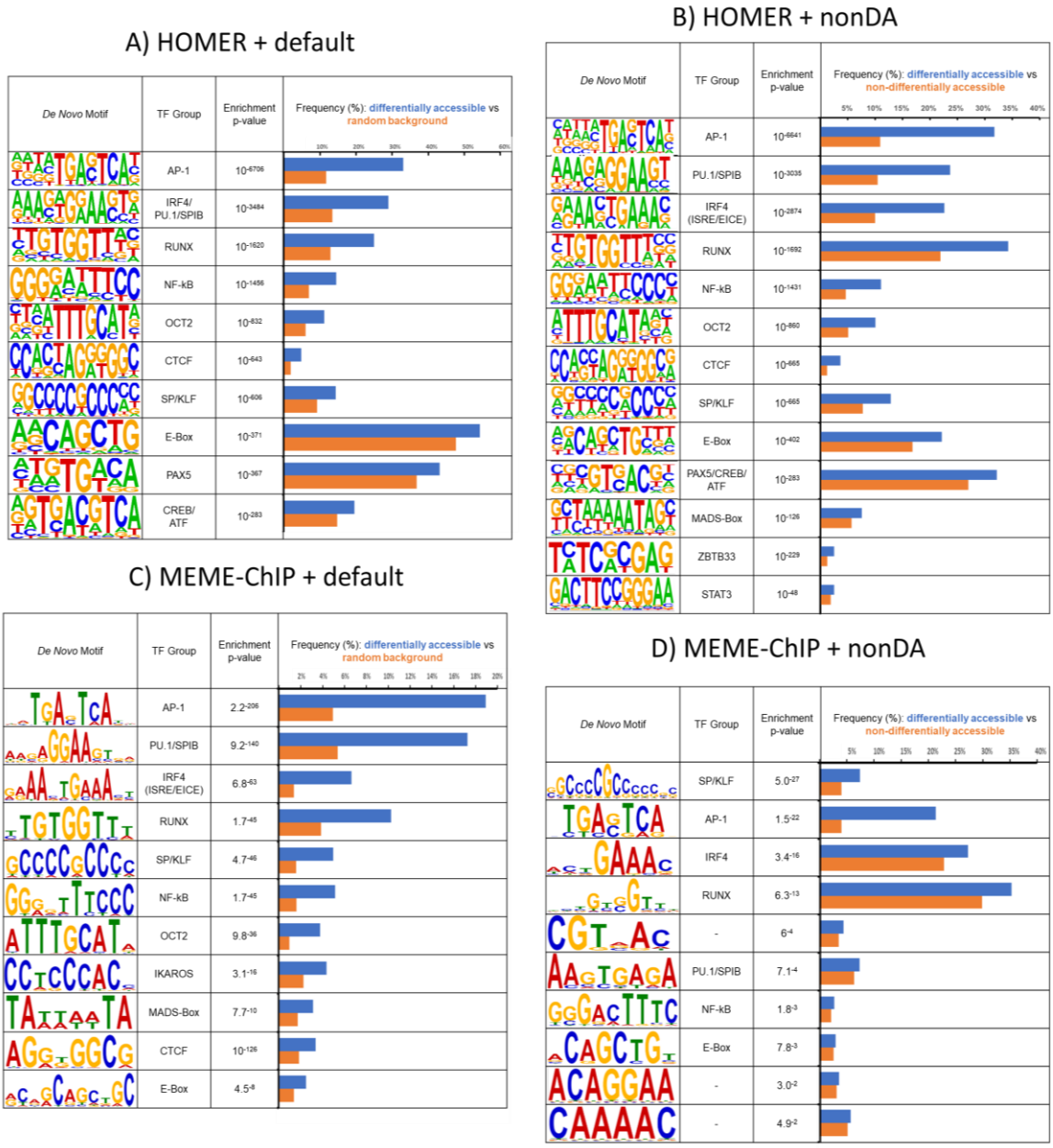


Figure 3.6 Motifs discovered through *de novo* discovery in differentially accessible regions using HOMER or MEME-ChIP. *De novo* motif discovery was performed using HOMER (A and B) and MEME-ChIP (C and D), using a background of software-generated sequences (A and C) or non-differentially accessible chromatin regions (B and D).

3.4.3 Motif occupancy can be predicted from ATAC-seq data

To identify binding sites for the 13 TF families represented in Figure 3.6B, motif occupancy was predicted within DARs using the footprint-based HINT-ATAC method and the footprint-independent BMO model. To assess the performance of each model, predicted binding in day 3 was compared with ChIP-seq measured binding in GM12878 – used as a surrogate for the day 3 activated B cell state. For each *de novo* motif, a

relevant factor (expressed in B cells and capable of recognising the motif), was selected for validation of predicted occupancy.

Figure 3.7A shows the predicted occupancy of differentially accessible motifs in day 3 activated B cells, and Figure 3.7B shows statistically significant overlaps ($p < 0.1$) between predicted occupancy and measured occupancy for the 13 TF-motif pairs.

For all factors, bar CTCF, BMO predicted a greater proportion of occupied motifs than HINT-ATAC (Figure 3.7A), with a mean predicted occupancy rate of 60% (BMO – excluding CTCF) compared to 30% (HINT-ATAC – all motifs). CTCF was the only motif for which BMO predicted fewer occupied sites, where only 63/3,516 (2%) differentially accessible motifs were predicted occupied. In contrast HINT-ATAC predicted 40% occupancy of CTCF motifs,

Both BMO and HINT-ATAC predicted binding sites significantly overlapped ChIP-seq binding sites for most factors, compared to motifs predicted to be unbound. A notable exception was for BMO and CTCF, where CTCF ChIP-seq peaks were significantly depleted from predicted binding sites. Aside from CTCF, BMO predictions showed greater enrichment for ChIP-seq binding sites than HINT-ATAC predictions, with increased odds ratios and smaller p-values (Figure 3.7B). This supported conclusions from the authors of BMO that their model better predicts ChIP-seq-measured transcription factor occupancy than footprint-based methods including HINT-ATAC (D'Oliveira Albanus et al., 2021). In contrast to the analysis here, the BMO authors found that BMO and HINT-ATAC showed similar performance in identifying CTCF bound motifs from GM12878 ATAC-seq. It is unclear why BMO performs poorly for CTCF on our *in vitro* B cell data.

Whist this analysis offered a useful indication of relative performance, there were several considerations. GM12878 B lymphoblastoid cells are transformed with the Epstein Barr Virus, and therefore differ from day 3 cells. Motifs may also be occupied by factors not considered in the comparison (e.g., Fos1 and Fra1 at AP-1 motifs), or be poorly recognised by their assigned factors due to PWM quality (e.g., PAX5/CREB/ATF). The quality of ChIP-seq datasets (e.g., sequencing depth, antibody quality) may also bias the motifs designated as transcription factor binding sites. Most importantly, this

exercise offered no indication of whether predicted binding sites were driving transcription.

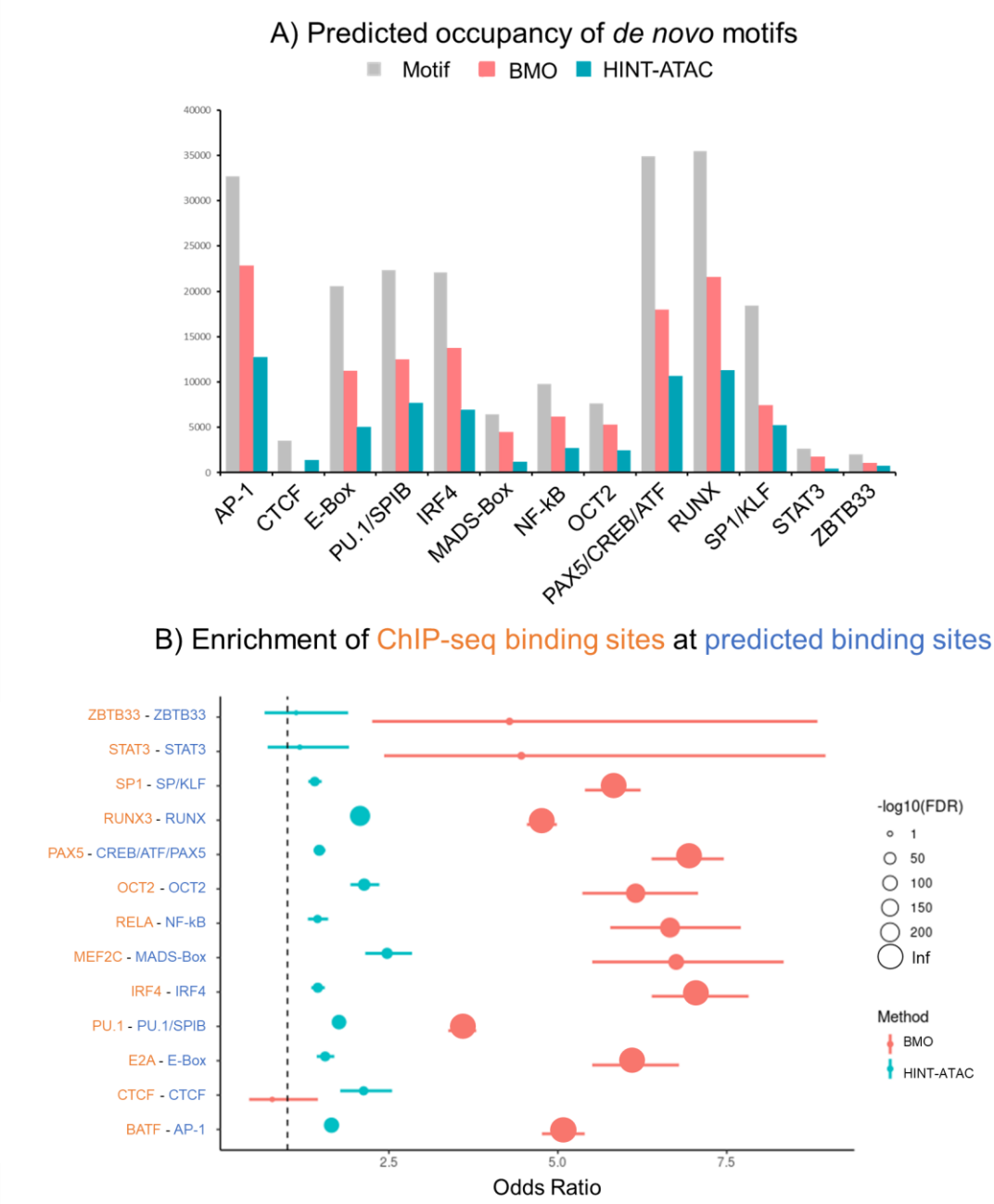


Figure 3.7 Predicted TF occupancy (using HINT-ATAC or BMO) compared to TF occupancy detected by ChIP-seq. A) Number of *de novo* motifs in DARs (grey) predicted to be occupied by BMO (teal) or HINT-ATAC (red). B) Enrichment of ChIP-seq binding sites in GM12878 for transcription factors at differentially accessible de novo motifs, predicted occupied by BMO (red) or HINT-ATAC (teal). The position of dots indicates the effect of enrichment (odds ratio), and their size indicates the statistical significance (BH-adjusted p value). Bars give the 95% confidence interval. Statistics (calculated by a two-sided Fisher test) describe the enrichment of ChIP-seq-detected binding sites for specific transcription factors (orange text) at *de novo* motifs with predicted occupancy (blue text)

3.4.4 cisREAD correctly predicts transcription factor target genes

Transcription factor binding at a cis-regulatory element alone is often insufficient to drive gene expression, which frequently requires combinatorial binding at promoters and enhancers. With this in mind, we next asked whether the binding sites predicted by BMO and HINT-ATAC can be linked to active transcription by the cisREAD method. To evaluate this, we tested the ability of cisREAD to identify confirmed transcription factor target genes, when using either BMO or HINT-ATAC in step 2. Here we focused on two transcription factors, NF-kB and IRF4, whose target genes have been extensively studied in a variety of B lineage cell types and stimulatory systems using chromatin immunoprecipitation and gene knockouts. Whilst these systems will all differ from the *in vitro* differentiation model, collectively this should indicate whether our method is able to identify transcription factor target genes from predicted binding sites.

In this exercise, predicted NF-kB/IRF4 targets were defined as genes with one or more cisREAD-selected CREs that were predicted bound by IRF4/NF-kB by BMO or HINT-ATAC. Gene set enrichment in predicted targets was compared to a background of predicted non-target genes. Predicted non-targets were defined as genes where no cisREAD-selected CREs were predicted bound by IRF4/NF-kB.

Figure 3.8 shows the enrichment of experimentally validated NF-kB(Figure 3.8A) and IRF4 targets (Figure 3.8B) in cisREAD-predicted target genes, using either BMO or HINT-ATAC, for 11 NF-kB gene sets and 16 IRF4 gene sets. In nearly all cases, cisREAD-predicted target genes were significantly enriched for IRF4 and NF-kB target gene sets ($p < 0.1$). This indicates that cisREAD paired IRF4 and NF-kB bound cis-regulatory elements to their correct target genes.

Altogether, Figure 3.8 shows no consistent advantage of BMO or HINT-ATAC for use within cisREAD. This could indicate that the additional binding sites detected by BMO (Figure 3.7) may be bound by transcription factors which do not drive expression. Importantly it suggests that the cisREAD method is robust to the choice of binding site prediction method.

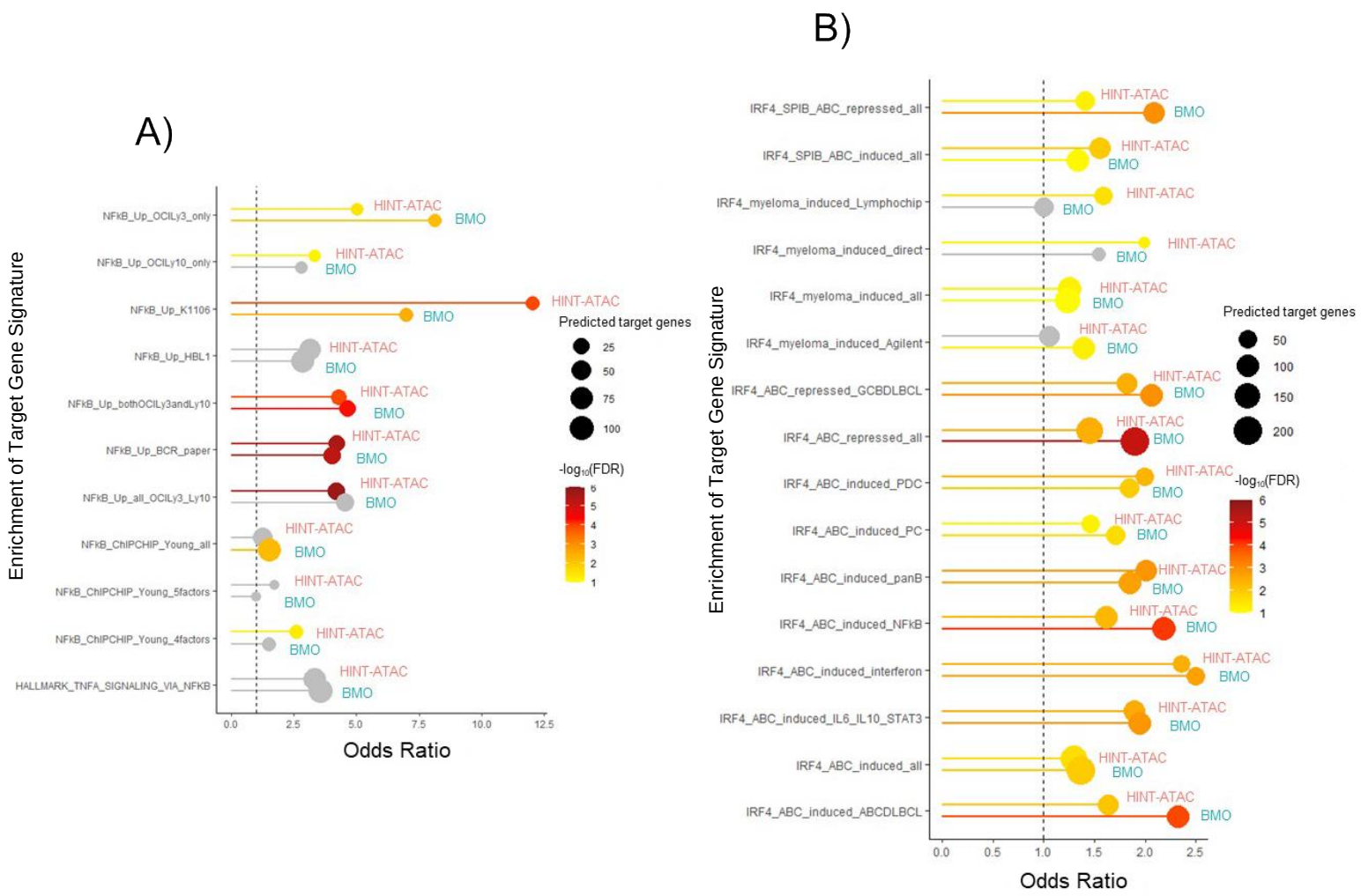


Figure 3.8 Enrichment of NF-kB and IRF4 target gene signatures in cisREAD-predicted target genes (using HINT-ATAC or BMO predicted binging sites). Enrichment was calculated (two-sided Fisher test) for experimentally determined TF target genes, in cisREAD-predicted TF-target genes, using either HINT-ATAC or BMO to predict TF occupancy at CREs. A) shows enrichment calculated for 11 NF-kB target gene sets, and B) for 16 IRF4 target gene sets from the Lymphochip and MSigDB databases. Each bar indicates enrichment results using HINT-ATAC (red text) or BMO (blue text). The position of each dot gives the effect size of enrichment (odds ratio), and the size gives the number of predicted target genes. Statistical significance ($-\log_{10}$ BH-adjusted p value) is given by colour, from yellow to red (greatest significance). Non-significant enrichment ($p > 0.1$) is shown in grey.

3.5 Conclusion

In this chapter we have introduced cisREAD as a method to identify gene-specific cis-regulatory elements across differentiation. The method is designed to prioritise regulatory elements downstream of core transcription factors, whose differential accessibility is linked to differential expression. cisREAD requires only ATAC-seq and RNA-seq data and makes use of commonly used bioinformatic tools and a bespoke R package. We have shown here how motif discovery and binding site prediction tools can be used in lieu of ChIP-seq to identify transcription factor binding sites. Importantly we found that cisREAD is robust to binding site prediction method and correctly identifies genes targeted by key transcription factors. Application of cisREAD for biological insight will be exemplified in chapter 4, which will focus on its use to predict gene-specific and global modes of regulation in B cell differentiation.

Chapter 4. Data integration with cisREAD identifies global and gene-specific mechanisms of transcriptional control in B cell differentiation

4.1. Introduction

Chapter 3 introduced the cisREAD method for predicting gene-specific cis-regulatory elements across differentiation, and it discussed cisREAD's development and application to the *in vitro* B cell time-course. This chapter leverages predictions from the cisREAD method, to identify global changes in gene regulation during B cell differentiation, and to generate hypotheses of gene-specific regulation. The results of both genome-wide and gene-specific analyses will be evaluated in the context of known regulatory mechanisms, to highlight the potential for cisREAD to predict new mechanisms of transcriptional control.

4.1.1 Gene regulatory networks during mature B cell differentiation

B cell differentiation (introduced in chapter 3)is driven through changes in gene regulation. Studies in mice and humans have uncovered how shifts in epigenetic remodelling and transcription factor activity rewire gene expression. This chapter will explore the dynamic networks of signal-inducible and lineage-specifying transcription factors, which coordinate activation upon immunisation, and guide B cells towards plasma cell fate. Particularly it will focus on the binding patterns and downstream transcriptional networks of key transcription factors, which occupy the *de novo* motifs presented in Figure 3.6.B of the previous chapter. To place these factors into context, this section will provide an overview of the established regulatory mechanisms which steer mature B cells through activation, germinal centre processes, and plasma cell differentiation.

4.1.1.1 Regulation of B cell activation

Intrinsic B cell transcription factors and signal-inducible germinal centre factors facilitate the transcriptional response to B cell activation (Figure 4.1). Mature B cell identity is upheld by a network of highly expressed transcription factors including PAX5, PU.1, SPIB, IRF8,and BACH2 (Nutt et al., 2015). These TFs, shown in blue in Figure 4.1, function to maintain B cell identity through upregulating B cell signalling machinery (PU.1, SPIB, PAX5 - Cobaleda *et al.*, 2007; Willis *et al.*, 2017). They also supress plasma cell (PC) differentiation, achieved through repression of plasma cell master regulator PRDM1 (PAX5, SPIB, PU.1-IRF8, and BACH2 - Delogu *et al.*, 2006; Schmidlin *et al.*, 2008; Carotta *et al.*, 2014; Ochiai and Igarashi, 2022).

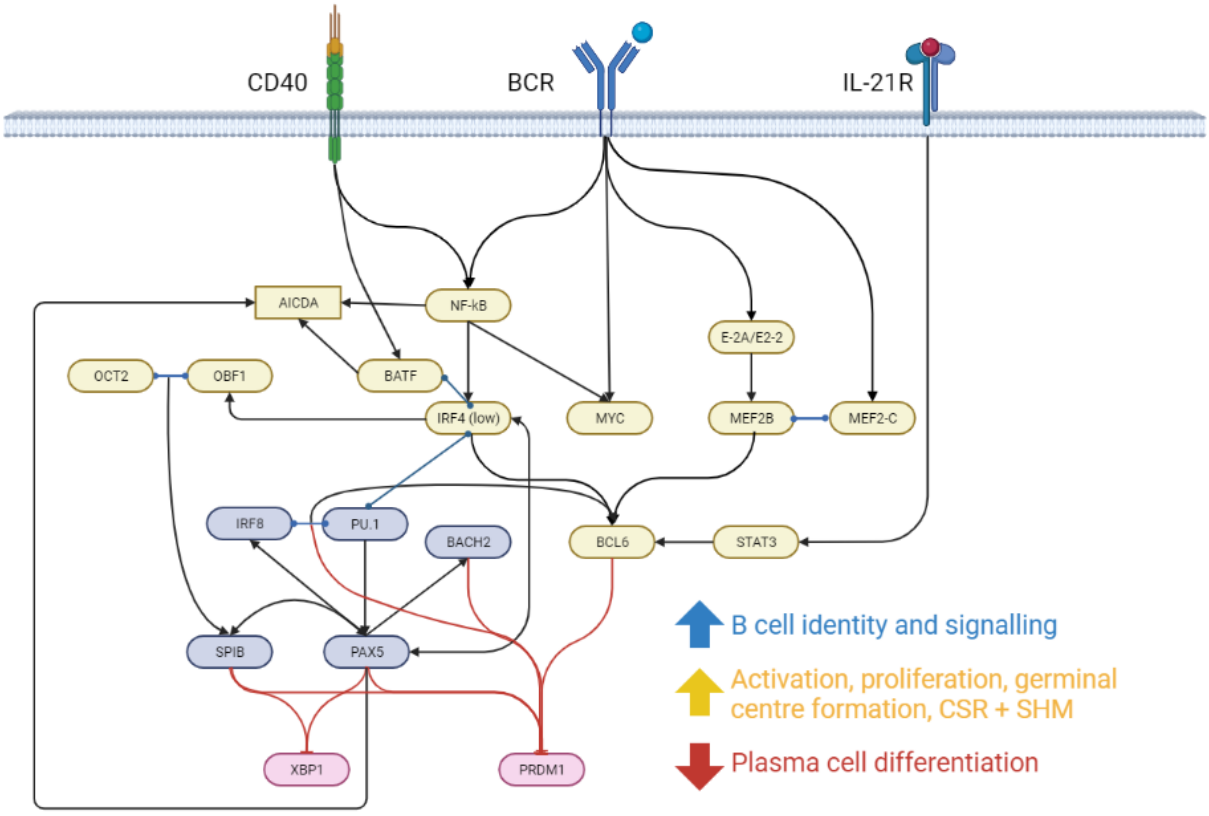


Figure 4.1 Transcriptional regulatory networks in B cell activation and germinal centre formation, induced by T cell dependent stimuli. Prior to activation, a network of B cell-specific transcription factors (in blue) maintains B cell identity through mutual activation (black arrows), and transcriptional repression (red arrows) of plasma cell regulators (in red). T cell-dependent activation induces signalling through CD40, BCR, and IL-21R which activate transcription factors (in yellow) responsible for co-ordinating B cell activation, proliferation, and germinal centre formation, as well as class switch recombination (CSR) and somatic hypermutation (SHM). Crucially NF-κB induces low levels of IRF4 which can co-regulate (blue arrows) with PU.1 or BATF. Low IRF4 upregulates GC master regulators OBF1-OCT2 and BCL6, which also blocks PC differentiation through PRDM1 repression. Figure created with BioRender.

Upon B cell activation under T cell dependent conditions, CD40 and BCR co-engagement stimulates signalling pathways, which lead to changes in gene regulation. These changes are coordinated by TFs shown in yellow in Figure 4.1. Crucially, NF-κB is induced through both canonical and non-canonical pathways to drive germinal centre development and proliferation, including through activation of MYC (De Silva et al., 2016; Heise et al., 2014; Luo et al., 2018). MYC expression is also induced through PI3K (Phosphoinositide 3-kinase) signalling downstream of BCR stimulation (Luo et al., 2018).

In vivo, BCR and CD40 signalling induce formation of germinal centres (GCs) through upregulation of BCL6. BCL6 coordinates germinal centre formation through its activity as a transcriptional repressor, and controls a wide processes including cell positioning, apoptosis, cellular signalling and T:B interactions (Basso and Dalla-Favera, 2012). Germinal centre commitment is also regulated by a range of other TFs including IRF4, OCT2 (co-binding with OBF1) and MADS-Box factors MEF2C and MEF2B (Laidlaw and Cyster, 2021).

In the initial activation, BCR and CD40 signalling transiently upregulate low levels of IRF4, through NF-κB. Low IRF4 leads to upregulation of BCL6 and OBF1 (Ochiai et al., 2013)o. IR4F alone has weak DNA binding affinity and forms heterodimers with either ETS-family TF PU.1, or AP-1 family TF BATF, when binding DNA. IRF4 co-binds with PU.1 at ETS-IRF composite elements (EICEs) and with BATF at AP1-IRF composite elements (AICEs) (Ochiai et al., 2013).

Alongside activation by IRF4, BCL6 is further upregulated by STAT3 downstream of IL-21 signalling, and also by PU.1-IRF8 and MEF2B (Brescia et al., 2018; Carotta et al., 2014; Diehl et al., 2008). MEF2B is a MADS-Box factor which is activated through E-Box transcription factors (Wöhner et al., 2016). The binding of E-Box factors, including the dominant E-protein E2A, is enhanced downstream of activation stimuli, through repression of their antagonist ID3 (Gloury et al., 2016). MEF2B can dimerise with fellow MADS-Box factor MEF2C, which is induced downstream of BCR engagement through MAPK (mitogen-activated protein kinase) signalling (Brescia et al., 2018; Khiem et al.,

2008). MEF2B and MEF2C act partially redundantly to promote germinal centre formation, proliferation and B cell survival (Brescia et al., 2018; Wilker et al., 2008). One study suggests that ZBTB33 may function to limit germinal centre proliferation through transcriptional repression of BCL6 and MYC (Koh et al., 2013).

The induction of OBF1 by low levels of IRF4 promotes GC B cell development through the activity its binding partner OCT2, which is expressed constitutively throughout the B lineage (Corcoran et al., 2014; Emslie et al., 2008; Ochiai et al., 2013). CD40/BCR signalling also induces the transcription factor BATF which is needed for germinal centre maintenance, alongside the mutagenic processes of somatic hypermutation (SHM) and class switch recombination (CSR) through regulation of AICDA (also controlled by NF-κB and PAX5) (Inoue et al., 2017; Ise et al., 2011; Zan and Casali, 2013). There is evidence that CD40 and BCR signalling also triggers a switch from RUNX1 to RUNX3 regulation, leading to upregulation of RUNX1-repressed cell cycle genes (Thomsen et al., 2021).

4.1.1.2 Regulation of germinal centre cycling and cell fate choice

As described in chapter 3, germinal centres (GCs) are divided into dark zones of proliferation and somatic hypermutation, and non-dividing light zones. Transcriptional regulation in the dark zone is controlled by FOXO1, which augments BATF expression through upregulation of CD40 and BCR signalling components (Inoue et al., 2017). After repeated rounds of SHM and proliferation, dark zone GC B cells enter the light zone. B cells with high affinity antibodies receive T cell help. This leads to induction of MYC, through CD40 and BCR, promoting entry to the cell cycle, and repression of FOXO1 by PI3K (Luo et al., 2018). This T cell help also refuels selected B cells through induction of MTORC1 (mammalian target of rapamycin complex 1), which promotes anabolic growth to sustain cell division and may contribute to PC differentiation through repression of BACH2 (Ersching et al., 2017; Kometani et al., 2013).

Evidence suggests a model where the extent of T cell help, determined by antibody affinity, affects the strength of BCR and CD40 signalling, and thereby determines memory B cell or plasma cell fate (Akkaya et al., 2020; Laidlaw and Cyster, 2021). B cells receiving weak T cell help maintain high BACH2 (possibly through insufficient

MTORC1), which sustains repression of plasma cell regulator PRDM1 and promotes differentiation into memory B cells (Shinnakasu et al., 2016).

4.1.1.3 Regulation of plasma cell differentiation and survival

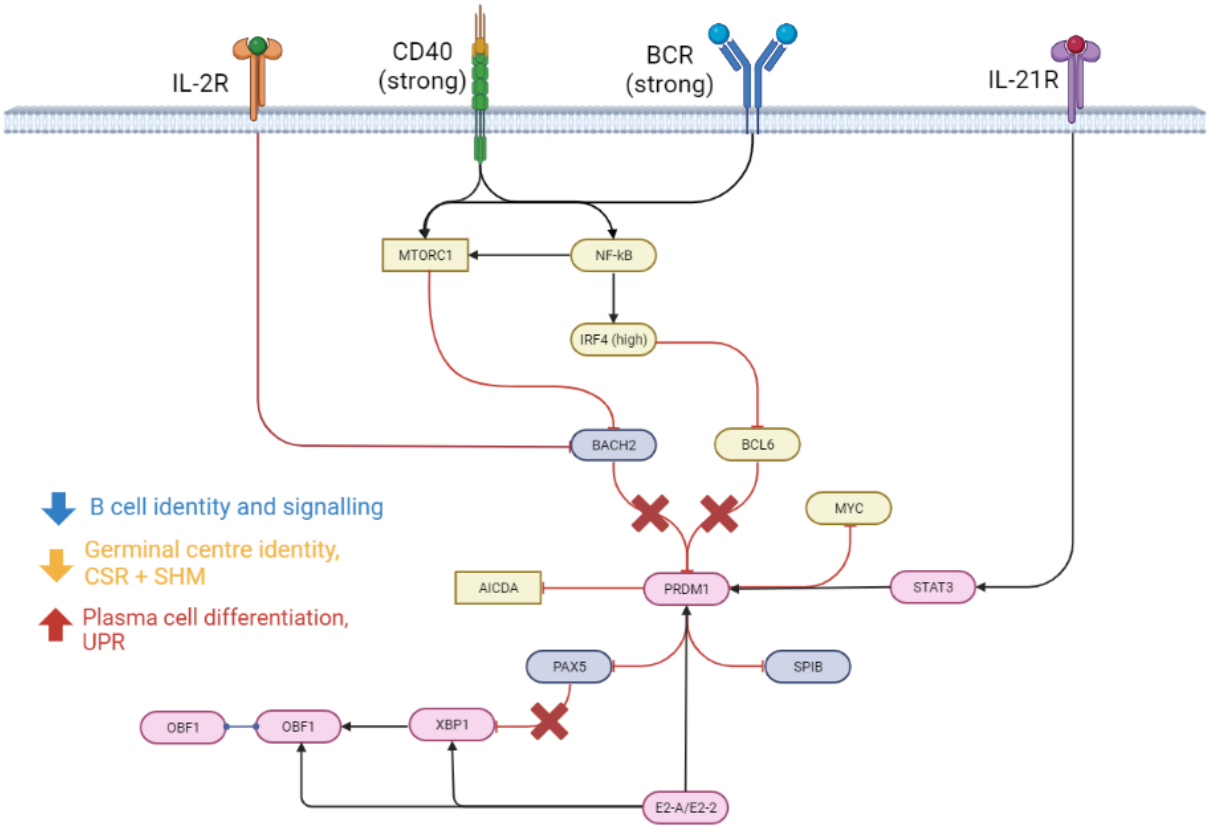


Figure 4.2 Transcriptional regulatory networks during plasma cell differentiation following T cell help. High affinity B cells receive strong T cell help and repress BCL6 and BACH2 to relieve repression of PRDM1 and induce expression by IRF4. Upregulation of PRDM1 leads to repression of germinal centre and B cell repression genes, including PAX5 which relieves repression on XBP1 and activates the unfolded protein response. Figure created with BioRender.

High affinity GC B cells, receiving strong T cell help, differentiate into plasma cells (Ise and Kurosaki, 2019; Laidlaw and Cyster, 2021). As shown in Figure 4.2, strong CD40 and BCR co-stimulation relieves repression of plasma cell genes (shown in red). Strong NF-κB signalling induces high levels of IRF4, which repress BCL6 and relieve PRDM1 from its repression (Sciammas et al., 2006). PRDM1 is also relieved of BACH2-mediated

repression, through repression by MTORC1 downstream of strong activation stimuli and T cell cytokine IL-2 (Hipp et al., 2017).

Following alleviation of repression, PRDM1 is upregulated by IRF4, E2A/E2-2, and STAT3 downstream of IL-21 (Kwon et al., 2009; Sciammas et al., 2006; Wöhner et al., 2016). At high levels, IRF4 shifts away from EICE and AICE sites towards ISRE (interferon specific response element) motifs, to which it binds as a homodimer (Ochiai et al., 2013). PRDM1 is also capable of binding ISREs in a mutually exclusive manner (Doody et al., 2010). IRF4 has also been noted to associate with architectural protein CTCF in plasmablasts and plasma cells (Cocco et al., 2020).

Following upregulation, PRDM1 acts to repress B cell (PAX5, SPIB) and germinal centre (AICDA and MYC) factors (Minnich et al., 2016). Repression of PAX5 leads to de-repression of XBP1, which co-ordinates antibody synthesis and secretion through regulation of the unfolded protein response (UPR) to endoplasmic reticulum (ER) stress (Shaffer et al., 2004).

Plasma cell longevity is supported by homing towards the bone marrow niche (Nutt et al., 2015; Tellier and Nutt, 2019). Plasmablasts are recruited to the bone marrow through the chemokine CXL12 (Hargreaves et al., 2001). Their retention and maturation is promoted through engagement of receptors and transcription factors including KLF2, which promotes quiescence (Winkelmann et al., 2011). The bone marrow stromal niche is home to eosinophils which secrete the B cell survival factor APRIL, alongside the cytokine IL-6. (Chu et al., 2011). APRIL signals through the BMCA (B cell maturation antigen) receptor to promote expression of the anti-apoptotic MCL1 protein, essential for PC survival (Peperzak et al., 2013), and induces NF-κB to prevent ER stress-associated cell death (Cornelis et al., 2020).

Whilst the modes of regulation in Figures 1 and 2 are well characterised, our understanding of B cell transcriptional regulation is still incomplete. Application of cisREAD to the *in vitro* B cell dataset (Figure 3.2) has identified transcription factors, chromatin regions and genes whose activity changes during human B cell differentiation; in a system driven by CD40 and BCR signalling, alongside IL-2, IL-21, IL-6 and APRIL. This first section of this chapter will harness these linkages to explore how these core TFs co-ordinate a dynamic transcriptional response to B cell differentiation

stimuli. This will fill in in gaps for the *cis* and *trans* acting factors which fine-tune transcription to orchestrate B cell maturation.

4.1.2 Control of master regulators *AICDA* and *PRDM1*

After determining transcriptional regulation on a genome-wide scale, this chapter will evaluate two gene-specific models to demonstrate recall of known regulatory elements and generate hypotheses of gene-specific transcriptional control. Models will be evaluated for two master-regulators, *AICDA* and *PRDM1*, which are essential to the generation of antibody diversity and the differentiation of plasma cells.

