An exploration of non-coding sequence variation across human erythroid differentiation

---

A thesis submitted for the degree of Doctor of Philosophy of the University of Oxford





Matthew E. Gosden
Balliol College

Doctor of Philosophy in Medical Sciences
WIMM Prize Studentship
MRC Weatherall Institute of Molecular Medicine
University of Oxford

Hilary Term 2024

Abstract

The ability to precisely regulate genes has been pivotal in the evolution of complex life and humanity. A cell's ability to dynamically control the 3D structure of the genome to physically bring together regulatory elements allows for the precise spatiotemporal expression of genes. Variation within the human genome is widespread and well documented, and accounts for the vast differences in phenotypes seen within a population. However, this variation can also lead to disease as the disruption of coding regions or a regulatory network can alter protein levels beyond a tolerable range leading to detrimental downstream effects.

In this thesis I have explored the genome wide variation present in a local population of 50 individuals by using a well-established erythroid differentiation system. By assessing open chromatin data and histone modifications, regions that are active in some individuals but not others have been found and defined. Many of these differential regions can be defined as a 'loss of function' variant, whereby a previously identified active region is shown to have its activity reduced. Additionally, with the use of allelic skew methods and haplotype analysis, we also demonstrate examples of the more elusive 'gain of function' variant, whereby a new element unique to a single haplotype is created.

To fully decipher the effects imparted by these variants, chromatin interaction data can be generated to uncover the regulatory networks of a specific cell type and stage. I have used MCC to generate high resolution interaction profiles that can unambiguously link regulatory elements whilst simultaneously generating transcription factor binding footprints. Moreover, technical advances of this assay have allowed the generation of allele specific interaction profiles from heterozygous variants, which has enabled the deciphering of variants and GWAS SNP mechanisms.

Due to the nature of cell differentiation, the precise genome regulation at each cell stage affects subsequent downstream cell stages. It is therefore important to consider the effects of these variants throughout a cell's lineage, as early disruption to processes in differentiation can impact terminally differentiated cells. By using the scMultiome system, I have been able to assess the changing genomic landscape across erythropoiesis to uncover variants that affect gene regulation early in differentiation. From these data it is clear that there is high enrichment for variant effects across all stages of erythropoiesis, but particularly at the later stages.

Acknowledgements

First and foremost I would like to express my tremendous gratitude to Professor Jim Hughes for almost nine years of support and supervision. Thank you for being so encouraging and understanding during what has been a pretty tumultuous four years, and for always helping to lead this project in a way that matched my interests and suited my skills. Thank you also for curating a welcoming lab environment that has always championed collaboration and has forged many cherished friendships for me. To my second supervisor Professor James Davies, thank you for breathing life into my project at a time when it needed it the most and for always being at hand when I was in need of technical support.

To all the members of the Hughes lab, both past and present, the lab has always been such a friendly and vibrant team, and it has been a pleasure to be a part of such an incredibly talented and lovely bunch of people. To Dr Damien Downes, despite only being physically present for the first six months of my DPhil, thank you for teaching me pretty much all the molecular biology and technical skills I needed to complete this project, and for putting up with years of my stupid questions; a true colleague. Thank you to Dr Caroline Scott for your invaluable support with the CD34 cultures, for checking on my cells whenever I was worried they were unhappy or when quarantines kept me at home, your immeasurable support helped generate the vital material that formed the bulk of this project. Thanks to Dr Matt Baxter, whose calm presence and sound advice has spanned all things technical and emotional whenever I needed it. And a big thank you to Nigel Roberts whose smooth management of the lab throughout all of the chaos has enabled us all to get on with our experiments. For when my pipettes could not solve my problems, thanks to all the computational members of the lab, especially Dr Edward Sanders and Simone Ravza, for your invaluable help with all things analysis. Special thanks in particular to Dr Ravza Gür, whose vital single cell expertise was matched only by her candid emotional support; thank you for being a friend, for emboldening and encouraging me to keep going when things got tough. And to my DPhil counterpart Jennifer Herrmann, with all your technical know-how, practical organisation and high energy I could not have asked for a better companion with which to go through the rollercoaster of emotions that is a DPhil. Thank you for enduring this madness with me, for teaching me "*Ich bin genug*", and for helping me whenever things went horizontal; I could not have done this without you.

I would also like to express my gratitude to my DGS Professor Marella De Bruijn for providing perfect pep talks and words of wisdom when the DPhil fog seemed impenetrable. Thank you also to the Oxford University Counselling Service, and in particular Ruth Sandford-Smith for helping to reinstate my self-confidence, and for providing the mental tools I needed to finish my DPhil.

I have relied heavily on the love and support of my friends and family throughout the years, and consider myself incredibly lucky to be surrounded by such wonderful, reliable people; to anyone who

has endured me blathering on about my project, the highs, lows and everything in between, or has simply made me smile over the course of my project, thank you for helping me make it this far. Thank you to my parents and family whose unconditional love and support is always appreciated and never goes unnoticed. My Mum's unwavering resilience in all that life can throw at one person continues to inspire me and has taught me more about courage than she could know. My Dad's level-headed pragmatic response to any and all problems has taught me the valuable skill of troubleshooting, and that with careful thought almost anything broken can be fixed. Thank you to my brother Sam for the consistent years of looking out for me, to my Aunty Lynn for guiding counsel, to my Nan who still doesn't know what genetics means, to Barney for unbridled enthusiasm, and to my nephews, Reuben and Rafe, whose unbounded energy I wish I could bottle. To Matt, thank you for all the cakes, cocktails, and critiques, call it cause or correlation but my mood and motivation soared when I moved to Reliance Way. To my lab friends who persuaded me to do a DPhil and then promptly left the WIMM, Lea, Hele, Rosa, Ron, thank you for your valued friendship and the South Park picnics. To Steph, thank you for your percipience, and for reminding me that it's ok to ask for help. To Caz, thank you for bearing witness to all the soul-searching conversations and for your impeccable talent of consistently up lifting people. And to Caz and Tom, an immense thank you for supporting me during the write up months, for providing a warm and bright home and for all your understanding words of affirmation; I promise I will move out of the spare room eventually.

And finally a profound thank you to Emma, who has been my rock and greatest confidant for over a decade. Thank you for providing a lifeline in the form of daily dialogue and distraction, often inane, occasionally heartfelt, but ever appreciated. Thank you for your unwavering support and for always being there for me when I need you, regardless of the time, day or night.

This work was generously funded by the Medical Research Council.

Table of Contents

Abstract **1**
Acknowledgements **2**
Table of Contents **4**
List of Figures **7**
List of Tables **10**
List of Abbreviations **11**
Chapter 1 – Introduction **13**
Cis-regulatory elements in the mammalian genome 13
Regulatory structures in the 3D genome 16
Variability in gene regulation in the human population 18
Evolution of regulatory elements 19
Conservation of regulatory elements versus species specificity 20
Origin of enhancer regulatory elements from transposable elements (TEs) 21
De novo formation of enhancer regulatory elements 22
Variation and determining its effect on the functional genome 24
Genome-wide association studies (GWAS) as a method to identify variants 25
Functional genomics analysis of regulatory GWAS variants 26
Intersection of haplotypes with markers of regulatory activity to prioritise SNPs 27
Identification of target cell types for non-coding genetic variation 30
Identification of effector genes linked to regulatory variation 32
Functional validation of the effect of regulatory variation on gene expression 36
Erythropoiesis as a cell model 39
Thesis aims **43**
Chapter 2 - Utilising epigenetic data from across erythroid differentiation to find and define variants within a human population **44**
Introduction **44**
Prioritising likely causal regulatory variants by mechanism of action. 45
Overall aims of the chapter 46
Results **48**
CD34 differentiation system 48
Pre-generated biobank data 48
Characterisation of additional donors for optimisation of downstream assays 49
Comparing ATAC peaks in the donors to find differential regions across individuals 51
Combining ATAC and H3K27ac to find differential peaks with associated activity 53
Determining the frequency of loss and gain of function variants 54
Allelic skew analysis of ATAC-seq data combined with WGS 57
The HBA locus demonstrates functional conservation and shows no variation across 50 individuals 58
A novel peak is unique to a single donor and represents a true gain of function variant 63
An enhancer proximal to SCARB2 shows variation in presence and activity across 50 individuals 68

The variable genomic landscape proximal to PPP1R3B exhibits both gain and loss of function enhancers 73
Variable regions can exhibit multiple peaks appearing simultaneously in certain individuals 81
Variants can also affect CTCF sites with reduced open chromatin reflecting a reduction of CTCF binding 86
A deletion variant relative to the reference genome creates a promoter-like region associated with active transcription 91
Discussion 97
Gains and losses of regulatory activity in the human population 97
Genomic mechanisms affected by variation 98
Acknowledgements 100
Chapter 3 - Using MCC to determine the mechanisms of loss and gain of function variants in late erythroblasts 101
Introduction 101
Performance of MCC 102
Overall aims of the chapter 105
Results 106
Initial optimisations 106
Designing MCC probes 112
Interaction profiles at the HBA locus confirms library quality 112
Using MCC data to determine the genetics of an uncharacterised donor 113
Utilising heterozygosity within capture regions to generate allele specific MCC profiles 114
Allele specific MCC profiles from a SCARB2 enhancer show different interactions with surrounding elements 117
Allele specific MCC profiles show a variant disrupting the binding of CTCF and subsequent interactions 123
Designing and validating MCC probes to variant regions 127
A variant in the R2 enhancer of HBA is a neutral SNP that does not affect interactions 127
Generating MCC libraries from characterised biobank donors 129
Generating MCC libraries from samples with a low cell input 129
Exploring interaction profiles at sites of variation in characterised individuals 130
MCC from a novel gain of function shows no interactions to surrounding landscape 133
MCC from co-appearing open chromatin peaks show weak local interactions 134
MCC from a site of novel transcription shows weak local interactions 136
MCC from the PPP1R3B enhancers show different interactions relative to genotype 137
Allelic MCC tracks from the PPP1R3B enhancers show minimal interactions between enhancers 143
Exploring the interaction profiles at allelic skewed open chromatin sites in characterised individuals 148
MCC profiles from regions of open chromatin skew also show interaction skew 148
MCC profiles from skewed CTCF sites show strong interaction skew 152
MCC profiles from a skewed enhancer containing a GWAS variant show differential interactions with JAK2 promoter 154
MCC profiles from an enhancer containing the erythroid GWAS variant rs696281 show strong allelic skew in a highly variable region 157
Discussion 165
Optimising and implementing MCC in primary human cells 165
MCC can be used to decipher the impact of GWAS variants 167
Acknowledgements 169

Chapter 4 - Assessing the changing genomic landscape of erythropoiesis 170
Introduction 170
Using scMultiome to map erythroid differentiation 171
Overall aims of the chapter 173
Results 175
Set up and optimisation of scMultiome in erythroid cells 175
Sequencing of the day 5 and day 10 libraries 176
Further lysis optimisation 177
Generating PBMC and day 0 libraries 178
Sequencing the PBMC and day 0 samples 180
10X Loupe analysis of erythroid differentiation 180
Defining an erythroid differentiation trajectory arc 184
Gene activity of erythroid markers validates the expected trajectory 187
Assessing genomic landscape changes at regions of variation 193
GWAS SNP enrichment across erythroid differentiation 200
Prioritisation of variants as causal during different differentiation stages using deep neural networks . 207
Regions of open chromatin are skewed in early populations and inactive in later stages 210
Discussion 213
Assessing allelic skew and GWAS intersection in early cell stages 214
Determining the mechanisms of early variants 215
Acknowledgements 217
Chapter 5 – Discussion 218
Identifying new variants in the genome, and their evolutionary and disease implications 219
Challenges in defining variants, interpretation and data visualisation 221
The human genome is highly variable 225
Mapping data to the right place at the right time 227
Novel techniques have helped define variant mechanisms and their timings 228
Conclusion 229
Chapter 6 - Methods 230
CD34 isolation 230
CD34 differentiation 232
ATAC-seq 232
ChIP-seq 234
ATAC and ChIP analysis 236
MCC 236
MCC Probe designing 238
Low cell MCC 258
MCC optimisations 261
MCC analysis 261
scMultiome 261
scMultiome analysis 267
Data analysis and bioinformatics 269
Supplementary Figures 271
References 287

List of Figures

Chapter 1

Figure 1.1: Epigenetic assays can determine genomic activity. 15
Figure 1.2: Genome organisation within a nucleus. 17
Figure 1.3: Transposable elements can impact the genome both directly and indirectly. 22
Figure 1.4: ATAC-seq is able to assay open chromatin with the use of a modified transposase. 29
Figure 1.5: 3C methods can determine the 3D structure of the genome. 33
Figure 1.6: Work flow of MCC. 36
Figure 1.7: Erythropoiesis results in the formation of mature red blood cells from HSCs. 41

Chapter 2

Figure 2.1: CD34^{+} culture media supplements and staging. 48
Figure 2.2: Combined epigenetic data from the 50 individuals, and donors 54 and 55. 51
Figure 2.3: Workflow to curate regions of differential open chromatin. 53
Figure 2.4: Distribution of loss and gain of function peaks within the biobank population. 56
Figure 2.5: Conservation of the regulatory elements at the *HBA* locus. 61
Figure 2.6: Dendrogram of haplotypes present at the R2 enhancer of *HBA.* 62
Figure 2.7: A gain of function variant found in *PTPRK* is only present in one donor. 66
Figure 2.8: Genomic landscape and histone modifications around a gain of function peak in an intron of *PTPRK*. 66
Figure 2.9: Dendrogram of the haplotypes at a gain of function variant at *PTPRK.* 68
Figure 2.10: Variation at the *SCARB2* locus. 71
Figure 2.11: Genomic landscape and histone modifications around a variable enhancer downstream of *SCARB2*. 71
Figure 2.12: Dendrogram of the haplotypes at a variable enhancer peak downstream of *SCARB2.* 72
Figure 2.13: Variation at the *PPP1R3B* locus. 76
Figure 2.14: Genomic landscape and histone modifications around two variable enhancers downstream of *PPP1R3B*. 76
Figure 2.15: The enhancers of *PPP1R3B* show complex inheritance across multiple haplotypes. 77
Figure 2.16: Dendrogram of the haplotypes at the loss of function peak downstream of *PPP1R3B.* 79
Figure 2.17: Dendrogram of the haplotypes at the gain of function peak downstream of *PPP1R3B.* 81
Figure 2.18: Multiple peaks arising simultaneously within the same haplotype. 84
Figure 2.19: Genomic landscape and histone modifications around variable peaks in an intron of *CLNK* occur simultaneously in donors. 84
Figure 2.20: Dendrogram of the haplotypes at the fourth peak of the simultaneously arising peaks at *CLNK.* .. 86
Figure 2.21: Variation at a CTCF site. 88
Figure 2.22: Genomic landscape and histone modifications around a variable CTCF site in an intron of *MYH15*. 89
Figure 2.23: Dendrogram of the haplotypes at the variable CTCF peak at *MYH15.* 90
Figure 2.24: Variation at a site downstream of *FERMT1* is associated with novel active transcription. 94

Figure 2.25: Genomic landscape and histone modifications around an active ATAC peak associated with novel transcription. 94
Figure 2.26: Dendrogram of the haplotypes at a site of novel transcription downstream of *FERMT1*. 96