4.1.2.1 Induction of AICDA drives antibody diversity downstream of activation signals

AICDA encodes the Activation-Induced Cytosine Deaminase (AID) enzyme, responsible for ensuring antibody diversity through somatic hypermutation and class-switch recombination of immunoglobulin genes (Muramatsu et al., 2000). *AICDA* is induced 48-60 hours following antigen encounter, through transcription factor binding at the gene promoter, and nearby regulatory elements downstream of activation stimuli (Pone et al., 2012; Zan and Casali, 2013). The TF-bound regulatory elements which control *AICDA* are well conserved between mice and humans and have been characterised in mice. CD40-ligation alongside TLR (Toll-like receptor) and BCR stimulation induces canonical and non-canonical NF-κB to upregulate *AICDA,* through direct binding at the gene promoter and a 5' enhancer (Tran et al., 2010). *AICDA* is also induced by BATF, acting at both at a 3' enhancer and the super-enhancer which spans *AICDA* and 5' gene *MFAP5* (Crouch et al., 2007; Ise et al., 2011; Lio et al., 2019). BATF has been shown to recruit TET, leading to demethylation of the super-enhancer and upregulation of *AICDA* (Lio et al., 2019). *AICDA* is also controlled in the B lineage by PAX5 and E2F at an intronic enhancer (Tran et al., 2010).

4.1.2.2 PRDM1 determines plasma cell differentiation downstream of activation induced IRF4 upregulation

PRDM1 encodes BLIMP1, the master regulatory of plasma cell differentiation. BLIMP1 promotes plasma cell fate primarily through transcriptional repression of B cell identity genes such as *BCL6*, *PAX5* and *SPIB* (Minnich et al., 2016; Shaffer et al., 2002; Turner et

al., 1994; Yu et al., 2000). *PRDM1* is repressed in B lymphocytes by B cell and GC transcription factors PAX5, SPIB, BACH2 and BCL6 (Calame, 2008; Nutt et al., 2011; Tellier and Nutt, 2019). BCL6 represses *PRDM1* directly, through binding intronic regulatory elements, and indirectly, through inhibiting the AP-1 activator which binds the promoter (Parekh et al., 2007; Shaffer et al., 2000; Tunyaplin et al., 2004; Vasanwala et al., 2002). The *PRDM1* promoter is also the site of direct transcriptional repression through PAX5 (Bullerwell et al., 2021; Mora-López et al., 2007); whereas SPIB and BACH2 supress transcription *via* proximal regulatory elements 5' of the gene (Ochiai et al., 2006; Schmidlin et al., 2008). Repression may also occur through the IRF8-PU.1 complex which binds the promoter and a 3' cis-regulatory element (Carotta et al., 2014).

As activated B cells differentiate to plasma cells, the *PRDM1* promoter and its regulatory elements are released from repression. The induction of IRF4, through CD40-dependent NF-κB activation and IL-21 dependent STAT3 activation, is a key event in the switch from PRDM1 repression to activation (Kwon et al., 2009; Saito et al., 2007). Transcription of *PRDM1* is driven through IRF4 activation at a proximal 3' enhancer, an enhancer in intron 5, and the *PRDM1* promoter (Klein et al., 2006; Kwon et al., 2009; Sciammas et al., 2006). SP1/SP3 and AP-1 also upregulate *PRDM1* at its promoter (Mora-López et al., 2008; Ohkubo et al., 2005).

Following IRF4-dependent induction, PRDM1 elevation is sustained in plasmablasts and plasma cells, where it functions in antigen presentation, antibody secretion and cellular stress responses including the unfolded protein response (Doody et al., 2007, 2006; Tellier et al., 2016). It has been shown that *PRDM1* maintenance occurs independent of IRF4 (Low et al., 2019); the transcription factors and cis-regulatory elements responsible for sustained expression remain unknown (Nutt et al., 2015). *PRDM1* is also considered a tumour suppressor gene in several B cell cancers including multiple myeloma and diffuse large B cell lymphoma (DLBCL). *PRDM1* has been associated with regulation by a downstream super-enhancer in a myeloma cell line (Lovén et al., 2013).

4.2 Aims and Objectives

This chapter aims to: 1, leverage cisREAD predictions to identify global changes in gene regulation which drive B cell differentiation; and 2, evaluate the utility of cisREAD to generate hypotheses of gene-specific transcriptional control. To achieve these aims, the following objectives will be met:

1.A to identify changes in transcription factor occupancy during B cell differentiation;

1.B to link dynamic transcription factor occupancy to changes in gene expression;

2.A to support predicted models of regulation for individual genes (AICDA and PRDM1) using transcription factor binding (ChIP-seq) and chromatin interaction (Hi-C) datasets; and

2.B to evaluate predictions in context with the literature.

4.3 Methods

4.3.1 Genome-wide analyses of transcriptional regulation

To meet aim 1, predictions from the cisREAD + HINT-ATAC run were subject to multiple global analyses, performed using command line tools and R.

4.3.1.1 Differential transcription factor footprinting

To identify changes in transcription factor binding site accessibility throughout the lineage, differential footprinting was performed with HINT-ATAC (Li et al., 2019). HINT-ATAC was supplied with PWMs for the 13 *de novo* transcription factors identified in chapter 3 (Figure 3.7.B), alongside differentiation stage specific BAM files for the 9 time-points. BED files were also supplied, giving the coordinates of differentially accessible regions (DARs). Differential footprinting was performed with the 'rgt-hint differential' command with the --bc option to perform bias correction relating to Tn5 cleavage. HINT-ATAC normalises ATAC-seq signal using DESeq2's median of the ratios methods, to control for differences in sequencing depth across samples (Love et al., 2014).

Following differential footprinting, TF activity scores were calculated by combining the 'protection scores' (differences in cleavage events between the footprints and flanking region) and 'openness-scores' (numbers of cleavage events around binding sites) from HINT-ATAC as described in Li *et al.*, 2019. Line-plots were also produced with HINT-ATAC, showing the mean ATAC-seq signal across all detected footprints in day 0, day3, day 6 and day 13 cell states.

To identify candidate regulators at *de novo* motifs which are common to multiple transcription factors, TF activity scores were tested for Pearson correlation with the $\log_2$ normalised expression of relevant genes, selected from the literature.

4.3.1.2 Transcription factor footprint enrichment in cis-regulatory element clusters

To investigate differentiation-stage specific transcription factor binding, k-means clustering of standardised $\log_2$ normalised chromatin accessibilities was performed for all cis-regulatory elements selected by LASSO regression to regulate a differentially expressed gene, with $k = 8$. k was chosen by incrementing the number of clusters until early and late ABC-expressed genes were separated. Following the identification of cell-stage specific regulatory clusters, TF occupancy enrichment was calculated for the 13 de novo motifs from Figure 3.6B (chapter 3). A cis-regulatory element was considered occupied if a transcription factor footprint was detected at any cell stage. Enrichment was calculated using a two-sided Fisher test for each TF-cluster combination, comparing whether the TF occupancy rate of a cluster is significantly greater than (enriched) or lesser than (depleted) the TF occupancy rate of all other clusters. To ensure robustness to TF-binding site prediction method, this analysis was repeated with gene-specific cis-regulatory elements from the cisREAD + BMO run.

De novo motifs enriched in each cluster were discovered using HOMER findMotifsGenome.pl, compared to a background of non-differentially accessible regions (Heinz et al., 2010). This was performed to identify motifs that are over-represented in each temporal cluster, in addition to those discovered across the whole process in chapter 3.

CREs in each cluster were annotated with HOMER annotatePeaks.pl to derive annotations and distance to nearest gene. GC content was calculated using the bedtools nuc command (Quinlan and Hall, 2010).

4.3.1.3 Transcription factor footprint enrichment in cis-regulatory elements, linked to gene co-expression modules

To explore gene expression dynamics across differentiation, a gene co-expression network was constructed by Dr Matthew Care using the Parsimonious Gene Correlation Network Analysis (PGCNA) method (Care et al., 2019) . PGCNA constructs gene co-expression networks, where genes (represented as nodes) are connected by edges (weighted by correlation) to the three most correlated genes. After network construction, correlated genes were grouped into 'Modules'. Modules are assumed to represent genes which are co-regulated by the same TF, functionally similar, or involved the same biological processes (van Dam et al., 2018).

PGCNA was applied to the gene expression data shown in Figure 3.2, alongside 3 additional samples (6h x 2, day 6 x1) without matching ATAC-seq. This dataset was analysed by Dr Matthew Care using DESeq2, identifying 16,296 differentially expressed genes (LRT; BH-FDR 0.01) (Love et al., 2014). The VST normalised expression data for the DEG were analysed with PGCNA2 [settings -n 1000, -f1, -b 100] (https://github.com/medmaca/PGCNA/tree/master/PGCNA2) selecting the best clustering using scaled cluster enrichment score. This gave a network with 16,296 nodes and 57,175 edges. A total of 23 modules were identified.

Module names were derived by Dr Matthew Care from gene-set over-representation analysis (FDR < 0.1), performed using 41,811 gene signatures curated from the Staudt lab, CORUM, MSigDB, UniProt, Gene Ontology and in-house databases (Ashburner et al., 2000; Liberzon et al., 2015; Shaffer et al., 2006). Enrichment of modules for signatures was assessed using a hypergeometric test, where the draw is the module genes, the successes were the signature genes, and the population were the genes in the mRNA count matrix.

Upon receipt of PGCNA module membership and names, transcription factors were linked to PGCNA modules by identifying TF footprints in cis-regulatory elements

selected to regulate expression of genes in each module. TF enrichment was then calculated using a two-sided Fisher test, comparing TF occupancy in CREs linked to that module to the occupancy of CREs not linked to that module. This analysis was also repeated with gene-specific cis-regulatory elements from the cisREAD + BMO dataset.

4.3.1.4 Exploration of PU.1/SPIB and AP-1 occupied cis-regulatory elements and target genes

To explore predicted regulation by PU.1 and AP-1 factors, occupancy at predicted binding sites was first confirmed using ChIP-seq. ChIP-seq peaks targeting PU.1 and BATF in GM12878 (Epstein-Barr Virus transformed B lymphoblastoid cells)were downloaded from ENCODE, under accessions ENCFF492ZRZ and ENCFF728KFD (ENCODE Project Consortium et al., 2020). SPIB binding sites were obtained from the union of ChIP-seq peaks called in diffuse large B cell lymphoma cell lines OCILy3 and OCILy10 (macs2 q value < 0.01) using data from Care et al. 2014. The SPIB datasets were realigned to hg38 and processed as described in (Care et al., 2014) by Dr Matthew Care. Binding sites for each factor were then intersected with DARs using bedtools intersect (Quinlan and Hall, 2010). Both GM12878 and OCILy cell lines are developmentally equivalent to day 3 activated B cells.

PU.1/SPIB and AP-1 targets were defined as genes with selected regulatory elements, with PU.1/SPIB and/or AP-1 footprints, whose accessibility was positively correlated with expression. These genes were k-means clustered by expression (*k*=5, incremented until early and late ABC clusters were separated). Enrichment of ChIP-seq PU.1, SPIB and BATF binding sites was calculated for cis-regulatory elements with PU.1/SPIB and/or AP-1 footprints, linked to each of the 5 expression clusters with a one-sided Fisher test. For each cluster, TF occupancy was compared at CREs with footprints for either PU.1/SPIB, PU.1/SPIB + AP-1 or AP-1, to all other DARs without the footprint(s).

Genes in each expression cluster, linked to either PU.1/SPIB, PU.1/SPIB + AP-1, or AP-1 footprints, were tested for enrichment of Gene Ontology (GO) biological processes (Ashburner et al., 2000; Carbon et al., 2021), Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto, 2000), Hallmark Molecular Signature DataBase (MSigDB) gene signatures (Liberzon et al., 2015) and Staudt lab gene signatures (Shaffer et al., 2006). Enrichment was first calculated using a one-sided

Fisher test, to test gene-set over-representation in TF targets in a cluster, compared to all other clusters (targeted by any of the 3 TFs.) Gene sets enriched with BH-adjusted p < 0.1 were considered significantly enriched. This was used to define potential 'temporally regulated' gene sets. Enrichment was also performed against genes with a similar expression profile without the given TF footprint(s) to identify gene sets preferentially regulated by a TF at a given differentiation stage ('TF-specific regulated').

To obtain background sets for gene set over-representation, similarly expressed genes were identified by training an XGBoost classifier (using the 'XGBoost' R package) on the five expression clusters (Chen and Guestrin, 2016). This was used to predict the cluster label for all other differentially expressed genes linked to CREs without the TF footprint(s). An XGBoost model, using the 'multi:softprob' objective function, was trained by five-fold cross validation (stratified folds), to identify the number of decision trees resulting in the lowest mean multi-classification error ('nrounds' = 47.) Genes were assigned a cluster label if the XGBoost-predicted probability of belonging to a class exceeded 0.9.

4.3.2 Evaluation of gene-specific models

To meet aim 2, gene-specific models from the cisREAD + HINT-ATAC set were evaluated for *AICDA* and *PRDM1* genes. To aid validation , ATAC-seq, ChIP-seq and Hi-C datasets were visualised using the WashU Epigenome browser (Li et al., 2022). ATAC-seq signal tracks were generated for the 9 *in vitro* B cell time points. BAMs were pooled across all replicates using samtools merge (Li et al., 2009), and bias-corrected, normalised bigwig signal tracks were produced using the 'rgt-hint tracks' command from HINT-ATAC (Li et al., 2019).

To evaluate *AICDA* chromatin contacts Knight-Ruiz (KR) normalised Hi-C data for GM12878 cells, aligned to hg19, was selected for visualised within the WashU epigenome browser (Knight and Ruiz, 2013; Rao et al., 2014). To evaluate *PRDM1* chromatin contacts, primary human plasma cell Hi-C data from Vilarrasa-Blasi et al., 2021 was obtained from the European Genome Archive (EGA). Filtered, valid read pairs, aligned to hg38 as described in Vilarrasa-Blasi *et al.*, 2021 were converted to .hic

files using HOMER (Heinz et al., 2010), and subject to Knight-Ruiz matrix balancing normalisation (Knight and Ruiz, 2013) within the WashU epigenome browser.

Additional ATAC-seq and ChIP-seq datasets for GM12878 cells were downloaded from the ENCODE consortium for evaluation of transcription factor occupancy (ENCODE Project Consortium et al., 2020). Details of all additional datasets used are shown in Table 4.1.

Table 4.1 Additional datasets used to evaluate gene-specific models. GM12878 B lymphoblastoid cells are used as surrogate for in vitro activated B cells.

Datatype	Sample	Target	Genome build	Source	Accession	Model evaluation
ATAC-seq	GM12878	-	hg38	ENCODE	ENCFF667MDI	AICDA, PRDM1
ChIP-seq	GM12878	RELA (NF-κB)	hg38	ENCODE	ENCFF149XND	AICDA
ChIP-seq	GM12878	RELB (NF-κB)	hg38	ENCODE	ENCFF936FAN	AICDA
ChIP-seq	GM12878	BATF	hg38	ENCODE	ENCFF716MIY	AICDA
ChIP-seq	GM12878	IRF4	hg38	ENCODE	ENCFF167KPF	AICDA, PRDM1
ChIP-seq	GM12878	PAX5	hg38	ENCODE	ENCFF886UGF	AICDA, PRDM1
ChIP-seq	GM12878	RUNX3	hg38	ENCODE	ENCFF857YLX	AICDA, PRDM1
ChIP-seq	GM12878	OCT2	hg38	ENCODE	ENCFF803HIP	PRDM1
Hi-C	GM12878	-	hg37	WashU (data from Rao et al. 2014)	GM12878_1 in_situ_1combined	AICDA
Hi-C	Primary plasma cells	-	hg38	EGA (data from Vilarrasa-Blasi *et al.*, 2021)	EGAD00001006486	PRDM1

4.4 Results and Discussion

4.4.1 Data integration with cisREAD reveals global changes in B cell gene regulation during differentiation

To identify global changes in gene regulation, genome-wide analyses were performed. These identified temporal patterns of transcription factor binding, and linked TF occupancy to differential accessibility, modular gene expression and functional pathways.

4.4.1.1 Dynamic occupancy of core transcription factors

During step 2 of cisREAD (described in chapter 3), transcription factor footprints were used to infer occupancy at 13 *de novo* motifs, which were enriched in differential chromatin regions (Figure 3.7.B). To assess global changes in TF occupancy at each motif through differentiation, differential footprinting was performed. This compared normalised footprint depth and accessibility between cell states. Differential footprints were visualised as line-plots, showing the chromatin accessibility of each binding site, averaged across all footprints in B cell (BC), activated B cell (ABC), plasmablast (PB) and plasma cell (PC) stages (Figure 4.3.A). Here transcription factor footprints were visible as dips in accessibility surrounding the motif where Tn5 cleavage was blocked by occupancy.

Stably binding factors with high DNA residency times (e.g., CTCF) exhibited strong footprints, with a marked difference in cleavages between the binding site itself, and flanking regions. Transiently binding factors (e.g., STAT3 and NF-κB) left much shallower footprints, where cleavage was only slightly depleted at the binding site.

For each motif, transcription factor activity scores, combining chromatin accessibility and transcription factor footprint strength, showed changes in occupancy across the time-course (Figure 4.3.B). Since many motifs are capable of occupancy by related transcription factors, TF activity was correlated with TF gene expression to identify candidate TFs which bind shared motifs (Figure 4.3.C).

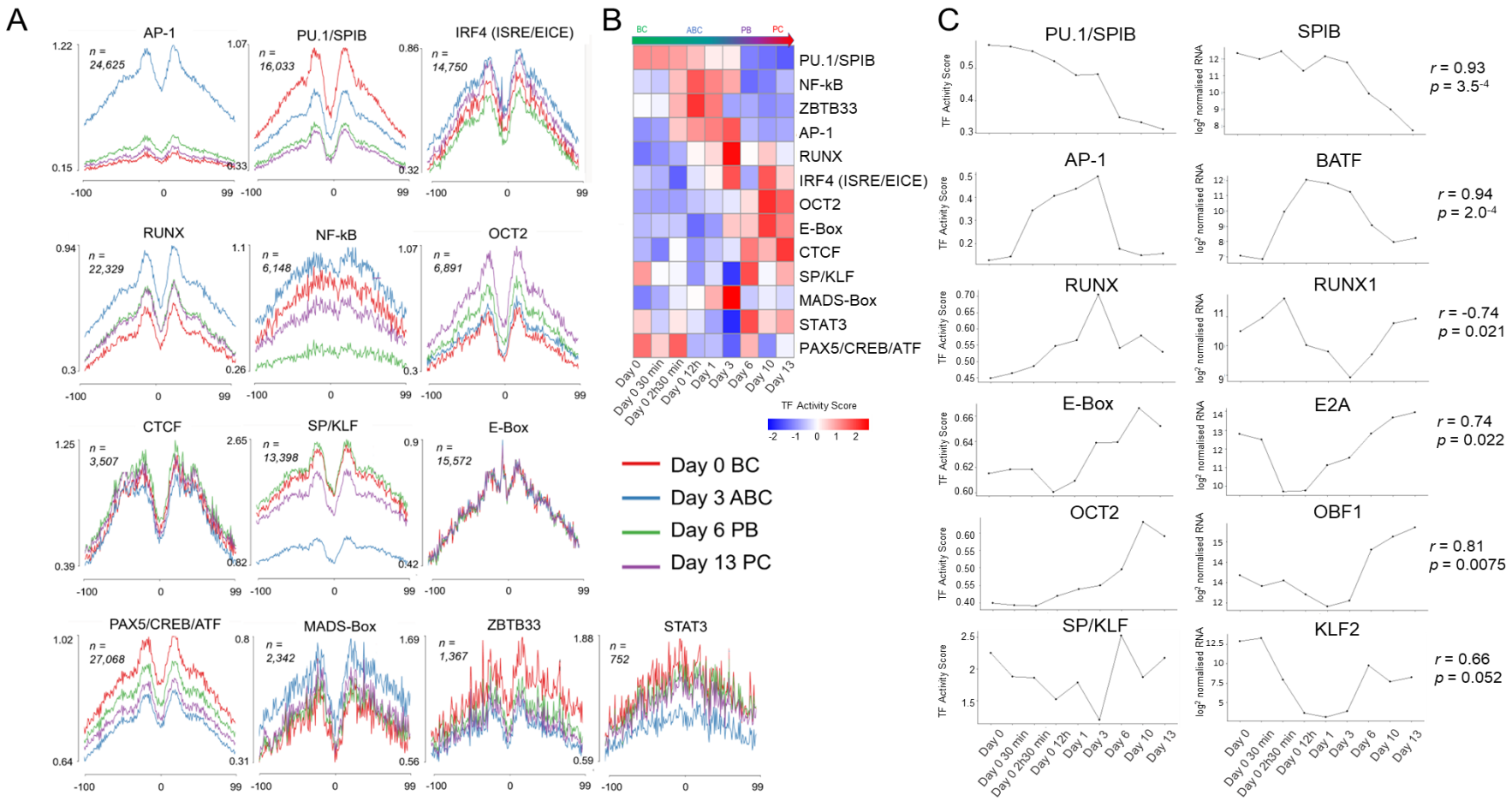


Figure 4.3 Differential transcription factor footprinting. A) TF footprint line-plots showing the mean normalised chromatin accessibility across all footprints in the 200bp window surrounding each de novo motif in day 0, day 3, day 6 and day 13 samples. n gives the total number of footprints called across the 4 time-points. Line-plots show changes in binding site accessibility and occupancy strength during transition from B cells, to activated B cells, then to plasmablasts and plasma cells. B) TF activity scores (z-score) at nine timepoints for each de novo motif, combining footprint accessibility and strength. C) Significant Pearson correlations ($p < 0.1$) between TF activity and gene expression of the TF, plots show mean accessibility and expression for each TF/gene across the time-course.

4.4.1.1.1 Changes in occupancy with B cell activation

Upon activation by CD40 and BCR stimulation at day 0, PU.1/SPIB footprints lost accessibility and strength (Figures 4.3.A and 4.3.B). The gradual loss of PU.1/SPIB activity between day 0 and day 3 was strongly correlated with the loss in *SPIB* expression (Figure 4.3.C). PU.1/SPIB loss coincided with gains in ZBTB33, NF-κB and AP-1 activity in the hours immediately following activation. ZBTB33 induction was transient, and declined after day 1, whilst NF-κB and AP-1 activity dropped upon withdrawal of CD40L at day 3. NF-κB accessibility increased slightly in plasma cells, following the addition of APRIL at day 6. The elevation of AP-1 activity in activated B cells was strongly correlated with the expression of AP-1 family member *BATF*.

MADs-Box, IRF4 and RUNX factors showed a delayed increase in response to activation stimuli, but all peaked sharply in day 3. The induction of RUNX activity was anticorrelated with *RUNX1* expression.

4.4.1.1.2 Changes in occupancy with plasma cell differentiation

Whilst the derivation of days 0-3 from total B cells, and days 6-13 from memory B cells, hindered direct comparison of regulation in the plasmablast transition, IRF4, E-Box, OCT2 and CTCF activities all increased with plasma cell differentiation. E-Box footprints showed no net change in accessibility, yet their TF activity score increased from day 3 due to increased footprint depth (Figure 4.3.B). The increase in E-Box activity was correlated with expression of *E2A*, noted to be the dominant E-protein associated with plasma cell regulation (Gloury et al., 2016). Elevated OCT2 activity in plasmablasts and plasma cells was strongly correlated with expression of the gene encoding the *OBF1* binding partner.

4.4.1.1.3 Occupancy linked to both B cells and plasma cells

SP/KLF, PAX5/CREB/ATF and STAT3 footprints lost accessibility upon activation and proliferation and gained accessibility upon the plasmablast transition and cell cycle exit. of SP/KLF activity was correlated with expression of the B cell quiescence factor KLF2 (Winkelmann *et al.*, 2011).

4.4.1.2 Cluster analysis of gene-specific cis-regulatory elements reveals TF binding dynamics across differentiation

To interrogate regulatory dynamics on a genome-wide scale, selected gene-specific CREs were divided into 8 regulatory clusters, based on differentiation-stage specific chromatin accessibility, using k-means clustering (Figure 4.4.A). *k* was incremented from 3 onwards until early-stage and late-stage ABC-specific clusters were separated. For each cluster, enrichment of core TF footprints (Figure 4.4.B) and *de novo* motifs (Figure 4.4.C) characterised transcriptional regulators which drove gene regulation at each stage.

4.4.1.2.1 Changes in accessibility and motif occupancy with B cell activation

Cluster 1 CREs were accessible prior to B cell activation, and lost accessibility at 2h30-24 hours post activation (Figure 4.4B). PU.1/SPIB footprints were uniquely enriched in this cluster (occupying 51% of CREs, Figure 4.4A)), alongside and a *de novo* motif matching the PAX5 half site (Cobaleda et al., 2007), discovered in 14% of CREs (Figure 4.4C). Cluster 2 CREs were also accessible in B cells, yet gained further accessibility immediately after stimulation, and remained accessible until the plasmablast transition at day 6 (Figure 4.4B). These B cell specific CREs were also enriched in PU.1/SPIB footprints, present in 39% of elements, but were additionally enriched in footprints for NF-κB (14%), ZBTB33 (4%) and AP-1 factors (32%) (Figure 4.4A), which were previously associated with activation in Figure 4.3

PU.1/SPIB occupancy was lost in Cluster 3 CREs, which were lowly accessible in B cells, and gained accessibility 2h30 post activation (Figure 4.4B). Instead, occupancy was dominated by AP-1 (65%), alongside regulation by ZBTB33 and NF-κB (Figure 4.4A). AP-1 occupancy (53%) continued into late-ABC cluster 4, which gained accessibility at 12h post activation (Figure 4.4B). *De novo* motif discovery found enrichment of a BATF-IRF4 composite in 17% of CREs, and RUNX footprints (33%) were also enriched (Figure 4.4C).

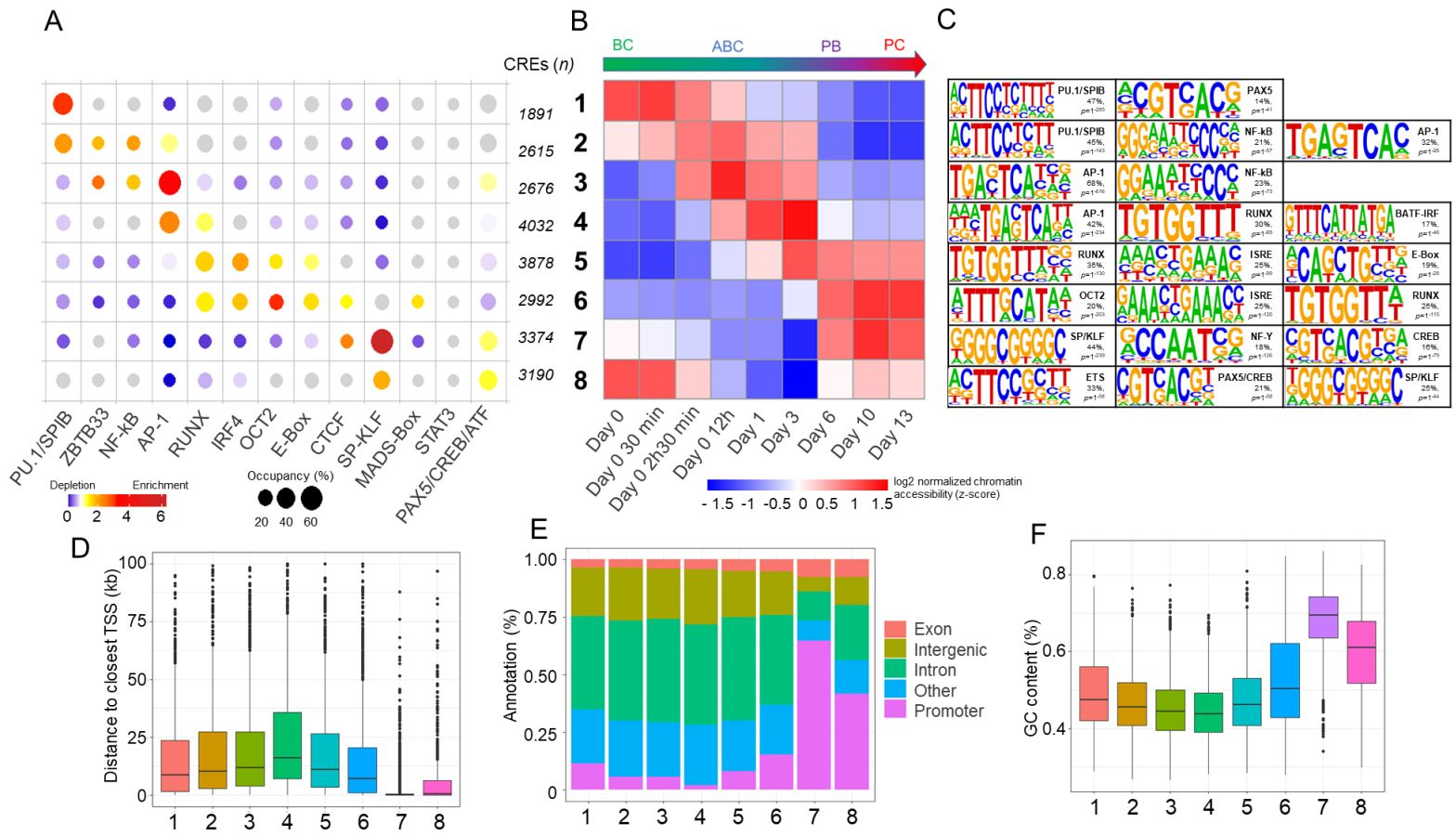


Figure 4.4 Enrichment of footprints and *de novo* motifs in cis-regulatory clusters. A) Bubbleplot showing enrichment of TF occupancy in each cluster. Size of bubbles gives the proportion of each cluster harbouring a TF footprint, colour shows significant ($p < 0.05$, two-sided Fisher test) enrichment (fold-change between cluster and other clusters > 1, red) or depletion (fold-change between cluster and other clusters < 1, blue), grey represents no significant enrichment. N gives the number of CREs in each cluster. B) Heatmap showing mean $\log_2$ normalised chromatin accessibility (z-score) of cis-regulatory elements significantly linked to gene expression, k-means clustered (k = 8). C) *De novo* motifs enriched in each cluster from HOMER ($p < 0.05$, occupancy > 10%). D) Boxplot showing distance of CREs to their nearest gene in each cluster. E) Genomic annotation of each CRE by HOMER. F) Boxplot showing GC content of CREs in each cluster.

4.4.1.2.2 Changes in accessibility and motif occupancy with plasma cell differentiation

Cluster 5 CREs gained accessibility around 1-3 days after activation, and sustained accessibility throughout plasma cell differentiation (Figure 4.4B). CREs in this cluster were enriched in occupancy by RUNX (41%), IRF4 (28%), E-Box (20%) and OCT2 (12%) footprints (Figure 4.4A). *De novo* motif analysis also reported significant enrichment of these factors and discovered an ISRE site for IRF4. RUNX (36%), IRF4 (25%), E-Box (24%) and OCT2 (20%) were also enriched in PB/PC specific cluster 6, alongside CTCF (6%) and MADS-Box factors (5%) (Figure 4.4C).

4.4.1.2.3 Accessibility and motif occupancy linked to both B cells and plasma cells

Cluster 7 CREs were accessible in plasmablasts and plasma cells yet exhibited moderate accessibility prior to activation (Figure 4.4B). Notably CREs in this cluster were close to their target gene (Figure 4.4.D), highly GC rich (Figure 4.4.E), enriched for gene promoters (Figure 4.4.F). This cluster exhibited strong enrichment for SP/KLF factors, which occupied 67% of CREs in this cluster, accompanied by occupancy by CTCF (8%) and PAX5/CREB/ATF (37%) (Figure 4.4A). *De novo* motif analysis revealed enrichment of additional promoter-associated motifs including NF-Y and CREB which conform to elements of previously defined XBP1 binding motifs (Figure 4.4C) (Acosta-alvear et al., 2007; Cocco et al., 2020).

The accessibility of cluster 8 CRE similarly declined in ABCs, however accessibility was greater in BCs and PCs (Figure 4.4B). Like cluster 7, this class was overrepresented for promoters and GC-rich elements, although to a lesser extent (Figures 4.4E and 4.4F),. SP/KLF (34%) and PAX5/CREB/ATF (40%) footprints were also enriched in this cluster (Figure 4.4A), and *de novo* motif enrichment showed enrichment of a PAX5/CREB binding site and an ETS motif, different to the PU.1/SPIB motif in earlier clusters (Figure 4.4C).

In order to alleviate concerns over the reliability of footprinting for transient binding factors, the clustering analysis was repeated with TF binding events predicted by a footprint independent method BMO (described in chapter 3) (D'Oliveira Albanus et al.,

2021). Replication with BMO binding sites showed highly concordant results (Appendix 4.1).

4.4.1.3 Parsimonious Gene Co-expression Network Analysis shows differential transcriptional regulation of temporally distinct biological pathways

To investigate how differential transcription factor occupancy controlled gene expression dynamics, gene-specific CREs were linked to co-expression modules identified using the Parsimonious Gene Co-Expression Network Analysis (PGCNA) method (Care et al., 2019). A total of 23 gene co-expression modules were identified, each enriched in differential biological processes, signalling pathways and transcription factor targets.

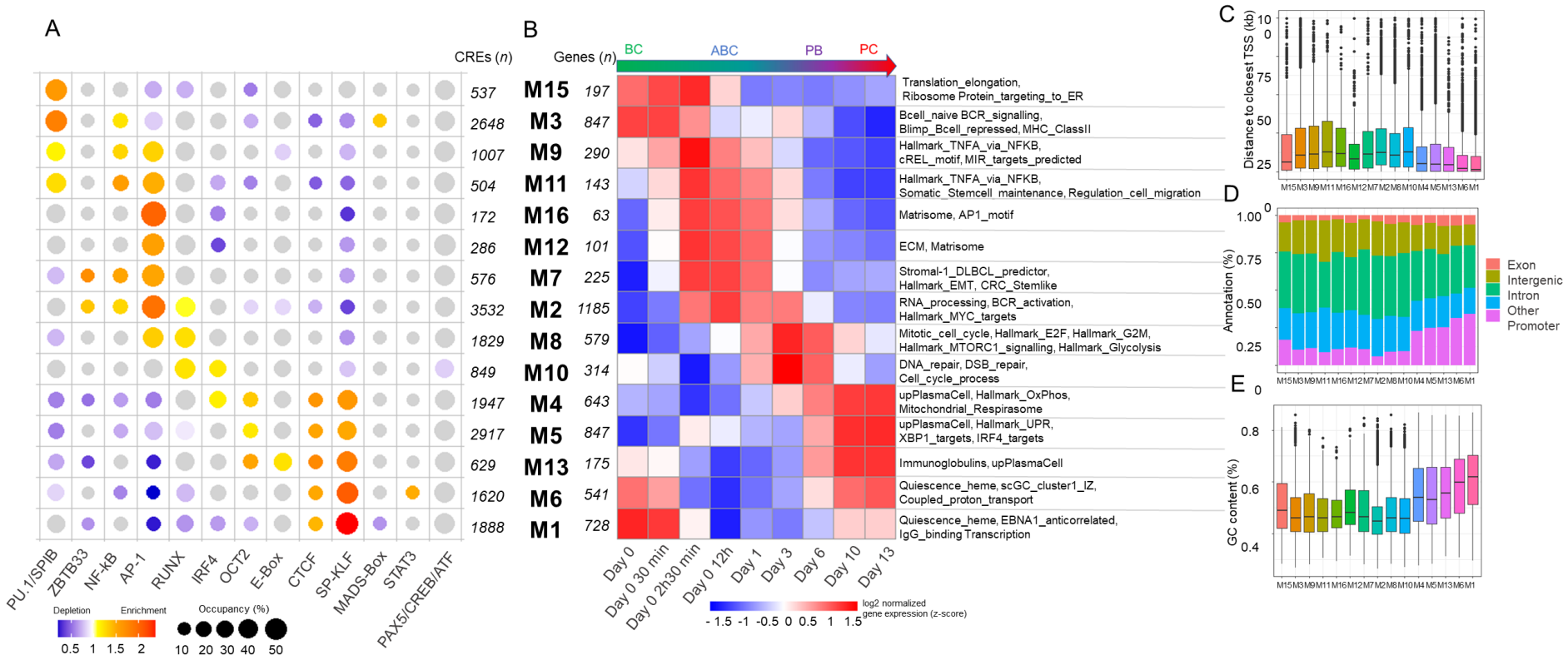


Figure 4.5 Enrichment of footprints in cis-regulatory elements linked to PGCNA co-expression modules M1-16. A) Bubbleplot showing enrichment of TF occupancy in cis-regulatory elements assigned to genes in PGCNA expression modules. Size of bubbles gives the proportion of each cluster harbouring a TF footprint, colour shows significant ($p < 0.05$, two-sided Fisher test) enrichment (fold-change between genes in module and genes not in module > 1, red) or depletion (fold-change < 1, blue), grey represents no significant enrichment. N gives the number of genes with significant CREs in each module. B) Heatmap showing mean $\log_2$ normalised gene expression (z-score) of genes with significantly linked CREs, module names reflect enriched gene sets in each module. Gene set names summarise enriched gene signatures determined by hypergeometric test (FDR < 0.1) C) Boxplot showing distance of CREs to their nearest gene in each cluster. D) Genomic annotation of each CRE by HOMER. E) Boxplot showing GC content of CREs in each cluster.