Chapter 3

Figure 3.1: MCC overview. 104
Figure 3.2: Tapestation profiles showing DNA patterns of a well digested and ligated MCC library. 107
Figure 3.3: Initial experiments showed digestion but failed ligations. 108
Figure 3.4: Bad digestions lead to bad ligations, regardless of temperature. 109
Figure 3.5: Shaking MCC libraries during digestion caused loss of material. 110
Figure 3.6: Successful MCC libraries of late erythroblasts derived from a blood cone sample. 112
Figure 3.7: MCC interactions from cone sample at *HBA* promoter showing expected profile. 113
Figure 3.8: Schematic of allelic MCC track generation. 116
Figure 3.9: Example of allelic split at *SCARB2* enhancer, with rs11097297. 119
Figure 3.10: Foot-printing data show rs11097297 disrupts protein binding. 122
Figure 3.11: *SCARB2* mature RNA allelic counts, split by rs11097297 genotype. 123
Figure 3.12: Example of allelic split at a CTCF site, with rs936266, in *MYH15I intron*. 125
Figure 3.13: Footprint and sasquatch data showing damage at rs936266 CTCF *MYH15* site. 126
Figure 3.14: Validating the second set of probes and demonstrating that a neutral SNP has no effect on allele specific tracks. 128
Figure 3.15: Validation of six MCC libraries at the *HBA* locus. 133
Figure 3.16: A Gain of Function variant unique to donor 1 shows no interactions. 134
Figure 3.17: Interactions at *CLNK* site of multiple novel peaks. 136
Figure 3.18: No interactions from active peak, despite active transcription. 137
Figure 3.19: Interactions at *PPP1R3B* from a C/C rs330946 genotype, and a G/G rs3748136 genotype. 139
Figure 3.20: Interactions at *PPP1R3B* from an A/C rs330946 genotype, and a G/G rs3748136 genotype. 140
Figure 3.21: Interactions at *PPP1R3B* from an A/C rs330946 genotype, and a G/A rs3748136 genotype. 142
Figure 3.22: Allelic MCC at *PPP1R3B* from an A/C rs330946 genotype, and a G/A rs3748136 genotype. 145
Figure 3.23: MCC foot-printing data at the *PPP1R3B* locus. 147
Figure 3.24: Allelic MCC from rs148015593 shows consistency across individuals. 150
Figure 3.25: Allelic MCC footprinting from rs148015593. 152
Figure 3.26: Allelic MCC tracks from rs897804 show differing CTCF interactions. 153
Figure 3.27: Allelic MCC footprint at rs897804. 154
Figure 3.28: Allelic MCC tracks from rs10758656 show differing enhancer interactions to *JAK2*. 156
Figure 3.29: Allelic MCC FP from rs10758656. 157
Figure 3.30: Allelic MCC tracks from rs696281 show differing enhancer interactions. 160
Figure 3.31: Allelic MCC FP from rs696281. 161
Figure 3.32: Close up of allelic MCC tracks from rs696281 show differing enhancer interactions. 162
Figure 3.33: Variation at the *BZW2* locus shows landscape pattern is independent of SNP. 164

Chapter 4

Figure 4.1: scMultiome uses GEMs and unique barcodes to label tagmented DNA and RNA from single cells. 172
Figure 4.2: Bioanalyser intermediate analysis of multiome libraries. 179
Figure 4.3: UMAP Loupe analysis of a single donor across erythropoiesis. 183
Figure 4.4: Changes in open chromatin can be seen when comparing ATAC from different erythroid stages. 184

Figure 4.5: Defining an erythroid trajectory and cell stage clusters. 187
Figure 4.6: Gene activity profiles show expected patterns across erythroid differentiation arc. 190
Figure 4.7: Gene activity profiles of late stage proteins show enrichment in later erythroid differentiation stages. 191
Figure 4.8: Using the trajectory clusters to visualise cell stage specific ATAC profiles at *HBA*. 192
Figure 4.9: ATAC accessibility of a *SCARB2* enhancer increases through erythroid differentiation. 195
Figure 4.10: Aggregated scATAC and bulk ATAC tracks are comparable. 196
Figure 4.11: ATAC accessibility of the PPP1R3B enhancers increase through erythroid differentiation and differ between donors. 197
Figure 4.12: ATAC accessibility of a CTCF site remain constantly open through erythroid differentiation. 198
Figure 4.13: ATAC accessibility changes through differentiation are complex over the variable CLNK region. 199
Figure 4.14: Using the Cell Type Prioritisation Platform to find GWAS intersection enrichment in cell populations. 201
Figure 4.15: Combined erythroid and PBMC UMAPs with overlaid GWAS enrichments. 203
Figure 4.16: Erythroid GWAS SNPs are enriched at all stages across erythropoiesis. 205
Figure 4.17: DeepHaem variant damage score predictions show changing values across erythropoiesis. 209
Figure 4.18: ATAC accessibility landscape at *ASB2* shows many peaks turning off simultaneously through erythropoiesis. 211
Figure 4.19: ATAC peaks spanning multiple genes show dynamic changing through erythropoiesis. 212
Figure 4.20: Tiled-C can detect chromatin interactions in as few as 2,000 cells. 217

Chapter 5

Figure 5.1: Loss and gain of function variants. 219
Figure 5.2: A SNP in an enhancer of *BCL11A* causes the derepression of *HBG*, and the subsequent production of foetal haemoglobin. 221

Chapter 6

Figure 6.1: FACS gating for removal of granulocytes and NETs prior to PBMC lysis. 262
Figure 6.2: Cell and nuclei integrity during scMultiome lysis optimisations. 263

Supplementary Figures

Supplementary Figure 1: TAD structure at *HBA.* 271
Supplementary Figure 2: TAD structure at *PTPRK.* 272
Supplementary Figure 3: TAD structure at *SCARB2.* 273
Supplementary Figure 4: TAD structures at *PPP1R3B.* 274
Supplementary Figure 5: TAD structures at *CLNK.* 275
Supplementary Figure 6: TAD structures at *MYH15.* 276
Supplementary Figure 7: TAD structures at *FERMT1.* 277
Supplementary Figure 8: Exemplar Mpileups. 278
Supplementary Figure 9: Exemplar of using BAMs to determine genotype. 279
Supplementary Figure 10: ATAC Skew at rs3748136 282
Supplementary Figure 11: MCC validation of probe set 3 at *HBA.* 283
Supplementary Figure 12: Phasing links SNPs on alleles at *PPP1R3B* locus. 284
Supplementary Figure 13: Open chromatin profiles from 96 cell types at the gain of function peak in PTPRK 286

List of Tables

Chapter 2

Table 2.1 Available data sets for the 50 individuals. 49

Chapter 3

Table 3.1: DNA yields of supernatants and samples from shaking experiment. 111
Table 3.2: Genotypes and activity at rs330946 and rs3748136 in six donors..... 138

Chapter 4

Table 4.1: Cell counts of optimisations, lysis efficiency and nuclei retention. 176
Table 4.2: Cell counts for day 5 and day 10 samples. 177
Table 4.3: cell counts and quality from the initial sequence of PBMC and day 0 samples. 180
Table 4.4: Nomenclature of UMAP clusters..... 185
Table 4.5: Regions open in early cell stages show allelic skew. 210

Chapter 6

Table 6.1: ATAC-seq custom primers. 234
Table 6.2: ChIP antibodies. 235
Table 6.3: MCC set 1 probes 243
Table 6.4: MCC set 2 probes 250
Table 6.5: MCC set 3 probes. 258
Table 6.6: MCC libraries generated 260
Table 6.7: MCC capture experiments and pools 261
Table 6.8: scMultiome ATAC indexing for donors 2 and 3, days 5 and 10. 264
Table 6.9: scMultiome ATAC indexing for donors 2 and 3, CD34 day 0 and PBMCs. 265
Table 6.10: scMultiome GEX indexing for donors 2 and 3, days 5 and 10. 265
Table 6.11: scMultiome GEX indexing for donors 2 and 3, CD34 day 0 and PBMCs. 266
Table 6.12: GWAS traits used in enrichment analysis. 269

Supplementary Tables

Supplementary Table 1: Summary of SNPs introduced in Chapter 2. 278
Supplementary Table 2: Nascent allelic RNA transcription of *SCARB2* across the 50 donor. 280
Supplementary Table 3: Mature allelic RNA transcription of *SCARB2* across the 50 donors. 282

List of Abbreviations

3C	Chromosome confirmation capture
3D	Three-dimensional
AO/PI	Acridine orange/ propidium iodide Stain
ATAC-seq	Assay for transposase-accessible chromatin sequencing
BFU-E	Burst forming unit-erythroid
bp	Base pairs
CFU-E	Colony forming unit-erythroid
CHEERS	Chromatin Element Enrichment Ranking by Specificity
ChIP-seq	Chromatin immunoprecipitation sequencing
CNN	Convolution neural network
CRE	*Cis*-regulatory element
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
DNase-seq	DNase I hypersensitive sites sequencing
EPO	Erythropoietin
eRNA	Enhancer RNA
GEM	Gel beads-in-emulsion
GoF	Gain of function
GRO-seq	Global run-on sequencing
GWAS	Genome-wide association studies
H3K4me1	Histone 3 lysine 4 mono-methylation
H4K4me3	Histone 4 lysine 4 tri-methylation
H3K27AC	Histone 4 lysine 27 acetylation
HDR	Homologous directed repair
HSC	Haematopoietic stem cell
HUDEP	Human Umbilical Cord Blood-Derived Erythroid Progenitor
IL-3	Interleukin-3
JAK2	Janus-kinase 2
kb	Kilobase pairs
KO	Knockout
LD	Linkage disequilibrium
LoF	Loss of function
MACS	Magnetic-activated cell sorting
Mb	Megabase pairs
MNase	Micrococcal nuclease
MCC	Micro Capture C
MLV	Multi Locus Viewer
PBMC	Peripheral blood mononuclear cell
PIC	Pre-initiation complex
Pol-II	RNA polymerase II
PolyA-	Non-polyadenylated

PolyA+	Polyadenylated
RE	Regulatory element
RNA-seq	RNA sequencing
scATAC-seq	Single cell ATAC-seq
SCF	Stem cell factor
scMultiome	Single cell Multiome
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TAD	Topologically associating domains
TE	Transposable element
TF	Transcription factor
TFBS	Transcription factor binding site
TSS	Transcriptional start site
TT-seq	Transient transcriptome sequencing
UTR	Untranslated region
WGS	Whole genome sequencing

Chapter 1 – Introduction

Cis-regulatory elements in the mammalian genome

The adult human body is a complex combination of hundreds of different cell types and tissues, with a total composition of around 37 trillion cells [1]. Different cell types are specialised in their function, morphology, and behaviour; yet despite this diversity, each cell is broadly encoded by the same genomic sequence, irrespective of function or tissue of origin. The human genome comprises over 3 billion base pairs (bp) [2], which contain the ~22,000 protein-coding genes utilised by all cell types from conception to adulthood. When compared with less complex multicellular organisms, our complexity cannot be attributed to an increase in the number of protein-coding genes. *C. elegans*, a simple roundworm of ~2,000 cells, has a comparable ~20,000 genes [3]. Instead, our increased complexity is largely explained by more flexibility in the use of existing genes. This is achieved via alternative splicing or the use of isoforms, and by the dynamic utilisation of *cis* regulatory elements (REs) to control gene expression.

Gene expression is regulated by both promoter and enhancer REs, which physically interact in close proximity to change the transcriptional output of a gene [4-8]. Promoters are specific regions adjacently located upstream of, or over lapping with, the 5' transcriptional start site (TSS) of coding sequences. Using binding motifs within their DNA sequence and under the influence of distal REs, promoters recruit the transcription factors (TFs) and complexes required for transcription initiation. This process starts with the formation of the pre-initiation complex (PIC), which in turn recruits RNA polymerase II (Pol-II) and initiates transcription. For the DNA to be accessible to these factors, active promoters are found within open chromatin sites that can be detected by assaying for DNA hypersensitivity with techniques such as assay for transposase-accessible chromatin sequencing (ATAC-seq). Promoters can be epigenetically defined using chromatin immunoprecipitation sequencing (ChIP-seq) for Pol-II binding and certain histone modifications, such as histone 4 lysine 4 tri-methylation (H3K4me3) or histone 4 lysine 27 acetylation (H3K27Ac). These modifications are associated with promoters and general transcriptional activity, respectively [9] (Figure 1.1).

Another main class of REs are referred to as enhancers. Enhancer REs are specific regions of DNA distinct from promoter REs that can upregulate or enhance the transcriptional output of target genes. In addition to H3K27Ac, they are also epigenetically marked by the histone 3 lysine 4 mono-methylation (H3K4me1) histone modification [10] (Figure 1.1). With an average sequence length of 100-1000bp, they contain TF binding sites that can recruit cell type-specific and general TFs responsible for the opening of chromatin at promoters, the recruitment of the PIC, and the initiation of transcription. As TFs are often cell type-specific, they can drive lineage-specific spatiotemporal gene expression. For factors to efficiently bind, active enhancers are dynamically nucleosome depleted to form open chromatin sites, facilitating further sequence recognition and binding of TFs. Unlike promoters, enhancers do not have a defined location relative to the target TSS or gene and can be located both up- and downstream of their associated genes, often found within introns of genes, including unrelated genes. Though usually proximal, they are regularly 10's to 100's of kilobase pairs (kb) away, and in some cases can be more than one megabase pairs (Mb) away from their target promoters [11]. Finally, genes are often controlled by multiple enhancers, especially highly expressed or highly regulated genes. When such enhancers are clustered together, they can form Locus Control Regions, also referred to as super enhancers [12].

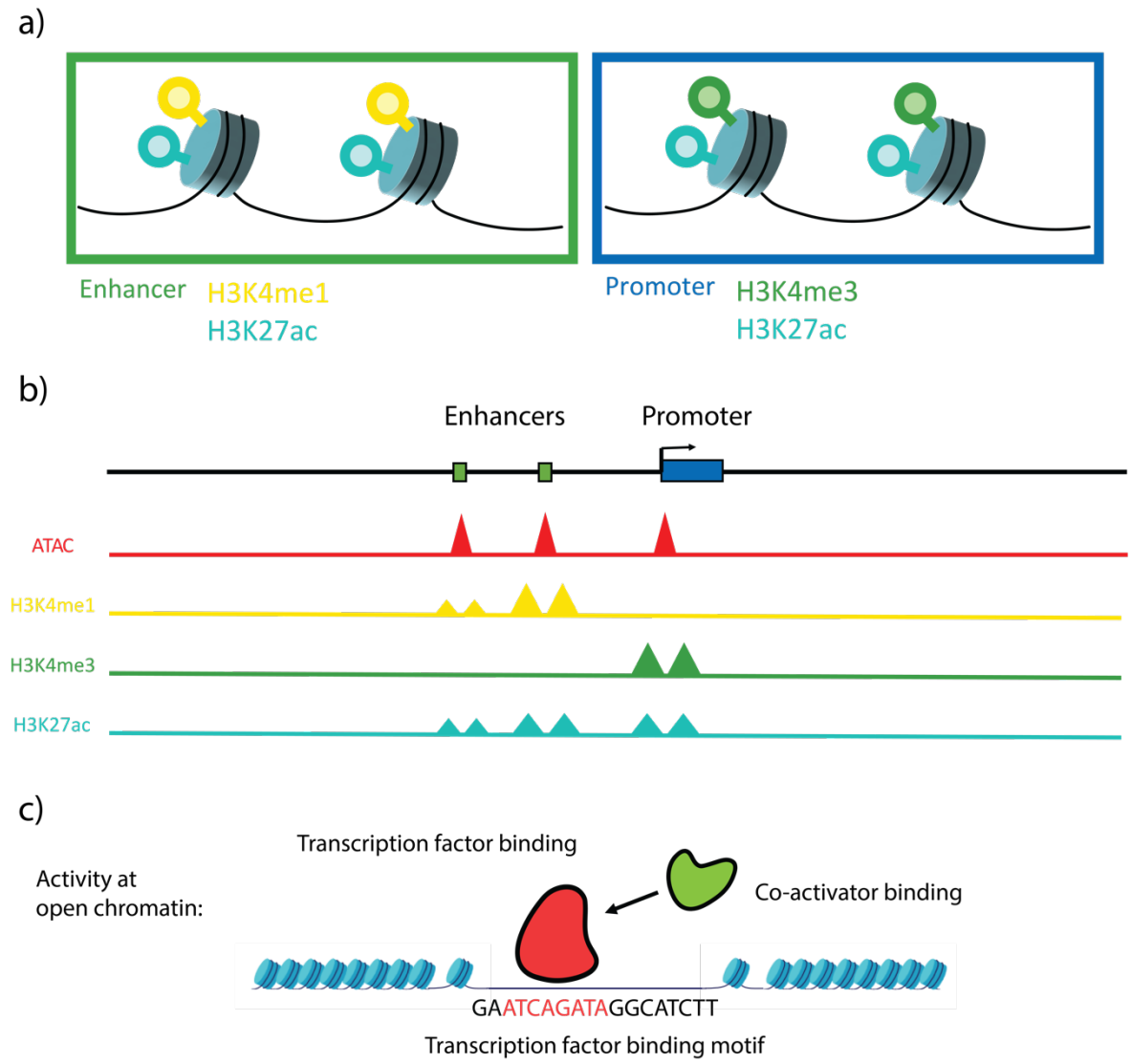


Figure 1.1: Epigenetic assays can determine genomic activity. a) Histone modifications H3K4me1 and H3K27ac mark active enhancers, H3K4me3 and H3K27ac mark active promoters. b) These regions are often also accessible, which can be detected with ATAC-seq. Histone modifications detectable with ChIP-seq. All these data can be aligned to the genome and visualised as tracks using the UCSC browser [13]. c) Regulatory elements at sites of open chromatin contain transcription factors binding motifs that bind transcription factors and can further recruit co-activators.

The third class of REs, boundary elements or CCCTC-binding factor (CTCF) binding sites, are thought to be more structural in nature in comparison with the active recruitment function of enhancer and promoter REs [14]. CTCF is a large protein containing multiple zinc fingers; these are DNA binding domains that recognise and bind an unusually large DNA sequence motif (from 12 to 165bp) [15, 16].

Unlike active enhancers and promoters, CTCF-bound sites are more consistent across different cell types with large numbers of binding sites persisting across all tested cell types. However, each cell type has a proportion of cell type-specific binding sites, which collectively across many cell types add up to large numbers of cell type specific CTCF binding sites. It has been shown conclusively that a CTCF binding site is capable of blocking interactions between active enhancers and promoters [14, 17], and so are thought to be a major mechanism in ensuring the correct cell type-specific interactions between REs. The ability of CTCF binding to block the interactions between REs is rooted in the three-dimensional (3D) structure of the genome [18].

Regulatory structures in the 3D genome

While distal in linear DNA, active enhancer and promoter REs are known to physically interact within the 3D nucleus. These interactions occur within 3D genomic structures called self-interacting domains, which are organised and demarcated by boundary elements, in particular CTCF-bound elements. Within these domains, a higher frequency of self-interactions have been observed compared with the rest of the genome [19], which suggested that genome organisation and gene expression are functionally linked. Genome organisation starts at a chromosomal level with interphase chromosomes occupying separate territories within the nucleus. Within these territories there is preferential interaction between active and inactive regions in A and B compartments, respectively. Topologically associating domains (TADs) and sub-TADs further segregate the genome into frequently interacting regions and self-interacting domains (Figure 1.2) [20]. Frequently interacting regions are genomic regions that are more often in close proximity to each other in relation to other genomic regions, regardless of proximity. Within these domains, RE proximity is both caused and strengthened by the interactions between DNA bound TFs and structural proteins. The formation of these domains and the subsequent interactions of REs within them is a highly dynamic process, which is thought to allow for the precise regulation of gene expression.

As cells differentiate or respond to environmental cues, genomic regions react accordingly: genes within B compartments may be permanently silenced, whereas active genes within A compartments may be differentially expressed. This precise pattern relies on the presence and combination of cell

type- and stage-specific TFs. By folding the genome in a certain way, RE proximity and interaction frequency can be modulated, allowing cells to regulate gene expression output.

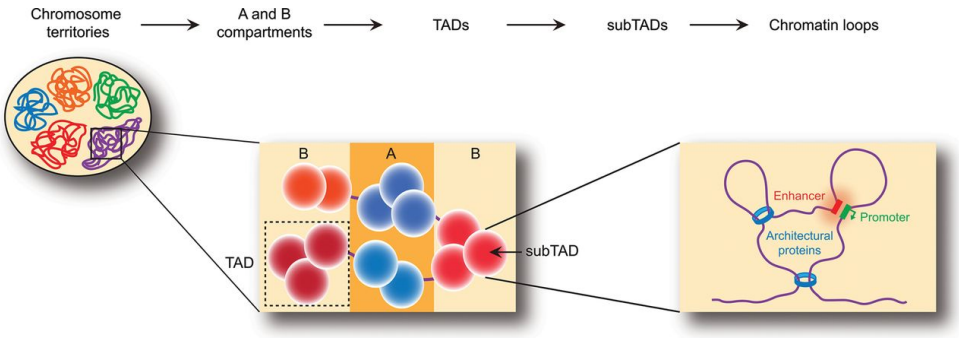


Figure 1.2: Genome organisation within a nucleus. The genome is organised from chromosome territories and A/B compartments, to TADS, sub-TADS, and chromatin loops. Self-interacting domains bring together REs and are demarcated by CTCF sites and architectural proteins (figure adapted from Denker & de Laat, 2016 [20]).

In the last decade, much effort has gone into trying to understand the processes that generate and dictate these 3D structures and regulatory interactions in the genome. From this work, the cohesin molecule has been identified as a key motive force in the 3D genome. Previously studied for its critical role in the cohesion of replicated sister chromatids in mitosis, it was realised that it had another nuclear role in interphase nuclei. A comprehensive body of work to date has shown that during interphase the cohesin complex is loaded onto chromatin and induces the chromatin to form an extruding loop, which progressively and systematically brings distal regions of chromatin in close contact; a process known as loop extrusion. Recent work has suggested that the activity of enhancers and/or promoters stimulate the loading of cohesin or its extrusion at active regions of the genome to presumably bring enhancers and promoters together efficiently [21, 22]. Further work has shown that CTCF, when bound to chromatin in a specific orientation, can block loop extrusion and so dictate which promoters and enhancers can interact. Therefore, while the binding of CTCF itself may not be cell type-specific, its position and collaboration with cohesin can dictate the cell type-specific interactions of REs by modulating loop extrusion. Additionally, studies have shown that CTCF sites at enhancers can increase the frequency of promoter interactions [23]. This is further supported by studies from CTCF sites associated with promoters that facilitate enhancer mediated gene regulation by increasing RE proximity [24]. It is therefore likely that mechanistically CTCF can facilitate loop extrusion, and that

the subset of enhancer and promoters that are also associated with CTCF are provided robustness in their long range interactions for gene expression [25]. The polymorphic binding of CTCF within a human population, in addition to differentially active enhancer and promoter REs, therefore also has the potential to be a powerful source of gene expression variability.

Current theories hypothesise that within larger scale TADs and sub-TADs, many more smaller scale chromatin loops also form and stack to build rosette like structures. Advances in sequencing methods, degrons systems, and super resolution microscopy have allowed the investigation of the effect of CTCF and cohesin depletion on chromatin architecture. Cohesin drives the formation of chromatin loops that facilitate <4Mb proximal interactions whilst also insulating distal (>4Mb) regions separated by several loops. Studies suggest that loop anchors are enriched for CTCF binding sites and form central hubs, and steric interference among cohesin molecules cause loops to stack radically forming characteristic 'rosettes'. These stacked loops increase the space between distal regions and are therefore able to prevent ectopic long-range interactions between REs. Any REs proximal to a CTCF binding site at the border of a loop are therefore likely to cluster together and interact more frequently, even across TAD boundaries [26]. The importance of these boundaries as insulators is demonstrated by disease phenotypes that can arise from the disruption of CTCF sites, leading to gene misregulation [27-29].

Variability in gene regulation in the human population

Due to our currently incomplete understanding of gene regulation, it is unclear how the system of distal REs controlling gene expression has evolved, and how variable they are in the normal human population. In contrast, variation in coding sequences has been extensively studied via initiatives such as the UK biobank, which has recently released the genome sequences of 500,000 individuals [30]. The reason for this disparity is that there exists an almost complete annotation of human genes, and we have a good understanding, at the sequence level, of coding and splicing potential. This means that with standard genome sequencing it is relatively straight forward to interpret coding variation and its effect. However, REs are much more challenging. Firstly, we do not have a complete map of their positions in the human genome due to their high levels of cell type specificity. Secondly, at present there exists limited predictive methods to understand which variants may affect their function

and which are neutral. This means it is extremely challenging to understand the potential for any sequence to alter function in the non-coding regions of the genome.

To that end, the current drive in the field is to generate open chromatin-derived, cell type-specific maps of all human cell types using single cell ATAC-seq (scATAC-seq) based approaches. Alongside this are the development of computational models, in particular, machine learning-based neural networks that can predict whether a sequence change may or may not change the activity of REs in a given cell type or situation [31, 32].

Evolution of regulatory elements

One long standing approach to identify functional elements within the genome is to use sequence conservation between sequences as a proxy for function [33]. Therefore, understanding the dynamics and mechanisms by which REs have evolved and are retained or lost between species can give us some indications of how reliable such methods are to determine regulatory variation between individuals.

When comparing across kingdoms, bacterial DNA loops that regulate gene expression have been observed [34], and yeast have prevalent Upstream Activating Sequences that are also able to act as enhancers in transgenic human cell models [35]. Despite these mechanisms being present in simpler organisms, it is their evolution and widespread abundance in mammals that is thought to have driven complex cellular diversity. Cellular plasticity and differentiation are driven by the acquisition of proteins with new functions or by the alteration and fine tuning of their spatiotemporal expression. This plasticity in utilising available proteins has shaped the ontogenetic and phylogenetic diversity of cell types and organism forms [36]. When considering REs, promoters have been shown to be more conserved with a slower evolution than enhancers, which are typically less conserved and more rapidly evolving [33, 37]. This is likely due to a promoter's central and non-redundant role in a gene's expression across many if not all cell types, so once evolved and functional, the evolutionary pressure on the gene's function will drive strong selection for the conservation of the promoter sequence.

In contrast to this, enhancers represent an extra layer of regulation which may be required in only a subset of cell types. Furthermore, as multiple enhancers can feed into a gene's regulation, both of

these features may lead to increased redundancy and therefore decrease sequence conservation for any given enhancer [38, 39]. This means that conservation is limited in its use to gauge regulatory variation, and at the extreme end, would completely fail to detect human-specific enhancers. However, due to their lower levels of conservation they potentially represent a more important source of population-level variation than promoter elements or coding sequences [12].

Conservation of regulatory elements versus species specificity

Despite enhancer REs being less conserved as a class than promoter REs, certain enhancers are conserved [40, 41]. Conservation is often linked to genes and enhancers of developmental function. Therefore, enhancers can be classified into highly conserved, weakly conserved, and species specific. One general trend seen in the literature is that enhancers linked to genes with key developmental roles, such as limb patterning or with basic organ/tissue formation, are much more likely to be conserved to much higher levels and shared across species [42-45].

The *Hox* genes are crucial in axial patterning for both invertebrates and vertebrates and are controlled by a series of enhancers [46]. Studies have shown that vertebrates separated by 500 million years of evolution still showed conservation at the hox enhancers [47]. An additional example of conservation is the *Sonic Hedgehog* gene, which is essential for correct limb bud formation and is controlled by a series of highly conserved enhancers that are found across all classes of vertebrates [43, 44]. These studies have also shown that the disruption of these enhancers affects gene expression and leads to a polydactyly phenotype. Interestingly, it has been shown that a single inactivating mutation in a highly conserved *Sonic Hedgehog* enhancer was responsible for the evolutionary loss of limbs during snake evolution, showing the potentially profound effects and evolutionary consequences of enhancer variation [48]. In contrast, human-specific enhancers are often associated with neural development, which reflects our increased cognition when compared with other great apes [36]. *De novo* human-specific enhancers associated with neural pathways are often fast evolving and associated with aging-related diseases [49, 50]. Moreover, similarly to limb development, disruption to the regulation of these pathways can have profound phenotypic effects [51]. SOX9 is a TF involved in a range of embryonic development processes, and mutations of its control region can cause a human craniofacial

disorder [11]. However, even strongly conserved enhancers can tolerate a certain level of mutation, demonstrating their robustness [52]. By tolerating a certain amount of change, they remain functional.

The evolution of enhancers has been integral to the evolution of mammals and humans, with the variation in their number and function accounting for vast species diversity. Several mechanisms of enhancer genesis and evolution have been proposed though they are not mutually exclusive, and a combination of factors are likely to drive their emergence and selection.

Origin of enhancer regulatory elements from transposable elements (TEs)

TEs are a major component of eukaryotic genomes, and broadly speaking are DNA sequences of viral origin with the ability to change their position within the genome (Figure 1.3). These elements have inherent regulatory potential due to their viral genome origin because they are enriched for transcription factor binding sites (TFBSs), and so are primed for potential enhancer function. Depending on where they integrate, they can interact with *cis*-regulatory networks and modify their output [53]. If they integrate into a locus and impart a beneficial gene expression output, this will be positively selected for. Such cases have been observed in a variety of cell types and transcriptional pathways [36]. However, more commonly they provide no change in fitness and instead will be eroded by point mutations until their transposable sequences are lost: 99.9% of the ~500,000 L1 copies are now fixed within a human genome with only ~100 young, active L1 elements present [53]. Ancestral or modified TEs undergoing exaptation is rare but present, and eroded TE sequences are widespread in the genome. When considering primate genomes, Trizzino *et al*. assessed liver-specific REs and showed broad conservation across genomic landscapes, with 63% conservation [54] likely due to the conserved nature of liver function. Newly evolved REs were found to be enriched in young TEs and were linked to immune response and neurodevelopmental genes reflecting the likely species specific adaptation of these systems. Conversely, REs contributing to core functions remained intact and only had 16% overlap with TEs [36]. This perhaps suggests their endurance over time has caused their drift away from discernible TE features, or that they originated via other means.

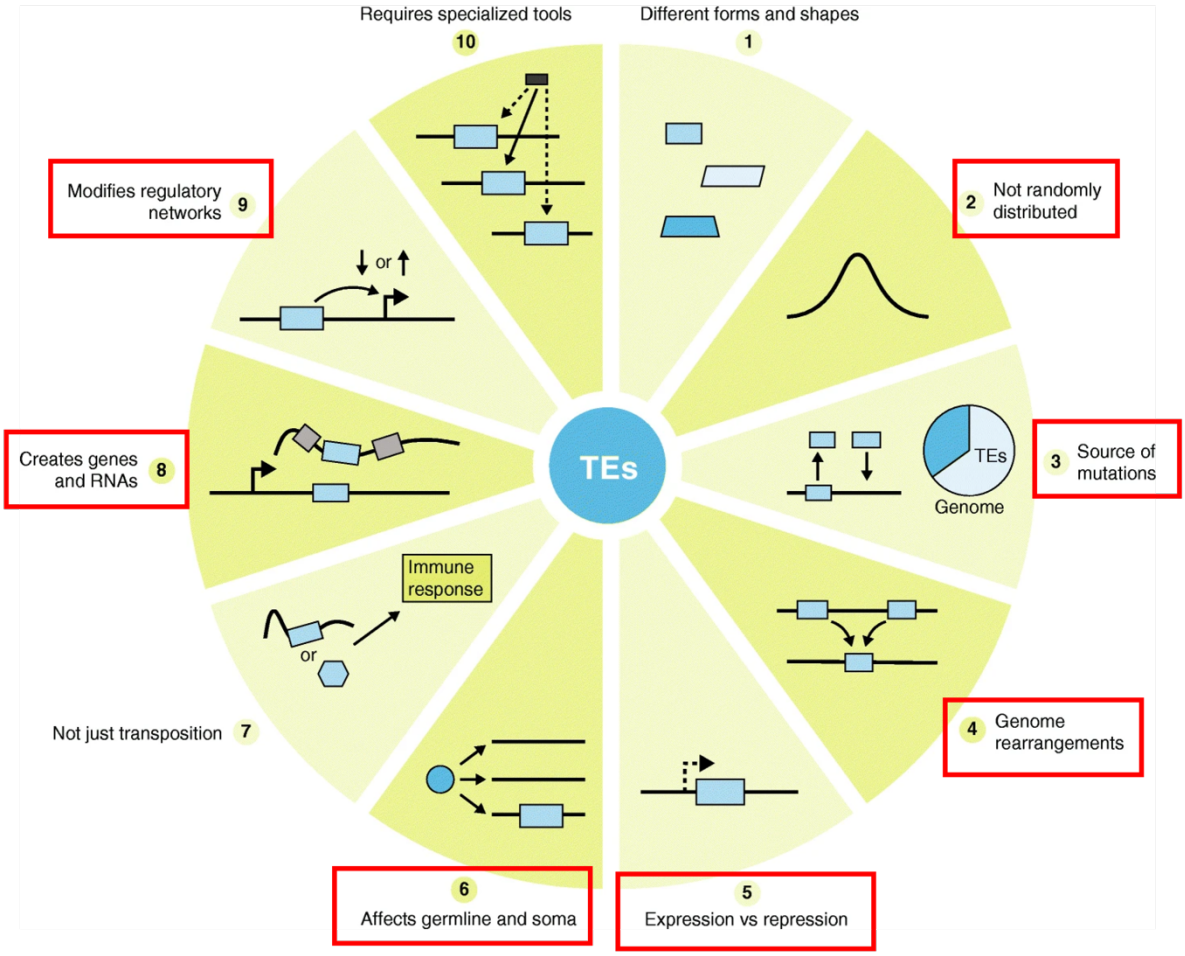


Figure 1.3: Transposable elements can impact the genome both directly and indirectly. Blue boxes represent TEs, grey boxes canonical exons, and the black box a sequencing read. TEs have many known characteristics and traits as depicted in the figure, several of which interplay with the regulatory network (red boxes). They are a source of mutation, influence genome rearrangements, and modify regulatory networks. One key trait is their ability to form regulatory elements. (Adapted from Bourque *et al.* [53]).