To identify regulatory inputs into temporal gene expression programs, TF footprint enrichment was calculated within cis-regulatory elements (Figure 4.5.A) that were linked to upregulation of gene co-expression modules (Figure 4.5.B). Only modules with significant TF enrichment are shown. TF enrichment was not observed in the remaining 8 modules due to the small number of DEGs in these modules.

4.4.1.3.1 Changes in RNA levels with B cell activation

Genes with B cell-specific expression, whose accessibility declined in the hours after activation, were placed in module 15 (M15). This module was enriched for ribosomal protein gene sets and was linked to CREs with unique, strong enrichment for PU.1/SPIB occupancy (35%). PU.1/SPIB was further enriched in CREs linked to module 3 (40%), which were highly expressed in B cells, and moderately expressed on activation. NF- κB (10%) and MADS-Box (4%) factors also contributed regulation to this module, which was enriched for gene signatures relating to naïve B cells, BCR signalling, repression by BLIMP1 and the major histocompatibility (MHC) Class II complex.

Genes in modules 9 and 11 were moderately expressed in B cells, but highly expressed immediately after activation. This module was enriched for gene sets relating to NF- κB signalling and linked to CREs with strong enrichment for NF- κB enrichment (11% and 13%) alongside PU.1/SPIB (25% and 27%) and AP-1 (33% and 38%).

AP-1 was the dominant regulator of activated B cell modules and was solely linked to matrisome and AP-1 motif modules 16 (52%) and 12 (40%), which were sharply upregulated 2h30-24 hours post activation. AP-1 was also enriched with 38% occupancy for the similarly expressed stromal module 7, alongside NF-κB at 13% and ZBTB33 at 4% of CREs. Enrichment of AP-1 (43%), NF-κB (10%) and ZBTB33 (33%) was also observed for the BCR activation/MYC target module 2, which was upregulated until day3, and enriched for RUNX occupancy (29%).

4.4.1.3.2 Changes in RNA levels with plasma cell differentiation

Modules 8 and 10 were upregulated in late-ABCs and plasmablasts and were enriched for cell cycle/MTORC1/glycolysis and cell cycle/DNA repair functions, respectively. These were linked to regulation by AP-1 (33% at module 8), RUNX (33% at modules 8 and 10) and IRF4 (20% module 10). IRF4, OCT2, E-Box CTCF and SP/KLF were associated with regulation of plasma cell genes in modules 4, 5 and 11. Specifically, IRF4 was linked to oxidative phosphorylation module 4 (18%), and E-Box was linked to immunoglobulin module 11 (21%).

4.4.1.3.3 RNA levels linked to both B cells and plasma cells

Modules 6 and 1 were downregulated with activation and proliferation and upregulated in both B cells and plasma cells. Both modules were enriched in gene sets linked to quiescence, and had more GC-rich, proximal and promoter elements than other modules (Figures 4.5.C, D and E). These genes were preferentially upregulated by SP/KLF (35% and 40%) and CTCF (6% for both), module 6 was also linked to STAT3 (2%). A repeat of the analysis with BMO-derived occupancy predictions yielded concordant results (Appendix 4.2).

4.4.1.4 Discussion of dynamic gene regulatory networks during B cell differentiation

Integration of epigenomic and transcriptomic datasets have been instrumental in uncovering how the chromatin environment shapes B cell populations, and responds to stimuli to determine cell fate on a genome wide scale (Barwick et al., 2018; Bunting et al., 2016; Chaudhri et al., 2020; Cocco et al., 2020; Scharer et al., 2018; Vilarrasa-Blasi et al., 2021). Studies into murine T cell-independent dynamics, have previously found increased accessibility and/or hypomethylation of NF-κB, AP-1, IRF4, OCT2, E-box and MADS-Box motifs upon activation with lipopolysaccharide (LPS), coinciding with downregulation of PU.1/SPIB sites (Barwick et al., 2018, 2016). The analyses of *in vitro* B cell regulation recapitulated many of the same observations under T cell dependent stimuli in human cells, whilst also characterising additional regulators including RUNX(3) and ZBTB33. Taken together, Figures 4.3, 4.4 and 4.5 suggest the

following regulatory dynamics (Figure 4.6).

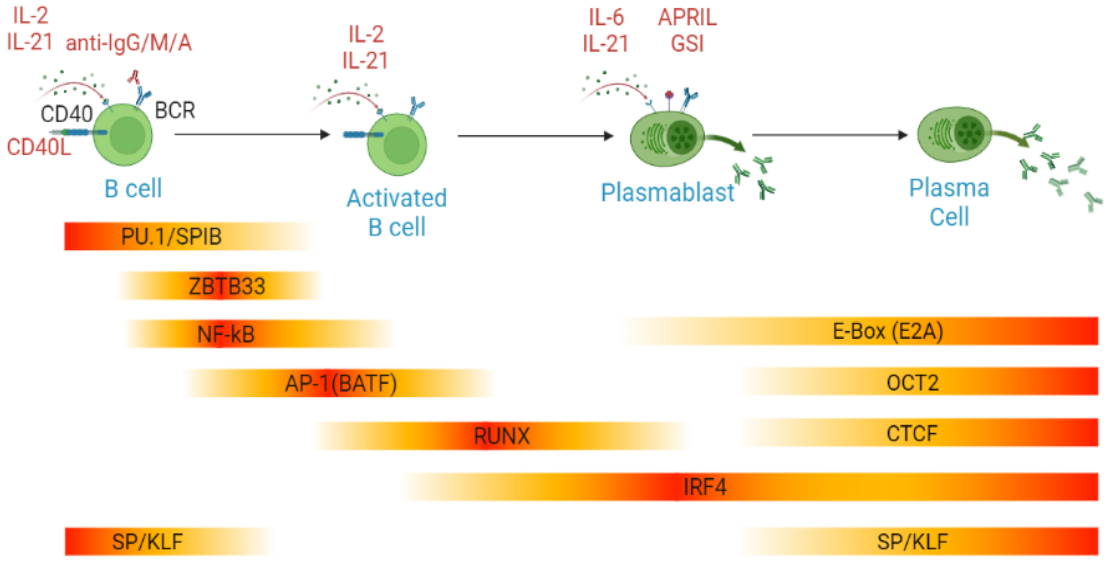


Figure 4.6 Regulatory dynamics for core transcription factors suggested by analysis of in vitro B cell differentiation ATAC-seq and RNA-seq. Diagram indicates temporal occupancy of key transcription factors in relation to differentiation stimuli. Figure created with BioRender.

The analysis supports a model where gene regulation shifts from PU.1/SPIB towards AP-1 upon B cell activation. Results suggest that AP-1 augments the initial regulatory response, propagated through NF-κB, by upregulation of NF-κB signalling, and ZBTB33 may also contribute regulation at early response elements (Figures 4.3 and 4.4), independent of NF-κB (Figure 4.5).

Whilst interpretation of the plasmablast transition is complicated by the differing start points of day 0-3 and day 6-13 samples, the data suggest that control shifts from AP-1 (alone and at AICE motifs) to RUNX and IRF4; co-ordinating sequential expression of MTORC1, cell cycle, and oxidative phosphorylation genes (Figure 4.5). There is evidence from the motifs in Figure 4.4.C that IRF4 occupancy shifts from AICE to ISRE sites during this transition. Regulation in plasmablasts and plasma cells is also contributed by OCT2, E-Box and CTCF factors.

The analysis supports the roles of KLF factors in B cell quiescence (Cao et al., 2010; Winkelmann et al., 2011), shown by the downregulation of SP/KLF motifs in cycling ABCs (Figures 4.3 and 4.4) and association with quiescence gene modules (Figure 4.5). However enrichment of the ubiquitous SP1 activator at highly accessible and GC rich promoter regions (such as those in SP/KLF enriched clusters/modules in Figures 4.4 and 4.5) might also contribute to this pattern (Hasegawa and Struhl, 2021).

4.4.1.5 Gene regulation shifts from PU.1/SPIB to AP-1 during B cell activation

The results in Figures 4.3, 4.4 and 4.5 revealed an overwhelming association between PU.1/SPIB with the B cell state, and AP-1 with the activated B cell state. To further investigate how the shift in regulatory inputs shapes B lineage expression programmes, gene-specific models were used to predict PU.1/SPIB and AP-1 target genes. ChIP-seq data was then indicated to implicate individual PU.1/SPIB and AP-1 family members with occupancy and gene regulation at distinct stages of differentiation.

4.4.1.5.1 A PU.1/SPIB-BATF gradient co-ordinates B cell activation

PU.1/SPIB and AP-1 targets were defined as genes with selected CREs, which were footprinted by PU.1/SPIB and/or AP-1, and with accessibility which positively correlated with gene expression. Predicted PU.1/SPIB and AP-1 target genes were collectively clustered by their expression using k-means clustering (k=5, incremented until early-stage and late-stage ABC expression clusters were separated) (Figure 4.7.A). Each cluster was then divided by predicted regulation; by linkage to CREs with PU.1/SPIB footprints, AP-1 footprints or both PU.1/SPIB and AP-1 footprints (either in the same CRE or in separate CREs). The chromatin accessibility of cisREAD-predicted CREs at each cluster was then visualised (Figure 4.7.B). Similarly, cis-elements were divided by the presence of footprints for PU.1/SPIB only, AP-1 only, and both PU.1/SPIB and AP-1 (at the same CRE).

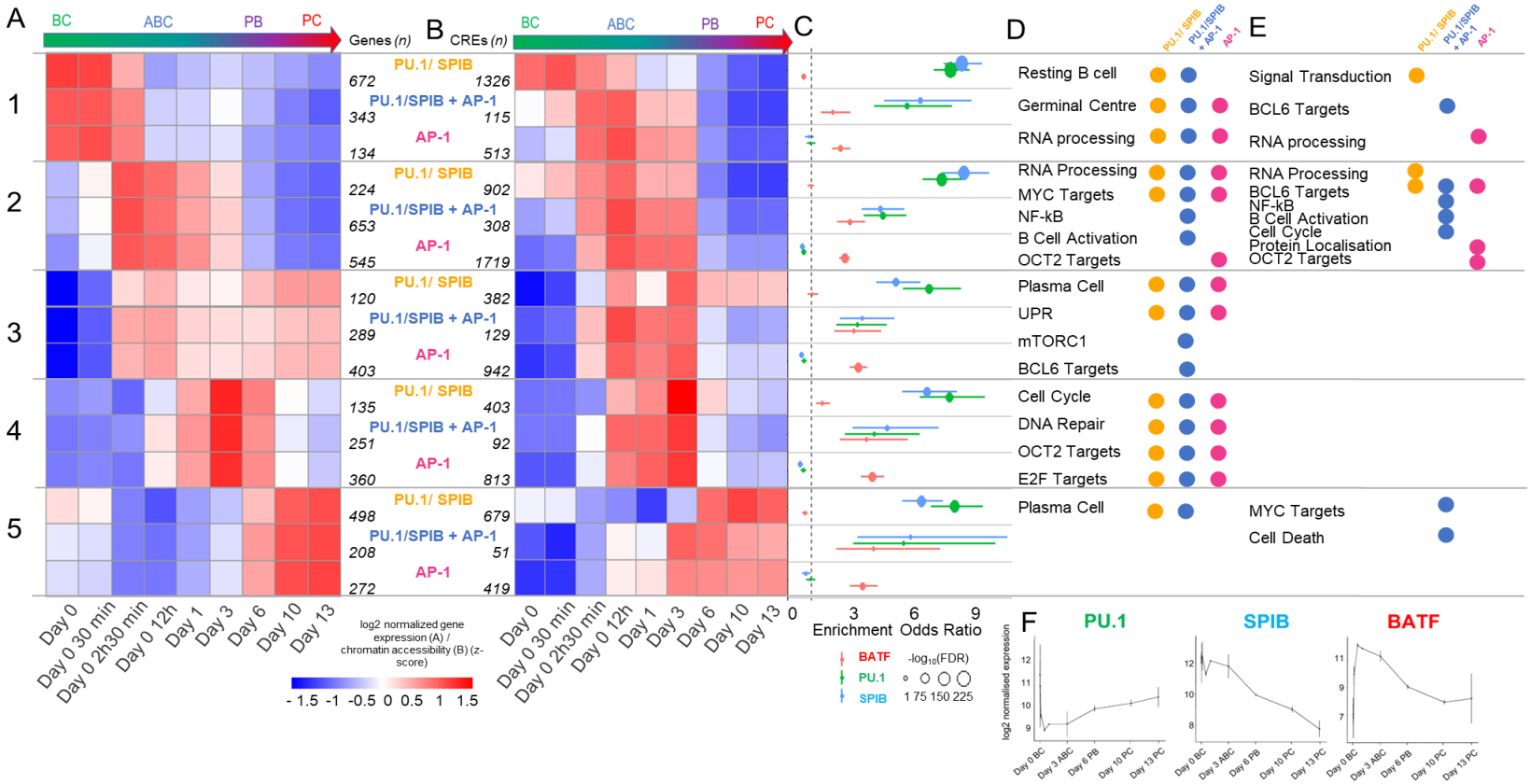


Figure 4.7 PU.1/SPIB and AP-1 target genes, linked to cis-regulatory elements, PU.1/SPIB and BATF occupancy and functional pathways. A) Heatmap showing mean log_2 normalised expression of genes linked by cisREAD to CREs with PU.1/SPIB and/or AP-1 footprints, k-means clustered by expression (k=5). Each cluster is separated by a line. Within each cluster genes are divided based on predicted regulation by PU.1/SPIB only, PU.1/SPIB and AP-1 (either in the same CRE or at different CREs) or AP-1 only. B) Heatmap showing the mean log_2 normalised chromatin accessibility of CREs linked to gene expression clusters (divided by lines), annotated by the presence of footprints for PU.1/SPIB only, PU.1/SPIB and AP-1 at the same CRE, or AP-1 only. C) Enrichment of ChIP-seq binding sites for SPIB (from OCILy-3/OCILy-10, blue), PU.1 (from GM12878, green), and BATF (an AP-1 factor, from GM12878, red) at CREs with footprints for PU.1/SPIB and/or AP-1, linked to each expression cluster, compared to all DARs without the given footprint. D) Gene sets enriched (FDR < 0.1) in genes in each cluster, linked to regulatory elements with PU.1/SPIB and/or AP-1 footprints, compared to genes in other clusters. E) Gene sets enriched (FDR < 0.1) in genes in each cluster, linked to regulatory elements with PU.1/SPIB and/or AP-1 footprints, compared to genes with similar expression patterns, not linked to given footprints F) log_2 normalised PU.1, SPIB and BATF expression.

Each cluster in Figure 4.7.A was assessed by the proportion of genes predicted to be upregulated by PU.1/SPIB or AP-1 uniquely or in combination. It was observed that most B cell upregulated genes were linked to PU.1/SPIB footprints, which were most accessible in B cell time-points (cluster 1). Conversely, most ABC upregulated genes were linked to AP-1 footprints which were accessible at ABC time-points (clusters 2,3 and 4). Almost half of genes upregulated in the hours following CD40 and BCR stimulation were predicted targets for both PU.1/SPIB and AP-1 cluster 2. PU.1/SPIB and AP-1 were predicted to regulate expression of these genes at different sets of regulatory elements, as indicated by the comparatively small set of cis-elements with both PU.1/SPIB and AP-1 footprints.

To confirm transcription factor occupancy at PU.1/SPIB footprints, existing ChIP-seq data was integrated (Figure 4.C). PU.1 ChIP-seq data came from lymphoblastoid cell line GM12878 (ENCODE Project Consortium et al., 2020), and SPIB ChIP-seq from DLBCL cell lines OCILy-3 and OCILy-10, which are surrogates for different activated B cell states (Care et al., 2014).

Both PU.1 and SPIB ChIP-seq binding sites significantly overlapped CREs with PU.1/SPIB footprints, but not CREs with only AP-1 footprints (Figure 4.7.D). Furthermore, the effect size of PU.1 and SPIB enrichment varied by cluster; PU.1 binding sites were most enriched at CREs with plasmablast/plasma cell accessibility (linked to clusters 3, 4 and 5) and SPIB binding sites were most enriched at CREs with activated B cell accessibility (linked to clusters 1 and 2). This may reflect increased *SPIB* in activated B cells and increased *PU.1* mRNA in PBs/PCs (Figure 4.7.F).

Amongst potential AP-1 binding transcription factors BATF has been previously identified as a regulator of germinal centre gene expression and a key driver of neoplastic B-cells related to the activated B-cell state, equivalent to day 3 of *in vitro* differentiation (Care et al., 2014). To test BATF enrichment at AP-1 footprints we employed existing BATF ChIP-seq data from GM12878 (ENCODE Project Consortium et al., 2020) and found significant enrichment of BATF binding sites at CREs with AP-1 footprints, but not at CREs with only PU.1/SPIB footprints. Mirroring transcription of *BATF* (Figure 4.7.F), BATF binding sites showed the greatest enrichment at CREs most accessible 12 hours to 3 days after activation (linked to cluster 4).

To suggest functions for PU.1, SPIB and AP-1 (BATF) in the B cell to activated B cell transition, gene set over-representation analyses was performed. This identified gene sets enriched in transcription factor target genes (PU.1/SPIB only, PU.1/SPIB and AP-1, and AP-1 only) in each cluster.

Firstly, to identify pathways enriched by differentiation stage, over-representation was performed against a background set of genes, which were targeted by any of the three factors in any other cluster. The resulting 'Temporally enriched' gene sets (FDR < 0.1) are summarised on Figure 4.7.D. Secondly, to identify pathways preferentially regulated by PU.1/SPIB and/or AP-1 at each stage a background set of similarly expressed genes, not linked to the given footprint(s) was used (Appendix 4.3). These 'TF-specific enriched' gene sets are summarised on Figure 4.7.E.

For cluster 1, PU.1/SPIB, but not AP-1, targets were enriched for B cell specific gene sets (Figure 4.7.D), showing TF-specific enrichment of genes involved in intracellular signal transduction (Figure 4.7.E). Both PU.1/SPIB and AP-1 were enriched for germinal centre expressed and RNA processing gene sets and AP-1 targets were enriched for RNA metabolism genes, compared to similar genes not linked to AP-1.

A variety of gene sets relating to RNA processing and MYC targets (PU.1/SPIB and AP-1), NF-κB and BCR activation (PU.1/SPIB in conjunction with AP-1) and OCT2 targets (AP-1 only) were enriched in cluster 2, over all over clusters. RNA-processing genes showed preferential regulation by PU.1/SPIB, and OCT2 targets showed preferential regulation by AP-1 (Figure 4.7.E). Several gene sets showed preferential co-regulation by PU.1/SPIB and AP-1, including NF-κB pathway, cell cycle and BCL6 target genes (also observed for cluster 1, Figure 4.7.F).

Genes expressed later in the activation process (cluster 4) were enriched for gene sets relating to cell cycle and DNA repair, alongside OCT2 and E2F targets, regardless of regulator. No transcription factor specific pathways were enriched for this cluster.

Overall, this combined view of PU.1/SPIB and AP-1 target genes suggests that gene regulation shifts from PU.1/SPIB to AP-1 during B cell activation, contingent on each factors' induction, with potential coregulation of activation pathways during the immediate response to CD40 and BCR engagement. Altogether the shift from PU.1/SPIB to AP-1 induces proliferation of B cells upon activation (cell cycle genes),

prepares ABCs for antibody-secretion through expansion of the endoplasmic reticulum (RNA processing genes); and promotes immunoglobulin gene mutation (DNA repair genes).

4.4.1.5.2 The shift from PU.1/SPIB to AP- 1 is a key determinant of B cell activation

The above analyses have highlighted how the activated B cell transition is defined by a shift from PU.1/SPIB-driven to AP-1-driven gene regulation. Whilst PU.1/SPIB and AP-1 factors have previously been linked to B cell differentiation and the activated B-cell state in mice and humans (Care et al., 2014; Ochiai et al., 2013; Scharer et al., 2018), this is the first detailed analysis of their temporal dynamics and downstream regulatory networks in human B cells. Crucially, the analysis revealed how B cell activation stimuli coordinate a shift in *PU.1*, *SPIB* and *BATF* mRNA, which induces the sequential co-expression of RNA processing, proliferation, and DNA repair genes to stimulate division and prepare for plasma cell identity. The analysis offers new insight on the transcriptional mechanisms driving differentiation, finding that the PU.1/SPIB – AP-1 axis is central to B cell activation and antibody secreting cell fate (Figure 4.8).

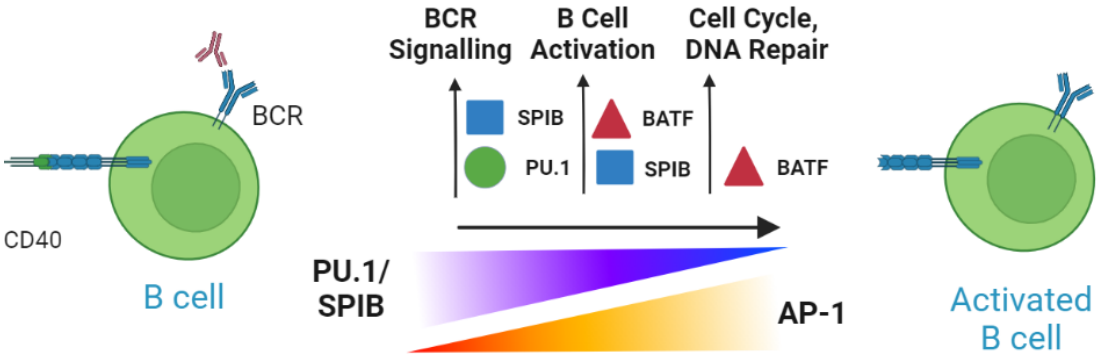


Figure 4.8 Diagram showing the shift from regulation at PU.1/SPIB to AP-1 motifs upon B cell activation by CD40 and BCR stimulation. In B cells PU.1 and SPIB upregulate BCR signalling, upon activation PU.1 is lost rapidly but SPIB is sustained, where it upregulates genes involved in B cell activation. This coincides with BATF upregulation, which first co-ordinates B cell activation with SPIB. BATF activity continues to increase with B cell activation, but SPIB expression is lost. BATF now upregulates cell cycle and DNA repair genes in activated B cells. Figure created with BioRender.

PU.1 and SPIB are two partially redundant transcription factors, shown to upregulate BCR signal transduction and receptors for CD40L, BAFF and TLR ligands and are required for B cell activation (Willis et al., 2017). PU.1 and SPIB are both downregulated in plasma cells, and SPIB over-expression and structural deregulation is associated with the ABC subtype of DLBCL (Care et al., 2014; Lenz et al., 2008).

Deregulation of SPIB in ABC-DLBCL may contribute to the differentiation block that characterises this tumour type.

PU.1/SPIB footprints were uniquely enriched in B cell specific cis-regulatory clusters (Figure 4.4) and linked to the expression of B cell specific gene modules expressed at these time-points (Figure 4.5). PU.1/SPIB enriched modules and PU.1/SPIB target genes (Figure 4.7) were enriched in pathways relating to B cell expression, BCR signalling and signal transduction. This is consistent with PU.1 and SPIB's complementary roles in environment sensing to facilitate B cell activation (Willis et al., 2017), and the importance of BCR signalling in the pathogenesis of ABC-DLBCL (Davis et al., 2010; Phelan et al., 2018).

Our data suggest that some PU.1/SPIB footprints sustain accessibility after activation and may contribute regulation to genes immediately induced by activation stimuli (Figure 4.7). Unlike *PU.1*, *SPIB* expression was maintained immediately post-activation (Figure 4.7.F) and was preferentially repressed after day 3. Thus, immediately after activation SPIB may preferentially occupy PU.1/SPIB motifs to modulate transcription. This is consistent with the established model of *SPIB* repression by PRDM1, which accumulates after day 3 as ABCs transition to the plasmablast and then PC states.

AP-1 occupied cis-regulatory elements open following CD40 and BCR engagement. AP-1 provides regulatory input from the onset of activation until the plasmablast transition, and gene set enrichment results suggest diverse functions controlling B cell activation, RNA processing and the cell cycle (Figures 4.5 and 4.7).

AP-1 subunits (including FOS, FRA1 or BATF partnered with JUN, JUNB or JUND) regulate temporally diverse processes in B cell maturation (Grötsch et al., 2014; Inada et al., 1998; Ise et al., 2011; Long et al., 2022; Ochiai et al., 2013; Ohkubo et al., 2005; Vasanwala et al., 2002). BATF is induced in a CD40, and MHC-II dependent manner (Inoue et al., 2017; Long et al., 2022), and is essential for co-ordinating class-switch recombination and germinal centre establishment (Ise et al., 2011; Morman et al., 2018). In germinal centres, BATF is induced as B cells transition from the light zone to dark zone upon selection by T cells and its expression is associated with cell cycle re-entry (Long et al., 2022). BATF over-expression is associated with ABC-subtype DLBCL (Care et al., 2014).

Our data show that BATF is induced following CD40 engagement, and BATF ChIP-seq binding sites for GM12878 overlap AP-1 footprints accessible in activated B cells (Figure 4.7). AP-1 functions suggested by gene set over-representation (Figures 4.5 and 4.7) support known roles for BATF in class switch recombination (DNA repair) and cell cycle entry, and indicate additional involvement in RNA processing, crosstalk with the NF-κB pathway and regulation of MYC, E2F and OCT2 target genes.

Figure 4.7 revealed that the transition from PU.1/SPIB-driven to AP-1 driven regulation is graded by expression of *PU.1*, *SPIB* and *BATF.* Our data suggest that the two alternate transcriptional programs intersect in the hours following activation, when NF-κB, MYC and BCL6 are induced (Calado et al., 2012; Dominguez-Sola et al., 2012; Gerondakis and Siebenlist, 2010; Robinson et al., 2020). Significant over-representation of NF- κB, MYC and BCL6 targets was observed in both cisREAD-predicted PU.1/SPIB and AP-1 targets. This suggests that, at this transitory stage, PU.1/SPIB and AP-1 regulatory networks are intertwined with those of other critical germinal centre factors. Overall, our data support a model where gene regulation gradually shifts from PU.1/SPIB towards AP-1 upon B cell activation, passing through an intermediate stage where SPIB and BATF may co-ordinate the expression of genes induced immediately upon activation. This gradient would also affect IRF4 binding partner choice and shift IRF4 occupancy from ETS-IRF4 composite elements (EICEs) and towards AP1-IRF4 composite elements (AICEs).

4.4.1.6 Limitations of the *in vitro* B cell system and genome-wide analyses of regulation

Altogether the work in this chapter demonstrated that cisREAD successfully prioritises differential transcription factor occupancy and chromatin accessibility with dynamic gene expression across B cell maturation. Applying this method to our *in vitro* system we were able to reveal new insight into transcriptional reprogramming during B cell activation. However, both the dataset and methods have their limitations.

Firstly, due to preferential differentiation of memory B cells into long-lived plasma cells, whilst day 0-3 time-points originate from total peripheral blood B cells, day 6-13 time-points originate from isolated memory B cell subpopulations (Cocco et al., 2012). Differences between total B cell derived and memory B cell derived gene regulation

between day 3 and day 6 do not affect the conclusion that gene regulation shifts from PU.1/SPIB to AP-1 upon B cell activation. Whilst this discrepancy may affect the types of regulation observed during the ABC-plasmablast transition, previous work in our *in vitro* B cell system found highly similar patterns of gene expression between total and memory B cell derived fractions in this transition (Cocco et al., 2020).

Secondly, in the absence of measured transcription factor binding, we use transcription factor footprinting, corrected for Tn5 cutting bias, as a proxy for TF occupancy at accessible regions (Li et al., 2019). Whilst many studies successfully employ ATAC-seq footprinting to interrogate TF binding dynamics (Barwick et al., 2018; R. Li et al., 2018; Vierstra et al., 2020), it has been noted that many transcription factors do not leave strong footprints, particularly those with short DNA residency times (Baek et al., 2017; D'Oliveira Albanus et al., 2021; Sung et al., 2014). This is particularly a concern for transiently binding signal-inducible TFs like STAT3 and NF-κB, for which we observe detectible yet 'shallow' footprints (Figure 4.3.A). The replication of our analysis with binding site predictions derived from the BMO model, based on motif accessibility and co-occurrence (D'Oliveira Albanus et al., 2021), suggests however that the cisREAD method and downstream analysis is robust to the use of footprints (Appendices 4.1 and 4.2).

Finally, the analysis has not been able to differentiate specific motifs or footprints for two key transcriptional regulators of terminal PC differentiation PRDM1 and XBP1. Both of these factors are expressed and occupy target sites in differentiating plasma cells (Cocco et al., 2020). PRDM1 motifs overlap with a subset of ISREs and may therefore be subsumed amongst a subset of accessible regions with an ISRE match. Additional binding motifs consistent with the XBP1 and associated NF-Y consensus sequences were identified in CRE clusters 7 and 8 (5-ACGTG-3/5-CACGT-3 and 5-CCAAT-3), which are associated specifically with genes induced at the plasma cell state (Acosta-alvear et al., 2007; Cocco et al., 2020). However, no sequence associated specifically with XBP1 appeared in the final set of *de novo motifs* and so XBP1-associated motif enrichment was not tested amongst PGCNA modules derived from the expression time course.

4.4.2 Gene-specific models recall known regulation and suggest new hypotheses of transcriptional control

The global analysis of gene regulation revealed how transcription factors orchestrate B cell differentiation on a genome-wide scale. These inferences were made through combined analyses of thousands of gene-specific predictions. In this next section, two gene-specific models are evaluated in depth: for *AICDA* and *PRDM1*. These two case studies exemplify how cisREAD can be used to generate hypotheses of gene-specific regulation, in line with aim 2. They demonstrate how cisREAD can identify both known regulatory relationships and suggest new mechanisms of transcriptional control.

4.4.2.1 AICDA model recalls validated regulatory elements, acting through NF-κB and AP-1 (BATF)

AICDA encodes the essential AID enzyme responsible for the mutagenic processes of CSR and SHM which underpin antibody diversity. Regulation of the *AICDA* gene has been extensively studied in murine systems and human cell lines by several groups. For these two reasons, the *AICDA* model was chosen for in depth evaluation.

13 Candidate CREs, within 100kb of the TSS, were considered to regulate *AICDA* (Table 4.2). After community detection, a LASSO model was constructed with 11 coCRE predictors. 8 of these predictors were selected by the model at the optimal λ, which minimised cross-validated error (Figure 4.9.A), Gene expression predicted by the *AICDA* model was highly correlated with actual expression (r=0.98), which increased from 2h30 post activation and peaked in day 3 activated B cells (Figure 4.9.B). 10 selected cis-regulatory elements, forming the 8 coCRE predictors, were selected by the *AICDA* model, however none were statistically significant (Table 4.2 and Figure 4.9.C).

Table 4.2 Detals of the 11 predictors in *AICDA* LASSO regression model. Each row represents one of 13 cis-regulatory elements, entered into the model alone or as part of a coCRE. TF footprints indicate *de novo* motif occupancy, predicted by HINT-ATAC, at any time-point. Four cis-regulatory elements (conserved between mice and humans) have been experimentally validated in mice.

coCRE / Predictor	CRE (hg38)	Transcription Factor Footprints	Distance to TSS (bp)	Pearson Correlation	Coefficient	*p* value	Annotation in Figure 4.9	Experimental Validation
1	chr12:8537820-8538220	PAX5/CREB/ATF	74,839	0.58	0.126	0.301	*I*	
2	chr12:8542248-8542648	AP-1, PAX5/CREB/ATF	70,411	0.36	-0.0803	0.43	*II*	
1	chr12:8578990-8579756	PAX5/CREB/ATF, RUNX	33,486	0.54	0.126	0.301	*III*	Crouch *et al.*, 2007
3	chr12:8591082-8591482	NF-κB, RUNX	21,577	0.51	0	NA		
4	chr12:8611137-8611537	E-Box, RUNX	1,522	0.90	0.428	0.0595	*IV*	Sayegh *et al.*, 2003; region 2 from Tran *et al.*, 2010
5	chr12:8613819-8614219	NF-κB, SP/KLF, ZBTB33	-1,160	0.09	-0.127	0.394	*V*	
6	chr12:8619101-8619501	E-Box, PU.1/SPIB	-6,442	0.52	0	NA		
7	chr12:8629441-8629841	NF-κB	-16,782	0.79	0.0864	0.708	*VI*	Tran *et al.*, 2010; TETE1 from Lio *et al.*, 2019
8	chr12:8644895-8645295	AP-1, E-Box	-32,236	0.82	0.388	0.2	*VII*	
8	chr12:8648613-8649013	AP-1, RUNX	-35,954	0.84	0.388	0.2	*VIII*	TETE2 from Lio *et al.*, 2019
9	chr12:8697669-8698069	SP/KLF	-85,010	-0.71	0	NA		
10	chr12:8698137-8698764	RUNX	-85,591	-0.77	-0.0885	0.603	*IX*	
11	chr12:8709466-8709866	IRF4, RUNX	-96,807	0.05	0.3	0.119	*X*	

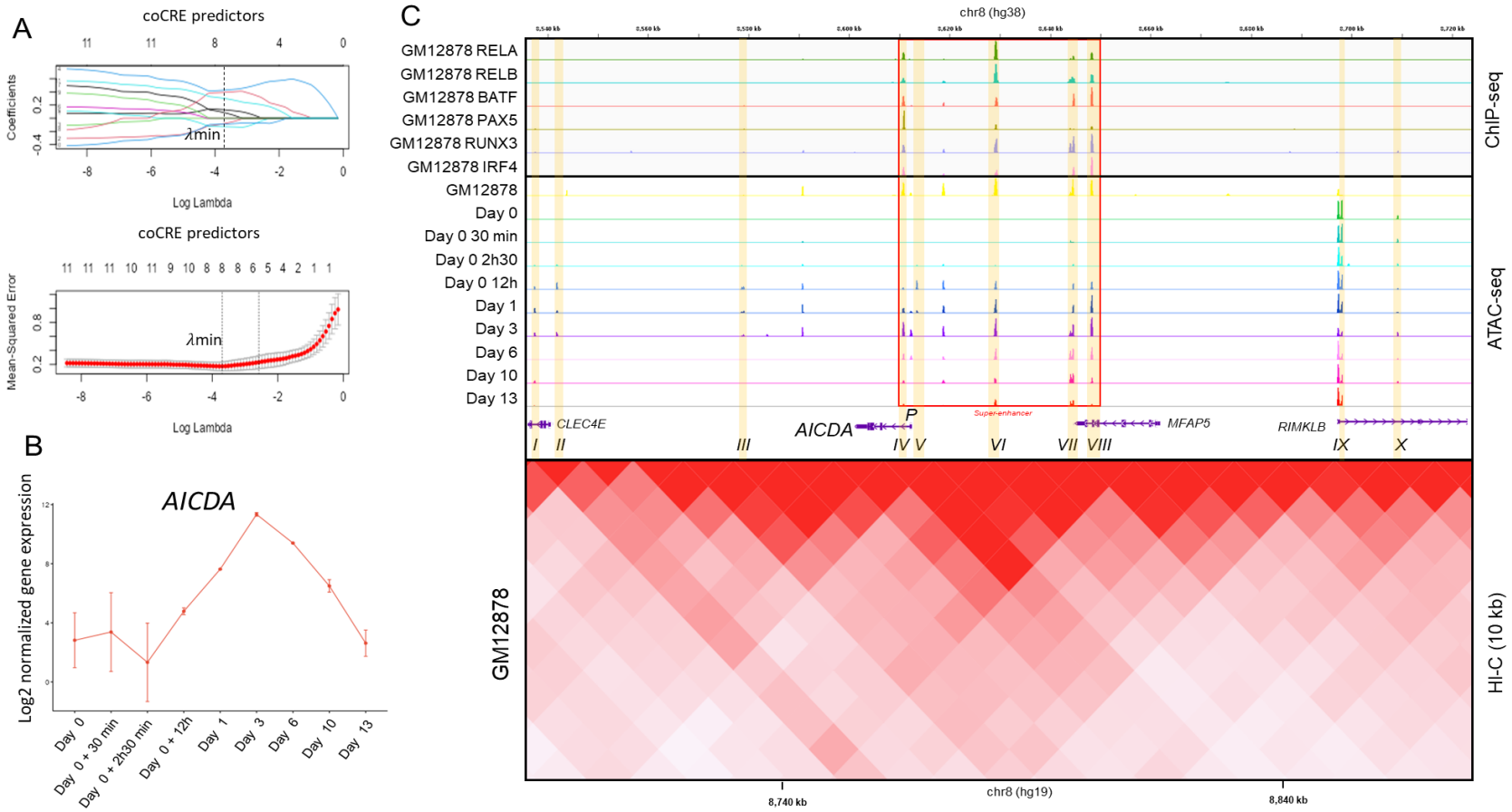


Figure 4.9 cisREAD predicted regulation for *AICDA*. A) The LASSO regression model selected 8 predictors at λmin. This model predicted *AICDA* expression, which was highly correlated with actual *AICDA* expression, shown in B) where points give mean across donors and bars give range. C) 10 selected cis-regulatory elements contributed to the *AICDA* regression model, thus were predicted to regulate the gene. These CREs, highlighted in yellow, were accessible in activated B cell states (as indicated by in vitro B cell ATAC-seq signal tracks) and occupied by relevant transcription factors (as indicated by GM12878 ChIP-seq tracks). Five of these overlapped a super-enhancer, defined by Lio et al. 2019, shown in red. GM12878 Hi-C (KR normalised, binned at 10kb) showed that the majority of selected CREs were observed to interact with the *AICDA* promoter, as evidenced by deep red squares, when tracing from the CRE to the promoter, indicating elevated numbers of chromatin contacts.