De novo formation of enhancer regulatory elements

Another proposed mechanism is the duplication of pre-existing enhancers that can then further change and evolve [55]. They may also emerge from *de novo* mutations in non-coding regions or within the extant ancestral 10,000-20,000 pseudogenes found within the genome [56, 57]. These regions that already contain binding sequences would require fewer alterations to become active. Several research groups have attempted to model the mutation rate required to create new promoters or enhancers. Stone *et al*. simulated the evolution of a promoter by looking at how long it would take to

develop a 6bp TFBS. They determined that each human gene would evolve a new 6bp binding site for a known TF somewhere within 2kb of its TSS at a rate of ~0.013 per genome. Within a population of 1,000,000 people this would equate to ~12,500 new 6bp binding sites during each generation [58]. However, forming a binding site and even recruiting factors does not necessarily equate to function as the genome context will determine if anything functional will happen. Therefore, since promoters often contain multiple TFBSs required for activation, a single 6bp site arising is unlikely to provide activity on its own or enough activity to produce a selectable change. Duque *et al*. take into consideration the complex nature of REs, and instead propose >0.5 million years of evolution are required to evolve a complex, functional, expression-driving promoter from genomic background [59].

Such estimates are even more complicated for enhancers, because their function is not tied to the proximity of a gene body and can work over large distances to influence gene expression. The composition of an enhancer is also highly plastic in nature with no rigidly fixed binding site composition. Therefore, the formation of an enhancer sequence can be arrived at by many routes. This means in practice that enhancers could arise from variation, utilising existing cryptic TFBSs, over a very large window of sequence around a gene and still potentially affect expression. Furthermore, when compared to other REs such as promoters, the plasticity of enhancers in their use of motifs and their ability to function without perfect sequence motifs suggests they are more likely to evolve [39, 52].

Ultraconserved enhancer function does not require perfect sequence conservation

With all these insights taken into account, the genesis and evolution of enhancers is still an emerging field, and these mechanisms are likely to interplay with multiple mechanisms converging. However, whatever the mechanism of enhancer generation, evolution can then either select for them and ultimately fix them in the species, select against them or they can remain as regulatory polymorphisms with the population.

Variation and determining its effect on the functional genome

Mutations or variants in the DNA, such as single nucleotide variants (SNVs) or single nucleotide polymorphisms (SNPs), are common and account for much of the inherit diversity within species. A SNV is any single base pair change and is classified as a SNP once present in >1% of a population. Mutations arise through multiple mechanisms, such as the failure of proofreading mechanisms to detect polymerase errors during DNA replication or by environmental mutagens [60]. Over time these errors will be passed down generations, accumulating in the population and accounting for genomic variation. In humans, SNPs are found at 1 in every ~1,000 nucleotides, equating to 4-5 million per individual [61]. Since variants alter the genetic code, they have the potential to influence the way in which a cell interprets that code, which may have downstream effects on functionality that lead to genetic traits. A well-studied class of variants are those which lie within the body of a gene. Here they have the potential to change the structure and/or stability of the encoded protein, either by directly changing codons or by altering splicing [62, 63]. Other variants can alter the gene transcript, such as those that effect the 3' untranslated region (UTR) sequence; altering polyadenylation, microRNA binding, or regulatory protein binding, all of which could alter the stability of the transcript and so alter levels of protein production [64].

A less studied class of genetic variants are those which alter distal REs and processes, as their location is less predictable than those within gene bodies. As described previously, the analysis of RE function is complicated by their unpredictable distribution and distance around their target genes and this make understanding the effect of variation within them similarly challenging. A further major complication of identifying such regulatory variants is that SNPs are often entirely silent even when they lie within REs [52]. However, when a SNP does alter TFBSs or disrupts enhancer grammar this can influence gene expression, leading to natural human variation or disease. Although, their effect on gene expression can often be subtle and hence the architecture of common genetic traits are often highly polygenic.

The subtlety and range of a SNPs impact can be related to the effects imparted on and around individual motifs within an element. It is known that with ample base changes to key sequences binding can be disrupted. However, the strength of binding to a motif is determined specifically by the

sequence, as TFs will bind with different affinities to motifs depending on their conservation. Because of this, motifs are tolerant to a certain level of change, hence why the frequency and position of a base change will determine the impact, with a gradual accumulation of SNPs altering a motifs strength until it is no longer recognisable to TFs [52]. Enhancer grammar is also an important factor in determining activity, and disruption to this organisation can alter the functionality of REs. The ordering, spacing and orientation of TF binding has been shown to be essential for certain regulatory loci [65]. Due to the precise interplay between TFs, whereby early pioneer factors are essential for opening chromatin before further TFs can bind and subsequently interact, the order in which TFs recognise a site is important. Once open, the spacing between TFBSs and their orientation will then determine how the TFs are able to interact with each other, and thus how they can influence genomic effects. However, every enhancer will behave differently, with some needing the precise occupation of TFs, and others being much more tolerant to changes in the spacing or occupancy of TFs [66]. The ways in which a SNP can effect an RE are broad and therefore so is the impact they can have on activity and function [67].

The intricate workings of an individual's unique combination of SNPs will subtly change the expression at many loci to increase or decrease an individual's propensity to a given trait. This is in contrast to SNPs within coding regions leading to monogenic diseases, which detrimentally alter protein structure and lead to diseases such as sickle cell disease [62] and cystic fibrosis [63]. An additional layer of complexity is that most common human diseases are caused by both polygenic interactions and environmental factors, hence the physiological effects of variants within REs are poorly predictive of an individual's outcome at present. However, despite these challenges several examples of monogenic disorders thought to be caused by regulatory mutations exist, such as the misregulation of *PAX6* causing aniridia [68] and the damage to an enhancer of *PTF1A* that leads to pancreas agenesis [69].

Genome-wide association studies (GWAS) as a method to identify variants

A common approach to overcome this complexity and lack of full predictability is to identify statistical enrichment of SNPs across the whole genome in a very large cohort (tens to hundreds of thousands) of affected individuals relative to matched unaffected individuals. Over the past two decades a multitude of GWAS studies have been performed to identify potentially causal SNPs in common

human traits, resulting in the identification of millions of disease- and trait-associated SNPs [70-72]. In contrast to monogenic disorders, in which the majority of these SNPs lie within coding regions, ~95% of SNPs associated with polygenic disorders are found to be intronic or intergenic, making the underlying biology and even the affected genes much more difficult to predict or interpret [73].

Moreover, due to high sequencing costs most GWAS are performed using microarrays to ascertain an individual's genotype rather than whole genome sequence. This means that only an incomplete representation of human variation can be sampled. Due to linkage disequilibrium (LD), SNPs will often exist in coinherited blocks that persist across individuals [74], which means that a SNP identified in GWAS may not be the true causal SNP and is instead merely representing a causative SNP within the same LD block. Therefore, statistical imputation should be used on GWAS-identified SNPs to get a more complete representation of the haplotypes of potential SNPs of interest [75]. Traditionally, imputation relies on using statistical analysis of the existing genome sequence data from thousands of individuals to determine the likelihood of a person with a given SNP coinheriting other SNPs as a block.

After imputation, to decode these genetic signals there are several interrelated genomic challenges that need to be solved.

- Which SNP(s) in a given haplotype has any effect?
- Which SNP(s) in a given haplotype is completely neutral?
- If a SNP does have an effect, what is the mechanism by which it exerts this effect (e.g. coding, splicing, regulatory)?
- If it is regulatory, in what class of regulatory element does it reside?
- If it is a cell type-specific mechanism in which cell type(s) is it acting?
- Which gene, genes or pathways are being affected in that cell type(s)?

Functional genomics analysis of regulatory GWAS variants

Once an imputed list of potentially causative SNPs has been established, how can it be determined which SNP(s) is having a causative effect? One method is to progress through each haplotype and try

to logically discount the mechanisms that it cannot be affecting, progressively annotating and reducing the possibilities. As stated previously, coding regions are well documented and tractable, therefore, haplotypes with potential causative SNPs within exons are straightforward to locate and follow up experimentally for protein function. Similarly, splice sites are, for the most part, well understood and annotated and so the effect of variants is predictable. These haplotypes can be prioritised for experimental analysis of transcript structure and stability. However, such transcript-affecting mechanisms are considered to constitute approximately a tenth of the genetic signals identified by GWAS, so other mechanisms must be contributing to the GWAS signals [76].

For the scope of this thesis, I will focus on variation within REs, which are considered a major mechanism of GWAS signals. This is due to the enriched coincidence between the genomic distribution of GWAS haplotypes and the position of distal REs.

Once focused on likely regulatory causal haplotypes, after discounting transcript-centric mechanisms, a similar but more refined list of interlinked challenges to solve GWAS signals emerge.

- In which cell type(s) are the haplotypes exerting their effect and so should act as the model system for their analysis?
- Which SNP(s) in the haplotype is causal?
- By which mechanism is it changing gene expression? What class of RE does it reside in?
- Which gene(s) is it changing the expression of?
- What is the direction and magnitude of the change?

This stepwise approach can be used to determine which SNP(s) has the potential to be causal and affect gene expression.

Intersection of haplotypes with markers of regulatory activity to prioritise SNPs

Since the functionality of REs is determined by sequences that allow the binding of TFs, it can be assumed that SNPs that alter these sites can also alter the activity of the element. A reasonable first assumption would therefore be that casual SNPs should lie within regions marked with signals

associated with TF binding. Although TF binding can be determined by ChIP-seq-based approaches, each TF active in the cell type must be individually tested to assess the full scope of TF binding. Due to both our incomplete knowledge of true repertoire of TFs active in any given cell type and incomplete tools (i.e. antibodies) for even the known TFs, this becomes a herculean task for even a single cell type. Therefore, assays which detect general properties associated with regulatory function, in particular open chromatin assays, are used as an initial proxy for TF binding and regulatory potential, to prioritise SNPs that lie in active regulatory regions in a cell type.

Historically, DNase I hypersensitive sites sequencing (DNase-seq) has been used to find open chromatin regions, as the enzyme can more efficiently cut active, accessible nucleosome-depleted DNA [77]. While technically difficult, when deeply sequenced this technique has the additional ability to detect more precise regions within the open chromatin peaks that are protected from the DNase I enzyme by bound TFs. These TF "footprints" can be highly informative in determining if a specific SNP lies within a region directly bound by TFs and can even be used allelically to determine if the footprint and TF binding is affected by a given SNP [78, 79]. More recently, open chromatin has been identified using ATAC-seq, which utilises a mutated Tn5 transposase that is hyperactive in open chromatin regions, similar to DNase I [80, 81] (Figure 1.4). An advantage of ATAC-seq is that it simultaneously incorporates sequencing adapters onto the fragments, thus producing an amplifiable sequencing library in one step. ATAC-seq is therefore a quicker and easier method than DNase-seq, but has the disadvantage that TF footprints are poorly detectable in ATAC-seq.

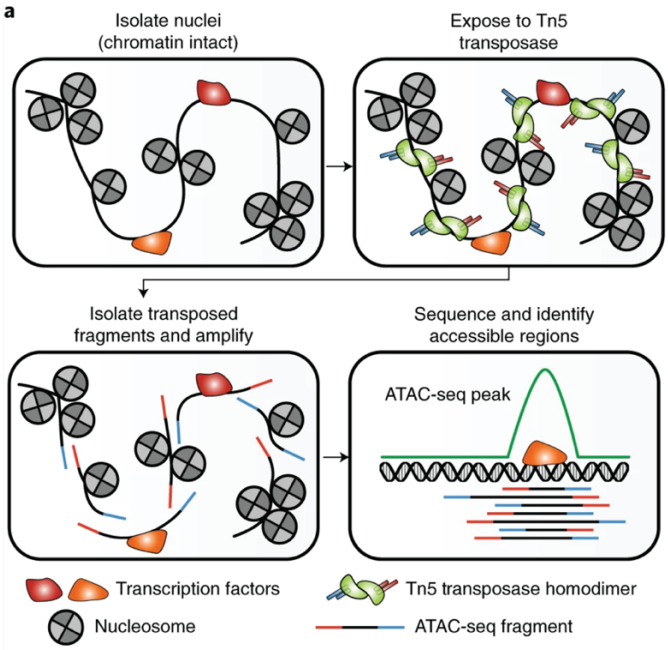


Figure 1.4: ATAC-seq is able to assay open chromatin with the use of a modified transposase. Chromatin can be fragmented and tagged at regions depleted of nucleosomes. A modified transposase is able to simultaneously fragment and tag DNA to form a library of open chromatin. These can be isolated and sequenced and once aligned to a genome will indicate genome wide accessibility. Adapted from Grandi *et al*. [82]

As both DNase-seq and ATAC-seq are sequencing-based approaches, the difference in open chromatin signals in individuals who are heterozygous for a SNP of interest can be bioinformatically ascertained [83]. This "allelic skew" can be used to validate the ability of a suspected haplotype to change the activity of a RE and can be generally applied to any sequence-based assay, in particular ChIP-seq. Furthermore, as the signals from both alleles will have come from the same assay, this removes any potential technical bias if mapping bias is accounted for. However, due to the difficulty of calling sequence variants from these data, it is most effectively applied when the individual's whole genome sequence has also been sequenced.

ChIP-seq provides valuable orthogonal data to the confirmation of causality and the understanding of the biology. ChIP-seq for chromatin marks (H3K4me1, H3K4me3 and H3K27ac) in addition to factors such as CTCF can assign a likely general function to the open chromatin site in question; primarily whether the RE is likely an enhancer, promoter, or boundary element [9, 10]. This is key to forming a reasonable hypothesis for how a variant could mediate a functional effect and guide downstream experiments. Similar to open chromatin, allelic skew analysis can also be used to help confirm causality

through genotype-specific changes in activity or binding [84]. ChIP-seq for TFs known to be active in a cell type can further characterise the *trans*-acting repertoire that act on the REs. This is of particular importance for enhancer REs, which are enriched for TF binding. Such methods can be used to try and determine which specific sequence motifs may be disrupted by SNPs within a haplotype, helping to support the causality of the sequence change. However, with over 1,600 predicted TFs in the human genome [85], the majority of which are poorly characterised, this is always an incomplete analysis and SNPs may affect functional motifs that cannot yet be interpreted. This limitation of ChIP-seq puts into context the potential of approaches such as DNase-seq footprint analysis, which can detect TF binding and changes in TF binding without knowledge of the TF involved.

Identification of target cell types for non-coding genetic variation

The effectiveness of the above approaches is highly dependent on having the correct cell type to analyse. Unfortunately, GWAS-enriched signals provide no indication of the target cell. Other approaches have been developed to identify likely target cell types from genetic signals.

Statistical models of enrichment

The first approach is statistical and based on the premise that a large majority of GWAS signals represent regulatory variants and the observation that the repertoire of active REs in a given cell type is highly specific. Therefore, the open chromatin maps of cell types which are the most relevant for a given trait should have a greater intersection with the GWAS signals of that trait; in essence, more SNPs in the haplotypes for the trait should lie within open chromatin sites in that cell type. This approach has been formalised within computational packages such as Chromatin Element Enrichment Ranking by Specificity (CHEERS) from the Trynka group [86], and the currently unpublished Cell Type Prioritisation Pipeline from the Hughes group. Such approaches are dependent on having maps of open chromatin sites from a comprehensive collection of human primary cell types, both to test each cell type against the genetics, but also to generate a well-sampled background of what represents a by-chance degree of intersection to judge enrichment against. Fortunately such approaches are becoming highly empowered by initiatives such as CATlas[87] and Descartes[88], which are progressively sampling primary human cell open chromatin landscapes using scATAC-seq approaches.

Neural network machine learning predictive models

While becoming increasingly powerful at identifying cell types enriched for GWAS signals, the above approach has two major limitations. Firstly, as it works based on an overall enrichment approach, it cannot confirm the causality of any SNP within a RE in that cell type. Secondly, while there exists a statistical enrichment of the whole data set's intersection with the whole genetics, it would be insensitive to subsets of haplotypes within the genetics set that work mechanistically in a completely different cell type to the majority of the genetics. This has fuelled interest in using neural network-based machine learning approaches that have the potential for overcoming these problems. Frequently, these approaches are based on a convolution neural network (CNN), which essentially learn to recognise features by the progressive adjustment of statistical weights within its network over many rounds of training.

The training is based on large feature sets of manually curated examples, and training progression is judged based on subsets of these manually curated examples that are withheld from the training set. The human genome is a great substrate for such networks as its functionality is encoded by patterns of a simple 4 letter code and modern genomics methods can produce millions of examples to train on. For example, a CNN can be trained to recognise the complex sequence patterns associated with open chromatin signals by showing it many examples of the sequence of open chromatin sites in a cell type and testing reliability using the sequence of sites that have been held out of the training [89].Typically, this is done simultaneously across hundreds of cell types and tens of thousands of open chromatin sites for each cell type, so that the network learns cell type specificity. As these networks learn from sequence, when judged accurate enough, they can be used to prioritise likelihood of causality of a specific SNP in any cell type as long as it was including in the training regime. In practice the prediction for wildtype sequence for an open chromatin site is compared with the prediction for the same sequence including the SNP of interest to assess any changes in prediction probability. These approaches have been developed relatively recently so they still require systematic validation, using ground truth causal examples across many cell types; however, such examples are sparse at the moment. Nonetheless, they can be powerful in forming hypotheses and guiding functional validation

of a predicted causal SNP and target cell type, such as its recent use to identify both the causal SNP in a haplotype associated with COVID mortality and the target cell type as lung epithelial cells [90].

Identification of effector genes linked to regulatory variation

The above approaches have been developed to address the key challenges of identifying and validating the cell types affected by causal regulatory variants within the haplotypes of a human trait. However, there remains the critical challenge of finding the gene(s) affected by this regulatory variation, imposed by the unpredictable linear relationship between REs and their target genes in the genome. This is challenging because many genes are controlled by several distal REs which can sometimes lie Mbs away in the linear sequence and even within the introns of unrelated genes. So, linking the gene(s) and REs based simply on linear proximity is highly unreliable.

A key feature of the biology of gene regulation can provide a potential solution to this problem. Namely that for an enhancer to exert its affect, it needs to frequently come into close proximity to the promoter it controls [7]. Chromosome conformation capture (3C) proximity-ligation assays, which can measure these events, can be used to link REs and genes. Although they are based on the same underlying principle, their ability to link genes and enhancers is highly dependent on the resolution of the assay used. In the context of GWAS, they are further limited by the throughput of the assay. A typical GWAS generates a large number of haplotypes that need to be interrogated and different 3C assays can assess a varying number of enhancers and genes in a single assay. These assays generally utilise chemically fixed material to generate snapshots of the 3D structure of the genome within a population of purified cells. The chromatin in the nuclei of these cells is then fragmented with endonucleases, typically a restriction endonuclease, and then after inactivation of the endonuclease, ligated back together using a DNA ligase. In this way the order of the fragments now represents their proximity in the nucleus rather than their original order in the linear genome (Figure 1.5). The chromatin is then de-crosslinked, and the DNA purified and used as the input for downstream analysis to detect the newly formed DNA junctions. Initially analysed using quantitative PCR, the read out is now typically sequencing based. Modern versions of the assay are more quantitative and highly multiplexed but come at the cost of complexity. Each nuclei represents a highly complex reordering

of the genome, which when combined with a 3.3 billion bp genome across hundreds of thousands or even millions of cells, means the output cannot just be simply sequenced but must be enriched.

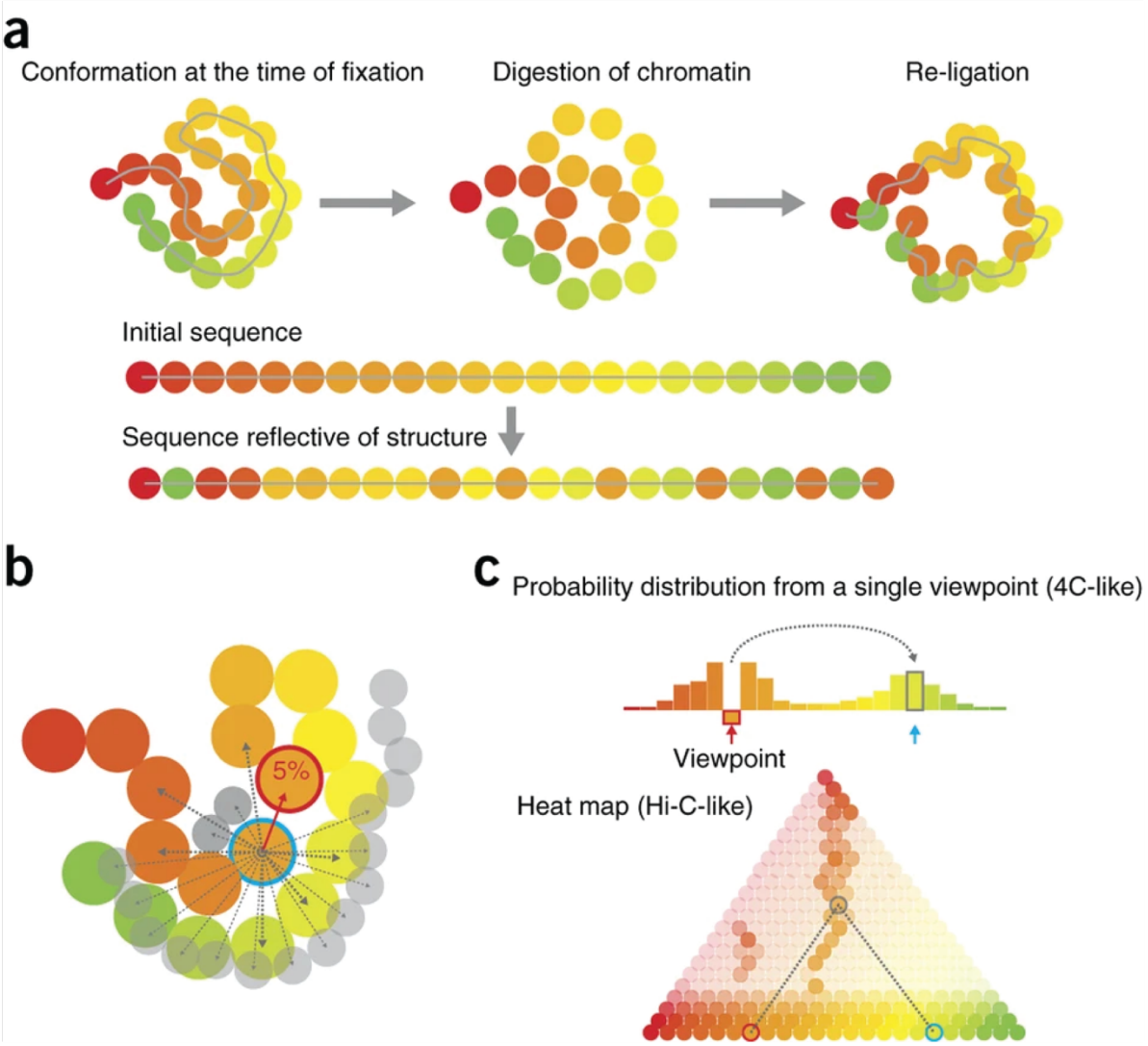


Figure 1.5: 3C methods can determine the 3D structure of the genome. a) Linear DNA is fixed, digested and religated. The new concatemer sequence will now reflect the 3D interactions. b) These interactions provide a snapshot of DNA interactions with few (~5%) reads ligating back to original proximal fragment. Each cut site in every cell has at least 20 potential reconnection partners both proximal (coloured fragments) and more distal (grey fragments). c) Depending on the technique, determined contact frequencies can be visualised from a one-vs-all viewpoint or as an all-vs-all heat map. (Adapted from Davies *et al*. [91])

3C assays

Multiple variants of these assays have been developed, from the original 3C [92] to its derivatives 4C [93], Hi-C [94], and Capture C [6]. Hi-C is a genome-wide assay where enrichment comes from the inclusion of a biotin between the ligated fragments so that the informative ligation junctions can be enriched before sequencing using a streptavidin pulldown. However, even with this enrichment these full genome data sets are of too low a resolution to reliably link enhancers and promoters, and instead give an overview of TADs and compartments. Adaptations of this protocol, such as Promoter Capture Hi-C [95], further enrich for interactions from promoter REs using RNA probe capture. However, these techniques are still mostly limited to low resolution analysis due to low levels of enrichment and large fragment sizes.

The Hughes and Davies laboratories have developed several Capture C technologies specifically to address the problem of GWAS interpretation [6, 96-98]. They combine high levels of enrichment using biotinylated DNA oligo capture with the use of frequent cutting restriction enzymes to generate high resolution data. This can then be combined with massive multiplexing of viewpoints. However, there are still some limitations with its resolution in certain circumstances, and the identification of statistically enriched interactions is highly computationally complex.

Micro Capture C (MCC)

MCC has recently been developed by the Davies Lab and builds upon the foundations of Capture C to improve data resolution [99, 100]. By utilising the non-sequence-specific cutting of micrococcal nuclease (MNase) instead of restriction enzymes, MCC vastly increases the resolution of interaction profiles generated. Instead of capturing sequence-dependant restriction fragments with an average size of 265bp [101], MNase fragments the genome into nucleosome-associated DNA and does not have a specific recognition sequence; this leads to much higher resolution data (Figure 1.6). When combined with the enrichment method used in Capture C, it retains the high levels of enrichment and signal-to-noise levels and is similarly massively multiplexed. This means that MCC can precisely analyse the regulatory interactions from thousands of enhancers or promoters at base pair resolution in a single assay. Combined with its superior signal-to-noise, MCC can conclusively detect these

interactions over any distance and can differentially detect interactions between elements within a few hundred bp from each other in the genome.

MCC also brings several unique but substantial advantages for the analysis of regulatory variation. Firstly, unlike restriction enzyme-based methods, MCC is not limited to positioning capture probes proximal to restriction enzyme sites, which are typically away from the element under analysis. Instead, MCC probes can be designed at an optimum position within the open chromatin site or target element of interest (Figure 1.6). Secondly, when performed at a site containing a heterozygous variant of interest in a genotyped individual, sequence reads from each allele can be split during the analysis to generate allele-specific interaction profiles from the two chromosomes within the same cells. And finally, the use of MNase, which similarly to DNase I, is blocked by bound TFs, means that TF footprint data can simultaneously be generated. This in turn can be analysed allelically to detect haplotype-specific changes in TF binding across different alleles. The combination of these MCC-specific properties makes it uniquely powerful for the analysis of regulatory variants, with changes in footprint data supporting causality of the variants. Additionally, allele-specific interaction profiles provide unparalleled insights into the gain or loss of regulatory interactions in the same assay.

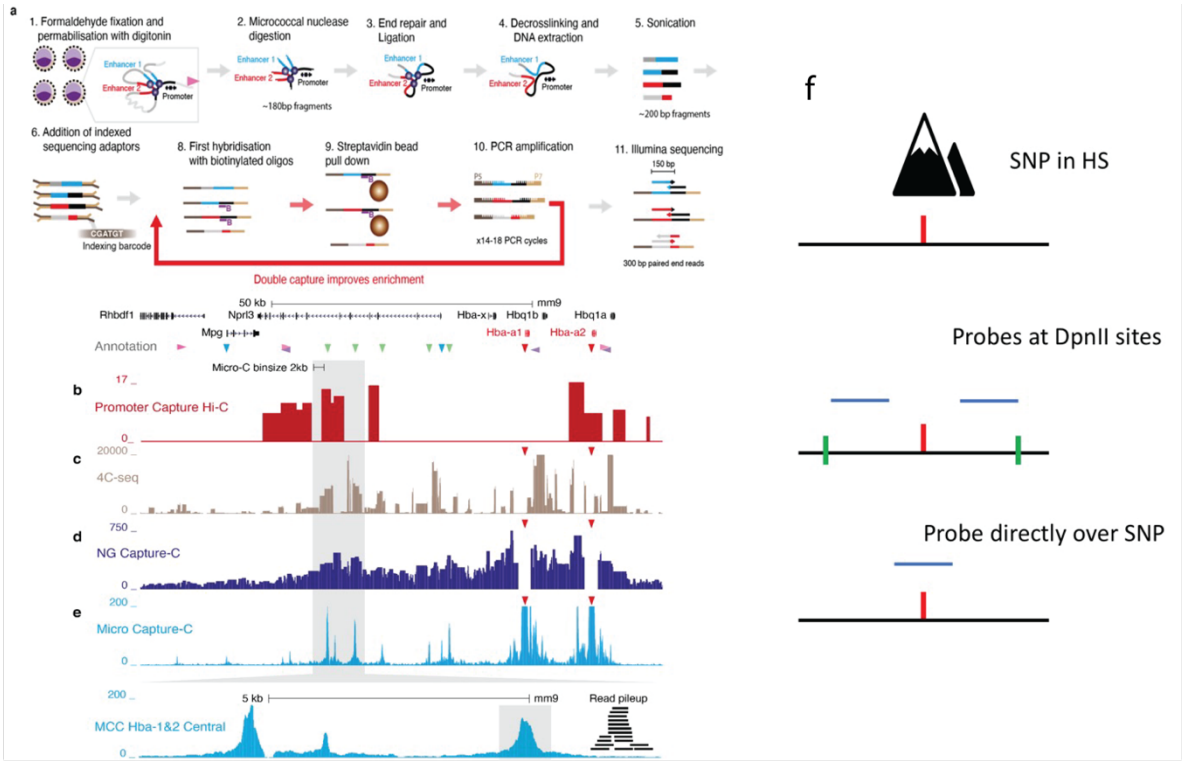


Figure 1.6: Work flow of MCC. a) Overview of the MCC process: Cells are formaldehyde fixed and digested with MNase, ligated, decrosslinked, sonicated, adapters are added and double capture enriches for regions of interest. Compared to other techniques such as b) Promoter Capture Hi-C, c) 4C-seq and d) NG Capture-C, e) MCC produces much higher resolution interaction profiles that precisely overlap with REs. f) When studying variants and designing capture probes, MCC allows for the design of probes directly over SNVs or SNPs. This is because of the MNase which is contrary to 3C that rely of restriction enzymes, at which cut sites determine where probes must be designed to. This means you can design over specific regions or SNPs of interest. (figure adapted from Hua *et al*. 2020 [99]).

Functional validation of the effect of regulatory variation on gene expression

Allelic analysis of open chromatin, ChIP-seq, and even 3C interactions can provide evidence that a SNP in a given haplotype is affecting the activity of a RE or its interactions with gene promoters; however, it does not ultimately prove a SNP has changed the expression of a target gene. While it is strongly suggestive, it is known that many elements that have similar epigenetic profiles to enhancer REs and even interact with promoters in 3D space have little to no effect on gene expression when removed by genome editing [102, 103]. Perhaps one of the best characterised models in this context is the

murine alpha globin locus, which contains five enhancer-like elements that interact with the alpha globin promoters. However, when these enhancer-like elements are deleted in mouse models, only two show substantial effect on transcription [102]. Interestingly, the same research group recently defined the remaining three elements as 'facilitating enhancers', which have a supportive but redundant role in gene expression [103]. Therefore, a variant affecting the activity of one of these redundant REs is likely to be less involved in a human trait compared with a variant affecting a classical, strong enhancer RE. It is therefore important to validate a detectable change in expression in the presence of the variant and there are two main approaches to do so.