4 of these selected CREs overlapped conserved regulatory elements, which have been functionally validated in mice. These include CRE *II*, a distal 3' enhancer required for *in vivo AICDA* expression in mice (Crouch et al., 2007); CRE *IV*, a B cell-specific intronic CRE bound by PAX5 and E-Box proteins which is required for efficient *AICDA* induction in mice (Sayegh *et al.*, 2003; Tran *et al.*, 2010); and CRE *VI*, a 5' enhancer found to drive basal and signal-inducible activity in mouse B lymphoma cells through factors including NF-κB (Tran *et al.*, 2010). CRE *VI*, alongside CRE *VII*, also overlapped two TET-responsive elements which have been shown to upregulate *AICDA* upon LPS-stimulation through demethylation of a super-enhancer by TET, which is recruited by BATF (Lio et al., 2019). Murine elements within this super-enhancer (spanning *IV* to *VIII*) have been shown to produce eRNAs (Meng et al., 2014) and upregulate *AICDA* through RNA polymerase mediated interactions (Kieffer-Kwon et al., 2013).

Integration with Hi-C data from GM12878, similar in differentiation-state to day 3 ABCs, showed evidence of chromatin interactions between *AICDA* and its super-enhancer (Figure 4.9.C). These were visible as the deep red triangle on the contact matrix, showing elevated chromatin contacts within this region. Chromatin looping interactions were also evident between the *AICDA* promoter and CRE *III*.

4 additional elements were selected by the *AICDA* model (*I* and *II* near *CLEC4E* and *IX* and *X* near *RIMKLB*) but Hi-C data only showed evidence of chromatin looping between *II* and the *AICDA* promoter. However, since these four regions differ in chromatin accessibility between in vitro ABCs and EBV-transformed GM12878 cells, a different chromatin topology could be present in vitro.

Overall, the *AICDA* model shows the cisREAD methodology can recall known regulatory elements and suggest additional, plausible candidate regulators.

4.4.2.2 PRDM1 model suggests upregulation through OCT2 and RUNX at distal plasma cell specific enhancers

PRDM1 encodes BLIMP1, a transcription factor described as the master regulatory of plasma cell differentiation. Whilst the transcription factor network governing its repression and induction have been well characterised (Figures 4.1 and 4.2), it has not been established how PRFM1 expression remains elevated in plasma cells independent of IRF4 (Low et al., 2019).

The *PRDM1* model at λmin (the value of λ which minimises the mean squared error of the model) selected a community of three distal co-acting enhancers located 3' of the gene and a lone intronic enhancer from a total of 21 predictors (Figure 4.10.A). The distal 3' community of cis-regulatory elements was a statistically significant predictor of *PRDM1* transcription, and expression predicted by the *PRDM1* model was strongly correlated with actual gene expression (Figure 4.10.B).

PRDM1 is induced upon the plasmablast transition (Figure 4.10.C), and the accessibility of the 4 selected CREs mirrored this expression pattern (Figure 4.10.D). The selected intronic CRE (*I*) was occupied by known *PRDM1* regulators PU.1/SPIB (repressively in B cells) and IRF4 (in plasma cells). The significant downstream coCRE was co-occupied by OCT2 and RUNX in plasmablasts and plasma cells, as well as PAX5/CREB/ATF in B cells (Figure 4.10.D.)

Whilst none of the selected CREs have undergone experimental validation, 2/3 members of the significant coCRE overlapped two distal super-enhancers, assigned to PRDM1 in multiple myeloma cells (Lovén et al., 2013). Hi-C in primary plasma cells showed a domain of elevated chromatin contacts spanning from the *PRDM1* promoter to 150kb downstream of the TSS, encompassing both super-enhancers (Figure 4.10.D).

Taken together, *in vitro* B cell ATAC-seq, the *PRDM1* model, GM12878 ChIP-seq and primary GC and PC Hi-C data suggest that a range of distal TF-bound regulatory elements downstream of the gene control *PRDM1* transcription. In B cells proximal and distal regulatory elements are occupied by repressive transcription factors including SPIB (possibly at CRE *I*) and PAX5 (at CRE *III*). *PRDM1* is elevated in the plasmablast transition by relief of repression and induction of IRF4 (including at CRE *I*). Following the plasmablast transition, several additional elements gain accessibility, including those selected by the LASSO regression model.

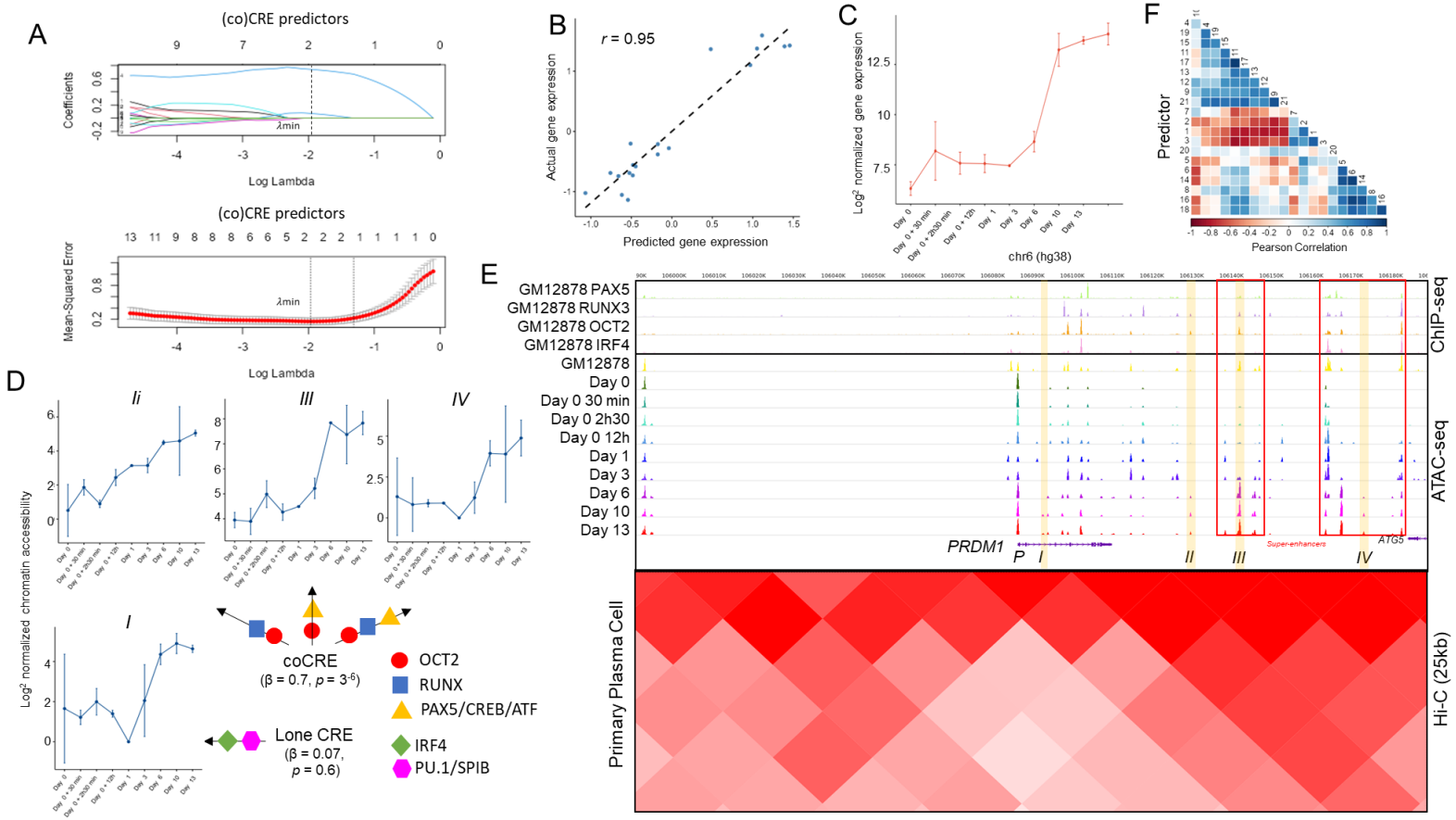


Figure 4.10 cisREAD predicted regulation for *PRDM1*. A) The LASSO regression model selected 2 predictors at λmin. B) This model predicted *PRDM1* expression that was highly correlated with actual PRDM1 expression, shown in C) where points give mean across donors and bars give range. D) The four selected CREs, one lone CRE and one coCRE, were similarly accessible in terminal differentiation. The coCRE was predicted co-bound by OCT2, RUNX and PAX5/CREB/ATF and the lone CRE was footprinted by IRF4 and PU.1/SPIB. E) These CREs, highlighted in yellow, were accessible in plasmablast and plasma cell states (as indicated by in vitro B cell ATAC-seq signal tracks) and occupied by relevant transcription factors (as indicated by GM12878 ChIP-seq tracks). Two of these overlapped super-enhancers, defined by Loven et al. 2013, shown in red. Primary plasma cell Hi-C (KR normalised, binned at 25kb) showed the distal 3' coCRE exists within a subdomain characterised by elevated chromatin contacts with the *PRDM1* promoter.

Footprints within these selected elements suggest that OCT2 and RUNX may be the factors which sustain *PRDM1* elevation in plasma cells, independent of IRF4 (Low et al., 2019). Whilst many other plausible regulators were rejected by the model, this example demonstrates that cisREAD can generate hypotheses of gene regulation to suggest new mechanisms of transcriptional control. The rejection of similarly accessible candidate regulators likely stems from issues associated with multicollinearity (observed in Figure 4.10.F). This limitation is discussed in the following section.

4.4.2.3 Gene-specific models highlight limitations the cisREAD method

The cisREAD methodology employs a correlation-based approach to link regulatory elements to target genes, with LASSO regression models selecting CREs whose accessibility positively or negatively predicts expression of a gene. This assumption enabled the successful prediction of validated *AICDA* enhancers, which were accessible only when *AICDA* was expressed.

Whilst correlation based methods have been shown to assign enhancers and silencers to genes to provide biological insight (Beekman et al., 2018; Huang et al., 2019; Vijayabaskar et al., 2019), they may overlook regulatory relationships where accessibility is not correlated with expression, such as priming (Moore et al., 2020), in which regulatory elements are accessible prior to expression. An example of this may be seen in the *PRDM1* model, where its intronic CREs, conserved and validated in mice, were not selected due to their prior accessibility in BC and ABC states.

In the *PRDM1* example, there are also cases where CREs whose accessibility is strongly correlated with expression (such as other elements within the super-enhancer regions) fail to be selected by the LASSO regression model. This is likely a result of multicollinearity (Figure 4.10.F), which LASSO handles by selecting one of the correlated variables, and rejecting the others since their inclusion would not improve model fit (Zou and Hastie, 2005). The community detection step aimed to alleviate multicollinearity, by entering correlated variables into the model as one predictor. However, the use of 'integrated similarity scores' to draw edges in the cis-regulatory networks (described in chapter 3) meant that correlated CREs were only grouped together if they shared TF footprints. The *PRDM1* example (Figure 4.10) showed that multiple similarly accessible CREs, components of the same super-enhancer, were

entered to the model as separate predictors due to disparate TF footprint profiles. Only one of these predictors was selected, and the others were eliminated from the model. To better capture super-enhancers, at the expense of TF-co-bound units, the community detection step could have been performed using only chromatin accessibility.

It should also be noted that in a minority of instances, selected regression coefficients reverse sign, with the model coefficient opposite between the correlation between accessibility and expression. This is due to the statistical paradox of suppression, which can arise upon addition of another correlated variable into the model (Tu et al., 2008). For this reason, the direction of regulation is best derived from correlation not coefficient.

4.5 Conclusion

Downstream analysis of the fine-grained map of chromatin accessibility and gene expression profiles across B cell differentiation, following application of the cisREAD method, has enabled the identification of stimuli-responsive shifts in transcription factor binding associated with distinct epigenetic programmes.

Application of our method to a model system of human B cell differentiation revealed how a core network of transcription factors exercise regulatory control over B cell differentiation, explicitly coupled to the CD40 and BCR activation stimuli administered to the *in vitro* cell system. This work has both identified established modes of transcriptional control, and uncovered new wires in the transcriptional circuitry which shapes mature B cell differentiation.

Crucially the analysis revealed unprecedented detail into the shift from PU.1/SPIB to AP-1 led regulation which co-ordinates T-cell dependent activation in humans. This finding may act as a springboard for future studies, which could validate the roles of PU.1, SPIB and BATF through experimentation. For example, the association with BATF/AP-1 with the ABC state suggested that BATF itself may be responsible for licencing ABC-specific regulatory elements. Observations here, and elsewhere, support a role for BATF as a pioneer factor in B cells (Morman et al., 2018). Recent

experimental work establishing BATFs pioneering function in T cells also supports investigation into BATF pioneering in the B lineage (Pham et al., 2019).

Key to these findings was the new cisREAD method; a data-driven approach which prioritises dynamically accessible regulatory elements, targeted by lineage-specific transcription factors and associated with the expression of differentiation associated genes. The global analyses presented in this chapter leveraged predictions made by gene-specific models, to make inferences on gene regulation on a genome-wide scale. In depth examination of two individual models showed that cisREAD was able to recall known modes of regulation and generate plausible new hypotheses, which could be experimentally validated with future work.

These case studies revealed both limitations and advantages of the method, however performance of cisREAD on the full dataset is still untested. This will be explored in the next chapter, where the performance of cisREAD will be evaluated at scale and benchmarked against alternative models.

Chapter 5. cisREAD identifies more regulatory chromatin interactions than alternative methods

5.1 Introduction

Chapter 3 described the development of cisREAD: a method to identify gene-specific cis-regulatory elements across differentiation from ATAC-seq and RNA-seq datasets. cisREAD was applied to datasets from an *in vitro* system of B cell differentiation, to identify regulatory elements which control differentiation-associated genes through the activity of lineage-specific transcription factors. Comparing predictions to experimentally derived gene sets, cisREAD correctly identified target genes for IRF4 and NF-kB transcription factors. Altogether this indicated that cisREAD assigned transcription factor-bound regulatory elements to their correct targets. In chapter 4 these CRE-gene linkages were used to gain insight into the transcription factor-led regulatory programmes which drive differentiation, and also predict hypotheses of gene-specific regulation.

Chapter 5 focuses on the validation of cisREAD-predicted gene regulation with publicly available datasets, benchmarking the method against alternative approaches. Since large-scale validation of cis-regulatory elements and their target genes is an ongoing area of research and debate, the chapter starts by outlining validation strategies common in the field.

5.1.1 Validating predicted regulatory interactions

Benchmarking methods to match regulatory elements to target genes is both difficult and contentious, as there is no agreed-upon set of 'ground truth' regulatory interactions (Moore et al., 2020). This owes to the difficulty of detecting regulatory interactions at scale using current techniques. Whilst low-throughput methods such as reporter assays provide functional validation for small numbers of gene-specific regulatory elements, evaluation of predictive models requires large, high-throughput datasets. Frequently used validation strategies include comparisons with chromatin interactions from Chromatin Conformation Capture (3C) based techniques, expression Quantitative Trait Loci (eQTLs), and CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) perturbation screens (Gasperini et al., 2020). This section will

highlight the advantages, limitations, and uses of each datatype.

5.1.2 Detecting regulatory chromatin interactions with chromosome conformation capture

Chromosome conformation capture (3C) techniques probe the topology of chromosomes, and can be used to identify topologically associated domains (TADs) and, at sufficient resolution, detect chromatin looping interactions between regulatory elements (McCord et al., 2020). A general introduction to this family of methods is provided in chapter 1.

Genome-wide 3C technologies can detect interactions between all chromosomal regions (Hi-C), or those mediated by a protein of interest (ChIA-PET) (Fullwood et al., 2009; Lieberman-Aiden et al., 2009). Hi-C offers an all-vs-all approach to capturing chromatin contacts between all regions of the genome, however large numbers of cells and high sequencing depths are required for resolution sufficient to identify individual chromatin loops (Rao et al., 2014). To increase resolution at the cost of scale, Hi-C libraries can be enriched for regions of interest. Through enriching for gene promoters, Promoter Capture Hi-C (PC Hi-C) can identify long-range interactions between distal chromatin regions and gene promoters at high-resolution (Mifsud et al., 2015). In a similar trade-off, ChIA-PET combines Chromatin Immunoprecipitation (ChIP) with Hi-C's proximity ligation step to identify chromatin interactions tethered by a protein. Through targeting proteins with known roles in cis-regulation or chromatin structure (e.g. RNA pol II or CTCF), ChIA-PET enables identification of chromatin interactions that delineate regulatory chromatin domains or enhance transcriptional activation (Tang et al., 2015).

Whilst Hi-C, ChIA-PET and PC Hi-C are powerful tools to study the spatial organisation of chromatin, data derived from 3C-based methods can suffer from technical artefacts and data-dependencies which result in noise and bias (Lajoie et al., 2015). The detection of chromatin looping interactions is complicated by the distance-dependent distribution of chromatin contacts, where the probability of contact decreases with genomic distance following a power-law relationship. This is accepted to arise from random collisions between chromatin regions due to the Brownian motion of chromosome polymers, which increase in frequency with proximity (McCord et al.,

2020). Alongside distance-dependent collisions, random ligations which occur in solution, and sequencing-based artefacts add to background levels of 3C contacts. Numerous methods have been developed to identify chromatin interactions by accounting for 'background' contact levels obtained by Hi-C, PC Hi-C or ChIA-PET (Cairns et al., 2016; He et al., 2015; Roayaei Ardakany et al., 2020).

Despite these technical limitations, 3C-based methods offer unparalleled insight into chromatin interactions between distal elements and gene promoters, and are a popular choice for both validating and training predictive methods (Cao et al., 2017; He et al., 2014; Roy et al., 2015). However, whilst chromatin interactions are widely accepted to necessitate regulation, not all interactions are regulatory. This must be considered when using interactions as surrogate for regulation.

5.1.3 Associating regulatory variants with gene expression through eQTLs

Expression Quantitative Trait Loci (eQTLs) are genetic variants associated with variation in expression across a population. As such, the presence of an eQTL within a cis-regulatory element provides evidence of gene-specific regulation *in vivo*. eQTL mapping requires genotype and expression data over large numbers of cells or tissues, and is performed by association testing similar in approach to GWAS studies (Nica and Dermitzakis, 2013). As such, these studies face limitations including causality and linkage disequilibrium.

eQTLs do not always have direct or independent roles. eQTLs may affect expression of a gene in cis – i.e., in a cis-regulatory element – or trans. Trans-regulatory effects on gene-expression can arise when a variant alters the expression of gene encoding a protein (e.g., a transcription factor, signalling molecule or receptor), which in turn alters expression of other genes. To disentangle the two mechanisms, eQTLs are divided into cis-eQTLs and trans-eQTLs based on proximity to the gene's TSS (Umans et al., 2021). Linkage disequilibrium (LD - where variants in a chromosomal region are co-inherited) means that multiple eQTLs may be associated with gene expression due to correlation with the true regulatory variant (Umans et al., 2021). In an attempt to separate 'causal' eQTLs from those in LD, a number of statistical fine-mapping

approaches have been developed and applied to eQTL analysis (The GTEx Consortium, 2021).

Due to the use of genotype data, only common variants are tested for association with expression. Many cis-elements do not harbour common variants, and thus would be undetectable by eQTL mapping. Furthermore, most cis-elements are highly cell-specific and the vast majority of eQTL analyses – such as those performed by the GTEx consortium – have been performed in tissues or whole blood. These samples are easily obtainable across large cohorts but comprise multiple, diverse cell types (Umans et al., 2021). In recent years, single-cell approaches to eQTL mapping using single-cell RNA-seq data, or deconvoluted bulk transcriptomes, are emerging to disentangle the cell-specific effects of eQTLs (Yazar et al., 2022).

Altogether, these limitations mean that eQTL mapping is a powerful but limited approach to *in vivo* validation of cis-regulatory effects on gene expression. Despite limitations, eQTLs have been used to validate and benchmark predicted CRE-gene relationships in a number of studies (Cao et al., 2017; Huang et al., 2019; Salviato et al., 2021; Wang et al., 2021).

5.1.4 Perturbing regulatory elements using CRISPR screens

Recently, experimental enhancer perturbation via CRISPR screens has enabled functional validation of enhancers and target genes endogenously. These experiments target guide RNAs to candidate enhancers and induce perturbations through CRISPR interference (CRISPRi) (Gasperini et al., 2020). Gene expression is then measured in cells with and without gRNAs using single cell RNA sequencing (Gasperini et al., 2019; Xie et al., 2017) or FlowFish (Fulco et al., 2019; Reilly et al., 2021). Whilst these assays can functionally validate gene-specific enhancers in their native chromatin context, widespread validation with these emerging techniques is hindered by a number of factors.

CRISPRi induced perturbation of non-coding DNA is expensive and difficult, and even when CRISPR screens are multiplexed, current techniques identify relatively few cis-regulatory relationships (Gasperini et al., 2020). The largest screen to date identified just 664 enhancer-gene pairs, with 90% of perturbed candidate enhancers yielding no

detectable change in expression (Gasperini et al., 2019). Whilst the technical challenges of CRISPRi may result in a high false negative rate, the true sensitivity of CRISPR screens is still unknown (Gasperini et al., 2020). Additionally, all large-scale CRISPR perturbation screens have so far been limited to the K562 myelogenous leukaemia cell line. Whilst this barrier has prevented validation in other cell types, predicted regulatory interactions in K562 have been evaluated using data from CRISPR screens (Fulco et al., 2019; Salviato et al., 2021).

5.2. Aims and Objectives

This chapter aimed to validate and benchmark the cisREAD method to identify gene-specific cis-regulatory elements. In order to achieve this aim the following objectives were set out:

1) to validate gene-specific cis-regulatory elements across B cell differentiation using publicly available datasets; and
2) to assess the relative performance of cisREAD through benchmarking against alternative methods.

5.3 Methods

This section describes the validation and benchmarking of cisREAD, using publicly available sets of 'gold standard' regulatory interactions to evaluate the performance of cisREAD, relative to alternative methods.

5.3.1 Validation Datasets

Due to the need to match validation data to the *in vitro* B cell system, predicted CRE-gene relationships were compared to chromatin interaction datasets from relevant related B cells. 4 chromatin interactions datasets, obtained by promoter capture Hi-C and RNA pol II ChIA-PET, were used for validation (Table 5.1).

Table 5.1 B cell specific chromatin interaction datasets used for validation.

Datatype	Cell type	Citation	Source	Genome Build
Promoter Capture Hi-C	GM12878	Cairns *et al.* 2016	https://osf.io/nemc6/	hg19
	Total Peripheral Blood B cells (CD19+)	Javierre *et al.* 2016	Javierre *et al.* 2016 Data S1	hg19
	Naïve Peripheral Blood B cells (CD19+, CD27-)	Javierre *et al.* 2016	Javierre *et al.* 2016 Data S1	hg19
ChIA-PET (RNA pol II)	GM12878	Dekker *et al.*, 2017	https://encodeproject.org – accession ENCFF913VWM	hg38

Significant PC Hi-C interactions in naïve (CD19+, CD27-) and total (CD19+) peripheral blood B cells were obtained from the supplementary material of Javierre *et al.*, 2016, and significant interactions in B cell lymphoblastoid cell line GM12878 were downloaded from the supplementary website associated with Cairns *et al.*, 2016. In all three cases, significant interactions were determined by the CHICAGO method from Cairns et al. 2016 and interactions with CHICAGO score > 5 were considered significant. RNA polymerase II ChIA-PET loops for GM12878, called following the ChIA-PIPE pipeline (Lee et al., 2020), were downloaded from the ENCODE portal (ENCODE Project Consortium et al., 2020).

5.3.2 Prediction Datasets

To perform benchmarking, a set of 6 prediction sets from 4 alternative methods were selected for validation alongside cisREAD (Table 5.2). These encompassed simpler approaches (the naïve nearest gene assumption and Pearson correlation between accessibility and expression) and published state-of-the-art models designed to predict enhancer-promoter interactions (JEME and Activity-by-Contact).

Table 5.2 B cell specific predicted regulatory interactions.

Model	Sample(s)	Citation	Source	Genome Build
cisREAD	*In vitro* B cells	-	-	hg38
Pearson correlation	*In vitro* B cells	-	-	hg38
Nearest Gene	*In vitro* B cells	-	-	hg38
JEME	GM12878	Cao *et al.* 2017	https://yiplab.cse.cuhk.edu.hk/jeme/ – ID 114	hg19
	Primary B Cells	Cao *et al.* 2017	https://yiplab.cse.cuhk.edu.hk/jeme/ – ID 032	hg19
Activity-by-Contact	GM12878	Fulco *et al.* 2019	https://osf.io/uhnb4/	hg19
	Primary B cells	Nasser et al. 2021	ftp://ftp.broadinstitute.org/outgoing/lincRNA/ABC/Nasser2021-Full-ABC-Output/	hg19

cisREAD predictions for the *in vitro* B cell dataset were taken from the cisREAD + Footprints set described in chapter 3. CREs were predicted to interact with the gene promoter if the CRE (either alone or as a coCRE constituent) was selected by a gene's LASSO regression model. Pearson correlation and nearest gene methods were also implemented on the *in vitro* B cell dataset. For Pearson correlation, DARs were assigned to DEGs if the Pearson correlation between chromatin accessibility and gene expression > 0.7. For nearest gene, DARs were assigned to DEGs if the TSS of the DEG was the closest TSS (annotated by HOMER annotatePeaks.pl) to the DAR's midpoint.

JEME and Activity-by-Contact were chosen for comparison due to their popularity and disparate methodologies. Cao *et al.*'s JEME model, described in detail in chapter 2, takes a supervised machine learning approach to predict enhancer-promoter interactions from gene expression, chromatin accessibility and histone-mark features. The model is trained on ChIA-PET chromatin interactions (Cao et al., 2017). In contrast, Fulco *et al.*'s Activity-By-Contact model predicts whether an enhancer E regulates a gene G through calculating its 'Activity-by-Contact' score. This is calculated from the product of the enhancer's 'Activity' A (geometric mean of H3K27 acetylation and chromatin accessibility signal) and 'Contact' C with the promoter (normalised Hi-C signal), divided by the 'Activity' and 'Contact' of all neighbouring enhancers within 5Mb (Equation 5.1). Fulco *et al.* found that their simple rule-based approach outperformed chromatin interaction-trained machine learning classifiers, including JEME, when benchmarking on a CRISPR perturbations (Fulco et al., 2019).

Equation 5.1 Activity-by-Contact model from Fulco et al. 2019

$$\text{Activity by Contact Score}_{E,G} = \frac{A_E\ C_{E,G}}{\sum_{e \text{ within 5Mb of G}} A_E\ X\ C_{E,G}}$$

Owing to input requirements (i.e., histone modification ChIP-seq, Hi-C), JEME and Activity-By-Contact models could not be implemented on the *in vitro* B cell dataset. Instead, predicted interactions from primary B cells and the GM12878 cell line were downloaded from supplementary websites associated with Cao *et al.*, 2017, Fulco *et al.*, 2019 and Nasser *et al.*, 2021.

5.3.3 Benchmarking strategy

Prediction datasets from the 5 models were compared to the 4 validation sets following the strategy outlined in Figure 5.1. Since JEME and Activity-by-Contact models predicted cell type-specific cis-regulatory interactions, the cell-types used for prediction were directly matched to the cell-types in the validation sets. Since cisREAD, Pearson correlation and nearest gene methods predicted cis-regulatory interactions across the whole *in vitro* B cell dataset (encompassing both naïve and activated B cell states) validation sets from both B cells and GM12878 (developmentally equivalent to ABCs) were used.

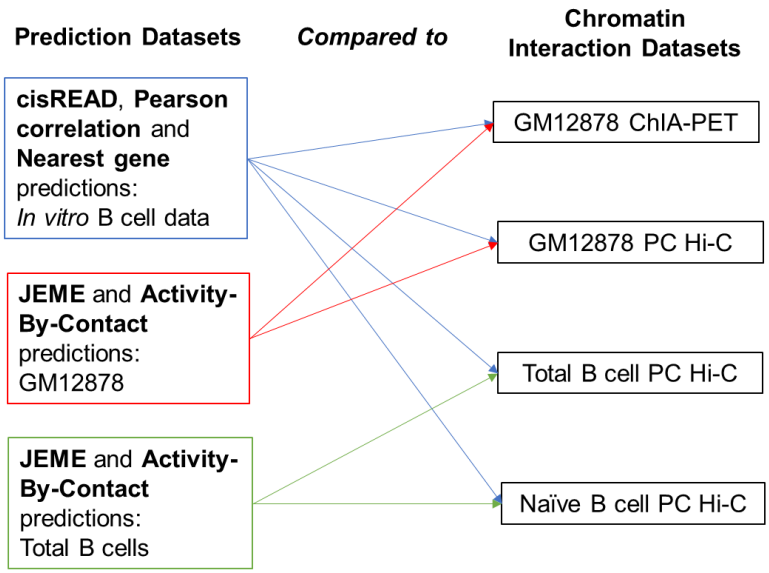


Figure 5.1 Benchmarking strategy. cisREAD, Pearson correlation and nearest gene methods made predictions for the in vitro B cell differentiation dataset. These predictions were compared to four chromatin interaction datasets: RNA polymerase II ChIA-PET in GM12878 cells, and promoter capture (PC) Hi-C in GM12878, total B cells and naïve Bs. JEME and Activity-By-Contact models made predictions for GM12878 and total B cells. For JEME and Activity-By-Contact, predictions in GM12878 were compared to GM12878 chromatin interaction datasets, and total B cell predictions were compared to total and naïve B cell chromatin interactions.

To facilitate comparisons, CRE and TSS coordinates for datasets aligned to hg38 (cisREAD, Pearson correlation, nearest gene, and RNA polymerase II ChIA-PET) were converted to hg19 using UCSC liftOver to match JEME, Activity-by-Contact and PC Hi-C datasets. For all 5 prediction datasets, CREs were resized to 1kb surrounding the region's midpoint. Interactions in both prediction and chromatin interaction datasets were limited to those spanning 2.5kb - 100kb from the CREs midpoint to the gene's TSS. In line with cisREAD, Pearson Correlation and Nearest Gene, JEME and Activity-By-Contact predictions were limited to those involving *in vitro* B cell DEGs.

Predicted-CRE gene linkages were intersected with chromatin interactions from validation datasets. A predicted interaction was considered validated if the gene's TSS overlapped (>= 1bp) a Hi-C/ChIA-PET fragment, and the CRE overlapped the interacting fragment. If multiple predicted and validated interactions overlapped (e.g., due to resolution), all were retained. True positive (TP) predictions were defined as interactions present in both prediction and validation datasets; false positives (FPs) were interactions in prediction but not validation datasets; and false negatives (FNs) were interactions in validation but not prediction datasets. To measure performance, Positive Predictive Value (PPV)/Precision, Recall/Sensitivity and F1 Score (balancing precision and recall) were calculated from these definitions.

Overlaps were also calculated between interaction sets for predictive methods. Dice co-efficients (Dice, 1945) – equivalent to F1 scores – were calculated to give the similarity between prediction sets, prior to hierarchical clustering.

Distance distribution plots were generated for prediction and validation datasets by kernel density estimation using the dist function in R.

5.4 Results

5.4.1 cisREAD better identified regulatory chromatin interactions than other methods

The performance of the 5 predictive methods when validated against chromatin interactions within 100kb is shown in Figure 5.2. Altogether, cisREAD outperformed alternative methods with the greatest mean F1 score (0.16) across the four tasks, indicating a good balance of PPV and recall (Figure 5.2.A).

On average, 12% of cisREAD-predicted regulatory interactions overlapped a chromatin contact (mean PPV). This was similar to the mean PPV of Pearson correlation, and JEME methods but lower than that of Fulco *et al.*'s Activity-By-Contact model (Figure 5.2.B).

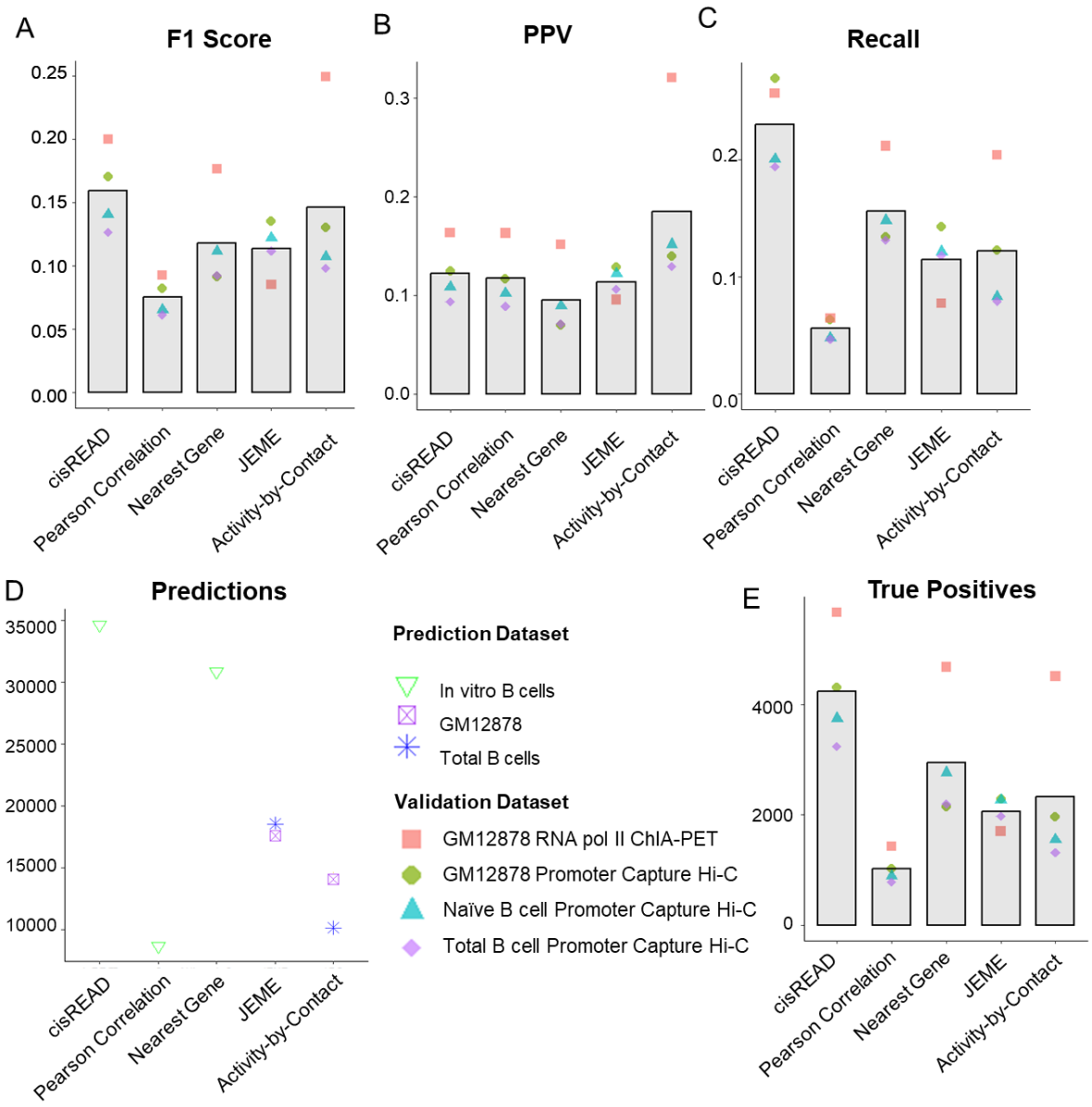


Figure 5.2 Performance of cisREAD benchmarked against other predictive methods using chromatin interaction datasets. A) F1-scores for each set of predictions compared to each chromatin interaction dataset (point) and on average (bar shows mean). B) Shows the positive predictive value and C) shows the recall of each method. D) Total number of predictions made by each method on the *in vitro* B cell dataset (cisREAD, Pearson correlation and nearest gene), or on GM12878 lymphoblastoids (JEME and Activity-By-Contact) or primary peripheral blood total B cells (JEME and Activity-By-Contact). E) Number of 'true positive' predictions made by each method validated in each chromatin interaction dataset (points) and on average (bar shows mean).

cisREAD recalled a mean of 23% of validated chromatin interactions involving a DEG and distal regulatory element (Figure 5.2.C). This exceeded the mean recall of all other methods and reflected how cisREAD predicted the most regulatory interactions (Figure 5.2.D) and identified the most validated 'true positive' interactions (Figure 5.2.E). Pearson correlation had the lowest recall of all methods, which may reflect the stringent threshold of $r > 0.7$.

5.4.2 Performance on PC Hi-C and ChIA-PET datatypes reflects distance distributions of regulatory interactions

Activity-By-Contact and nearest gene approaches better predicted RNA polymerase II ChIA-PET loops than promoter capture Hi-C interactions (Figure 5.2). This may reflect how Activity-By-Contact and nearest gene identified more proximal interactions, with a similar distance distribution to the ChIA-PET interaction dataset. In contrast, cisREAD, Pearson correlation and JEME predicted a wider range of proximal and distal interactions, with a similar distance distribution to promoter-capture Hi-C datasets (Figure 5.3). In conjunction with the reported performance metrics, these distance distributions indicated that cisREAD was more capable of detecting distal interactions than other models, including the state-of-the-art Activity-By-Contact model, or by simply assigning CREs to the nearest gene.

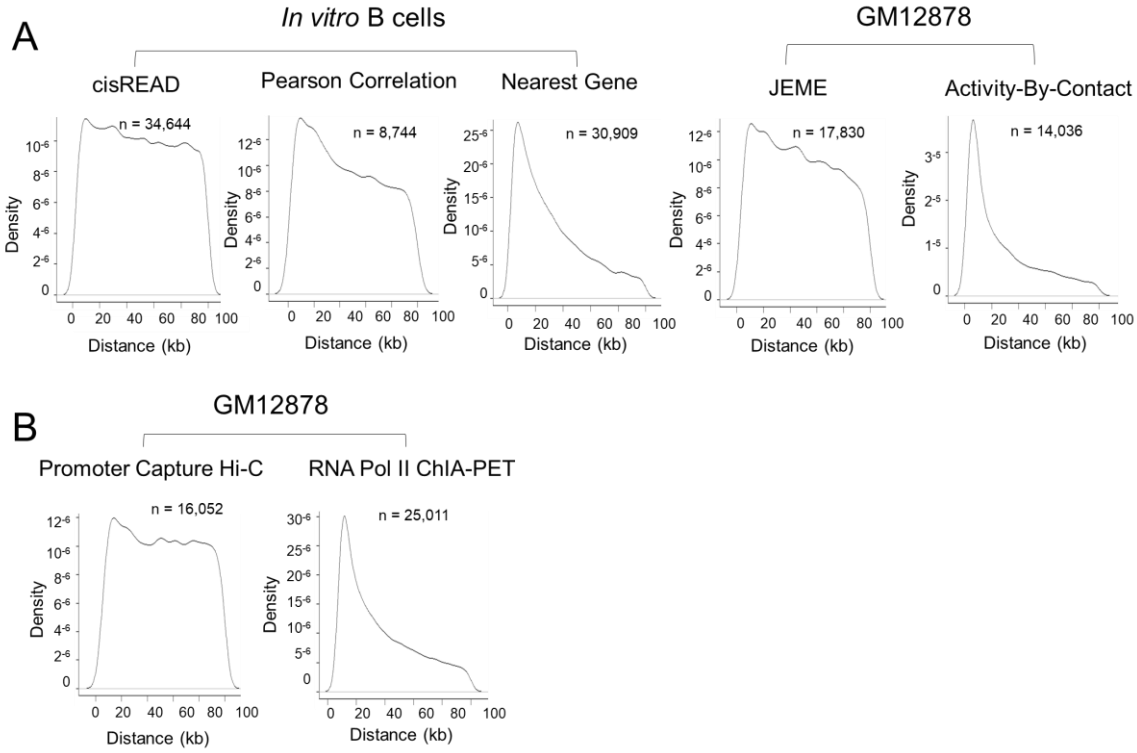


Figure 5.3 Density plots showing the CRE-gene distance distributions of predictive methods and chromatin interaction datasets. Interactions were limited to those spanning 2.5-100kb and involving the TSS for a differentially expressed gene during B cell differentiation. *n* gives the total number of interactions in the dataset. A) shows distance distributions for the five predictive methods, in either *in vitro* B cells or GM12878. B) shows distance distributions for significant promoter capture Hi-C and RNA pol II ChIA-PET interactions in GM12878.

5.4.3 Methods predict different sets of regulatory interactions

Additionally, prediction datasets were overlapped with each other to assess similarity between methods. Hierarchical clustering of Dice similarity coefficients (Figure 5.4) showed that JEME and Activity-by-Contact prediction sets in GM12878 and total B cells clustered by model not sample, and the three methods implemented on *in vitro* B cells clustered together. Interestingly, cisREAD predictions were more similar to nearest gene predictions than Pearson correlation predictions, with 34% of cisREAD-assigned gene targets overlapping the nearest gene. Activity-by-Contact predictions were also most similar to nearest gene predictions, whilst JEME-predicted interactions showed little overlap with predictions made by other methods (0.03-0.09).

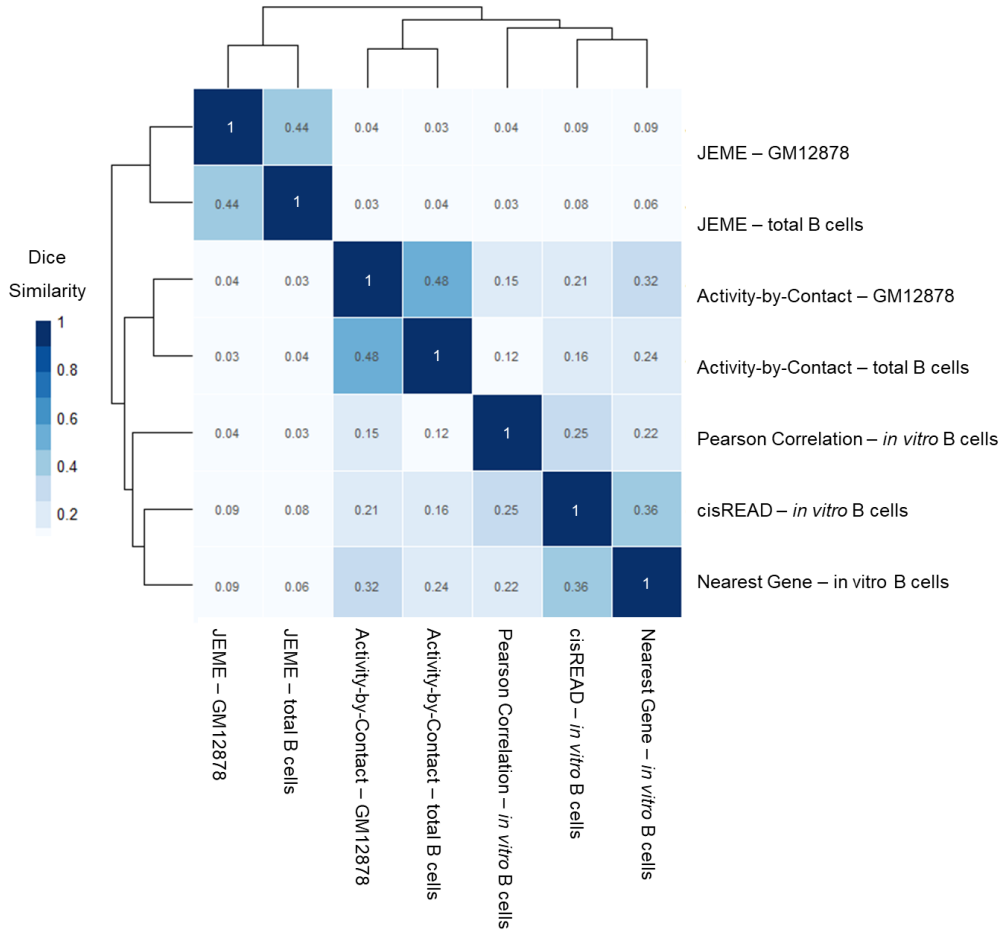


Figure 5.4 Hierarchical clustering heatmap of Dice similarity coefficients between prediction datasets.

5.5 Discussion

5.5.1 The best of a bad bunch? cisREAD outperformed alternative methods, but identified few validated interactions

Altogether the benchmarking results indicate that overall cisREAD best predicted regulatory interactions, driven by high recall of distal regulatory elements. The improved performance on the more distal Hi-C datatype may offer an advantage in assigning target genes to distally-binding factors, such as the AP-1 complex (Bejjani et al., 2019). The data showed that cisREAD better assigned CREs to genes than either of the simpler methods which could have been implemented on the *in vitro* B cell dataset (Pearson correlation at $r > 0.7$ and nearest gene) and may outperform the state-of-the-art Activity-by-Contact method. These results highlight the advantage of cisREAD and support its use in chapter 4 to infer regulatory mechanisms which drive B cell differentiation.

In contrast to results reported in chapter 2, cisREAD outperformed the JEME model with similar PPV but greater recall. In chapter 2, JEME was retrained on murine Hi-C data and shown to perform slightly, but not conclusively, better than the related Vijayabaskar et al. method (also using community detection and LASSO regression). The difference in performance may relate to the limitations of the small reporter assay validation dataset used in chapter 2, or alterations to the JEME model through retraining. In the results presented here, cisREAD showed a clear advantage over JEME despite closer sample matching between prediction and validation samples (which was also the case for Activity-by-Contact).

Despite showing relative superior performance, the overlap between cisREAD predictions and validated chromatin interactions was low. The performance metrics reported here were similar in value to those reported by other groups when following similar benchmarking strategies (Hait and Elkon, 2022; Salviato et al., 2021). Some researchers claim these low across-the-board performance metrics indicate the inadequacy of current methods to predict cis-regulation (Moore et al., 2020). Others suggest that it is methods of validation, not prediction, which require improvement (Hoellinger et al., 2023).

The degree to which chromatin interactions validate regulatory interactions is debated, with potential for false positives and false negatives. Support from other datatypes comes from cell-imaging studies which have found that sustained proximity between enhancers and promoters is necessary for transcriptional activation (Chen et al., 2018). However other groups have evidence of cis-regulation occurring in the absence of contact (Alexander et al., 2019; Benabdallah et al., 2019), possibly through the proposed model of liquid-liquid phase separation (Hnisz et al., 2017). In addition chromatin interactions do not necessary reflect regulation, as stable loops can be maintained even when genes and enhancers are inactive (Ghavi-Helm et al., 2014). Chromatin interaction datasets are also dependent on the methodology used to identify looping interactions. For example, a benchmarking study found the functional relevance of ChIA-PET loops (determined through overlaps with eQTLs and CRISPR perturbations) varies widely depending on loop calling method (Tang et al., 2022). Since the limitations of chromatin interaction datasets complicate benchmarking,

future work validating against additional datatypes could better evaluate model performance.

5.5.2 cisREAD and Activity-by-Contact methods differentially predicted PC Hi-C and ChIA-PET interactions

Whilst cisREAD performed better than the Activity-by-Contact model, when averaged across the four datasets, it was less clear whether cisREAD or Activity-by-Contact was the best performing model overall. This was shown by the disparity of performance on PC Hi-C and ChIA-PET validation datatypes (Figure 5.2). These results, and the distance distributions in Figure 5.3, suggest that Activity-by-Contact may better predict proximal interactions (such as those identified by ChIA-PET) and cisREAD may better predict distal interactions (such as those identified by PC Hi-C).

The distance distributions of PC Hi-C and ChIA-PET datasets may relate to the methods used to obtain each set of chromatin interactions. All three PC Hi-C datasets use the CHiCAGO method to identify statistically significant interactions, where contact counts are elevated above background levels, which are dependent on genomic distance and technical noise. This method aimed to separate robust chromatin loops, presumed to be regulatory interactions stabilised by transcription factor binding, from those occurring by random collisions, between proximal elements, or technical artefacts (Cairns et al., 2016). Whilst interaction-calling methods which adjust for the same biases have been developed for ChIA-PET, ENCODE called ChIA-PET loops using the ChIA-PIPE method which does not account for genomic distance (Lee et al., 2020).

It is unclear to what extent the association between genomic distance and chromatin contacts stems from random non-functional collisions or proximal regulatory interactions. Statistical methods designed to overcome the distance-dependence bias of 3C data may be too stringent and remove true regulatory interactions between proximal elements. However, failure to account for distance-bias may inflate the true number of short-range interactions. Activity-by-Contact employs Hi-C data in its 'Contact' feature, and due to its distance-dependence, the model is implicitly biased towards proximal predictions. This may explain its poorer performance on the distal PC Hi-C interaction datatype. Future validation with eQTL or CRISPR perturbations, whilst

not currently available for B cells, may better evaluate performance in the absence of distance bias.

5.6 Conclusion

Despite limitations, the benchmarking exercise showed that cisREAD better identified regulatory chromatin interactions than alternative methods, including both simpler approaches and the published JEME model. The data also indicated that cisREAD best identified distal interactions, such as those captured by PC Hi-C, over all other tested models. These results defended the use of cisREAD to assign CREs to genes, in order to understand genet regulation in B cells. In future, cisREAD's performance could be explored further through extensive benchmarking against other predictive methods using additional validation datatypes.

Additional support for predictive methods could also come from the presence of disease-associated regulatory variants, matched to relevant target genes. The potential for cisREAD to characterise regulatory variants will be explored in the final discussion chapter.

Chapter 6. Discussion

Understanding gene regulation is essential to understanding the development and maintenance of healthy cellular identity. The next generation sequencing revolution has produced masses of genomics, epigenomics, transcriptomics and 3D genomics data. There is an outstanding need for applicable, implementable, and interpretable methods which integrate multi-omics data to yield insight into the biological role of gene regulation. This final chapter will discuss the contributions of this thesis to the fields of bioinformatics, gene regulation and immunology. This will be done by reviewing the findings and discussion points of each chapter, reflecting on the work performed, and looking ahead to future avenues of research. In addition, it will evaluate to what extent the aims of the thesis, set out in the introductory chapter, were met. Finally, it will recommend future avenues of research into cutting-edge single cell methods, and clinically translatable work on variant annotation.

The work presented in this thesis has contributed new computational methodology and knowledge of regulatory mechanisms during the essential immune process of B cell differentiation.

6.1 Discussion of comparative evaluation of Vijayabaskar et al. and JEME methods

Chapter 1 introduced the unsupervised Vijayabaskar et al. method, developed within the Westhead group (Vijayabaskar et al., 2019). This approach identified 'communities' of co-regulating cis-elements (which are co-accessible and bound by common TFs) nearby differentiation-specific genes. These cis-regulatory communities were then assigned to genes, by performing variable-selection using gene-specific LASSO regression models. This approach worked to select cis-regulatory elements which were active when the gene was expressed. The Vijayabaskar et al. method held several advantages over alternative approaches. These included the ability to operate without expensive 3D genomics data, highly interpretable methodology, the avoidance of pitfalls associated with supervised learning in this field, and the ability to prioritise TF-CRE-gene relationships central to a given system.

Chapter 2 set out to evaluate the Vijayabaskar et al. method, in comparison to the supervised JEME model (Cao et al., 2017). This involved reapplying JEME to the murine haematopoietic dataset, used by Vijayabaskar et al, then validating performance using a reporter-assay dataset. This task was challenging due to numerous reasons. These included design differences between the two methods; Vijayabaskar et al. predicts gene-specific CREs across a system and JEME predicts enhancer-promoter interactions in a sample. Furthermore, the JEME model was unavoidably altered in its retraining and reapplication, and the validation dataset was limited by its small size, biased selection of regions, and inability to detect chromatin-dependent regulatory mechanisms.

Evaluated against this dataset, JEME identified slightly more true positive interactions, and slightly fewer false positive interactions than the Vijayabaskar et al. method. However, JEME also made far more untested predictions, many of which lacked chromatin or TF features suggestive of regulation. Whilst it was difficult to determine the 'best' performing method, the analysis found a large, statistically significant, overlap between Vijayabaskar et al. and JEME methods. These findings served to support the publication of Vijayabaskar et al. 2019.

6.2 Discussion of the development of cisREAD

Despite the advantages of the Vijayabaskar et al. approach, the method required an abundance of histone and TF ChIP-seq datasets for each sample, in addition to chromatin accessibility and gene expression datasets. These requirements limit the application of the Vijayabaskar et al. method to other datasets. In chapter 3, we were presented with a dataset spanning human *in vitro* B cell differentiation. This dataset comprised sample-matched chromatin accessibility and gene expression datasets but lacked the ChIP-seq data required by the Vijayabaskar et al. method.

Using this dataset, we developed the cisREAD method: integrating ATAC-seq and RNA-seq datasets to identify cis-Regulatory Elements Across Differentiation. cisREAD built on the core community detection and LASSO regression mechanism, from Vijayabaskar et al., to generalise to minimal data inputs. This involved: 1) differential analysis to identify candidate CREs and genes important to differentiation; 2) identification of

transcription factor binding sites through *de novo* motif discovery and binding site prediction; 3) for each gene, detection of co-accessible and co-bound cis-regulatory element communities; and 4) for each gene, selection of cis-regulatory elements (which are active when a gene is expressed) using LASSO regression.

Different methods to predict TF binding sites were explored during the development of cisREAD. This involved exploring different motif discovery tools and methods of binding site prediction. The footprint-dependent HINT-ATAC tool was compared to the footprint-independent BMO method, over concerns that some transiently binding transcription factors do not leave detectable footprints (Baek et al., 2017; Sung et al., 2016). When evaluating methods of binding site prediction, BMO was found to best predict binding sites detected by ChIP-seq. However, both BMO and HINT-ATAC were found to similarly predict TF-target genes when used within cisREAD. This suggested the cisREAD method was robust to choice of binding site predictions.

Whilst considerable adaptations were made to the Vijayabaskar et al. method, some elements were retained. This included the core community detection and LASSO regression mechanism, and the 100kb distance threshold used to assign candidate cis-regulatory elements to genes. Whilst cis-regulatory elements can operate from over a megabase away (Lettice et al., 2003), there were concerns that raising the threshold would lead to '$p >> n$' situations, which could increase the instability of LASSO beyond alleviation by community detection (Zou and Hastie, 2005). However CRISPR studies had since confirmed that most cis-regulatory elements regulate genes within 100kb. Fulco et al., 2019 reported that 84% of CREs which altered gene expression when peturbed operated within 100kb of their target gene, and Gasperini et al., 2019 reported a median distance of 24kb between a perturbed CRE and its altered target gene. These findings suggest that the 100kb threshold is appropriate, and would only miss a minority of cis-regulatory interactions.

Validation datasets (such as those used in chapter 5) could have been used to optimise genomic distance and other parameters, although with the risk of overfitting. Parameters which could have been optimised include the 'similarity score' threshold for community detection (default 0.3), and the use of λ_{min} or λ_{SE} in LASSO regression.

These are all provided as tunable parameters within the cisREAD R package. Furthermore, alternative methods of variable selection could have been explored. These could have included selection by elastic-net regression, or through feature importance scores from other machine learning models.

6.3 Discussion of biological interpretation of cisREAD results

Following its development and application to the B lineage dataset, chapter 4 saw the results of cisREAD interpreted in the context of B cell biology. This involved performing global analyses of predicted regulation (i.e., clustering, network analysis, ChIP-seq integration) and evaluating gene-specific models against the literature. The chapter started with a succession of analyses centered around the roles of key transcription factors. These were identified by the *de novo* motif analysis of the *in vitro* B cell ATAC-seq dataset, which was described in chapter 3.

Firstly, key transcription factor motifs were analyzed at the binding sites level: TF footprints were found to exhibit differential activity, increasing in strength and accessibility at temporally distinct stages. Secondly, transcription factor motifs were analyzed at the cis-regulatory element level: TF footprints were found to be differentially enriched in cis-regulatory elements, clustered by temporal accessibility. Finally, transcription factor motifs were analyzed at the gene level: mirroring the prior analysis, TF footprints were found to be differentially enriched in cis-regulatory elements linked to gene co-expression modules, each associated with functional pathways and processes. Together, these analyses identified shifts TF-led regulation which drive transitions between cell states. Crucially it revealed that a shift from PU.1/SPIB led regulation to AP-1 (BATF) led regulation was a key determinant of B cell activation, which was explored in the subsequent analysis. This involved integrating relevant ChIP-seq datasets to attribute cis-regulatory elements, genes, and pathways to specific transcription factors at PU.1/SPIB and AP-1 binding sites. These results corroborated similar findings for PU.1 and SPIB in murine B cells stimulated with T-independent stimuli (Willis et al. 2017), and assigned novel roles to the AP-1 transcription factor BATF. Altogether the work provided new detail into the timings and transcriptional effects of the shift from ETS to AP-1 factors, which has been previously reported in B cell differentiation (Ochiai et al., 2013, Scharer et al., 2018). In

addition, it highlighted importance of understudied factors RUNX3 (Thomsen et al., 2021) and ZBTB33 (Koh et al., 2013) in B cell activation.

The global analyses in chapter 4 formed a core section of the Emmett et al. publication. They demonstrated that cisREAD can be used to identify known and novel regulatory mechanisms. Furthermore, they contributed the first detailed description of the shift between PU.1, SPIB and BATF (AP-1) transcription factors in human B cell activation. This finding provided mechanistic insight into the dysregulation of SPIB and BATF in activated B cell subtype diffuse large B cell lymphoma; where higher SPIB expression is associated with an earlier cell of origin and increased survival, and higher BATF expression is associated with a later cell of origin and decreased survival (Care et al., 2014).

The global analyses in this chapter were performed on cisREAD-predicted 'enhancer' relationships. These were defined as regulatory elements with non-zero β coefficients, where accessibility was positively correlated with expression. The analyses were limited to positive regulatory elements for ease of interpretation. However many of the 'key' transcription factors (e.g. SPIB) also operate as transcriptional repressors (Schmidlin et al., 2008).

The extent that cisREAD-predicted, negatively correlated elements represent transcriptional silencers has yet to be investigated. These putative silencers could be characterised by future analysis of histone modifications, transcription factor motifs, and chromatin interactions. Whilst negative correlation has previously been used for computational prediction of silencers (Doni Jayavelu et al., 2020), most silencer elements are understood to be bifunctional (Segert et al., 2021). This means that cisREAD's correlation approach may not be identify silencers which act as enhancers in different cell stages. Similarly, cisREAD may miss enhancer relationships which do not exhibit correlation, such as priming elements.

Alongside the global analyses in chapter 4, two models for well-studied genes were evaluated with additional ChIP-seq and Hi-C data in context with the literature. Whilst not capable of evaluating performance at scale, these examples demonstrated that

cisREAD could recall both experimentally validated regions and suggest new modes of transcriptional control. These models also revealed limitations of cisREAD, such as the potential for sparse LASSO models to overlook super-enhancers, where multiple closely spaced, yet separate, regulatory elements control transcription of cell identity genes (Whyte et al., 2013). This happens when cisREAD places co-accessible super-enhancer constituents into different communities (due to binding site differences) and drops true predictors due to multicollinearity. This could be combatted by decreasing the similarity score threshold for community detection, or by detecting communities using only co-accessibility. Both the community detection threshold and score-type are provided as adjustable parameters in the cisREAD R package.

6.4 Discussion of benchmarking

Finally, chapter 5 saw cisREAD benchmarked against two published models (JEME and Activity-by-Contact) and two easily implementable methods (Pearson correlation and nearest gene). Benchmarking was performed by validation of B cell-specific predictions against B cell-specific chromatin interactions. These were obtained from promoter capture Hi-C and RNA polymerase II ChIA-PET datasets. The exercise revealed that, on average, cisREAD best predicted regulatory chromatin interactions, with the greatest F1-score averaged across four datasets. Considering each datatype separately, it was observed that cisREAD best predicted promoter capture Hi-C datasets (occurring over longer genomic distances) but performed second to Activity-by-Contact when validating against the ChIA-PET interaction set (occurring over shorter genomic distances). These differences could be explained by the distance bias of 3C based data, which is used within the Activity-by-Contact model, and not corrected for in the ChIA-PET dataset. Alternatively, distance may have been 'over-corrected' in the PC Hi-C dataset. Evaluation against additional validation datatypes like eQTLs, not biased by distance, would better evaluate the comparative performance of predictive models.

Alongside relative performance, the benchmarking exercise reported low performance metrics for all tested methods. This could represent the inadequacy of 3C-derived chromatin interactions to act as surrogate for cis-regulatory relationships. Again, more extensive validation using additional datatypes would offer a more comprehensive view of performance.

Finally, the benchmarking exercise in this chapter only evaluated pairwise interactions between promoter regions and distal CREs. A compelling feature of cisREAD is the detection of cis-regulatory element communities, designed to reflect co-regulation through multi-way interactions. In this thesis, interactions between members of cis-regulatory communities (coCREs) were not evaluated. Additional work looking at chromatin contacts within coCREs could provide evidence for or against the existence of predicted cis-regulatory communities.

6.5 Evaluation of Aims

This thesis addressed the problem of predicting gene regulation from multi-omics data. As introduced in chapter 1, dozens of researchers have addressed this challenge with a range of bioinformatics methods. Whilst there are many existing methods to predict gene regulation, this thesis argues there is still an unmet need for methods which are 1) applicable, 2) implementable and 3) interpretable. Therefore, the overarching aim of this thesis was to develop a method which: 1) required minimal data inputs, and thus could be applied to a range of datasets; 2) could be implemented through open-source software; and 3) was easy to interpret by biologists. By development of cisREAD, all three criteria were met. These are discussed as follows:

1. cisREAD is applicable to systems without ChIP-seq or 3D genomics data

Firstly, the significant adaptation of the Vijayabaskar et al. method to chromatin accessibility and gene expression inputs reduced the datasets required per sample. The original method required H3K27Ac and ChIP-seq for multiple TFs, in addition to these two datatypes, which were not available in the B cell differentiation dataset. Due to its streamlined input requirements, cisREAD is widely applicable to other systems without an abundance of high-throughput datatypes. This is a considerable advantage, as it means cisREAD can be applied to a wide range biological datasets with only one epigenomic and transcriptomic datatype. The benchmarking exercise in chapter 5 suggested that cisREAD performed favourably to published models (JEME and Activity-by-Contact) which used additional ChIP-seq or 3D genomics features.

2. cisREAD can be implemented through its R package

Secondly, to promote reproducibility, an R package was developed to detect cis-regulatory communities near a gene and predict those which control its transcription. This R package is freely available (and documented) at www.github.com/AmberEmmett/cisREAD. The cisREAD R package requires users to supply a chromatin accessibility matrix, gene expression matrix and TF binding matrix across multiple samples. It is recommended that the accessibility and expression matrices are subsetted for DARs/DEGs following step 1 of the cisREAD framework. The TF binding matrix should be generated following cisREAD step 2. This involves 1) selecting TF motifs through *de novo* motif discovery, 2) detecting binding sites through footprinting (or a footprint independent method), and 3) summarising predicted binding sites in a binary matrix.

Whilst these steps can be easily reproduced using popular bioinformatics tools, this part of the workflow was not implemented in the cisREAD R package. To improve reproducibility, these first two steps, could have been integrated by writing a wrapper for these tools, or integrating them into a reproducible pipeline with workflow managers like Nextflow or Snakemake. Other improvements could have included containerisation with Docker or Singularity, where software and dependencies are deployed together in a virtual environment; or development of a graphical user interface with RShiny to remove the barrier of programming skills.

3. The cisREAD method is interpretable to biologists studying gene regulation

Finally, cisREAD, like the previous Vijayabaskar et al. method, is informed by established mechanisms of transcriptional regulation. The concept of gene-specific coCREs reflects regulation within transcriptional hubs, where multiple co-accessible regulatory elements (including both promoters and enhancers) interact in 3D space to control transcription through transcription factor binding. The cisREAD method is therefore highly interpretable to biologists studying gene regulation, and the

framework facilitates further analysis of transcription factor-led regulation, as demonstrated in chapter 4.

Importantly, the cisREAD method (like Vijayabaskar et al.) prioritises the importance of regulatory interactions to systems of differentiation. The system-centric approach separates our cisREAD method from alternative models which predict cell-specific interactions (e.g., Activity-by-Contact and JEME), or interactions across large sets of unrelated samples (e.g., Thurman et al. or FOCS). Particularly, the method is better suited to small biological systems than other correlation approaches (like FOCs) as the community detection step of cisREAD avoids removing correlated features from the predictive set. This is important when the variety of conditions in the data is not high. This unique focus on systems biology enables users to identify the cis and trans regulators which drive progression through a set of cell-states or conditions. Altogether, this means cisREAD is highly useful to researchers studying gene regulation across systems of differentiation or disease.

Overall, the new cisREAD method fulfils the goal of the thesis. Its application to B cell differentiation has uncovered new regulatory mechanisms, and the method can be applied by other researchers to uncover further insights into gene regulation.

6.6 Directions for future research

The work in this thesis has contributed new methodology and knowledge of B cell differentiation. Future work could further develop the cisREAD methodology or explore gene-dysregulation in diseases of B cells. This section will recommend three avenues for future research: experimental testing of cis-regulatory mechanisms, adapting cisREAD for single cell datasets and using cisREAD-derived regulatory annotations for variant annotation.

6.6.1 Experimental validation of predicted gene regulation

In chapter 4, cisREAD was used to assign TFs and CREs to genes to identify transcription factor led regulation of B cell differentiation. Experimental validation of these global regulatory mechanisms (exemplified by the shift from PU.1/SPIB to AP-1

in B cell activation) could include: ChIP-seq to confirm binding of relevant transcription factors in different stages of B cell differentiation; and gene knockdown/knockout followed by RNA-seq to validate effects of transcription factor expression on target gene expression.

These techniques could be employed for experimental validation of the PU.1/SPIB-AP-1 shift. ChIP-seq for PU.1, SPIB and AP-1 factors (e.g. FOS, FRA1, BATF) spanning the B cell time-course would confirm temporal occupancy of transcription factor motifs, evaluating whether the same binding site (or cis-regulatory element) may be utilised by different factors at different time-points. ChIP-seq detected TF occupancy at promoters could also validate the predicted cis-regulatory interactions used to assign TFs to target genes. The effect of TFs on target gene expression could be investigated by RNAi (RNA interference, silencing transcription of the TF gene) or CRISPRi (CRISPR interference, repressing the transcription of the gene) followed by RNA-seq to detect transcriptomic changes in the absence /reduction of TF expression (Martinez et al. 2002, Qi et al. 2013).

cisREAD could aid future research by suggesting hypotheses of gene-specific regulation, enabling researchers to prioritise gene-specific candidate CREs for experimental testing. This application is exemplified by the AICDA and PRDM1 models in chapter 4, which evaluated cisREAD as a method for hypothesis generation. Experimentation could test the functionality of predicted gene-specific CREs using techniques including:

1. CRISPR-Cas9 perturbation of the candidate CRE (or a constituent motif) followed by RT-qPCR (quantitative polymerase chain reaction) of the target gene, to test the effect of the candidate CRE (or binding site) on RNA levels of the gene;
2. chromatin immunoprecipitation of candidate binding TFs, to test occupancy of predicted binding sites;
3. reporter assays, where the candidate CRE is placed upstream of a reporter gene with a measurable product, to test the ability of a CRE to enhance transcription; and

4. Chromosome Conformation Capture (3C) to test for chromatin contacts between distal CREs and the target gene promoter.

Applied to a large set of genes, cisREAD could suggest candidates for Massively Parallel Reporter Assays (MPRAs) or CRISPR screens (Arnold et al. 2013, Gasperini et al. 2019).

6.6.2 Adaptation to single cell multi-omics data

Over the period this research was conducted (2018-2023), biological research has shifted away from bulk sequencing towards single cell sequencing. The development of single cell 'omics techniques means that researchers can now probe the DNA, RNA, epigenome, and chromatin conformation of individual cells. The move towards single cell sequencing has revolutionised many fields, including the study of gene regulation. Importantly, this has led to the development of new tools, designed to account for the unique biases and statistical challenges of single cell data. Popular tools for studying gene regulation include Cicero and ArchR (Granja et al., 2021; Pliner et al., 2018). Future work could involve the adaptation of cisREAD to single cell multi-omics data, developing methodology capable of handling the sparsity of single cell datasets whilst retaining the core predictive workflow developed in this thesis.

6.6.3 Annotation of regulatory variants associated with B cell-specific diseases

Future research could also focus on applying B cell regulatory datasets, including ATAC-seq peaks and cisREAD-linked target genes, to the annotation of non-coding variants. This task is vital for the interpretation of clinical genomes and statistical genetics studies, where it can be difficult to prioritise non-coding variants and their target genes. As such, the work in this thesis could inform the characterisation of disease-associated variants, which exert their effects through gene regulation.

The regulatory datasets produced in this thesis could be used to annotate regulatory variants associated with diseases of B cells. These include somatic variants, found in cancer genomes, and germline variants, associated with disease phenotypes through genome-wide association studies. Characterising regulatory variants with B lineage annotations, could help elucidate the genetic underpinnings of B cell neoplasms (such as chronic lymphocytic leukaemia, B cell lymphomas and multiple myeloma) and B cell-

mediated autoimmune diseases. Previous studies have successfully utilised B cell-specific epigenomics, and 3D genomics, datasets to characterise regulatory drivers of B cell cancers (Arthur et al., 2018; Puente et al., 2015), and causal variants in autoimmune diseases (Farh et al., 2015; Javierre et al., 2016). Ultimately, the work presented in this thesis could lay the foundation for clinically translatable research into the genetic mechanisms underlying cancer and heritable disease.

References

Acosta-alvear, D., Zhou, Y., Blais, A., Tsikitis, M., Lents, N.H., Arias, C., Lennon, C.J., Kluger, Y., Dynlacht, B.D., 2007. XBP1 Controls Diverse Transcriptional Regulatory Networks. Mol. Cell 27, 53–66.

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

Akkaya, M., Kwak, K., Pierce, S.K., 2020. B cell memory: building two walls of protection against pathogens. Nat. Rev. Immunol. 20, 229–238.

Alexander, J.M., Guan, J., Li, B., Maliskova, L., Song, M., Shen, Y., Huang, B., Lomvardas, S., Weiner, O.D., 2019. Live-cell imaging reveals enhancer-dependent sox2 transcription in the absence of enhancer proximity. Elife 8, 1–42.

Amemiya, H.M., Kundaje, A., Boyle, A.P., 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. 9, 1–5.

Andersson, R., 2015. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. BioEssays 37, 314–323.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F.O., Jørgensen, M., Andersen, P.R., Bertin, N., Rackham, O., Burroughs, A.M., Baillie, J.K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C.J., Meehan, T.F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C.O., Heutink, P., Hume, D.A., Jensen, T.H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A.R.R., Carninci, P., Rehli, M., Sandelin, A., 2014. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461.

Andersson, R., Sandelin, A., 2020. Determinants of enhancer and promoter activities of regulatory elements. Nat. Rev. Genet.

Andrews, S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., Stark, A., 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science (80-.). 339, 1074–1077.

Arthur, S.E., Jiang, A., Grande, B.M., Alcaide, M., Cojocaru, R., Rushton, C.K., Mottok, A., Hilton, L.K., Lat, P.K., Zhao, E.Y., Culibrk, L., Ennishi, D., Jessa, S., Chong, L., Thomas, N., Pararajalingam, P., Meissner, B., Boyle, M., Davidson, J., Bushell, K.R., Lai, D., Farinha, P., Slack, G.W., Morin, G.B., Shah, S., Sen, D., Jones, S.J.M., Mungall, A.J., Gascoyne, R.D., Audas, T.E., Unrau, P., Marra, M.A., Connors, J.M., Steidl, C., Scott, D.W., Morin, R.D., 2018. Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. Nat. Commun. 9.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: Tool for the unification of biology. Nat. Genet. 25, 25–29.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R., 2021. Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods 18, 1196–1203.

Badia-i-Mompel, P., Wessels, L., Müller-Dott, S., Trimbour, R., Ramirez Flores, R.O., Argelaguet, R., Saez-Rodriguez, J., 2023. Gene regulatory network inference in the era of single-cell multi-omics. Nat. Rev. Genet.

Baek, S., Goldstein, I., Hager, G.L., 2017. Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. Cell Rep. 19, 1710–1722.

Bailey, T. and, Elkan, C., 1994. Fitting a mixture model by expectation maximization. AAAI Press 3–9.

Bailey, T.L., 2021. Sequence analysis STREME : accurate and versatile sequence motif discovery 37, 2834–2840.

Banerji, J., Rusconi, S., Schaffner, W., 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299–308.

Bannister, A.J., Kouzarides, T., 2011. Regulation of chromatin by histone modifications. Cell Res. 21, 381–395.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A., 2013. NCBI GEO: Archive for functional genomics data sets - Update. Nucleic Acids Res. 41, D991–D995.

Barwick, B.G., Scharer, C.D., Bally, A.P.R., Boss, J.M., 2016. Plasma cell differentiation is coupled to division-dependent DNA hypomethylation and gene regulation. Nat. Immunol. 17, 1216–1225.