Genome editing

The first approach is genome editing of the enhancer RE. Since the deletion of an entire enhancer does not prove the causality of the SNP within it, and instead merely proves that the whole RE was important, more precise edits are needed. Therefore, homologous directed repair (HDR) editing should be used to either add or remove the SNP in the appropriate cell type without altering the adjacent sequence [104, 105]. Although this should optimally be performed in primary cells due the highly specific nature of gene regulation, HDR is difficult to achieve in primary cells. The use of CRISPR editing to stimulate HDR of a template fragment with the required edit incorporated can improve efficiency but are still extremely challenging. Therefore, HDR editing is mostly limited to cell lines. However, due to cellular differences and even large-scale genomic alterations incurred during immortalisation and culturing, it cannot be assumed that the regulation of a locus in a primary cell type is precisely replicated in a model cell line [106, 107]. Similarly, the clone-to-clone variation after editing can cause substantially general changes in gene expression, which limits the sensitivity with which gene expression changes can be detected. It is therefore preferable, where possible, to examine gene expression changes linked to regulatory variation in primary cells. Even though next generation editing approaches, such as base editing, are becoming more efficient, they still cause major alterations to the cell state via the experimental processes, such as transfection or lipofection, used to generate the edit and are still limited in the range of edits they can produce [108].

Analysis of heterozygous individuals

An alternative approach is to again take advantage of heterozygous individuals. This is done by assessing the allelic skew in gene expression, similar as discussed in the context of open chromatin, ChIP-seq and MCC, to detect genotype-specific changes in gene expression. This has the advantage of analysing expression effects in primary cells and in the same sample; comparing the activity of one chromosome to the other in the same cell. In addition, this analysis minimises experimental noise caused by editing or batch effects. Finally, it allows for allelic skew to be detected genome-wide across many loci in a single assay. However, this does dictate several aspects of the overall experimental design.

Firstly, as the effects of REs on active transcription act almost exclusively at the nascent transcription stage, the RNA-seq method used should detect as much nascent transcription as possible. However, many nascent transcription methods, such as transient transcriptome sequencing (TT-seq) [109] and global run-on sequencing (GRO-seq) [110], are typically laborious and demanding of cell numbers and so unsuited to primary cells. An alternative is to use RNA sequencing (RNA-seq) methods that retain large amounts of intron signal, such as non-polyadenylated (PolyA-) RNA-seq, which uses total RNA that has been depleted of both polyadenylated (PolyA+) and ribosomal RNA. PolyA- RNA-seq only requires standard RNA extraction and library preparation [111]. This also has the advantage that intronic regions in genes contain many more SNPs that can be used to rigorously determine allelic skew in expression.

Secondly, the causal haplotype is frequently not in the gene that is affected and can be millions of bases away, and to be able to analyse any RNA-seq data, there must first be a genotype to align the sequencing reads to. Due to short-read technologies, standard genome sequence output is incapable of determining which SNPs are on the same chromosome beyond ~200bp apart. Therefore, the whole genome sequencing (WGS) method used to determine the genotype of the individual should be capable of phasing the genome sequence over distances that span large regulatory domains to link the variant under analysis with polymorphisms in the gene to be tested. While substantially more expensive than standard genome sequencing, the platforms for this are expanding, each with their own pros and cons. For example, Nanopore sequencing can sequence extremely long fragments to

phase SNPs, but sequence quality has traditionally been poor for calling variation [112]. PACBIO platforms have superior read quality but shorter fragment size, resulting in shorter regions of phased genome [113]. Until recently, the 10x linked-read platform provided an attractive combination of length of phasing, sequence quality, and cost [114]. The technology works with standard Illumina sequencing platforms and so has high sequence quality but uses the 10x gel beads-in-emulsion (GEM) library generation technology to specifically barcode >100kb stretches of input DNA. The high complexity of the barcode design means that most of the reads, and therefore the variants detected in them, can be confidently assigned to the same chromosome over Mb distances.

Thirdly, there are many reasons, such as imprinting, that may cause a strongly skewed expression of a transcript in a single individual's cells [115, 116]. Therefore, it is key to show that the coherence between the presence of a causal haplotype and gene expression skew on the same chromosome is reproducible and robust. To this end, it is important to perform the above-described analysis on enough individuals to show that the effect is consistent and reproducible. This thesis will depend on an in-house biobank of 50 individuals, from which all have deep, phased 10x linked WGS. Furthermore, the erythroid genomic landscape of these individuals has been characterised with ATAC-seq, ChIP-seq, and RNA-seq, and erythropoiesis will be used as a model system for this thesis.

Erythropoiesis as a cell model

Erythrocytes are the most abundant cell type in adult blood, and with a concentration of ~$5x10^6$/ml, they account for around 40-45% of blood volume. Their distinctive red colour is produced by the oxygen carrying haemoglobin that makes up ~35% of a cell's weight, a tetramer of alpha globin, beta globin, and associated haem groups. A terminally differentiated erythrocyte in circulation has an average life span of only 120 days, and so a high production and turnover is required to maintain enough cells for efficient oxygen transport; an adult human will produce ~$2.4x10^6$ new erythrocytes each second [117]. The specific form of these cells varies throughout development, with embryonic and adult erythrocytes utilising structurally different forms of haemoglobin [118]. Despite these differences, they share a common differentiation lineage, and their utilisation of different haemoglobin subunits is instead a result of different gene regulation and expression [119]. The blood

cell lineage starts with haematopoietic stem cells (HSCs) (Figure 1.7): multipotent cells that can develop into all types of blood cells, which originate from endothelial cells that retain haemopoietic potential [120, 121]. Erythropoiesis can be broken down into several distinct stages, starting with HSCs and ending with terminally differentiated erythrocytes; the burst forming unit-erythroid (BFU-E) cells mark the first stage of erythroid-specific differentiation (Figure 1.7).

In an adult human, the bone marrow is responsible for the genesis of the vast majority of erythrocytes, with the spleen providing extra support in times of hypoxic or anaemic stress. The bone marrow is a complex niche of vascular tissue, osteoblasts, and stromal cells that support the HSCs and growth of blood cells by exposing developing cells to growth factors and cytokines. Erythropoiesis is promoted at erythroblastic islands that are composed of a central macrophage nurse cell and the surrounding developing erythrocytes at differing stages. The supporting macrophage ensures the early provision of nutrients to the cells, as well as the required proliferative and survival signals. During this differentiation, BFU-E cells progress to colony forming unit-erythroid (CFU-E) cells, before undergoing four to five rapid cell divisions, advancing through the basophilic, polychromatic, and orthochromatic erythroblast stages. These divisions are marked by the substantial accumulation of haemoglobin and the condensation of nuclear structures, as well as an expansion of cell number and a decrease in cell size. The resultant reticulocyte cell enucleates, with the nurse macrophage phagocytosing the ejected nuclei, before entering the blood circulation where it will fully mature into a terminally differentiated erythrocyte over the course of ~1 week [122].

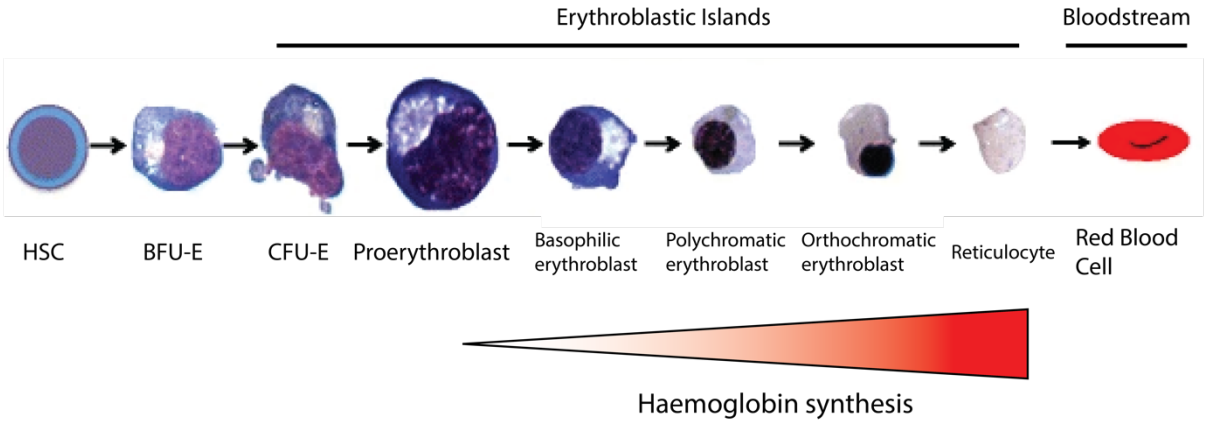


Figure 1.7: Erythropoiesis results in the formation of mature red blood cells from HSCs. HSCs mark the start of erythropoiesis with BFU-Es marking the first erythroid specific cell stage. Cells then progress through several cell stages in the erythroblastic islands before finalising maturation in the bloodstream (Adapted from Zivot *et al*. [123])

HSCs are also found at low concentrations in circulating blood, where they are thought to have a role in vascular maintenance and neoangiogenesis [124]. HSCs are marked by CD34, a transmembrane glycoprotein that is present on early cells then lost as they differentiate [125]. It has been shown that the rate at which HSCs are released into circulation varies throughout the day, with a 6-fold higher yield of CD34+ cells extracted during the daytime and late afternoon, compared with a lower yield during the night [126].

Because these early HSCs can be extracted from adult peripheral blood, several model systems have been developed for the *ex vitro* differentiation of erythrocytes [127-133] using supplementation with key factors in the culture system, such as stem cell factor (SCF), interleukin-3 (IL-3), and erythropoietin (EPO), to drive expansion and differentiation [134].

Utilising erythropoiesis to study genomics has several advantages. These are listed below.

- Ease of acquisition: unlike many other adult tissues, blood is very easily accessible via a minimally invasive procedure. Fortunately, despite most blood maturation being localised within the bone marrow, there are enough circulating HSCs in adult blood to start a successful culture. By using common blood separation techniques and magnetic bead selection columns, ample CD34+ cells can be extracted from 50-100ml of blood.

- Cell number: as many genomic assays require millions of cells, often with several replicates or different conditions such as ChIP-seq antibodies, having large numbers of cells allows for the acquisition of multiple data sets from the same differentiation. The natural expansion potential of erythrocytes *in vivo* is extremely high and this is also reflected in their *ex vivo* culture, with successful differentiations yielding tens of millions of terminally differentiated cells. The obvious caveat of studying genomics through erythropoiesis is the enucleation prior to full maturation, but since this step happens in the late reticulocyte stage, all erythroblasts and prior progenitors are suitable for study, and are characterised by extensive programmed gene regulatory changes.
- Verified differentiation system: many examples of *ex vivo* differentiation systems exist, and reproducibility is essential; the Scott *et al*. [127]protocol has been verified and reliably produces terminally differentiated erythrocytes. When differentiating and expanding cells to acquire sufficient cell numbers for data generation, confidence that the populations taken at certain stages will be comparable across differentiations and healthy individuals is imperative. This confidence allows for the subsequent comparison of data sets and builds a greater epigenetic profile of the model system.
- Well-documented loci: in order to validate the reproducibility of the system, certain genomic regions with well-defined profiles and activity can be used as controls. In erythropoiesis, the loci of the alpha and beta globin genes are exemplary as they are highly upregulated and are uniquely well characterised in terms of their regulatory elements. Due to the robustness of the globin genes and their regulatory landscapes through erythropoiesis, the extensive and thorough study of these loci have contributed greatly to the field of gene regulation [102, 103, 135, 136].

Thesis aims

Several studies have demonstrated the general enrichment of common genetics in regulatory elements and the mechanistic effect of loss of function GWAS variants in several cell models [11, 43, 137-142]. I therefore set out to explore the prevalence and patterns of variation within a specified cell type across multiple normal individuals. The main aims will be explored over 3 results chapters.

First results chapter (Chapter 2): Our initial aim was to use a large sampling of natural variation (50 individuals) to understand the genetic variability of gene regulation in the normal population in a single cell type. Subsequently, we aimed to then leverage the large GWAS datasets available for variation in erythroid parameters in the human population, to understand the links between this regulatory variation and human genetics. We have used haplotype analysis to highlight the existence and prevalence of a poorly considered mechanism, where regulatory activity exists only on a single haplotype background which we refer to as gain-of-function.

Second results chapter (Chapter 3): The overarching aim is to begin to contribute to a generalisable systematic approach to decode human non-coding genetics. To do this we have used the fact that the samples we have used are from individuals which have had their genomes well sequenced and phased. In this chapter I will optimise the cutting-edge 3C method Micro Capture-C to work in primary late erythroid cells and adapt it to generate allele specific interactions using individuals heterozygous for variants of interest. This approach will be integrated into a generalisable allele-specific approach that combines the allele-specific Micro Capture-C data with allele-specific open chromatin and transcription to confirm the causality, mechanisms of action and effector genes.

Third results chapter (Chapter 4): The previous chapters investigated regulatory variation and genetics in a single late stage of erythroid differentiation. However, it seemed feasible that other GWAS variants act at earlier stages of differentiation to affect the physiological parameters of the latter stage of differentiation. In this chapter we perform a pilot study on two individuals, using single cell multiome analysis to capture regulatory variation across differentiation to test this concept.

Chapter 2 - Utilising epigenetic data from across erythroid differentiation to find and define variants within a human population

Introduction

Genetic variation drives diversity within a species and is shaped by both genetic drift and selection [143]. Though often silent, variants that impact genes or regulatory networks can have profound effects either via the action of a single variant or via polygenic traits [11, 44]. Determining which variants are functional, along with the effect they impart, remains challenging, but can be achieved with a systematic combined computational and experimental analytic approach.

Using a variety of assays, we are able to study the structure and function of the genome and epigenome. WGS ascertains the genomic sequence, including variants, of an individual. Epigenetic assays such as ATAC-seq reveal open chromatin structure, while ChIP-seq detects protein binding and histone modifications. Different RNA-seq approaches can quantitatively capture both mature and nascent RNA as an indication of transcription levels and splicing for isoform structure. A wide variety of 3C methods detect physical interactions between chromatin domains and TAD boundaries. At high resolution 3C assays can also detect the physical association between enhancers, promoters and boundary elements, linking enhancer and gene activity.

By obtaining and overlaying these data sets from the same cell type from characterised individuals, these assays become much more informative. The combination of such assays allows for the coherent analysis of genotype specific changes in epigenetic activity, 3C interactions and transcriptional output. It is also well documented that different cell types exhibit characteristic epigenetic profiles, depending

on their required gene expression profiles, and so choosing the right cell type to study the effects of regulatory variation is crucial [144].

The model system used throughout this project is of human erythroid differentiation which generates polychromatic erythroblast cells (figure 1.7). This model takes $CD34^+$ HSCs from adult human blood, and using a combination of factors, differentiates these cells down the erythroid lineage [127]. This system has several advantages: compared to other human tissues, blood is easily accessible; the differentiation protocol to generate mature erythrocytes is well established; the gene expression profile of certain genes, such as the globin genes, are specific and well documented and act as valuable controls; and the later stages can yield millions of cells for downstream analysis. Combined, these make the $CD34^+$ system an excellent model for studying human variation and its effect on gene regulation during the differentiation process.

Prioritising likely causal regulatory variants by mechanism of action.

Variation within the genome can alter gene function via several mechanisms. When variants reside in coding regions the effects are often easier to interpret than those in REs. The codon lexicon is well established and easily readable, and an altered transcript can have a profound impact on protein structure and function [62, 63]. However, with modern advances in genomic assays, causal variants outside of the coding regions have also been discovered and are shown to be prevalent in common human disease genetics [73]. Approaches to understand their mechanisms are being developed and they are now thought to act predominately through effecting gene regulation. Due to the complexity of gene regulation, there are multiple ways a variant can affect gene expression. However, broadly speaking as the DNA sequence underlying an element contains protein binding motifs that determine the interacting transcription factors, changing these recognition sites can in turn change what binds, and therefore the activity at this site.