Barwick, B.G., Scharer, C.D., Martinez, R.J., Price, M.J., Wein, A.N., Haines, R.R., Bally, A.P.R., Kohlmeier, J.E., Boss, J.M., 2018. B cell activation and plasma cell differentiation are inhibited by de novo DNA methylation. Nat. Commun. 9.

Basso, K., Dalla-Favera, R., 2012. Roles of BCL6 in normal and transformed germinal center B cells. Immunol. Rev. 247, 172–183.

Baubec, T., Schübeler, D., 2014. Genomic patterns and context specific interpretation of DNA methylation. Curr. Opin. Genet. Dev. 25, 85–92.

Beekman, R., Chapaprieta, V., Russiñol, N., Vilarrasa-blasi, R., Verdaguer-dot, N., Martens, J.H.A., Duran-ferrer, M., Kulis, M., Serra, F., Javierre, B.M., Wingett, S.W., Clot, G., Queirós, A.C., Castellano, G., Blanc, J., Gut, M., Merkel, A., Heath, S., Siebert, R., Martí-renom, M.A., Puente, X.S., López-otín, C., Gut, I., 2018. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nat. Med. 24.

Bejjani, F., Evanno, E., Zibara, K., Piechaczyk, M., Jariel-Encontre, I., 2019. The AP-1 transcriptional complex: Local switch or remote command? Biochim. Biophys. Acta - Rev. Cancer 1872, 11–23.

Benabdallah, N.S., Williamson, I., Illingworth, R.S., Kane, L., Boyle, S., Sengupta, D.,

Grimes, G.R., Therizols, P., Bickmore, W.A., 2019. Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. Mol. Cell 76, 473-484.e7.

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 57 , No . 1 (1995), Publi. J. R. Stat. Soc. 57, 289–300.

Berglund, L.J., Avery, D.T., Ma, C.S., Moens, L., Deenick, E.K., Bustamante, J., Boisson-dupuis, S., Wong, M., Adelstein, S., Arkwright, P.D., Bacchetta, R., Bezrodnik, L., Dadi, H., Roifman, C.M., Fulcher, D.A., Ziegler, J.B., Smart, J.M., Kobayashi, M., Picard, C., Durandy, A., Cook, M.C., Casanova, J., Uzel, G., Tangye, S.G., 2013. IL-21 signalling via STAT3 primes human na ¨ ı ve B cells to respond to IL-2 to enhance their differentiation into plasmablasts 122, 3940–3950.

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A. V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., Abraham, B.J., Afeyan, L.K., Guo, Y.E., Rimel, J.K., Fant, C.B., Schuijers, J., Lee, T.I., Taatjes, D.J., Young, R.A., 2018. Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell 175, 1842-1855.e16.

Bonifer, C., Cockerill, P.N., 2017. Chromatin priming of genes in development: Concepts, mechanisms and consequences. Exp. Hematol. 49, 1–8.

Bose, D.A., Donahue, G., Reinberg, D., Shiekhattar, R., Bonasio, R., Berger, S.L., 2017. RNA Binding to CBP Stimulates Histone Acetylation and Transcription. Cell 168, 135-149.e22.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., Crawford, G.E., 2008. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell 132, 311–322.

Brand, A.H., Breeden, L., Abraham, J., Sternglanz, R., Nasmyth, K., 1985. Characterization of a "silencer" in yeast: A DNA sequence with properties opposite to those of a transcriptional enhancer. Cell 41, 41–48.

Breiman, L., 2001. Random Forests 1–32.

Brescia, P., Schneider, C., Holmes, A.B., Shen, Q., Hussein, S., Pasqualucci, L., Basso, K., Dalla-Favera, R., 2018. MEF2B Instructs Germinal Center Development and Acts as an Oncogene in B Cell Lymphomagenesis. Cancer Cell 34, 453-465.e9.

Broad Institute, 2019. Picard Toolkit:A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218.

Buenrostro, J.D., Wu, B., Chang, H.Y., Greenleaf, W.J., 2015. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr. Protoc. Mol. Biol. 2015, 21.29.1-21.29.9.

Bullerwell, C.E., Robichaud, P.P., Deprez, P.M.L., Joy, A.P., Wajnberg, G., D'Souza, D., Chacko, S., Fournier, S., Crapoulet, N., Barnett, D.A., Lewis, S.M., Ouellette, R.J., 2021. EBF1 drives hallmark B cell gene expression by enabling the interaction of PAX5 with the MLL H3K4 methyltransferase complex. Sci. Rep. 11, 1–14.

Bunting, K.L., Soong, T.D., Singh, R., Jiang, Y., Béguelin, W., Poloway, D.W., Swed, B.L., Hatzi, K., Reisacher, W., Teater, M., Elemento, O., Melnick, A.M., 2016. Multi-tiered Reorganization of the Genome during B Cell Affinity Maturation Anchored by a Germinal Center-Specific Locus Control Region. Immunity 45, 497–512.

Cai, Y., Zhang, Y., Loh, Y.P., Tng, J.Q., Lim, M.C., Cao, Z., Raju, A., Lieberman Aiden, E., Li, S., Manikandan, L., Tergaonkar, V., Tucker-Kellogg, G., Fullwood, M.J., 2021. H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nat. Commun. 12.

Cairns, J., Freire-Pritchett, P., Wingett, S.W., Várnai, C., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.M., Osborne, C., Fraser, P., Spivakov, M., 2016. CHiCAGO: Robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 17, 1–17.

Calado, D.P., Sasaki, Y., Godinho, S.A., Pellerin, A., Köchert, K., Sleckman, B.P., De

Alborán, I.M., Janz, M., Rodig, S., Rajewsky, K., 2012. The cell-cycle regulator c-Myc is essential for the formation and maintenance of germinal centers. Nat. Immunol. 13, 1092–1100.

Calame, K., 2008. Activation-dependent induction of Blimp-1. Curr. Opin. Immunol. 20, 259–264.

Calo, E., Wysocka, J., 2013. Modification of Enhancer Chromatin: What, How, and Why? Mol. Cell 49, 825–837.

Cano-Gamez, E., Trynka, G., 2020. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front. Genet. 11, 1–21.

Cao, F., Fullwood, M.J., 2019. Inflated performance measures in enhancer–promoter interaction-prediction methods. Nat. Genet. 51, 1196–1204.

Cao, Q., Anyansi, C., Hu, X., Xu, L., Xiong, L., Tang, W., Mok, M.T.S., Cheng, C., Fan, X., Gerstein, M., Cheng, A.S.L., Yip, K.Y., 2017. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat. Genet. 49, 1428–1436.

Cao, Z., Sun, X., Icli, B., Wara, A.K., Feinberg, M.W., 2010. Role of Krüppel-like factors in leukocyte development, function, and disease. Blood 116, 4404–4414.

Carbon, S., Douglass, E., Good, B.M., Unni, D.R., Harris, N.L., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., Hartline, E., Fey, P., Thomas, P.D., Albou, L.P., Ebert, D., Kesling, M.J., Mi, H., Muruganujan, A., Huang, X., Mushayahama, T., LaBonte, S.A., Siegele, D.A., Antonazzo, G., Attrill, H., Brown, N.H., Garapati, P., Marygold, S.J., Trovisco, V., dos Santos, G., Falls, K., Tabone, C., Zhou, P., Goodman, J.L., Strelets, V.B., Thurmond, J., Garmiri, P., Ishtiaq, R., Rodríguez-López, M., Acencio, M.L., Kuiper, M., Lægreid, A., Logie, C., Lovering, R.C., Kramarz, B., Saverimuttu, S.C.C., Pinheiro, S.M., Gunn, H., Su, R., Thurlow, K.E., Chibucos, M., Giglio, M., Nadendla, S., Munro, J., Jackson, R., Duesbury, M.J., Del-Toro, N., Meldal, B.H.M., Paneerselvam, K., Perfetto, L., Porras, P., Orchard, S., Shrivastava, A., Chang, H.Y., Finn, R.D., Mitchell, A.L., Rawlings, N.D., Richardson, L., Sangrador-Vegas, A., Blake, J.A., Christie, K.R., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D.M., Harris, M.A., Oliver, S.G., Rutherford, K., Wood, V., Hayles, J., Bähler, J., Bolton,

E.R., de Pons, J.L., Dwinell, M.R., Hayman, G.T., Kaldunski, M.L., Kwitek, A.E., Laulederkind, S.J.F., Plasterer, C., Tutaj, M.A., Vedi, M., Wang, S.J., D'Eustachio, P., Matthews, L., Balhoff, J.P., Aleksander, S.A., Alexander, M.J., Cherry, J.M., Engel, S.R., Gondwe, F., Karra, K., Miyasato, S.R., Nash, R.S., Simison, M., Skrzypek, M.S., Weng, S., Wong, E.D., Feuermann, M., Gaudet, P., Morgat, A., Bakker, E., Berardini, T.Z., Reiser, L., Subramaniam, S., Huala, E., Arighi, C.N., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Bateman, A., Blatter, M.C., Boutet, E., Bowler, E., Breuza, L., Bridge, A., Britto, R., Bye-A-Jee, H., Casas, C.C., Coudert, E., Denny, P., Es-Treicher, A., Famiglietti, M.L., Georghiou, G., Gos, A.N., Gruaz-Gumowski, N., Hatton-Ellis, E., Hulo, C., Ignatchenko, A., Jungo, F., Laiho, K., Le Mercier, P., Lieberherr, D., Lock, A., Lussi, Y., MacDougall, A., Ma-Grane, M., Martin, M.J., Masson, P., Natale, D.A., Hyka-Nouspikel, N., Orchard, S., Pedruzzi, I., Pourcel, L., Poux, S., Pundir, S., Rivoire, C., Speretta, E., Sundaram, S., Tyagi, N., Warner, K., Zaru, R., Wu, C.H., Diehl, A.D., Chan, J.N., Grove, C., Lee, R.Y.N., Muller, H.M., Raciti, D., van Auken, K., Sternberg, P.W., Berriman, M., Paulini, M., Howe, K., Gao, S., Wright, A., Stein, L., Howe, D.G., Toro, S., Westerfield, M., Jaiswal, P., Cooper, L., Elser, J., 2021. The Gene Ontology resource: Enriching a GOld mine. Nucleic Acids Res. 49, D325–D334.

Care, M.A., Cocco, M., Laye, J.P., Barnes, N., Huang, Y., Wang, M., Barrans, S., Du, M., Jack, A., Westhead, D.R., Doody, G.M., Tooze, R.M., 2014. SPIB and BATF provide alternate determinants of IRF4 occupancy in diffuse large B-cell lymphoma linked to disease heterogeneity. Nucleic Acids Res. 42, 7591–7610.

Care, M.A., Westhead, D.R., Tooze, R.M., 2019. Parsimonious Gene Correlation Network Analysis (PGCNA): a tool to define modular gene co-expression for refined molecular stratification in cancer. npj Syst. Biol. Appl. 5, 1–17.

Carotta, S., Willis, S.N., Hasbold, J., Inouye, M., Pang, S.H.M., Emslie, D., Light, A., Chopin, M., Shi, W., Wang, H., Morse, H.C., Tarlinton, D.M., Corcoran, L.M., Hodgkin, P.D., Nutt, S.L., 2014. The transcription factors IRF8 and PU.1 negatively regulate plasma cell differentiation. J. Exp. Med. 211, 2169–2181.

Cedar, H., Sabag, O., Reizel, Y., 2022. The role of DNA methylation in genome-wide gene regulation during development. Dev. 149, 1–6.

Chaudhri, V.K., Dienger-Stambaugh, K., Wu, Z., Shrestha, M., Singh, H., 2020. Charting the cis-regulome of activated B cells by coupling structural and functional genomics. Nat. Immunol. 21, 210–220.

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., Gregor, T., 2018. Dynamic interplay between enhancer–promoter topology and gene activity. Nat. Genet. 50, 1296–1303.

Chen, M.J., Yokomizo, T., Zeigler, B., Dzierzak, E., Speck, A., 2009. Runx1 is required for the endothelial to hematopoietic cell transition but not thereafter. Nature 457, 887–891.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13-17-Augu, 785–794.

Cho, W.K., Spille, J.H., Hecht, M., Lee, C., Li, C., Grube, V., Cisse, I.I., 2018. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science (80-.). 361, 412–415.

Chu, V.T., Fröhlich, A., Steinhauser, G., Scheel, T., Roch, T., Fillatreau, S., Lee, J.J., Löhning, M., Berek, C., 2011. Eosinophils are required for the maintenance of plasma cells in the bone marrow. Nat. Immunol. 12, 151–159.

Clauset, A., Newman, M.E.J., Moore, C., 2004. Finding community structure in very large networks. Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top. 70, 6.

Cobaleda, C., Schebesta, A., Delogu, A., Busslinger, M., 2007. Pax5: The guardian of B cell identity and function. Nat. Immunol. 8, 463–470.

Cocco, M., Care, M.A., Saadi, A., Al-maskari, M., Doody, G., Tooze, R., 2020. A dichotomy of gene regulatory associations during the activated B-cell to plasmablast transition. Life Sci. Alliance 3, 1–20.

Cocco, M., Stephenson, S., Care, M.A., Newton, D., Barnes, N.A., Davison, A., Rawstron, A., Westhead, D.R., Doody, G.M., Tooze, R.M., 2012. In Vitro Generation of Long-lived Human Plasma Cells. J. Immunol. 189, 5773–5785.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S.,

Nelson, S.F., Pellegrini, M., Jacobsen, S.E., 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452, 215–219.

Consortium, T.G., 2021. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science (80-.). 369, 1318–1330.

Corcoran, L., Emslie, D., Kratina, T., Shi, W., Hirsch, S., Taubenheim, N., Chevrier, S., 2014. Oct2 and Obf1 as facilitators of B: T cell collaboration during a humoral immune response. Front. Immunol. 5, 1–13.

Core, L.J., Waterfall, J.J., List, J.T., 2008. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. Science (80-.). 322, 1845–1848.

Cornelis, R., Hahne, S., Taddeo, A., Petkau, G., Malko, D., Durek, P., Thiem, M., Heiberger, L., Peter, L., Mohr, E., Klaeden, C., Tokoyoda, K., Siracusa, F., Hoyer, B.F., Hiepe, F., Mashreghi, M.F., Melchers, F., Chang, H.D., Radbruch, A., 2020. Stromal Cell-Contact Dependent PI3K and APRIL Induced NF-κB Signaling Prevent Mitochondrial- and ER Stress Induced Death of Memory Plasma Cells. Cell Rep. 32.

Crouch, E.E., Li, Z., Takizawa, M., Fichtner-Feigl, S., Gourzi, P., Montaño, C., Feigenbaum, L., Wilson, P., Janz, S., Papavasiliou, F.N., Casellas, R., 2007. Regulation of AID expression in the immune response. J. Exp. Med. 204, 1145–1156.

Cruz-Molina, S., Respuela, P., Tebartz, C., Kolovos, P., Nikolic, M., Fueyo, R., van Ijcken, W.F.J., Grosveld, F., Frommolt, P., Bazzi, H., Rada-Iglesias, A., 2017. PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. Cell Stem Cell 20, 689-705.e9.

Cutter, A.R., Hayes, J.J., 2015. A brief review of nucleosome structure. FEBS Lett. 589, 2914–2922.

Cyster, J.G., Allen, C.D.C., 2019. B Cell Responses: Cell Interaction Dynamics and Decisions. Cell 177, 524–540.

D'Oliveira Albanus, R., Kyono, Y., Hensley, J., Varshney, A., Orchard, P., Kitzman, J.O.,

Parker, S.C.J., 2021. Chromatin information content landscapes inform transcription factor and DNA interactions. Nat. Commun. 12, 1–12.

Dao, L.T.M., Spicuglia, S., 2018. Transcriptional regulation by promoters with enhancer function. Transcription 9, 307–314.

Davidson, I.F., Peters, J.M., 2021. Genome folding through loop extrusion by SMC complexes. Nat. Rev. Mol. Cell Biol. 22, 445–464.

Davis, R.E., Ngo, V.N., Lenz, G., Tolar, P., Young, R.M., Romesser, P.B., Kohlhammer, H., Lamy, L., Zhao, H., Yang, Y., Xu, W., Shaffer, A.L., Wright, G., Xiao, W., Powell, J., Jiang, J.K., Thomas, C.J., Rosenwald, A., Ott, G., Muller-Hermelink, H.K., Gascoyne, R.D., Connors, J.M., Johnson, N.A., Rimsza, L.M., Campo, E., Jaffe, E.S., Wilson, W.H., Delabie, J., Smeland, E.B., Fisher, R.I., Braziel, R.M., Tubbs, R.R., Cook, J.R., Weisenburger, D.D., Chan, W.C., Pierce, S.K., Staudt, L.M., 2010. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. Nature 463, 88–92.

de Almeida, B.P., Reiter, F., Pagani, M., Stark, A., 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat. Genet. 54, 613–624.

de Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., Natoli, G., 2010. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. PLoS Biol. 8.

De Silva, N.S., Anderson, M.M., Carette, A., Silva, K., Heise, N., Bhagat, G., Klein, U., 2016. Transcription factors of the alternative NF-kB pathway are required for germinal center B-cell development. Proc. Natl. Acad. Sci. U. S. A. 113, 9063–9068.

Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'Shea, C.C., Park, P.J., Ren, B., Ritland Politz, J.C., Shendure, J., Zhong, S., 2017. The 4D nucleome project. Nature 549, 219–226.

Dekker, J., Rippe, K., Dekker, M., Kleckner, N., 2002. Capturing Chromosome Conformation. Science (80-.). 295, 1306–1311.

Delogu, A., Schebesta, A., Sun, Q., Aschenbrenner, K., Perlot, T., Busslinger, M., 2006.

Gene repression by Pax5 in B cells is essential for blood cell homeostasis and is reversed in plasma cells. Immunity 24, 269–281.

Di Giammartino, D.C., Polyzos, A., Apostolou, E., 2020. Transcription factors: building hubs in the 3D space. Cell Cycle 19, 2395–2410.

Dice, L.R., 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26, 297–302.

Diehl, S.A., Schmidlin, H., Nagasawa, M., van Haren, S.D., Kwakkenbos, M.J., Yasuda, E., Beaumont, T., Scheeren, F.A., Spits, H., 2008. STAT3-Mediated Up-Regulation of BLIMP1 Is Coordinated with BCL6 Down-Regulation to Control Human Plasma Cell Differentiation. J. Immunol. 180, 4805–4815.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380.

Dominguez-Sola, D., Victora, G.D., Ying, C.Y., Phan, R.T., Saito, M., Nussenzweig, M.C., Dalla-Favera, R., 2012. The proto-oncogene MYC is required for selection in the germinal center and cyclic reentry. Nat. Immunol. 13, 1083–1091.

Doni Jayavelu, N., Jajodia, A., Mishra, A., Hawkins, R.D., 2020. Candidate silencer elements for the human and mouse genomes. Nat. Commun. 11, 1–15.

Doody, G.M., Care, M.A., Burgoyne, N.J., Bradford, J.R., Bota, M., Bonifer, C., Westhead, D.R., Tooze, R.M., 2010. An extended set of PRDM1/BLIMP1 target genes links binding motif type to dynamic repression. Nucleic Acids Res. 38, 5336–5350.

Doody, G.M., Stephenson, S., McManamy, C., Tooze, R.M., 2007. PRDM1/BLIMP-1 Modulates IFN-γ-Dependent Control of the MHC Class I Antigen-Processing and Peptide-Loading Pathway. J. Immunol. 179, 7614–7623.

Doody, G.M., Stephenson, S., Tooze, R.M., 2006. BLIMP-1 is a target of cellular stress and downstream of the unfolded protein response. Eur. J. Immunol. 36, 1572–1582.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B.,

Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B.R., Landt, S.G., Lee, B.K., Pauli, F., Rosenbloom, K.R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J.M., Song, L., Trinklein, N.D., Altshuler, R.C., Birney, E., Brown, J.B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T.S., Gerstein, M., Giardine, B., Greven, M., Hardison, R.C., Harris, R.S., Herrero, J., Hoffman, M.M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G.K., Merkel, A., Mortazavi, A., Parker, S.C.J., Reddy, T.E., Rozowsky, J., Schlesinger, F., Thurman, R.E., Wang, J., Ward, L.D., Whitfield, T.W., Wilder, S.P., Wu, W., Xi, H.S., Yip, K.Y., Zhuang, J., Bernstein, B.E., Green, E.D., Gunter, C., Snyder, M., Pazin, M.J., Lowdon, R.F., Dillon, L.A.L., Adams, L.B., Kelly, C.J., Zhang, J., Wexler, J.R., Good, P.J., Feingold, E.A., Crawford, G.E., Dekker, J., Elnitski, L., Farnham, P.J., Giddings, M.C., Gingeras, T.R., Guigó, R., Hubbard, T.J., Kent, W.J., Lieb, J.D., Margulies, E.H., Myers, R.M., Stamatoyannopoulos, J.A., Tenenbaum, S.A., Weng, Z., White, K.P., Wold, B., Yu, Y., Wrobel, J., Risk, B.A., Gunawardena, H.P., Kuiper, H.C., Maier, C.W., Xie, L., Chen, X., Mikkelsen, T.S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M.J., Durham, T., Ku, M., Truong, T., Eaton, M.L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B.A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O.J., Park, E., Preall, J.B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K.S., Schaeffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, Huaien, Hayashizaki, Y., Reymond, A., Antonarakis, S.E., Hannon, G.J., Ruan, Y., Carninci, P., Sloan, C.A., Learned, K., Malladi, V.S., Wong, M.C., Barber, G.P., Cline, M.S., Dreszer, T.R., Heitner, S.G., Karolchik, D., Kirkup, V.M., Meyer, L.R., Long, J.C., Maddren, M., Raney, B.J., Grasfeder, L.L., Giresi, P.G., Battenhouse, A., Sheffield, N.C., Showers, K.A., London, D., Bhinge, A.A., Shestak, C., Schaner, M.R., Kim, S.K., Zhang, Z.Z., Mieczkowski, P.A., Mieczkowska, J.O., Liu, Z., McDaniell, R.M., Ni, Y., Rashid, N.U., Kim, M.J., Adar, S., Zhang, Zhancheng, Wang, T., Winter, D., Keefe, D., Iyer, V.R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E.C., Varley, K.E., Gasper,

C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K.M., Anaya, M., Cross, M.K., Muratet, M.A., Newberry, K.M., McCue, K., Nesmith, A.S., Fisher-Aylor, K.I., Pusey, B., DeSalvo, G., Parker, S.L., Balasubramanian, Sreeram, Davis, N.S., Meadows, S.K., Eggleston, T., Newberry, J.S., Levy, S.E., Absher, D.M., Wong, W.H., Blow, M.J., Visel, A., Pennachio, L.A., Petrykowska, H.M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J.M., Griffiths, E., Harte, R., Hendrix, D.A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M.F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J.M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M.L., Van Baren, M.J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Zhengdong, Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R.P., Auerbach, R.K., Balasubramanian, Suganthi, Bettinger, K., Bhardwaj, N., Boyle, A.P., Cao, A.R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J.D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V.X., Karczewski, K.J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X.J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K.K., Yang, X., Struhl, K., Weissman, S.M., Penalva, L.O., Karmakar, S., Bhanvadia, R.R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D.L., Byron, R., Canfield, T.K., Diegel, M.J., Dunn, D., Ebersol, A.K., Frum, T., Garg, K., Gist, E., Hansen, R.S., Boatman, L., Haugen, E., Humbert, R., Johnson, A.K., Johnson, E.M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M.T., Neph, S.J., Neri, F. V., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Rynes, E., Sanchez, M.E., Sandstrom, R.S., Shafer, A.O., Stergachis, A.B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, Hao, Weaver, M.A., Yan, Y., Zhang, M., Akey, J.M., Bender, M., Dorschner, M.O., Groudine, M., MacCoss, M.J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N.M., Sobral, D., Vaquerizas, J.M., Batzoglou, S., Sidow, A.,

Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M.W., Schaub, M.A., Miller, W., Bickel, P.J., Banfai, B., Boley, N.P., Huang, H., Li, J.J., Noble, W.S., Bilmes, J.A., Buske, O.J., Sahu, A.D., Kharchenko, P. V., Park, P.J., Baker, D., Taylor, J., Lochovsky, L., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Emmett, A.M.L., Saadi, A., Care, M.A., Doody, G., Tooze, R.M., Westhead, D.R., 2023. Integration of chromatin accessibility and gene expression data with cisREAD reveals a switch from PU.1/SPIB-driven to AP-1-driven gene regulation during B cell activation. bioRxiv

Emslie, D., 'Costa, K.D., Hasbold, J., Metcalf, D., Takatsu, K., Hodgkin, P.O., Corcoran, L.M., 2008. Oct2 enhances antibody-secreting cell differentiation through regulation of IL-5 receptor α chain expression on activated B cells. J. Exp. Med. 205, 409–421.

ENCODE Project Consortium, Abascal, F., Acosta, R., Addleman, N.J., Adrian, J., Afzal, V., Aken, B., Akiyama, J.A., Jammal, O. Al, Amrhein, H., Anderson, S.M., Andrews, G.R., Antoshechkin, I., Ardlie, K.G., Armstrong, J., Astley, M., Banerjee, B., Barkal, A.A., Barnes, I.H.A., Barozzi, I., Barrell, D., Barson, G., Bates, D., Baymuradov, U.K., Bazile, C., Beer, M.A., Beik, S., Bender, M.A., Bennett, R., Bouvrette, L.P.B., Bernstein, B.E., Berry, A., Bhaskar, A., Bignell, A., Blue, S.M., Bodine, D.M., Boix, C., Boley, N., Borrman, T., Borsari, B., Boyle, A.P., Brandsmeier, L.A., Breschi, A., Bresnick, E.H., Brooks, J.A., Buckley, M., Burge, C.B., Byron, R., Cahill, E., Cai, L., Cao, L., Carty, M., Castanon, R.G., Castillo, A., Chaib, H., Chan, E.T., Chee, D.R., Chee, S., Chen, Hao, Chen, Huaming, Chen, J.Y., Chen, S., Cherry, J.M., Chhetri, S.B., Choudhary, J.S., Chrast, J., Chung, D., Clarke, D., Cody, N.A.L., Coppola, C.J., Coursen, J., D'Ippolito, A.M., Dalton, S., Danyko, C., Davidson, C., Davila-Velderrain, J., Davis, C.A., Dekker, J., Deran, A., DeSalvo, G., Despacio-Reyes, G., Dewey, C.N., Dickel, D.E., Diegel, M., Diekhans, M., Dileep, V., Ding, B., Djebali, S., Dobin, A., Dominguez, D., Donaldson, S., Drenkow, J., Dreszer, T.R., Drier, Y., Duff, M.O., Dunn, D., Eastman, C., Ecker, J.R., Edwards, M.D., El-Ali, N., Elhajjajy, S.I., Elkins, K., Emili, A., Epstein, C.B., Evans, R.C., Ezkurdia, I., Fan, K., Farnham, P.J., Farrell, N.P., Feingold, E.A., Ferreira, A.M., Fisher-Aylor, K., Fitzgerald, S., Flicek, P., Foo, C.S., Fortier, K., Frankish, A., Freese, P., Fu, S., Fu, X.D., Fu, Y., Fukuda-

Yuzawa, Y., Fulciniti, M., Funnell, A.P.W., Gabdank, I., Galeev, T., Gao, M., Giron, C.G., Garvin, T.H., Gelboin-Burkhart, C.A., Georgolopoulos, G., Gerstein, M.B., Giardine, B.M., Gifford, D.K., Gilbert, D.M., Gilchrist, D.A., Gillespie, S., Gingeras, T.R., Gong, P., Gonzalez, A., Gonzalez, J.M., Good, P., Goren, A., Gorkin, D.U., Graveley, B.R., Gray, M., Greenblatt, J.F., Griffiths, E., Groudine, M.T., Grubert, F., Gu, M., Guigó, R., Guo, H., Guo, Yu, Guo, Yuchun, Gursoy, G., Gutierrez-Arcelus, M., Halow, J., Hardison, R.C., Hardy, M., Hariharan, M., Harmanci, A., Harrington, A., Harrow, J.L., Hashimoto, T.B., Hasz, R.D., Hatan, M., Haugen, E., Hayes, J.E., He, P., He, Y., Heidari, N., Hendrickson, D., Heuston, E.F., Hilton, J.A., Hitz, B.C., Hochman, A., Holgren, C., Hou, L., Hou, S., Hsiao, Y.H.E., Hsu, S., Huang, H., Hubbard, T.J., Huey, J., Hughes, T.R., Hunt, T., Ibarrientos, S., Issner, R., Iwata, M., Izuogu, O., Jaakkola, T., Jameel, N., Jansen, C., Jiang, L., Jiang, P., Johnson, A., Johnson, R., Jungreis, I., Kadaba, M., Kasowski, M., Kasparian, M., Kato, M., Kaul, R., Kawli, T., Kay, M., Keen, J.C., Keles, S., Keller, C.A., Kelley, D., Kellis, M., Kheradpour, P., Kim, D.S., Kirilusha, A., Klein, R.J., Knoechel, B., Kuan, S., Kulik, M.J., Kumar, S., Kundaje, A., Kutyavin, T., Lagarde, J., Lajoie, B.R., Lambert, N.J., Lazar, J., Lee, A.Y., Lee, D., Lee, E., Lee, J.W., Lee, K., Leslie, C.S., Levy, S., Li, B., Li, H., Li, N., Li, X., Li, Y.I., Li, Ying, Li, Yining, Li, Yue, Lian, J., Libbrecht, M.W., Lin, S., Lin, Y., Liu, D., Liu, J., Liu, P., Liu, T., Liu, X.S., Liu, Yan, Liu, Yaping, Long, M., Lou, S., Loveland, J., Lu, A., Lu, Y., Lécuyer, E., Ma, L., Mackiewicz, M., Mannion, B.J., Mannstadt, M., Manthravadi, D., Marinov, G.K., Martin, F.J., Mattei, E., McCue, K., McEown, M., McVicker, G., Meadows, S.K., Meissner, A., Mendenhall, E.M., Messer, C.L., Meuleman, W., Meyer, C., Miller, S., Milton, M.G., Mishra, T., Moore, D.E., Moore, H.M., Moore, J.E., Moore, S.H., Moran, J., Mortazavi, A., Mudge, J.M., Munshi, N., Murad, R., Myers, R.M., Nandakumar, V., Nandi, P., Narasimha, A.M., Narayanan, A.K., Naughton, H., Navarro, F.C.P., Navas, P., Nazarovs, J., Nelson, J., Neph, S., Neri, F.J., Nery, J.R., Nesmith, A.R., Newberry, J.S., Newberry, K.M., Ngo, V., Nguyen, R., Nguyen, T.B., Nguyen, T., Nishida, A., Noble, W.S., Novak, C.S., Novoa, E.M., Nuñez, B., O'Donnell, C.W., Olson, S., Onate, K.C., Otterman, E., Ozadam, H., Pagan, M., Palden, T., Pan, X., Park, Y., Partridge, E.C., Paten, B., Pauli-Behn, F., Pazin, M.J., Pei, B., Pennacchio, L.A., Perez, A.R., Perry, E.H., Pervouchine, D.D., Phalke, N.N., Pham, Q., Phanstiel, D.H., Plajzer-Frick, I., Pratt, G.A., Pratt, H.E., Preissl, S., Pritchard, J.K., Pritykin, Y.,

Purcaro, M.J., Qin, Q., Quinones-Valdez, G., Rabano, I., Radovani, E., Raj, A., Rajagopal, N., Ram, O., Ramirez, L., Ramirez, R.N., Rausch, D., Raychaudhuri, S., Raymond, J., Razavi, R., Reddy, T.E., Reimonn, T.M., Ren, B., Reymond, A., Reynolds, A., Rhie, S.K., Rinn, J., Rivera, M., Rivera-Mulia, J.C., Roberts, B.S., Rodriguez, J.M., Rozowsky, J., Ryan, R., Rynes, E., Salins, D.N., Sandstrom, R., Sasaki, T., Sathe, S., Savic, D., Scavelli, A., Scheiman, J., Schlaffner, C., Schloss, J.A., Schmitges, F.W., See, L.H., Sethi, A., Setty, M., Shafer, A., Shan, S., Sharon, E., Shen, Q., Shen, Y., Sherwood, R.I., Shi, M., Shin, S., Shoresh, N., Siebenthall, K., Sisu, C., Slifer, T., Sloan, C.A., Smith, A., Snetkova, V., Snyder, M.P., Spacek, D. V., Srinivasan, S., Srivas, R., Stamatoyannopoulos, G., Stamatoyannopoulos, J.A., Stanton, R., Steffan, D., Stehling-Sun, S., Strattan, J.S., Su, A., Sundararaman, B., Suner, M.M., Syed, T., Szynkarek, M., Tanaka, F.Y., Tenen, D., Teng, M., Thomas, J.A., Toffey, D., Tress, M.L., Trout, D.E., Trynka, G., Tsuji, J., Upchurch, S.A., Ursu, O., Uszczynska-Ratajczak, B., Uziel, M.C., Valencia, A., Biber, B. Van, van der Velde, A.G., Van Nostrand, E.L., Vaydylevich, Y., Vazquez, J., Victorsen, A., Vielmetter, J., Vierstra, J., Visel, A., Vlasova, A., Vockley, C.M., Volpi, S., Vong, S., Wang, H., Wang, M., Wang, Q., Wang, R., Wang, T., Wang, W., Wang, X., Wang, Y., Watson, N.K., Wei, X., Wei, Z., Weisser, H., Weissman, S.M., Welch, R., Welikson, R.E., Weng, Z., Westra, H.J., Whitaker, J.W., White, C., White, K.P., Wildberg, A., Williams, B.A., Wine, D., Witt, H.N., Wold, B., Wolf, M., Wright, J., Xiao, R., Xiao, X., Xu, Jie, Xu, Jinrui, Yan, K.K., Yan, Y., Yang, H., Yang, X., Yang, Y.W., Yardımcı, G.G., Yee, B.A., Yeo, G.W., Young, T., Yu, T., Yue, F., Zaleski, C., Zang, C., Zeng, H., Zeng, W., Zerbino, D.R., Zhai, J., Zhan, L., Zhan, Y., Zhang, B., Zhang, Jialing, Zhang, Jing, Zhang, K., Zhang, L., Zhang, P., Zhang, Q., Zhang, X.O., Zhang, Y., Zhang, Z., Zhao, Y., Zheng, Y., Zhong, G., Zhou, X.Q., Zhu, Y., Zimmerman, J., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., Halow, J., Van Nostrand, E.L., Freese, P., Gorkin, D.U., Shen, Y., He, Y., Mackiewicz, M., Pauli-Behn, F., Williams, B.A., Mortazavi, A., Keller, C.A., Zhang, X.O., Elhajjajy, S.I., Huey, J., Dickel, D.E., Snetkova, V., Wei, X., Wang, X., Rivera-Mulia, J.C., Rozowsky, J., Zhang, Jing, Chhetri, S.B., Zhang, Jialing, Victorsen, A., White, K.P., Visel, A., Yeo, G.W., Burge, C.B., Lécuyer, E., Gilbert, D.M., Dekker, J., Rinn, J., Mendenhall, E.M., Ecker, J.R., Kellis, M., Klein, R.J., Noble, W.S., Kundaje, A., Guigó, R., Farnham, P.J., Cherry, J.M., Myers, R.M., Ren,

B., Graveley, B.R., Gerstein, M.B., Pennacchio, L.A., Snyder, M.P., Bernstein, B.E., Wold, B., Hardison, R.C., Gingeras, T.R., Stamatoyannopoulos, J.A., Weng, Z., 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710.

Erceg, J., Saunders, T.E., Girardot, C., Devos, D.P., Hufnagel, L., Furlong, E.E.M., 2014. Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer's Activity. PLoS Genet. 10.

Ernst, J., Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 9, 215–216.

Ernst, J., Kellis, M., 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol. 33, 364–376.

Ersching, J., Efeyan, A., Mesin, L., Jacobsen, J.T., Pasqual, G., Grabiner, B.C., Dominguez-Sola, D., Sabatini, D.M., Victora, G.D., 2017. Germinal Center Selection and Affinity Maturation Require Dynamic Regulation of mTORC1 Kinase. Immunity 46, 1045-1058.e6.

Erwin, G.D., Oksenberg, N., Truty, R.M., Kostka, D., Murphy, K.K., Ahituv, N., Pollard, K.S., Capra, J.A., 2014. Integrating Diverse Datasets Improves Developmental Enhancer Prediction. PLoS Comput. Biol. 10, 1–20.

Farh, K.K.H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., Hatan, M., Carrasco-Alfonso, M.J., Mayer, D., Luckey, C.J., Patsopoulos, N.A., De Jager, P.L., Kuchroo, V.K., Epstein, C.B., Daly, M.J., Hafler, D.A., Bernstein, B.E., 2015. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343.