Loss of function variants, where a recognition site in a defined peak is damaged by a base change preventing or reducing a TFs ability to bind, are currently the most widely recognised and studied[145]. This is largely because, while examining epigenetic profiles, a SNP residing in an open chromatin site is a strong indicator of causality, and so it can be prioritised for investigation. However,

variants can also have a gain of function effect, whereby a recognition site is created or altered, so that protein binding is enabled or enhanced, which in turn can alter gene expression [90, 146]. When such an event occurs in an existing element, the fact that the variant lies in an open chromatin site will equally prioritise it for investigation.

However, such changes have the potential to create a *de novo* open chromatin site where none previously existed, so that the activity of the region is unique to that haplotype or variant[146, 147]. As such epigenetic prioritisation of variants is generally performed using generic sets of open chromatin data from a handful of individuals, if these individuals do not contain the correct haplotype, no open chromatin site will be detected in the region of the variant. This will therefore mean such variants are inevitably not prioritised as likely causal. When looking for potential causative SNPs, GWAS will often provide many thousands of associated variants after imputation, and so using open chromatin as a guide to which are likely causative is a common first step. However, depending on the genotype of the cells from which the open chromatin data is derived, this may well lead to such gain of function SNPs being systematically disregarded, making this a poorly considered mechanism.

When studying the effects of human variation, there are two main ways to validate their effects. The first involves finding SNPs, usually from a GWAS, and editing these into primary cells or cell lines to see their effect on chromatin architecture, gene expression and protein levels. The other method, which will be used in this project, is to assess in the appropriate cell systems genotype specific variation across multiple individuals. From data of regulatory and transcriptomic signals, those that vary consistently with genotype can be found.

In this project, a cohort of 50 individuals was used to provide a pool of variation to study, using both pre-existing and newly generated ATAC-seq data from these individuals. By searching for differential regions, these were then characterised as either loss or gain of function variants.

Overall aims of the chapter

In this chapter, I will describe the erythroid differentiation system that will enable the reproducible study of genomic variants in polychromatic erythroblasts derived from this cohort. I will make

extensive use of a large dataset generated by myself and other members of the Hughes lab prior to the start of this project, that has characterised several genomic features across 50 local sequenced individuals (individuals from the Oxford Biobank, details in methods). These features include open chromatin (ATAC-seq), active chromatin marks (ChIP-seq for H3K4me1, H3K4me3 and H3K27ac), both stable (PolyA+ selected) and pseudo-nascent/intron retaining (Poly+ and Ribosomal depleted, referred to as PolyA-) transcription and binding of structural proteins (ChIP-seq for CTCF).

From this population, I will show clear examples of regulatory variation, and have attempted to define and characterise these variants by using the haplotypes in which they reside. From this I present and highlight the presence and importance of cryptic gain of function variants, which are currently underappreciated in the field but have the potential to be an important component in the understanding of human evolution and disease.

Results

CD34 differentiation system

In vitro erythroid differentiation starts with the isolation of CD34$^+$ cells via the separation of peripheral blood mononuclear cells (PBMCs) from adult blood by histopaque layering and centrifugation. PBMCs are then washed and CD34$^+$s are isolated with magnetic-activated cell sorting (MACS) beads and magnetic column separation. Once purified, cells can be frozen, assayed or cultured. When culturing, phase one lasts seven days, and culture media is supplemented with SCF and IL-3 to promote proliferation and expansion. The second phase removes the IL-3 to promote differentiation up until day 11, where SCF is removed, and holo-transferrin levels are increased to promote enucleation (Figure 2.1). Genomic assays can be performed at any day prior to full enucleation, with data in this project collected at days 0, 5, 7, 10 and 13. To check the differentiation has progressed as predicted, cytospins can be used to assess cell morphology, and ATAC can be compared with prior differentiations to check open chromatin signatures.

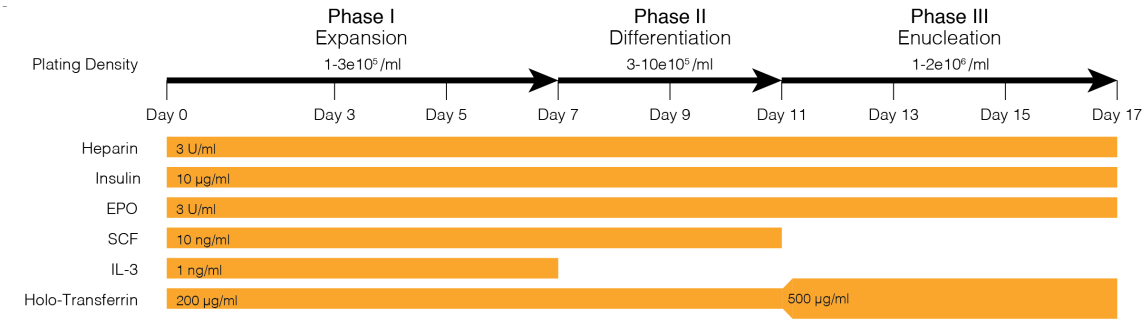


Figure 2.1: CD34$^+$ culture media supplements and staging. Adapted from Scott et.al [127]

Pre-generated biobank data

As stated, for the following variation analysis I have used a large data set that was generated by myself and other members of the Hughes lab prior to starting my DPhil project. This data set is composed of 50 local individuals, 47 of which are from the Oxford Biobank. For all these individuals, erythroid differentiation and genomic assays were performed at day 13 of the differentiation (Table 2.1). The reliability of the erythroid differentiation system in producing late erythroblasts (predominately

polychromatic erythroblasts) after a set number of days in specific media means we can be confident that these data across the individuals reflect the same cell type. All the donors were selected without any selection criteria, and so represent a random sample of the Oxford population. This comes with its own caveats; as a small population of mostly European descendants, this population cannot be expected to reflect all human variation, but instead acts a good workable population with inherent regulatory variation.

Data type		Data sets	Donors
ATAC-seq		50	All
ChIP-Seq	CTCF	50	All
	H3K4me1	5	1, 2, 15, 19, 30
	H3K4me3	5	1, 2, 15, 19, 30
	H3K27Ac	50	All
RNA	Nascent RNA	50	All
	Mature RNA	50	All
10X Whole Genome seq		50	All

Table 2.1 Available data sets for the 50 individuals.

Characterisation of additional donors for optimisation of downstream assays

Central to the planned logistics of this project was the idea of recalling the characterised individuals from the local biobank to generate cells as and when needed. However, due to the COVID-19 pandemic and its disruption, recalling these individuals became challenging. As several of the assays used throughout this project need a large amount of cells for optimisation and set up, it was decided to add two extra donors recruited locally to overcome this logistical block.

To determine the epigenetic profiles of these individuals, designated as donors 54 and 55, and to check the differentiation, I performed ATAC and ChIP-seq on CD34 differentiation-derived samples (Figure 2.2). The ATAC-seq data for donor 54 did not show the expected coverage despite ample sequencing, and was not repeated due to sample limitation; the ChIP-seq data for H3K27ac however matches with both donor 54 and the existing biobank data sets, and so the differentiation can be

assumed to be successful. By amalgamating all the available day 13 data sets from table 2.1 into combined tracks, these provide an extremely deep high quality data set that I use throughout this project. These act as a benchmark to which I can compare newly generated data for reproducibility and to help determine activity at any region in the erythroid genome (Figure 2.2).

Figure 2.2 shows the alpha globin (*HBA*) locus, which houses the two functional paralogues of the alpha globin genes (*HBA* and *HBA2*) which are highly upregulated during erythroid development; these therefore act as a useful control locus in our erythroid system. In this locus, the alpha gene is duplicated, with two copies laying adjacent to each other (locus TAD structure shown in supplementary Figure 1). Upstream of these promoters are four enhancers: R1, R2, R3 and R4. The *HBA* enhancers can be deemed a super enhancer as they are a strong group of putative enhancers in close genomic proximity that also exhibit high levels of mediator binding, further emphasising their importance [148]. This organisation is mostly conserved in the mouse, and knockout (KO) enhancer models of all the enhancer regions have shown that the R1 and R2 are the main drivers of gene expression [102, 103]. Three of these enhancers lie within the introns of the housekeeping gene *NPRL3*. Additionally, strongly bound CTCF sites exist upstream of the enhancer regions, and also adjacent to the most downstream globin gene (*HBA1*). 3C experiments have confirmed these as boundary elements for the locus.

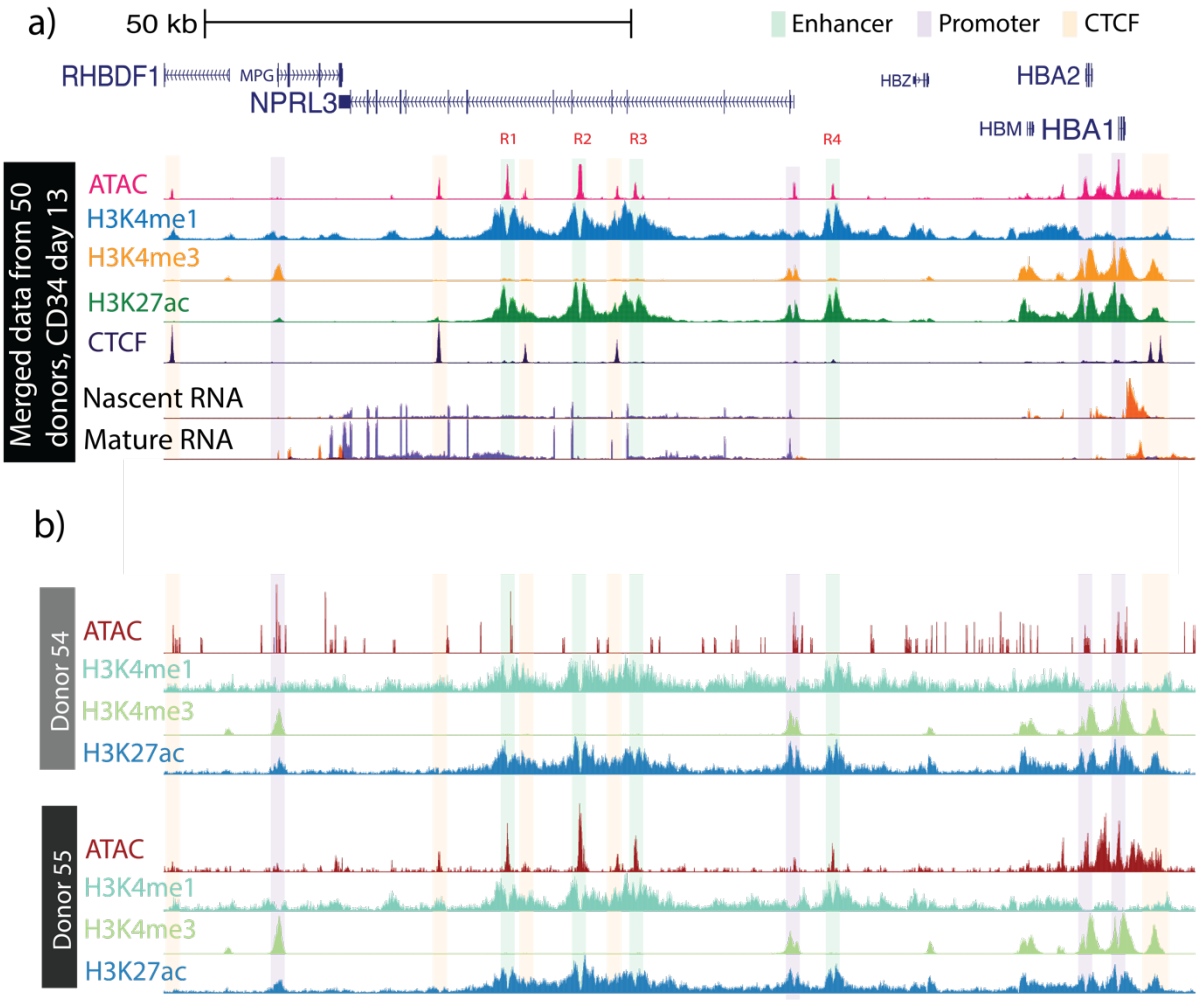


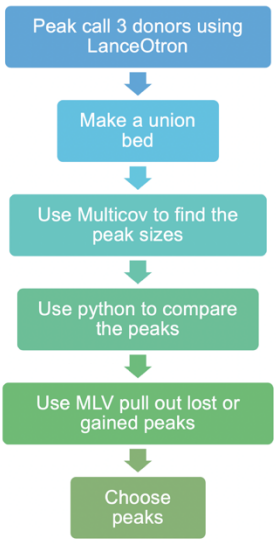
Figure 2.2: Combined epigenetic data from the 50 individuals, and donors 54 and 55. a) Epigenetic profile at the *HBA* locus, track heights auto scaled to data view. Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP-seq from the 5 available sets (Table 2.1). Nascent RNA (PolyA-) and Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. b) Same *HBA* locus with epigenetic data from the newly characterised donors 54 and 55. ATAC (red), H3K4me1 (light blue), H3K4me3 (green) and H3K27ac (dark blue), with the exception of the donor 54 ATAC all show comparable profiles to the merged 50 donor data sets and match the expected landscape at the *HBA* locus.

Comparing ATAC peaks in the donors to find differential regions across individuals

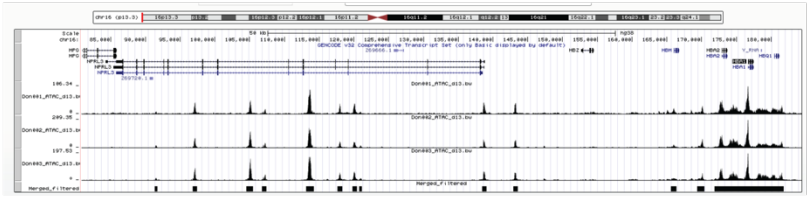
To find differential hypersensitive regions within our biobank donors that may be linked to variants, I compared the ATAC tracks across individuals (Figure 2.3). I first looked at the variation between the

primary three donors (1 to 3), as these were individuals with which I planned to do more extensive MCC and single cell analysis in the following chapters. To do this, I first "peak called" their ATAC bigwig files using the peak calling algorithm LanceOtron[149] and combined the coordinates to generate a union bed file that contained all present ATAC peaks. Next, I used Multicov[150] on the individual samples to find the height of the peaks at all the locations in the union bed file. Using python, I then compared the peak sizes at these regions, pulling out the minimum and maximum peak heights, then subtracted these and generated a percentage difference value. By sorting on this percentage difference, I was able to select the regions with the greatest change between the three individuals. By inputting this data as a bed file into the Multi Locus Viewer (MLV) data analysis platform, I was able to easily sort the data, and view the corresponding ATAC tracks. From this, I could manually go through the regions and curate a list that contained truly differential ATAC peaks.

a)



b)



c)

Chr	Start	Stop	Don001	Don002	Don003	minPeak	maxPeak	peakDiff	%decrease
chr1	10004	10324	531	872	872	531	872	341	39.10550458715596
chr1	11138	11459	318	725	703	318	725	407	56.137931034482754
chr1	28860	29678	583	1722	1774	583	1774	1191	67.13641488162345
chr1	32437	32923	895	1740	1369	895	1740	845	48.56321839080459
chr1	38560	38912	584	847	934	584	934	350	37.473233404710925
chr1	181348	182040	733	1370	1479	733	1479	746	50.439486139283304
chr1	199187	200199	983	2704	2702	983	2704	1721	63.64644970414201
chr1	202955	203691	1293	2030	1883	1293	2030	737	36.30541871921182
chr1	778310	779207	2599	3677	3977	2599	3977	1378	34.64923309026905
chr1	826916	828061	1336	2166	2482	1336	2482	1146	46.17244157937147
chr1	842804	843132	302	670	651	302	670	368	54.92537313432836
chr1	959025	959487	375	859	1017	375	1017	642	63.12684365781711
chr1	1013137	1013756	366	1171	1303	366	1303	937	71.91097467382963
chr1	1019176	1019917	622	1089	1157	622	1157	535	46.24027657735523
chr1	1115834	1116725	1782	4115	4640	1782	4640	2858	61.5948275862069
chr1	1231659	1232525	1319	3144	3261	1319	3261	1942	59.55228457528365
chr1	1273573	1274312	560	921	1149	560	1149	589	51.261966927763275
chr1	1307649	1308835	1190	2417	2824	1190	2824	1634	57.86118980169972
chr1	1324341	1325098	714	1593	1842	714	1842	1128	61.237785016286644

d)