Farnung, L., Vos, S.M., 2022. Assembly of RNA polymerase II transcription initiation complexes. Curr. Opin. Struct. Biol. 73, 102335.

Fernández, M., Miranda-Saavedra, D., 2012. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. Nucleic Acids Res. 40.

Filippakopoulos, P., Knapp, S., 2012. The bromodomain interaction module. FEBS Lett.

586, 2692–2704.

Firpi, H.A., Ucar, D., Tan, K., 2010. Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics 26, 1579–1586.

Fornes, O., Castro-Mondragon, J.A., Khan, A., Van Der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W.W., Mathelier, A., 2020. JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. Nucleic Acids Res. 48, D87–D92.

Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., Nguyen, T.H., Kane, M., Perez, E.M., Durand, N.C., Lareau, C.A., Stamenova, E.K., Aiden, E.L., Lander, E.S., Engreitz, J.M., 2019. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat. Genet. 51, 1664–1669.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G.Y., Huang, P.Y.H., Welboren, W.J., Han, Y., Ooi, H.S., Ariyaratne, P.N., Vega, V.B., Luo, Y., Tan, P.Y., Choy, P.Y., Wansa, K.D.S.A., Zhao, B., Lim, K.S., Leow, S.C., Yow, J.S., Joseph, R., Li, H., Desai, K. V., Thomsen, J.S., Lee, Y.K., Karuturi, R.K.M., Herve, T., Bourque, G., Stunnenberg, H.G., Ruan, X., Cacheux-Rataboul, V., Sung, W.K., Liu, E.T., Wei, C.L., Cheung, E., Ruan, Y., 2009. An oestrogen-receptor-α-bound human chromatin interactome. Nature 462, 58–64.

Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., Trapnell, C., Ahituv, N., Shendure, J., 2019. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell 176, 377-390.e19.

Gasperini, M., Tome, J.M., Shendure, J., 2020. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat. Rev. Genet. 21, 292–310.

Gerondakis, S., Siebenlist, U., 2010. Roles of the NF-kappaB pathway in lymphocyte development and function. Cold Spring Harb. Perspect. Biol. 2, 1–29.

Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., Furlong,

E.E.M., 2014. Enhancer loops appear stable during development and are associated with paused polymerase. Nature 512, 96–100.

Gisselbrecht, S.S., Palagi, A., Kurland, J. V., Rogers, J.M., Ozadam, H., Zhan, Y., Dekker, J., Bulyk, M.L., 2020. Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. Mol. Cell 77, 324-337.e8.

Gloury, R., Zotos, D., Zuidscherwoude, M., Masson, F., Liao, Y., Hasbold, J., Corcoran, L.M., Hodgkin, P.D., Belz, G.T., Shi, W., Nutt, S.L., Tarlinton, D.M., Kallies, A., 2016. Dynamic changes in Id3 and E-protein activity orchestrate germinal center and plasma cell development. J. Exp. Med. 213, 1095–1111.

Goode, D.K., Obier, N., Vijayabaskar, M.S., Lie-A-Ling, M., Lilly, A.J., Hannah, R., Lichtinger, M., Batta, K., Florkowska, M., Patel, R., Challinor, M., Wallace, K., Gilmour, J., Assi, S.A., Cauchy, P., Hoogenkamp, M., Westhead, D.R., Lacaud, G., Kouskoff, V., Göttgens, B., Bonifer, C., 2016. Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Dev. Cell 36, 572–587.

Gorbovytska, V., Kim, S.K., Kuybu, F., Götze, M., Um, D., Kang, K., Pittroff, A., Brennecke, T., Schneider, L.M., Leitner, A., Kim, T.K., Kuhn, C.D., 2022. Enhancer RNAs stimulate Pol II pause release by harnessing multivalent interactions to NELF. Nat. Commun. 13.

Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., Greenleaf, W.J., 2021. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. 53, 935.

Grötsch, B., Brachs, S., Lang, C., Luther, J., Derer, A., Schlötzer-Schrehardt, U., Bozec, A., Fillatreau, S., Berberich, I., Hobeika, E., Reth, M., Wagner, E.F., Schett, G., Mielenz, D., David, J.P., 2014. The AP-1 transcription factor Fra1 inhibits follicular B cell differentiation into plasma cells. J. Exp. Med. 211, 2199–2212.

Gusmao, E.G., Dieterich, C., Zenke, M., Costa, I.G., 2014. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics 30, 3143–3151.

Hait, T.A., Amar, D., Shamir, R., Elkon, R., 2018. FOCS: A novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map.

Genome Biol. 19, 1–14.

Hait, T.A., Elkon, R., 2022. CT-FOCS : a novel method for inferring cell type-specific enhancer – promoter maps 50.

Hargreaves, D.C., Hyman, P.L., Lu, T.T., Ngo, V.N., Bidgol, A., Suzuki, G., Zou, Y.R., Littman, D.R., Cyster, J.G., 2001. A coordinated change in chemokine responsiveness guides plasma cell movements. J. Exp. Med. 194, 45–56.

Hasegawa, Y., Struhl, K., 2021. Different SP1 binding dynamics at individual genomic loci in human cells. Proc. Natl. Acad. Sci. U. S. A. 118.

He, B., Chen, C., Teng, L., Tan, K., 2014. Global view of enhancer-promoter interactome in human cells. Proc. Natl. Acad. Sci. 111, E2191–E2199.

He, C., Zhang, M.Q., Wang, X., 2015. MICC: An R package for identifying chromatin interactions from ChIA-PET data. Bioinformatics 31, 3832–3834.

He, Y., Gorkin, D.U., Dickel, D.E., Nery, J.R., Castanon, R.G., Lee, A.Y., Shen, Y., Visel, A., Pennacchio, L.A., Ren, B., Ecker, J.R., 2017. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. Proc. Natl. Acad. Sci. 114, E1633–E1640.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E., Ren, B., 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. 39, 311–318.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell 38, 576–589.

Heise, N., de Silva, N.S., Silva, K., Carette, A., Simonetti, G., Pasparakis, M., Klein, U., 2014. Germinal center B cell maintenance and differentiation are controlled by distinct NF-κB transcription factor subunits. J. Exp. Med. 211, 2103–2118.

Hipp, N., Symington, H., Pastoret, C., Caron, G., Monvoisin, C., Tarte, K., Fest, T., Delaloy, C., 2017. IL-2 imprints human naive B cell fate towards plasma cell

through ERK/ELK1-mediated BACH2 repression. Nat. Commun. 8.

Hnisz, D., Schuijers, J., Lin, C.Y., Weintraub, A.S., Abraham, B.J., Lee, T.I., Bradner, J.E., Young, R.A., 2015. Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. Mol. Cell 58, 362–370.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., Sharp, P.A., 2017. A Phase Separation Model for Transcriptional Control. Cell 169, 13–23.

Hoellinger, T., Mestre, C., Aschard, H., Goff, W. Le, Foissac, S., Faraut, T., Djebali, S., 2023. Enhancer / gene relationships : Need for more reliable genome-wide reference sets 1–12.

Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Proble. Technometrics 12, 55–67.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., Noble, W.S., 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods 9, 473–476.

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., Hardison, R.C., Dunham, I., Kellis, M., Noble, W.S., 2013. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41, 827–841.

Hu, X., Hu, Y., Wu, F., Leung, R.W.T., Qin, J., 2020. Integration of single-cell multi-omics for gene regulatory network inference. Comput. Struct. Biotechnol. J. 18, 1925–1938.

Huang, D., Ovcharenko, I., 2022. Enhancer – silencer transitions in the human genome. Genome Res. 32, 437–448.

Huang, D., Petrykowska, H.M., Miller, B.F., Elnitski, L., Ovcharenko, I., 2019. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. Genome Res. 29, 657–667.

Hutter, C., Zenklusen, J.C., 2018. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell 173, 283–285.

Hyman, A.A., Weber, C.A., Jülicher, F., 2014. Liquid-liquid phase separation in biology.

Annu. Rev. Cell Dev. Biol. 30, 39–58.

Hyun, K., Jeon, J., Park, K., Kim, J., 2017. Writing, erasing and reading histone lysine methylations. Exp. Mol. Med. 49.

Inada, K., Okada, S., Phuchareon, J., Hatano, M., Sugimoto, T., Moriya, H., Tokuhisa, T., 1998. c-Fos induces apoptosis in germinal center B cells. J. Immunol. 161.

Inoue, F., Ahituv, N., 2015. Decoding enhancers using massively parallel reporter assays. Genomics 106, 159–164.

Inoue, T., Shinnakasu, R., Ise, W., Kawai, C., Egawa, T., Kurosaki, T., 2017. The transcription factor Foxo1 controls germinal center B cell proliferation in response to T cell help. J. Exp. Med. 214, 1181–1198.

Ise, W., Kohyama, M., Schraml, B.U., Zhang, T., Schwer, B., Basu, U., Alt, F.W., Tang, J., Oltz, E.M., Murphy, T.L., Murphy, K.M., 2011. The transcription factor BATF controls the global regulators of class-switch recombination in both B cells and T cells. Nat. Immunol. 12, 536–543.

Ise, W., Kurosaki, T., 2019. Plasma cell differentiation during the germinal center reaction. Immunol. Rev. 288, 64–74.

Javierre, B.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., Freire-Pritchett, P., Spivakov, M., Fraser, P., Burren, O.S., Cutler, A.J., Todd, J.A., Wallace, C., Wilder, S.P., Kreuzhuber, R., Kostadima, M., Zerbino, D.R., Stegle, O., Burden, F., Farrow, S., Rehnström, K., Downes, K., Grassi, L., Ouwehand, W.H., Frontini, M., Hill, S.M., Wang, F., Stunnenberg, H.G., Martens, J.H., Kim, B., Sharifi, N., Janssen-Megens, E.M., Yaspo, M.L., Linser, M., Kovacsovics, A., Clarke, L., Richardson, D., Datta, A., Flicek, P., 2016. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167, 1369-1384.e19.

Jenuwein, T., Allis, C.D., 2001. Translating the histone code. Science (80-.). 293, 1074–1080.

Jin, F., Selvaraj, S., Yen, C.-A., Li, Y., Schmitt, A.D., Ye, Z., Dixon, J.R., Lee, A.Y., Ren, B., Espinoza, C.A., 2013. A high-resolution map of the three-dimensional chromatin

interactome in human cells. Nature 503, 290–294.

Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B., 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science (80-.). 316, 1497–1502.

Joyner, C.J., Ley, A.M., Nguyen, D.C., Ali, M., Corrado, A., Tipton, C., Scharer, C.D., Mi, T., Woodruff, M.C., Hom, J., Boss, J.M., Duan, M., Gibson, G., Roberts, D., Andrews, J., Lonial, S., Sanz, I., Lee, F.E.H., 2022. Generation of human long-lived plasma cells by developmentally regulated epigenetic imprinting. Life Sci. Alliance 5, 1–15.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30.

Karolchik, D., 2003. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493–D496.

Keene, M.A., Corces, V., Lowenhaupt, K., Elgin, S.C., 1981. DNase I hypersensitive sites in Drosophila chromatin occur at the 5’ ends of regions of transcription. Proc. Natl. Acad. Sci. U. S. A. 78, 143–146.

Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., Snoek, J., 2018. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 28, 739–750.

Khiem, D., Cyster, J.G., Schwarz, J.J., Black, B.L., 2008. A p38 MAPK-MEF2C pathway regulates B-cell proliferation. Proc. Natl. Acad. Sci. U. S. A. 105, 17067–17072.

Kieffer-Kwon, K.R., Tang, Z., Mathe, E., Qian, J., Sung, M.H., Li, G., Resch, W., Baek, S., Pruett, N., Grøntved, L., Vian, L., Nelson, S., Zare, H., Hakim, O., Reyon, D., Yamane, A., Nakahashi, H., Kovalchuk, A.L., Zou, J., Joung, J.K., Sartorelli, V., Wei, C.L., Ruan, X., Hager, G.L., Ruan, Y., Casellas, R., 2013. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell 155, 1507–1520.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P.F., Kreiman, G., Greenberg, M.E., 2010. Widespread

transcription at neuronal activity-regulated enhancers. Nature 465, 182–187.

Kim, T.K., Shiekhattar, R., 2015. Architectural and Functional Commonalities between Enhancers and Promoters. Cell 162, 948–959.

Kleftogiannis, D., Kalnis, P., Bajic, V.B., 2015. DEEP: A general computational framework for predicting enhancers. Nucleic Acids Res. 43, 1–14.

Kleftogiannis, D., Kalnis, P., Bajic, V.B., 2016. Progress and challenges in bioinformatics approaches for enhancer identification. Brief. Bioinform. 17, 967–979.

Klein, U., Casola, S., Cattoretti, G., Shen, Q., Lia, M., Mo, T., Ludwig, T., Rajewsky, K., Dalla-Favera, R., 2006. Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination. Nat. Immunol. 7, 773–782.

Knight, P.A., Ruiz, D., 2013. A fast algorithm for matrix balancing. IMA J. Numer. Anal. 33, 1029–1047.

Koh, D.I., Yoon, J.H., Kim, M.K., An, H., Kim, M.Y., Hur, M.W., 2013. Kaiso is a key regulator of spleen germinal center formation by repressing Bcl6 expression in splenocytes. Biochem. Biophys. Res. Commun. 442, 177–182.

Koh, K.P., Rao, A., 2013. DNA methylation and methylcytosine oxidation in cell fate decisions. Curr. Opin. Cell Biol. 25, 152–161.

Kometani, K., Nakagawa, R., Shinnakasu, R., Kaji, T., Rybouchkin, A., Moriyama, S., Furukawa, K., Koseki, H., Takemori, T., Kurosaki, T., 2013. Repression of the Transcription Factor Bach2 Contributes to Predisposition of IgG1 Memory B Cells toward Plasma Cell Differentiation. Immunity 39, 136–147.

Kouno, T., Moody, J., Kwon, A.T.J., Shibayama, Y., Kato, S., Huang, Y., Böttcher, M., Motakis, E., Mendez, M., Severin, J., Luginbühl, J., Abugessaisa, I., Hasegawa, A., Takizawa, S., Arakawa, T., Furuno, M., Ramalingam, N., West, J., Suzuki, H., Kasukawa, T., Lassmann, T., Hon, C.C., Arner, E., Carninci, P., Plessy, C., Shin, J.W., 2019. C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. Nat. Commun. 10.

Kreuger, F., 2012. Trim Galore! A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra

functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries.

Kulakovskiy, I. V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., Kolpakov, F.A., Makeev, V.J., 2018. HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 46, D252–D259.

Kwon, H., Thierry-Mieg, D., Thierry-Mieg, J., Kim, H.P., Oh, J., Tunyaplin, C., Carotta, S., Donovan, C.E., Goldman, M.L., Tailor, P., Ozato, K., Levy, D.E., Nutt, S.L., Calame, K., Leonard, W.J., 2009. Analysis of Interleukin-21-Induced Prdm1 Gene Regulation Reveals Functional Cooperation of STAT3 and IRF4 Transcription Factors. Immunity 31, 941–952.

Lacaud, G., Kouskoff, V., 2017. Hemangioblast , hemogenic endothelium , and primitive versus definitive hematopoiesis. Exp. Hematol. 49, 19–24.

Laidlaw, B.J., Cyster, J.G., 2021. Transcriptional regulation of memory B cell differentiation. Nat. Rev. Immunol. 21, 209–220.

Lajoie, B.R., Dekker, J., Kaplan, N., 2015. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Methods 72, 65–75.

Lam, D.D., de Souza, F.S.J., Nasif, S., Yamashita, M., López-Leal, R., Otero-Corchon, V., Meece, K., Sampath, H., Mercer, A.J., Wardlaw, S.L., Rubinstein, M., Low, M.J., 2015. Partially Redundant Enhancers Cooperatively Maintain Mammalian Pomc Expression Above a Critical Functional Threshold. PLoS Genet. 11, 1–21.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., Weirauch, M.T., 2018. Review The Human Transcription Factors. Cell 172, 650–665.

Lancrin, C., Sroczynska, P., Stephenson, C., Allen, T., Kouskoff, V., Lacaud, G., 2009. The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. Nature 457, 892–895.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat.

Methods 9, 357–359.

Lebien, T.W., Tedder, T.F., 2008. B lymphocytes: How they develop and function. Blood 112, 1570–1580.

Lee, B., Wang, J., Cai, L., Kim, M., Namburi, S., Tjong, H., Feng, Y., Wang, P., Tang, Z., Abbas, A., Wei, C.L., Ruan, Y., Li, S., 2020. ChIA-PIPE: A fully automated pipeline for comprehensive ChIA-PET data analysis and visualization. Sci. Adv. 6.

Lee, R., Kang, M.K., Kim, Y.J., Yang, B., Shim, H., Kim, S., Kim, K., Yang, C.M., Min, B.G., Jung, W.J., Lee, E.C., Joo, J.S., Park, G., Cho, W.K., Kim, H.P., 2022. CTCF-mediated chromatin looping provides a topological framework for the formation of phase-separated transcriptional condensates. Nucleic Acids Res. 50, 207–226.

Leinonen, R., Sugawara, H., Shumway, M., 2011. The sequence read archive. Nucleic Acids Res. 39, D19–D21.

Lenz, G., Wright, G.W., Emre, N.C.T., Kohlhammer, H., Dave, S.S., Davis, R.E., Carty, S., Lam, L.T., Shaffer, A.L., Xiao, W., Powell, J., Rosenwald, A., Ott, G., Muller-Hermelink, H.K., Gascoyne, R.D., Connors, J.M., Campo, E., Jaffe, E.S., Delabie, J., Smeland, E.B., Rimsza, L.M., Fisher, R.I., Weisenburger, D.D., Chan, W.C., Staudt, L.M., 2008. Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. Proc. Natl. Acad. Sci. U. S. A. 105, 13520–13525.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., de Graaff, E., 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. 12, 1725–1735.

Li, D., Purushotham, D., Harrison, J.K., Hsu, S., Zhuo, X., Fan, C., Liu, S., Xu, V., Chen, S., Xu, J., Ouyang, S., Wu, A.S., Wang, T., 2022. WashU Epigenome Browser update 2022. Nucleic Acids Res. 50, W774–W781.

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

Li, R., Cauchy, P., Ramamoorthy, S., Boller, S., Chavez, L., Grosschedl, R., 2018. Dynamic

EBF1 occupancy directs sequential epigenetic and transcriptional events in B-cell programming. Genes Dev. 32, 96–111.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., Oh, S., Kim, H.S., Glass, C.K., Rosenfeld, M.G., 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498, 516–520.

Li, Y., Chen, C. yu, Kaye, A.M., Wasserman, W.W., 2015. The identification of cis-regulatory elements: A review from a machine learning perspective. BioSystems 138, 6–17.

Li, Y., Shi, W., Wasserman, W.W., 2018. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics 19, 1–14.

Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M., Costa, I.G., 2019. Identification of transcription factor binding sites using ATAC-seq. Genome 20, 1–21.

Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16, 321–332.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., Tamayo, P., 2015. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 1, 417–425.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-.). 326, 289–293.

Lim, L.W.K., Chung, H.H., Chong, Y.L., Lee, N.K., 2018. A survey of recently emerged genome-wide computational enhancer predictor tools. Comput. Biol. Chem. 74, 132–141.

Lio, C.W.J., Shukla, V., Samaniego-Castruita, D., González-Avalos, E., Chakraborty, A., Yue, X., Schatz, D.G., Ay, F., Rao, A., 2019. TET enzymes augment activation-induced deaminase (AID) expression via 5-hydroxymethylcytosine modifications

at the Aicda superenhancer. Sci. Immunol. 4, 1–15.

Liu, F., Li, H., Ren, C., Bo, X., Shu, W., 2016. PEDLA: Predicting enhancers with a deep learning-based algorithmic framework. Sci. Rep. 6, 1–14.

Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R., 2014. A significance test for the lasso. Ann. Stat. 42, 413–468.

Long, D.X., Joanna, D.X., 2018. RUNX1 and the endothelial origin of blood. Exp. Hematol. 68, 2–9.

Long, Z., Phillips, B., Radtke, D., Meyer-Hermann, M., Bannard, O., 2022. Competition for refueling rather than cyclic reentry initiation evident in germinal centers. Sci. Immunol. 7, eabm0775.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 1–21.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., Young, R.A., 2013. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell 153, 320–334.

Low, M.S.Y., Brodie, E.J., Fedele, P.L., Liao, Y., Grigoriadis, G., Strasser, A., Kallies, A., Willis, S.N., Tellier, J., Shi, W., Gabriel, S., O'Donnell, K., Pitt, C., Nutt, S.L., Tarlinton, D., 2019. IRF4 Activity Is Required in Established Plasma Cells to Regulate Gene Transcription and Mitochondrial Homeostasis. Cell Rep. 29, 2634-2645.e5.

Lu, Y., Qu, W., Shan, G., Zhang, C., 2015. DELTA: A distal enhancer locating tool based on adaboost algorithm and shape features of chromatin modifications. PLoS One 10, 1–20.

Luo, W., Weisel, F., Shlomchik, M.J., 2018. B Cell Receptor and CD40 Signaling Are Rewired for Synergistic Induction of the c-Myc Transcription Factor in Germinal Center B Cells. Immunity 48, 313-326.e5.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann,

B., Hecht, J., Spielmann, M., Visel, A., Mundlos, S., 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161, 1012–1025.

Machanick, P., Bailey, T.L., 2011. MEME-ChIP: Motif analysis of large DNA datasets. Bioinformatics 27, 1696–1697.

Maleki, F., Ovens, K., Hogan, D.J., Kusalik, A.J., 2020. Gene Set Analysis: Challenges, Opportunities, and Future Research. Front. Genet. 11, 1–16.

Mark, H., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., Witten, I., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explor. 11, 10–18.

Marsman, J., Thomas, A., Osato, M., O'sullivan, J.M., Horsfield, J.A., 2017. A DNA contact map for the mouse runx1 gene identifies novel haematopoietic enhancers. Sci. Rep. 7, 1–11.

Martens, J.H.A., Stunnenberg, H.G., 2013. BLUEPRINT: Mapping human blood cell epigenomes. Haematologica 98, 1487–1489.

Martinez, J., Patkaniowska, A., Urlaub, H., Lührmann, R., Tuschl, T. 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell 110, 563-574

Mayran, A., Drouin, J., 2018. Pioneer transcription factors shape the epigenetic landscape. J. Biol. Chem. 293, 13795–13804.

McCord, R.P., Kaplan, N., Giorgetti, L., 2020. Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function. Mol. Cell 77, 688–708.

McGhee, J.D., Wood, W.I., Dolan, M., Engel, J.D., Felsenfeld, G., 1981. A 200 base pair region at the 5′ end of the chicken adult β-globin gene is accessible to nuclease digestion. Cell 27, 45–55.

Mcgrath, K.E., Frame, J.M., Palis, J., 2015. Early hematopoiesis and macrophage development. Semin. Immunol. 27, 379–387.

Medvinsky, A., Rybtsov, S., Taoudi, S., 2011. Embryonic origin of the adult hematopoietic system: advances and questions. Development 138, 1017–1031.

Meng, F.L., Du, Z., Federation, A., Hu, J., Wang, Q., Kieffer-Kwon, K.R., Meyers, R.M., Amor, C., Wasserman, C.R., Neuberg, D., Casellas, R., Nussenzweig, M.C., Bradner, J.E., Liu, X.S., Alt, F.W., 2014. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. Cell 159, 1538–1548.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., Herman, B., Happe, S., Higgs, A., Leproust, E., Follows, G.A., Fraser, P., Luscombe, N.M., Osborne, C.S., 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. 47, 598–606.

Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I.E., Males, M., Viales, R.R., Furlong, E.E.M., 2018. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev. 32, 42–57.

Minnich, M., Tagoh, H., Bönelt, P., Axelsson, E., Fischer, M., Cebolla, B., Tarakhovsky, A., Nutt, S.L., Jaritz, M., Busslinger, M., 2016. Multifunctional role of the transcription factor Blimp-1 in coordinating plasma cell differentiation. Nat. Immunol. 17, 331–343.

Mir, M., Bickmore, W., Furlong, E.E.M., Narlikar, G., 2019. Chromatin topology, condensates and gene regulation: Shifting paradigms or just a phase? Dev. 146, 1–6.

Moore, J.E., Pratt, H.E., Purcaro, M.J., Weng, Z., 2020. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. Genome Biol. 21, 1–16.

Moore, L.D., Le, T., Fan, G., 2013. DNA methylation and its basic function. Neuropsychopharmacology 38, 23–38.

Mora-López, F., Pedreño-Horrillo, N., Delgado-Pérez, L., Brieva, J.A., Campos-Caro, A., 2008. Transcription of PRDM1, the master regulator for plasma cell differentiation, depends on an SP1/SP3/EGR-1 GC-box. Eur. J. Immunol. 38, 2316–2324.

Mora-López, F., Reales, E., Brieva, J.A., Campos-Caro, A., 2007. Human BSAP and BLIMP1 conform an autoregulatory feedback loop. Blood 110, 3150–3157.

Mora, A., Sandve, G.K., Gabrielsen, O.S., Eskeland, R., 2015. In the loop: promoter–enhancer interactions and bioinformatics. Brief. Bioinform. 17, 980–995.

Morman, R.E., Schweickert, P.G., Konieczny, S.F., Taparowsky, E.J., 2018. BATF regulates the expression of Nfil3, Wnt10a and miR155hg for efficient induction of antibody class switch recombination in mice. Eur. J. Immunol. 48, 1492–1505.

Moroney, J.B., Vasudev, A., Pertsemlidis, A., Zan, H., Casali, P., 2020. Integrative transcriptome and chromatin landscape analysis reveals distinct epigenetic regulations in human memory B cells. Nat. Commun. 11, 1–18.

Muramatsu, M., Kazuo, K., Sidonia, F., Shuichi, Y., Yoichi, S., Tasuku, H., 2000. Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. Cell 102, 553–563.

Nakato, R., Sakata, T., 2021. Methods for ChIP-seq analysis: A practical workflow and advanced applications. Methods 187, 44–53.

Nasser, J., Bergman, D.Y., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., Mualim, K., Natri, H.M., Weeks, E.M., Munson, G., Kane, M., Kang, H.Y., Cui, A., Ray, J.P., Eisenhaure, T.M., Collins, R.L., Dey, K., Pfister, H., 2021. Genome-wide enhancer maps link risk variants to disease genes. Nature.

Ngan, C.Y., Wong, C.H., Tjong, H., Wang, W., Goldfeder, R.L., Choi, C., He, H., Gong, L., Lin, J., Urban, B., Chow, J., Li, M., Lim, J., Philip, V., Murray, S.A., Wang, H., Wei, C.L., 2020. Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. Nat. Genet. 52, 264–272.

Nica, A.C., Dermitzakis, E.T., 2013. Expression quantitative trait loci: Present and future. Philos. Trans. R. Soc. B Biol. Sci. 368.

Nutt, S.L., Hodgkin, P.D., Tarlinton, D.M., Corcoran, L.M., 2015. The generation of antibody-secreting plasma cells. Nat. Rev. Immunol. 15, 160–171.

Nutt, S.L., Taubenheim, N., Hasbold, J., Corcoran, L.M., Hodgkin, P.D., 2011. The genetic network controlling plasma cell differentiation. Semin. Immunol. 23, 341–349.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745.

Ochiai, K., Igarashi, K., 2022. Exploring novel functions of BACH2 in the acquisition of antigen-specific antibodies. Int. Immunol. 35, 257–265.

Ochiai, K., Katoh, Y., Ikura, T., Hoshikawa, Y., Noda, T., Karasuyama, H., Tashiro, S., Muto, A., Igarashi, K., 2006. Plasmacytic transcription factor Blimp-1 is repressed by Bach2 in B cells. J. Biol. Chem. 281, 38226–38234.

Ochiai, K., Maienschein-Cline, M., Simonetti, G., Chen, J., Rosenthal, R., Brink, R., Chong, A.S., Klein, U., Dinner, A.R., Singh, H., Sciammas, R., 2013. Transcriptional Regulation of Germinal Center B and Plasma Cell Fates by Dynamical Control of IRF4. Immunity 38, 918–929.

Ohkubo, Y., Arima, M., Arguni, E., Okada, S., Yamashita, K., Asari, S., Obata, S., Sakamoto, A., Hatano, M., O-Wang, J., Ebara, M., Saisho, H., Tokuhisa, T., 2005. A Role for c- fos /Activator Protein 1 in B Lymphocyte Terminal Differentiation . J. Immunol. 174, 7703–7710.

Oudelaar, A.M., Beagrie, R.A., Gosden, M., de Ornellas, S., Georgiades, E., Kerry, J., Hidalgo, D., Carrelha, J., Shivalingam, A., El-Sagheer, A.H., Telenius, J.M., Brown, T., Buckle, V.J., Socolovsky, M., Higgs, D.R., Hughes, J.R., 2020. Dynamics of the 4D genome during in vivo lineage specification and differentiation. Nat. Commun. 11.

Pal, K., Forcato, M., Ferrari, F., 2019. Hi-C analysis: from data generation to integration. Biophys. Rev. 11, 67–78.

Palacio, M., Taatjes, D.J., 2022. Merging Established Mechanisms with New Insights: Condensates, Hubs, and the Regulation of RNA Polymerase II Transcription. J. Mol. Biol. 434, 167216.

Pang, B., Snyder, M.P., 2020. Systematic identification of silencers in human cells. Nat. Genet. 52, 254–263.

Pang, B., van Weerd, J.H., Hamoen, F.L., Snyder, M.P., 2023. Identification of non-coding silencer elements and their regulation of gene expression. Nat. Rev. Mol. Cell Biol. 24, 383–395.

Panigrahi, A.K., Foulds, C.E., Lanz, R.B., Hamilton, R.A., Yi, P., Lonard, D.M., Tsai, M.J., Tsai, S.Y., O'Malley, B.W., 2018. SRC-3 Coactivator Governs Dynamic Estrogen-Induced Chromatin Looping Interactions during Transcription. Mol. Cell 70, 679-694.e7.

Parekh, S., Polo, J.M., Shaknovich, R., Juszczynski, P., Lev, P., Ranuncolo, S.M., Yin, G., Klein, U., Cattoretti, G., Dalla Favera, R., Shipp, M.A., Melnick, A., 2007. BCL6 programs lymphoma cells for survival and differentiation through distinct biochemical mechanisms. Blood 110, 2067–2074.

Pasini, D., Malatesta, M., Jung, H.R., Walfridsson, J., Willer, A., Olsson, L., Skotte, J., Wutz, A., Porse, B., Jensen, O.N., Helin, K., 2010. Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. Nucleic Acids Res. 38, 4958–4969.

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

Peperzak, V., Vikström, I., Walker, J., Glaser, S.P., Lepage, M., Coquery, C.M., Erickson, L.D., Fairfax, K., MacKay, F., Strasser, A., Nutt, S.L., Tarlinton, D.M., 2013. Mcl-1 is essential for the survival of plasma cells. Nat. Immunol. 14, 290–297.

Pereira, R., Oliveira, J., Sousa, M., 2020. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. J. Clin. Med. 9.

Pérez-García, A., Marina-Zárate, E., Álvarez-Prado, Á.F., Ligos, J.M., Galjart, N., Ramiro, A.R., 2017. CTCF orchestrates the germinal centre transcriptional program and prevents premature plasma cell differentiation. Nat. Commun. 8.

Pervez, M.T., Hasnain, M.J.U., Abbas, S.H., Moustafa, M.F., Aslam, N., Shah, S.S.M., 2022. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. Biomed Res. Int. 2022.

Pham, D., Moseley, C.E., Gao, M., Savic, D., Winstead, C.J., Sun, M., Kee, B.L., Myers, R.M., Weaver, C.T., Hatton, R.D., 2019. Batf Pioneers the Reorganization of Chromatin in Developing Effector T Cells via Ets1-Dependent Recruitment of Ctcf. Cell Rep. 29, 1203-1220.e7.

Phelan, J.D., Young, R.M., Webster, D.E., Roulland, S., Wright, G.W., Kasbekar, M., Shaffer, A.L., Ceribelli, M., Wang, J.Q., Schmitz, R., Nakagawa, M., Bachy, E., Huang, D.W., Ji, Y., Chen, L., Yang, Y., Zhao, H., Yu, X., Xu, W., Palisoc, M.M., Valadez, R.R., Davies-Hill, T., Wilson, W.H., Chan, W.C., Jaffe, E.S., Gascoyne, R.D., Campo, E., Rosenwald, A., Ott, G., Delabie, J., Rimsza, L.M., Rodriguez, F.J., Estephan, F., Holdhoff, M., Kruhlak, M.J., Hewitt, S.M., Thomas, C.J., Pittaluga, S., Oellerich, T., Staudt, L.M., 2018. A multiprotein supercomplex controlling oncogenic signalling in lymphoma. Nature 560, 387–391.

Pilzecker, B., Jacobs, H., 2019. Mutating for good: DNA damage responses during somatic hypermutation. Front. Immunol. 10, 1–13.

Piper, J., Elze, M.C., Cauchy, P., Cockerill, P.N., Bonifer, C., Ott, S., 2013. Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res. 41.

Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A.C., Steemers, F.J., Shendure, J., Trapnell, C., 2018. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell 71, 858-871.e8.

Pone, E.J., Zhang, J., Mai, T., White, C.A., Li, G., Sakakura, J.K., Patel, P.J., Al-Qahtani, A., Zan, H., Xu, Z., Casali, P., 2012. BCR-signalling synergizes with TLR-signalling for

induction of AID and immunoglobulin class-switching through the non-canonical NF-κB pathway. Nat. Commun. 3.

Pratt, H., Weng, Z., 2018. Decoding the non-coding genome: Opportunities and challenges of genomic and epigenomic consortium data. Curr. Opin. Syst. Biol. 11, 82–90.

Price, M.J., Scharer, C.D., Kania, A.K., Randall, T.D., Boss, J.M., 2021. Conserved Epigenetic Programming and Enhanced Heme Metabolism Drive Memory B Cell Reactivation. J. Immunol. 206, 1493–1504.

Puente, X.S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J.I., Munar, M., Rubio-Pérez, C., Jares, P., Aymerich, M., Baumann, T., Beekman, R., Belver, L., Carrio, A., Castellano, G., Clot, G., Colado, E., Colomer, D., Costa, D., Delgado, J., Enjuanes, A., Estivill, X., Ferrando, A.A., Gelpí, J.L., González, B., González, S., González, M., Gut, M., Hernández-Rivas, J.M., López-Guerra, M., Martín-García, D., Navarro, A., Nicolás, P., Orozco, M., Payer, Á.R., Pinyol, M., Pisano, D.G., Puente, D.A., Queirós, A.C., Quesada, V., Romeo-Casabona, C.M., Royo, C., Royo, R., Rozman, M., Russiñol, N., Salaverría, I., Stamatopoulos, K., Stunnenberg, H.G., Tamborero, D., Terol, M.J., Valencia, A., López-Bigas, N., Torrents, D., Gut, I., López-Guillermo, A., López-Otín, C., Campo, E., 2015. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 526, 519–524.

Qi, L.S., Larson, M.H, Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., Lim, W.A, 2021. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152, 1173-1183

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

Rachakonda, S., Hoheisel, J.D., Kumar, R., 2021. Occurrence, functionality and abundance of the TERT promoter mutations. Int. J. Cancer 149, 1852–1862.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., Wysocka, J., 2011. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279–285.

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., Ren, B., 2013. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. PLoS Comput. Biol. 9, 1–14.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., Manke, T., 2016. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44, W160–W165.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., Aiden, E.L., 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–80.

Rauluseviciute, I., Drabløs, F., Rye, M.B., 2019. DNA methylation data by sequencing: Experimental approaches and recommendations for tools and pipelines for data analysis. Clin. Epigenetics 11, 1–13.

Ray-Jones, H., Spivakov, M., 2021. Transcriptional enhancers and their communication with gene promoters, Cellular and Molecular Life Sciences. Springer International Publishing.

Reilly, S.K., Gosai, S.J., Gutierrez, A., Mackay-Smith, A., Ulirsch, J.C., Kanai, M., Mouri, K., Berenzy, D., Kales, S., Butler, G.M., Gladden-Young, A., Bhuiyan, R.M., Stitzel, M.L., Finucane, H.K., Sabeti, P.C., Tewhey, R., 2021. Direct characterization of cis-regulatory elements and functional dissection of complex genetic associations using HCR–FlowFISH. Nat. Genet. 53, 1166–1176.

Rimel, J.K., Taatjes, D.J., 2018. The essential and multifunctional TFIIH complex. Protein Sci. 27, 1018–1037.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Quon, G., Sandstrom, R.S., Eaton, M.L., Wu, Y.C., Pfenning, A.R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R.A., Shoresh, N., Epstein, C.B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R.D., Lister, R., Hong, C., Gascard, P., Mungall, A.J., Moore, R., Chuah, E., Tam, A., Canfield, T.K., Hansen, R.S., Kaul, R., Sabo, P.J., Bansal, M.S.,

Carles, A., Dixon, J.R., Farh, K.H., Feizi, S., Karlic, R., Kim, A.R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R.C., Siebenthall, K.T., Sinnott-Armstrong, N.A., Stevens, M., Thurman, R.E., Wu, J., Zhang, B., Zhou, X., Beaudet, A.E., Boyer, L.A., De Jager, P.L., Farnham, P.J., Fisher, S.J., Haussler, D., Jones, S.J.M., Li, W., Marra, M.A., McManus, M.T., Sunyaev, S., Thomson, J.A., Tlsty, T.D., Tsai, L.H., Wang, W., Waterland, R.A., Zhang, M.Q., Chadwick, L.H., Bernstein, B.E., Costello, J.F., Ecker, J.R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J.A., Wang, T., Kellis, M., 2015. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–329.

Roayaei Ardakany, A., Gezer, H.T., Lonardi, S., Ay, F., 2020. Mustache: Multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. Genome Biol. 21, 1–17.

Robinson, J., 2012. Integrated Genomics Viewer. Nat. Biotechnol. 29, 24–26.

Robinson, M.J., Ding, Z., Pitt, C., Brodie, E.J., Quast, I., Tarlinton, D.M., Zotos, D., 2020. The Amount of BCL6 in B Cells Shortly after Antigen Engagement Determines Their Representation in Subsequent Germinal Centers. Cell Rep. 30, 1530-1541.e4.

Roco, J.A., Mesin, L., Binder, S.C., Nefzger, C., Gonzalez-Figueroa, P., Canete, P.F., Ellyard, J., Shen, Q., Robert, P.A., Cappello, J., Vohra, H., Zhang, Y., Nowosad, C.R., Schiepers, A., Corcoran, L.M., Toellner, K.M., Polo, J.M., Meyer-Hermann, M., Victora, G.D., Vinuesa, C.G., 2019. Class-Switch Recombination Occurs Infrequently in Germinal Centers. Immunity 51, 337-350.e7.

Rojano, E., Seoane, P., Ranea, J.A.G., Perkins, J.R., 2019. Regulatory variants: From detection to predicting impact. Brief. Bioinform. 20, 1639–1654.

Rosvall, M., Bergstrom, C.T., 2008. Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. U. S. A. 105, 1118–1123.

Roy, S., Siahpirani, A.F., Chasman, D., Knaack, S., Ay, F., Stewart, R., Wilson, M., Sridharan, R., 2015. A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Res. 43, 8694–8712.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A. V., Manteiga, J.C., Li, C.H., Guo, Y.E., Day, D.S., Schuijers, J., Vasile, E., Malik, S., Hnisz, D., Lee, T.I., Cisse, I.I., Roeder, R.G., Sharp, P.A., Chakraborty, A.K., Young, R.A., 2018. Coactivator condensation at super-enhancers links phase separation and gene control. Science (80-.). 361, eaar3958.

Saito, M., Gao, J., Basso, K., Kitagawa, Y., Smith, P.M., Bhagat, G., Pernis, A., Pasqualucci, L., Dalla-Favera, R., 2007. A Signaling Pathway Mediating Downregulation of BCL6 in Germinal Center B Cells Is Blocked by BCL6 Gene Alterations in B Cell Lymphoma. Cancer Cell 12, 280–292.

Salviato, E., Djordjilović, V., Hariprakash, J.M., Tagliaferri, I., Pal, K., Ferrari, F., 2021. Leveraging three-dimensional chromatin architecture for effective reconstruction of enhancer-target gene regulatory interactions. Nucleic Acids Res. 49, 1–22.

Sanborn, A.L., Rao, S.S.P., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K.P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E.K., Lander, E.S., Aiden, E.L., 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U. S. A. 112, E6456–E6465.

Sayegh, C.E., Quong, M.W., Agata, Y., Murre, C., 2003. E-proteins directly regulate expression of activation-induced deaminase in mature B cells. Nat. Immunol. 4, 586–593.

Scharer, C.D., Barwick, B.G., Guo, M., Bally, A.P.R., Boss, J.M., 2018. Plasma cell differentiation is controlled by multiple cell division-coupled epigenetic programs. Nat. Commun. 9.

Schier, A.C., Taatjes, D.J., 2020. Structure and mechanism of the RNA polymerase II transcription machinery. Genes Dev. 34, 465–488.

Schmidlin, H., Diehl, S.A., Nagasawa, M., Scheeren, F.A., Schotte, R., Uittenbogaart, C.H., Spits, H., Blom, B., 2008. Spi-B inhibits human plasma cell differentiation by repressing BLIMP1 and XBP-1 expression. Blood 112, 1804–1812.

Schmidt, F., Marx, A., Baumgarten, N., Hebel, M., Wegner, M., Kaulich, M., Leisegang, M.S., Brandes, R.P., Göke, J., Vreeken, J., Schulz, M.H., 2021. Integrative analysis

of epigenetics data identifies gene-specific regulatory elements. Nucleic Acids Res. 49, 10397–10418.

Schneider, T.D., Stephens, R.M., 1990. Sequence logos: Nucleic Acids Res. 18, 6097–6100.

Schoenfelder, S., Fraser, P., 2019. Long-range enhancer–promoter contacts in gene expression control. Nat. Rev. Genet.

Schrider, D.R., Kern, A.D., 2018. Supervised Machine Learning for Population Genetics : A New Paradigm. Trends Genet. 34, 301–312.

Schütte, J., Wang, H., Antoniou, S., Jarratt, A., Wilson, N.K., Riepsaame, J., Calero-Nieto, F.J., Moignard, V., Basilico, S., Kinston, S.J., Hannah, R.L., Chan, M.C., Nürnberg, S.T., Ouwehand, W.H., Bonzanni, N., de Bruijn, M.F., Göttgens, B., 2016. An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. Elife 5, 1–27.

Sciammas, R., Shaffer, A.L., Schatz, J.H., Zhao, H., Staudt, L.M., Singh, H., 2006. Graded Expression of Interferon Regulatory Factor-4 Coordinates Isotype Switching with Plasma Cell Differentiation. Immunity 25, 225–236.

Segert, J.A., Gisselbrecht, S.S., Bulyk, M.L., 2021. Transcriptional Silencers: Driving Gene Expression with the Brakes On. Trends Genet. 37, 514–527.

Sellars, M., Kastner, P., Chan, S., 2011. Ikaros in B cell development and function. World J. Biol. Chem. 2, 132.

Shaffer, A.L., Lin, K.-I., Kuo, T.C., Hurt, E.M., Rosenwald, A., Giltnane, J.M., Yang, L., Zhao, H., Calame, K., Staudt, L.M., 2002. Blimp-1 Orchestrates Plasma Cell Differentiation by Extinguishing the Mature B Cell Gene Expression Program. Immunity 17, 51–62.

Shaffer, A.L., Shapiro-Shelef, M., Iwakoshi, N.N., Lee, A.H., Qian, S.B., Zhao, H., Yu, X., Yang, L., Tan, B.K., Rosenwald, A., Hurt, E.M., Petroulakis, E., Sonenberg, N., Yewdell, J.W., Calame, K., Glimcher, L.H., Staudt, L.M., 2004. XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation. Immunity 21, 81–93.

Shaffer, A.L., Wright, G., Yang, L., Powell, J., Ngo, V., Lamy, L., Lam, L.T., Davis, R.E., Staudt, L.M., 2006. A library of gene expression signatures to illuminate normal and pathological lymphoid biology. Immunol. Rev. 210, 67–85.

Shaffer, A.L., Yu, X., He, Y., Boldrick, J., Chan, E.P., Staudt, L.M., 2000. BCL-6 represses genes that function in lymphocyte differentiation, inflammation, and cell cycle control. Immunity 13, 199–212.

Shar, N.A., Vijayabaskar, M.S., Westhead, D.R., 2016. Cancer somatic mutations cluster in a subset of regulatory sites predicted from the ENCODE data. Mol. Cancer 15, 1–9.

Sheffield, N.C., Thurman, R.E., Song, L., Safi, A., Stamatoyannopoulos, J.A., Lenhard, B., Crawford, G.E., Furey, T.S., 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. 23, 777–788.

Shen, Y., Yue, F., Mc Cleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Ren, B., Lobanenkov, V. V., 2012. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–120.

Shinnakasu, R., Inoue, T., Kometani, K., Moriyama, S., Adachi, Y., Nakayama, M., Takahashi, Y., Fukuyama, H., Okada, T., Kurosaki, T., 2016. Regulated selection of germinal-center cells into the memory B cell compartment. Nat. Immunol. 17, 861–869.

Shlyueva, D., Stampfel, G., Stark, A., 2014. Transcriptional enhancers: From properties to genome-wide predictions. Nat. Rev. Genet. 15, 272–286.

Sikorski, T.W., Buratowski, S., 2009. The basal initiation machinery: beyond the general transcription factors. Curr. Opin. Cell Biol. 21, 344–351.

Simcha, D., Price, N.D., Geman, D., 2012. The Limits of De Novo DNA Motif Discovery. PLoS One 7.

Slatko, B.E., Gardner, A.F., Ausubel, F.M., 2018. Overview of Next-Generation Sequencing Technologies. Curr. Protoc. Mol. Biol. 122.

Soutourina, J., 2018. Transcription regulation by the Mediator complex. Nat. Rev. Mol.

Cell Biol. 19, 262–274.

Spitz, F., Furlong, E.E.M., 2012. Transcription factors: From enhancer binding to developmental control. Nat. Rev. Genet. 13, 613–626.

Srivastava, A., Malik, L., Sarkar, H., Zakeri, M., Almodaresi, F., Soneson, C., Love, M.I., Kingsford, C., Patro, R., 2020. Alignment and mapping methodology influence transcript abundance estimation. Genome Biol. 21, 1–29.

Stark, R., Brown, G., 2011. DiffBind: differential binding analysis of ChIP-Seq peak data.

Stark, R., Grzelak, M., Hadfield, J., 2019. RNA sequencing: the teenage years. Nat. Rev. Genet. 20, 631–656.

Stephenson, S., Care, M.A., Doody, G.M., Tooze, R.M., 2022. APRIL Drives a Coordinated but Diverse Response as a Foundation for Plasma Cell Longevity. J. Immunol. 209, 926–937.

Stoltzfus, J.C., 2011. Logistic regression: A brief primer. Acad. Emerg. Med. 18, 1099–1104.

Stormo, G.D., 2000. DNA binding sites: Representation and discovery. Bioinformatics 16, 16–23.

Stunnenberg, H.G., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., Amin, V., Amit, I., Antonarakis, S.E., Aparicio, S., Arima, T., Arrigoni, L., Arts, R., Asnafi, V., Badosa, M.E., Bae, J.B., Bassler, K., Beck, S., Berkman, B., Bernstein, B.E., Bilenky, M., Bird, A., Bock, C., Boehm, B., Bourque, G., Breeze, C.E., Brors, B., Bujold, D., Burren, O., Bussemakers, M.J., Butterworth, A., Campo, E., Carrillo-de-Santa-Pau, E., Chadwick, L., Chan, K.M., Chen, W., Cheung, T.H., Chiapperino, L., Choi, N.H., Chung, H.R., Clarke, L., Connors, J.M., Cronet, P., Danesh, J., Dermitzakis, M., Drewes, G., Durek, P., Dyke, S., Dylag, T., Eaves, C.J., Ebert, P., Eils, R., Eils, J., Ennis, C.A., Enver, T., Feingold, E.A., Felder, B., Ferguson-Smith, A., Fitzgibbon, J., Flicek, P., Foo, R.S.Y., Fraser, P., Frontini, M., Furlong, E., Gakkhar, S., Gasparoni, N., Gasparoni, G., Geschwind, D.H., Glažar, P., Graf, T., Grosveld, F., Guan, X.Y., Guigo, R., Gut, I.G., Hamann, A., Han, B.G., Harris, R.A., Heath, S., Helin, K., Hengstler, J.G., Heravi-Moussavi, A., Herrup, K., Hill, S., Hilton, J.A., Hitz, B.C., Horsthemke, B., Hu, M., Hwang, J.Y., Ip, N.Y., Ito, T., Javierre, B.M., Jenko, S.,

Jenuwein, T., Joly, Y., Jones, S.J.M., Kanai, Y., Kang, H.G., Karsan, A., Kiemer, A.K., Kim, S.C., Kim, B.J., Kim, H.H., Kimura, H., Kinkley, S., Klironomos, F., Koh, I.U., Kostadima, M., Kressler, C., Kreuzhuber, R., Kundaje, A., Küppers, R., Larabell, C., Lasko, P., Lathrop, M., Lee, D.H.S., Lee, S., Lehrach, H., Leitão, E., Lengauer, T., Lernmark, Å., Leslie, R.D., Leung, G.K.K., Leung, D., Loeffler, M., Ma, Y., Mai, A., Manke, T., Marcotte, E.R., Marra, M.A., Martens, J.H.A., Martin-Subero, J.I., Maschke, K., Merten, C., Milosavljevic, A., Minucci, S., Mitsuyama, T., Moore, R.A., Müller, F., Mungall, A.J., Netea, M.G., Nordström, K., Norstedt, I., Okae, H., Onuchic, V., Ouellette, F., Ouwehand, W., Pagani, M., Pancaldi, V., Pap, T., Pastinen, T., Patel, R., Paul, D.S., Pazin, M.J., Pelicci, P.G., Phillips, A.G., Polansky, J., Porse, B., Pospisilik, J.A., Prabhakar, S., Procaccini, D.C., Radbruch, A., Rajewsky, N., Rakyan, V., Reik, W., Ren, B., Richardson, D., Richter, A., Rico, D., Roberts, D.J., Rosenstiel, P., Rothstein, M., Salhab, A., Sasaki, H., Satterlee, J.S., Sauer, S., Schacht, C., Schmidt, F., Schmitz, G., Schreiber, S., Schröder, C., Schübeler, D., Schultze, J.L., Schulyer, R.P., Schulz, M., Seifert, M., Shirahige, K., Siebert, R., Sierocinski, T., Siminoff, L., Sinha, A., Soranzo, N., Spicuglia, S., Spivakov, M., Steidl, C., Strattan, J.S., Stratton, M., Südbeck, P., Sun, H., Suzuki, N., Suzuki, Y., Tanay, A., Torrents, D., Tyson, F.L., Ulas, T., Ullrich, S., Ushijima, T., Valencia, A., Vellenga, E., Vingron, M., Wallace, C., Wallner, S., Walter, J., Wang, H., Weber, S., Weiler, N., Weller, A., Weng, A., Wilder, S., Wiseman, S.M., Wu, A.R., Wu, Z., Xiong, J., Yamashita, Y., Yang, X., Yap, D.Y., Yip, K.Y., Yip, S., Yoo, J. Il, Zerbino, D., Zipprich, G., Hirst, M., 2016. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell 167, 1145–1149.

Sung, M.H., Baek, S., Hager, G.L., 2016. Genome-wide footprinting: Ready for prime time? Nat. Methods 13, 222–228.

Sung, M.H., Guertin, M.J., Baek, S., Hager, G.L., 2014. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol. Cell 56, 275–285.

Tang, L., Hill, M.C., Ellinor, P.T., Li, M., 2022. Bacon: a comprehensive computational benchmarking framework for evaluating targeted chromatin conformation capture-specific methodologies. Genome Biol. 23, 1–21.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A.,

Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S.Z., Penrad-Mobayed, M., Sachs, L.M., Ruan, X., Wei, C.L., Liu, E.T., Wilczynski, G.M., Plewczynski, D., Li, G., Ruan, Y., 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell 163, 1611–1627.

Taylor, J., Tibshirani, R., 2018. Post-selection inference for l1-penalized likelihood models. Can. J. Stat. 46, 41–61.

Tellier, J., Nutt, S.L., 2019. Plasma cells: The programming of an antibody-secreting machine. Eur. J. Immunol. 49, 30–37.

Tellier, J., Shi, W., Minnich, M., Liao, Y., Crawford, S., Smyth, G.K., Kallies, A., Busslinger, M., Nutt, S.L., 2016. Blimp-1 controls plasma cell function through the regulation of immunoglobulin secretion and the unfolded protein response. Nat. Immunol. 17, 323–330.

Thomsen, I., Kunowska, N., de Souza, R., Moody, A.-M., Crawford, G., Wang, Y.-F., Khadayate, S., Whilding, C., Strid, J., Karimi, M.M., Barr, A.R., Dillon, N., Sabbattini, P., 2021. RUNX1 Regulates a Transcription Program That Affects the Dynamics of Cell Cycle Entry of Naive Resting B Cells. J. Immunol. 207, 2976–2991.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Zhancheng, Zhang, Zhuzhu, Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E., Stamatoyannopoulos, J.A., 2012. The accessible chromatin landscape of the human genome. Nature 489, 75–82.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. 58, 267–288.

Tran, T.H., Nakata, M., Suzuki, K., Begum, N.A., Shinkura, R., Fagarasan, S., Honjo, T.,

Nagaoka, H., 2010. B cell-specific and stimulation-responsive enhancers derepress Aicda by overcoming the effects of silencers. Nat. Immunol. 11, 148–154.

Trezise, S., Nutt, S.L., 2021. The gene regulatory network controlling plasma cell function. Immunol. Rev. 303, 23–34.

Tsai, A., Alves, M.R.P., Crocker, J., 2019. Multi-enhancer transcriptional hubs confer phenotypic robustness. Elife 8, 1–17.

Tsai, P.F., Dell'Orso, S., Rodriguez, J., Vivanco, K.O., Ko, K.D., Jiang, K., Juan, A.H., Sarshad, A.A., Vian, L., Tran, M., Wangsa, D., Wang, A.H., Perovanovic, J., Anastasakis, D., Ralston, E., Ried, T., Sun, H.W., Hafner, M., Larson, D.R., Sartorelli, V., 2018. A Muscle-Specific Enhancer RNA Mediates Cohesin Recruitment and Regulates Transcription In trans. Mol. Cell 71, 129-141.e8.

Tu, Y.K., Gunnell, D., Gilthorpe, M.S., 2008. Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon - The reversal paradox. Emerg. Themes Epidemiol. 5, 1–9.

Tunyaplin, C., Shaffer, A.L., Angelin-Duclos, C.D., Yu, X., Staudt, L.M., Calame, K.L., 2004. Direct Repression of prdm1 by Bcl-6 Inhibits Plasmacytic Differentiation . J. Immunol. 173, 1158–1165.

Turner, C.A., Mack, D.H., Davis, M.M., 1994. Blimp-1, a novel zinc finger-containing protein that can drive the maturation of B lymphocytes into immunoglobulin-secreting cells. Cell 77, 297–306.

Umans, B.D., Battle, A., Gilad, Y., 2021. Where Are the Disease-Associated eQTLs? Trends Genet. 37, 109–124.

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., de Magalhães, J.P., 2018. Gene co-expression analysis for functional classification and gene-disease predictions. Brief. Bioinform. 19, 575–592.

Van Den Berge, K., Hembach, K.M., Soneson, C., Tiberi, S., Clement, L., Love, M.I., Patro, R., Robinson, M.D., 2019. RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. Annu. Rev. Biomed. Data Sci. 2, 139–173.

Vandereyken, K., Sifrim, A., Thienpont, B., Voet, T., 2023. Methods and applications for

single-cell and spatial multi-omics. Nat. Rev. Genet. 24.

Vasanwala, F.H., Kusam, S., Toney, L.M., Dent, A.L., 2002. Repression of AP-1 Function: A Mechanism for the Regulation of Blimp-1 Expression and B Lymphocyte Differentiation by the B Cell Lymphoma-6 Protooncogene. J. Immunol. 169, 1922–1929.

Vastenhouw, N.L., Schier, A.F., 2012. Bivalent histone modifications in early embryogenesis. Curr. Opin. Cell Biol. 24, 374–386.

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., Stamatoyannopoulos, J.A., 2020. Global reference mapping of human transcription factor footprints. Nature 583, 729–736.

Vijayabaskar, M.S., Goode, D.K., Obier, N., Lichtinger, M., Emmett, A.M.L., Abidin, F.N.Z., Shar, N., Hannah, R., Assi, S.A., Lie-A-Ling, M., Gottgens, B., Lacaud, G., Kouskoff, V., Bonifer, C., Westhead, D.R., 2019. Identification of gene specific cis-regulatory elements during differentiation of mouse embryonic stem cells: An integrative approach using high-throughput datasets. PLoS Comput. Biol. 15, e1007337.

Vilarrasa-Blasi, R., Soler-Vila, P., Verdaguer-Dot, N., Russiñol, N., Di Stefano, M., Chapaprieta, V., Clot, G., Farabella, I., Cuscó, P., Kulis, M., Agirre, X., Prosper, F., Beekman, R., Beà, S., Colomer, D., Stunnenberg, H.G., Gut, I., Campo, E., Marti-Renom, M.A., Martin-Subero, J.I., 2021. Dynamics of genome architecture and chromatin function during human B cell differentiation and neoplastic transformation. Nat. Commun. 12, 1–18.

Visel, A., Minovitsky, S., Dubchak, I., Pennacchio, L.A., 2007. VISTA Enhancer Browser - A database of tissue-specific human enhancers. Nucleic Acids Res. 35, D88–D92.

Voelkerding, K. V., Dames, S.A., Durtschi, J.D., 2009. Next-generation sequencing:from basic research to diagnostics. Clin. Chem. 55, 641–658.

Wang, D., Rendon, A., Wernisch, L., 2013. Transcription factor and chromatin features predict genes associated with eQTLs. Nucleic Acids Res. 41, 1450–1463.

Wang, H., Huang, B., Wang, J., 2021. Predict long-range enhancer regulation based on protein-protein interactions between transcription factors. Nucleic Acids Res. 49, 10347–10368.

Wasserman, W.W., Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. 5, 276–287.

Watanabe, K., Sugai, M., Nambu, Y., Osato, M., Hayashi, T., Kawaguchi, M., Komori, T., Ito, Y., Shimizu, A., 2010. Requirement for Runx Proteins in IgA Class Switching Acting Downstream of TGF-β1 and Retinoic Acid Signaling. J. Immunol. 184, 2785–2792.

Wei, Y., Zhang, S., Shang, S., Zhang, B., Li, S., Wang, X., Wang, F., Su, J., Wu, Q., Liu, H., Zhang, Y., 2016. SEA: A Super-enhancer Archive. Nucleic Acids Res. 44, D172–D179.

Whalen, S., Truty, R.M., Pollard, K.S., 2016. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet. 48, 488–496.

Whitaker, J.W., Nguyen, T.T., Zhu, Y., Wildberg, A., Wang, W., 2015. Computational schemes for the prediction and annotation of enhancers from epigenomic assays. Methods 72, 86–94.

Whyte, W.A., Bilodeau, S., Orlando, D.A., Hoke, H.A., Frampton, G.M., Foster, C.T., Cowley, S.M., Young, R.A., 2012. Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. Nature 482, 221–225.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., Young, R.A., 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319.

Wilker, P.R., Kohyama, M., Sandau, M.M., Albring, J.C., Nakagawa, O., Schwarz, J.J., Murphy, K.M., 2008. Transcription factor Mef2c is required for B cell proliferation and survival after antigen receptor stimulation. Nat. Immunol. 9, 603–612.

Willis, S.N., Tellier, J., Liao, Y., Trezise, S., Light, A., O'Donnell, K., Garrett-Sinha, L.A., Shi, W., Tarlinton, D.M., Nutt, S.L., 2017. Environmental sensing by mature B cells

is controlled by the transcription factors PU.1 and SpiB. Nat. Commun. 8.

Wilson, N.K., Foster, S.D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P.M., Kinston, S., Ouwehand, W.H., Dzierzak, E., Pimanda, J.E., De Bruijn, M.F.T.R., Göttgens, B., 2010. Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell 7, 532–544.

Wingender, E., 2008. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief. Bioinform. 9, 326–332.

Winkelmann, R., Sandrock, L., Porstner, M., Roth, E., Mathews, M., Hobeika, E., Reth, M., Kahn, M.L., Schuh, W., Jäck, H.M., 2011. B cell homeostasis and plasma cell homing controlled by Krüppel-like factor 2. Proc. Natl. Acad. Sci. U. S. A. 108, 710–715.

Wöhner, M., Tagoh, H., Bilic, I., Jaritz, M., Poliakova, D.K., Fischer, M., Busslinger, M., 2016. Molecular functions of the transcription factors E2A and E2-2 in controlling germinal center B cell and plasma cell development. J. Exp. Med. 213, 1201–1221.

Wu, C., M. Bingham, P., Livak, K.J., Holmgren, R., Elgin, S.C.R., 1979. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell 16, 797–806.

Xi, W., Beer, M.A., 2018. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. PLoS Comput. Biol. 14, 1–12.

Xie, S., Duan, J., Li, B., Zhou, P., Hon, G.C., 2017. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Mol. Cell 66, 285-299.e5.

Yan, F., Powell, D.R., Curtis, D.J., Wong, N.C., 2020. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol. 21, 1–16.

Yang, Z., Algesheimer, R., Tessone, C.J., 2016. A comparative analysis of community detection algorithms on artificial networks. Sci. Rep. 6.

Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M.G., Andersen, S., Lu, Q., Rowson, A., Taylor, T.R.P., Clarke, L., Maccora, K., Chen, C., Cook, A.L., Ye,

C.J., Fairfax, K.A., Hewitt, A.W., Powell, J.E., 2022. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. Science (80-.). 376.

Young, C., Brink, R., 2021. The unique biology of germinal center B cells. Immunity 54, 1652–1664.

Yu, J., Angelin-Duclos, C., Greenwood, J., Liao, J., Calame, K., 2000. Transcriptional Repression by Blimp-1 (PRDI-BF1) Involves Recruitment of Histone Deacetylase. Mol. Cell. Biol. 20, 2592–2603.

Zan, H., Casali, P., 2013. Regulation of Aicda expression and AID activity. Autoimmunity 46, 83–101.

Zaret, K.S., 2020. Pioneer Transcription Factors Initiating Gene Network Changes.

Zeng, W., Chen, S., Cui, X., Chen, X., Gao, Z., Jiang, R., 2021. SilencerDB: A comprehensive database of silencers. Nucleic Acids Res. 49, D221–D228.

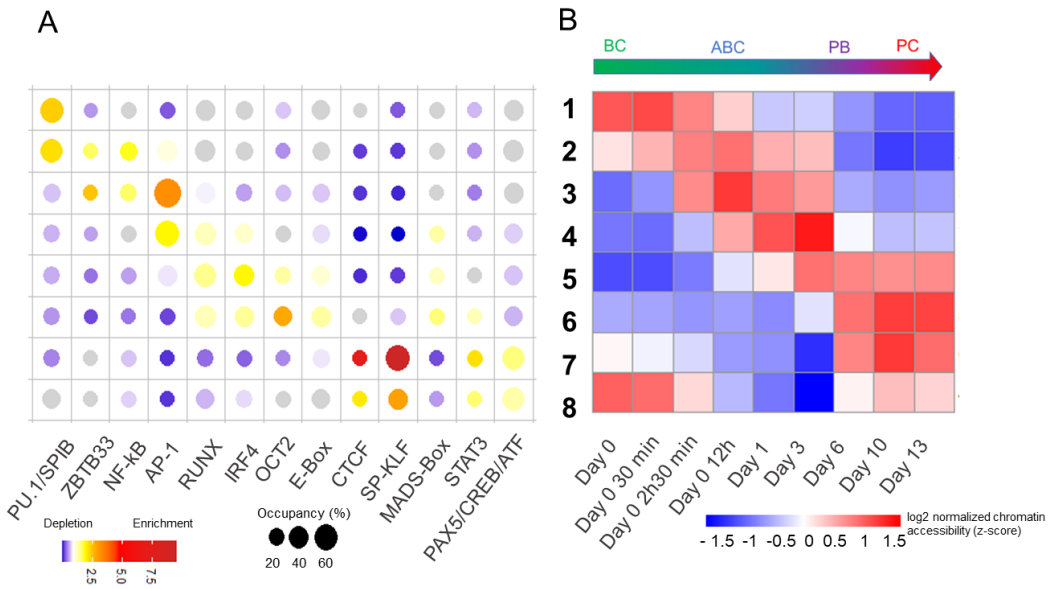
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., Shirley, X.S., 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9.

Zhu, I., Song, W., Ovcharenko, I., Landsman, D., 2021. A model of active transcription hubs that unifies the roles of active promoters and enhancers. Nucleic Acids Res. 49, 4493–4505.

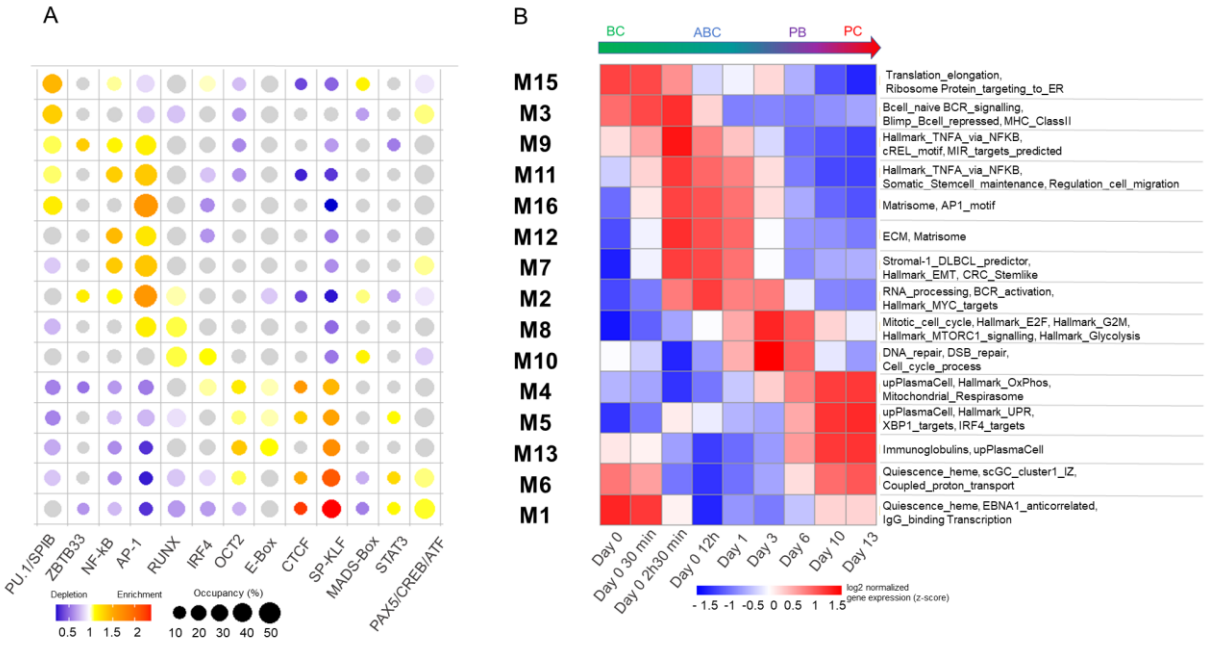
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 768.

Zuin, J., Roth, G., Zhan, Y., Cramard, J., Redolfi, J., Piskadlo, E., Mach, P., Kryzhanovska, M., Tihanyi, G., Kohler, H., Eder, M., Leemans, C., van Steensel, B., Meister, P., Smallwood, S., Giorgetti, L., 2022. Nonlinear control of transcription through enhancer–promoter interactions. Nature 604, 571–577.

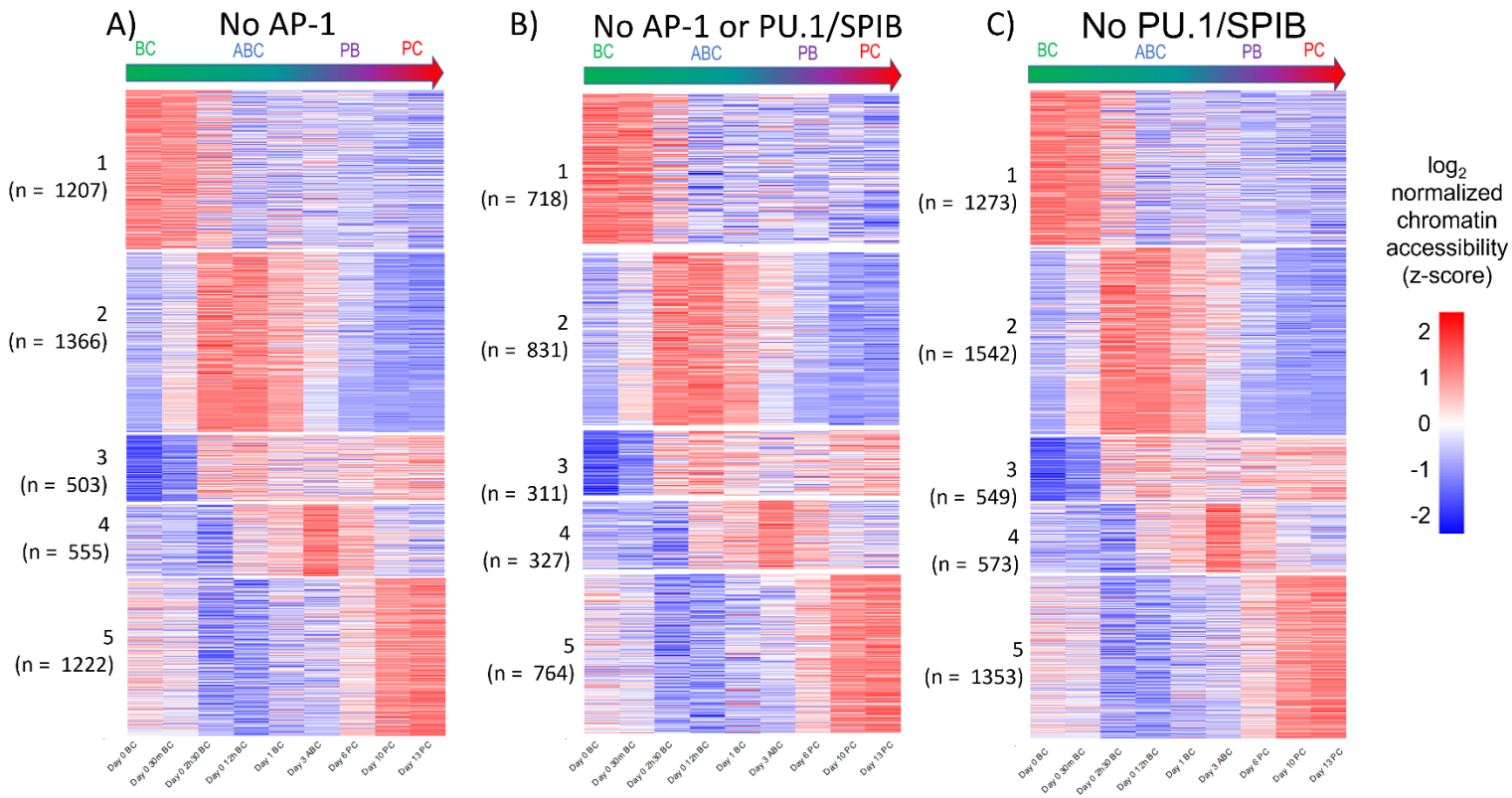
Appendices



Appendix 4.1 Figure showing replication of TF enrichment in cis-regulatory clusters, using predictions from cisREAD + BMO. A) Bubbleplot showing enrichment of TF occupancy in each cluster. Size of bubbles gives the proportion of each cluster harbouring a TF footprint, colour shows significant ($p < 0.05$, two-sided Fisher test) enrichment (fold-change between cluster and other clusters > 1, red) or depletion (fold-change between cluster and other clusters < 1, blue), grey represents no significant enrichment. B) Heatmap showing mean log2 normalised chromatin accessibility (z-score) of cis-regulatory elements significantly linked to gene expression, k-means clustered (k = 8).



Appendix 4.2 A) Bubbleplot showing enrichment of TF occupancy in each cluster. Size of bubbles gives the proportion of each cluster harbouring a TF footprint, colour shows significant ($p < 0.05$, two-sided Fisher test) enrichment (fold-change between genes in module and genes not in module > 1, red) or depletion (fold-change < 1, blue), grey represents no significant enrichment.B) Heatmap showing mean $\log_2$ normalised gene expression (z-score) of genes with significantly linked CREs, module names reflect enriched gene sets in each module.



Appendix 4.3 Figure showing background genes for gene set over-representation analysis of PU.1/SPIB and/or AP-1 target gene clusters, against similarly expressed genes not linked to the factor(s). For each analysis, similarly, expressed genes were obtained by training a machine learning classifier on five expression clusters, and predicting the cluster label for all other differentially expressed genes not linked to CREs with given footprint(s). N gives number of genes, not linked to the factor(s) predicted to belong to each cluster. Heatmaps show z-score log2 normalized gene expression for: A) Similar expressed genes not linked to AP-1, B) similarly expressed genes not linked AP-1 or PU.1/SPIB, C) similarly expressed genes not linked to PU.1/SPIB