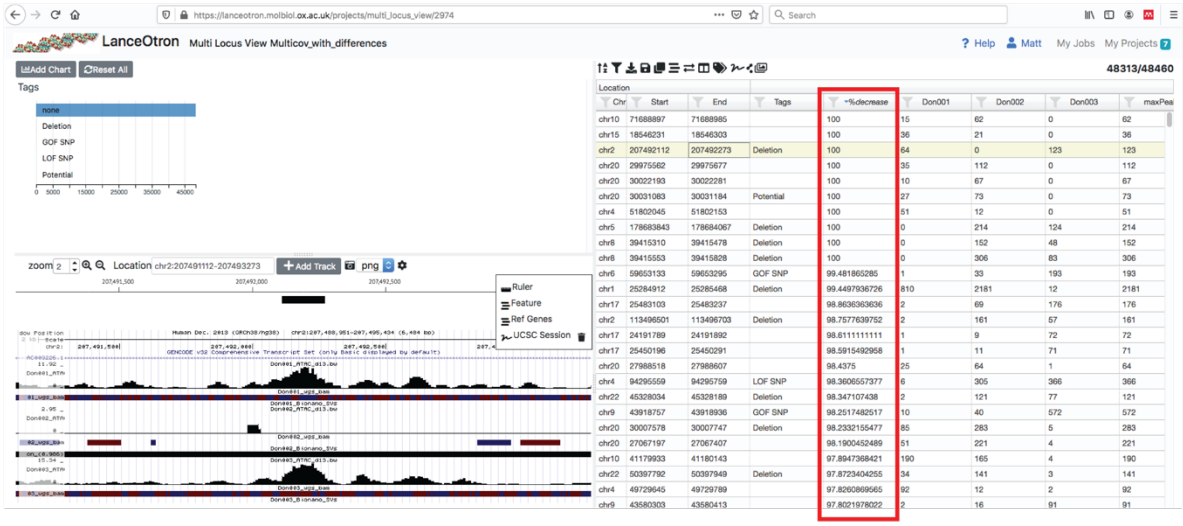


Figure 2.3: Workflow to curate regions of differential open chromatin. a) Workflow shows the steps taken to compare the differential regions within the ATAC tracks of three individuals. b) UCSC browser tracks show example of the ATAC data used, and the results of the union bed of LanceOtron peak calls. c) Table shows the output of the python script composed to determine the difference in peak heights as a sortable percentage. d) This table was then put into MLV for easier visualisation, filtering and sorting of the data. Regions of interest were manually selected from this view.

Combining ATAC and H3K27ac to find differential peaks with associated activity

Although open chromatin is a good initial indication of activity, chromatin marks can be used in conjunction with these data to give a better indication of how active an open chromatin region is in a

particular cell type. As a variant causing change in a highly active open chromatin region is more likely to be having a downstream effect, it is more pertinent to also look for differential chromatin marks associated with differential open chromatin regions. Therefore, to improve the differential peak search, the H3K27ac data was incorporated using search criteria developed by Jeanne Inchauspé in the Hughes group. Broadly, a metric was computed using both ATAC and H3K27ac signals, which when sorted and filtered using the MLV platform could find extreme examples of variation.

This enabled the discovery of regions with a decreased activity, that can hypothesised to be loss of functions (LoFs). Moreover, this analysis also allowed for the detection of unique variants that are only present in single donors, that we refer to as gain of functions (GoFs). Since these GoFs are only present in a single individual, often a single haplotype, it can be assumed that this is a newly created open chromatin site, and not damage to an existing site. This could be achieved by filtering the peaks and isolating the ones that were only present in one donor.

By using this combined analysis, these differential regions are also associated with activity and provide more cases of variation across the genome. Since these regions are specifically active open chromatin, they are more likely to be having an effect on regulatory networks. This initial analysis identified multiple areas with extreme regulatory variation, which are further characterised in both this and subsequent chapters. However, it is difficult from a small sampling to fully characterise them as what we have defined as LoF or GoF variants. To do this, we need to expand our analysis and introduce analysis of the haplotype structures associated with activity.

Determining the frequency of loss and gain of function variants

To determine the frequency at which these LoFs and GoFs were present within our biobank population, meta-analysis was done by Edward Sanders (Figure 2.4). Peak calling of the ATAC data for each donor was undertaken using LanceOtron, and peak confidence scores were assigned to each peak from the algorithm's trained neural network. A union of all the ATAC peaks found in the biobank population was then made (total=39,146), and the individuals compared back to the union to determine which sites were differentially present. Peaks were considered confident with a score of

either <= 0.05 (for absent) or >= 0.95 (for present) by the peak caller. Therefore, confident regions were either confidently present or confidently absent in the donors.

The plot in Figure 2.4 represents the distribution and variability of peaks across the 50 donors. The simplest cases are those that appear on the far righthand side and are consistently called across all 50 genomes tested (total = 12,820). These are regions such as those in the *HBA* locus that do not change across the individuals (Figure 2.5). Unsurprisingly, due to the general conservation of function in the human genome this group contains the greatest number of the total peaks. Moving left from this peak, the next bin represents complete LoF events detected in a single individual. From here there is a sharp drop in frequency, which fits with the necessity for both alleles to be simultaneously inactivated to completely lose a detectable peak. Due to the nature of this analysis, in order for these LoF events to be detected, they must therefore be homozygous. The far most left peaks represent events that would match the proposed pattern for true GoF events, infrequent enough in this population to be found in only one individual (total = 6,616). Importantly, from this analysis these are much more common as a class of events across the population than the loss of a typically constant peak. Since humans are diploid organisms, this observation is likely due to the fact that in this peak based analysis only one allele needs to gain activity to produce a variable peak. From this observation, it can be assumed that these GoFs are likely to be heterozygous, but due to the nature of the analysis, could also represent homozygous peaks. However, to lose a peak on the right hand side of the distribution both alleles need to be inactivated simultaneously to cause a detectable loss of the peak. It was possible that such events were enriched for on the sex chromosomes due to the single copy of the Y chromosome and the inactivation of the X chromosomes in female donors. However, exclusion of variable events on the sex chromosomes makes little difference to the overall pattern seen in Figure 2.4.

Due to the large number of variable regions detected even in a simple peak-based analysis, it is not possible to comprehensively cover each of these events. This endeavour instead represents a large body of work for the Hughes lab going forward. Therefore, I have chosen examples from the above distribution to showcase the different types of events detected and these will be followed up in the subsequent chapters.

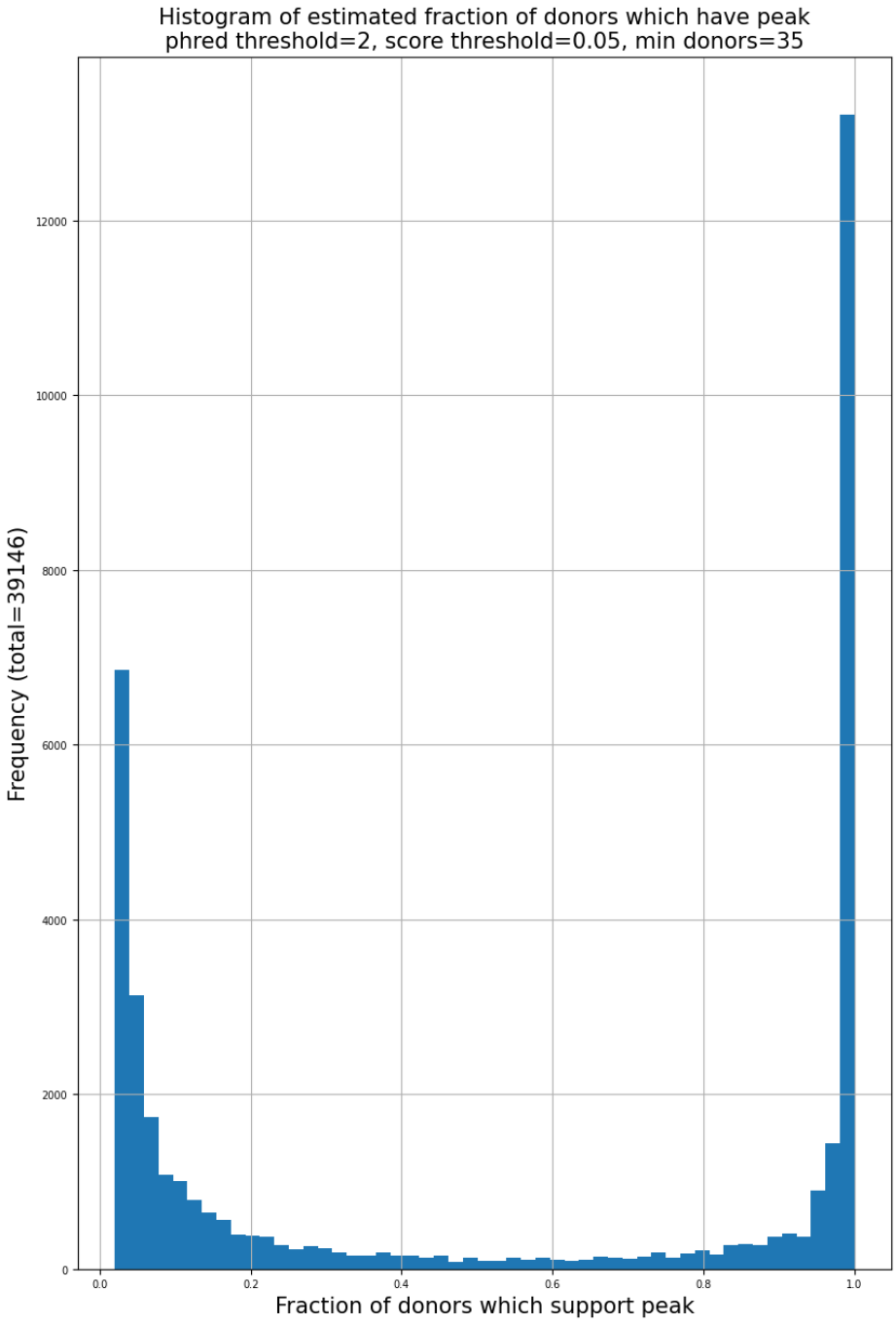


Figure 2.4: Distribution of loss and gain of function peaks within the biobank population. Graph depicts a pile up of peaks against how many donors they are present in. A total of 39,146 confident peaks were detected across 50 donors and used for this analysis. All peaks were called as either confidently present or confidently absent by the LanceOtron training algorithm. Left most distribution are peaks that are confidently present in one individual only and absent in all others, therefore representing a gain of function. Conversely, right most distribution are peaks that are confidently present in all 50 individuals. The middle distribution represents peaks

that are variable across the donors. Peaks to the immediate left of the far most right distribution represent the homozygous loss of a single peak in a single donor, and therefore represent loss of function peaks.

Allelic skew analysis of ATAC-seq data combined with WGS

When defining a variant and its function, it is important to consider the entire genomic context in which it resides. For diploid chromosomes, each allele can carry a different variant, and when viewing genomic data this will generally be an amalgamation of the reads and therefore activity of both alleles. In our examples of variation, it is therefore also important to determine whether these are heterozygous or homozygous, and if heterozygous whether the two different haplotypes are differentially active. Since many variants are silent or lie outside of TF binding motifs, two alleles can have different sequences yet remain equally active. We also know that completely breaking or creating an element are likely to be relatively rare events, and therefore novel examples of these occurring simultaneously on two alleles, as detected in our initial peak-based analysis, has a low probability. It is therefore important to increase the sensitivity of detection of regulatory variation by using an allelic approach.

When analysing sequencing data, reads are aligned to a reference sequence and piled up to form visually accessible tracks. As an individual's genome will not exactly match the reference genome, and because of random sequencing errors, this alignment process must not be too stringent as to not throw away all the data; since each allele will have a unique sequence of variants, this means that variant data is still present in alignment files and can be used to determine which allele a read came from. By applying this logic to ATAC data, when looking at a peak that contains a variant, it is possible to determine which allele a read came from, and therefore which allele was more open than the other. There are several patterns this can take: the alleles under the peaks have the same sequence and are therefore indistinguishable; one allele contains a silent variant that doesn't affect open chromatin and so each behaves similarly; a variant can be damaging to one site and therefore reducing its open chromatin; or a variant is enhancing a regions open chromatin and therefore skewing upwards. One caveat of this method is the risk of over stringent aligners losing data that has not aligned properly, with more reads this would therefore skew the data toward the allele that better matches the reference.

In order to look at this, an analytical pipeline was generated by Edward Sanders of the Hughes group for the analysis of the 50 individuals. This pipeline provided a list of all the regions that show allelic skew in my primary three donors. Allelic skew in the ATAC data was calculated in all donors, and a score generated. These regions were then sorted by their p value and whether they had a SNP that was also found in GWAS. The p value was calculated by comparing the degree of allelic skew in the samples, and so the lower the value, the greater the difference between each allele. Potential for technical skew in the alignment procedure was normalised for using a "Fake FASTQ" based approach. Here, synthetic reads with specific variants were generated and aligned using the same aligners and parameters to detect variance in the ability of the aligners to align reads in the presence of sequence variants.

From this list of statistically significant skewed open chromatin, I was able to select those that were consistently skewed across multiple carriers, that had a GWAS associated variant within the region, as well as those with a high p value but no associated GWAS hits. However, similarly to all the other examples presented in this chapter, the true effect of these variants on the regulatory network in which they reside is currently unknown. I therefore took these presented variants, as well as the new list of allelically skewed regions forward and will present these in the next chapter for analysis with MCC.

The HBA locus demonstrates functional conservation and shows no variation across 50 individuals

This analysis showed that across a sampling of normal individuals a spectrum of variation can be detected, with certain essential enhancers and promoters being present in all healthy individuals, to variants that are only present in a single individual. The *HBA* locus (Figure 2.2) represents a key example of functional conservation in the predominantly European population sampled here. These elements would be represented in the extreme right hand distribution of Figure 2.4. The integrity of this locus can be seen by comparing the individuals and observing that all peaks are present in all profiles, showing that there is no variation within this locus or its elements. Instead, this locus and all its elements are highly conserved both in ATAC-seq and H3K27ac signals across the 50 donors (Figure 2.5).

The elements of the *HBA* locus represent examples of the peaks on the right hand side of Figure 2.4, a highly stable distribution that are consistently accessible and active across the population sampled. Though consistent in their activity, across a population this activity is expected to exist on a range of haplotypes over these open chromatin regions. By assessing the haplotypes from the WGS data at the R2 enhancer of *HBA*, many haplotypes can be observed, and all are associated with a strong open chromatin signal (Figure 2.6). This demonstrates that the sequence changes associated with this diversity do not lead to a change in activity in healthy individuals, and instead would be expected to be silent alterations. Any variant that would cause a dramatic phenotype change would be selected against in a population and would therefore be lost. This is consistent with the fact that our biobank cohort consists of a healthy adult population, where any dramatic changes to an essential process such as globin production would likely have an observable phenotype.

Going forward we have combined integrated haplotype analysis with allelic skew analysis to determine the haplotype structure of active alleles across the 100 haplotypes available from the 50 donors. The allelic analysis uses the presence of mapped variants from the individuals WGS to quantitatively assess the relative ATAC-seq signal in heterozygous individuals. In this way we can increase our ability to detect regulatory variation compared to a simple peak calling-based analysis and simultaneously determine the haplotype structure or structures associated with regulatory activities. The resultant dendrograms (e.g. Figure 2.6) are able to show the diversity of haplotypes present at a region as well as the relative activity associated with each haplotype. These figures do not show evolutionary history of the haplotypes, but instead show the relative activity at each haplotype, and indicate the frequency of each haplotype within this population.

From our peak based analysis (Figure 2.4) we have detected a sizeable population of peaks that appear highly variable. Following this observation we were particularly interested in the examples where activity was only associated with a single haplotype on a background of inactive haplotypes. These regions would be consistent with and defined as a GoF haplotype. Conversely, where activity was associated with a range of haplotypes, but a lack of activity was associated with a specific haplotype, this would be consistent with a LoF haplotype.

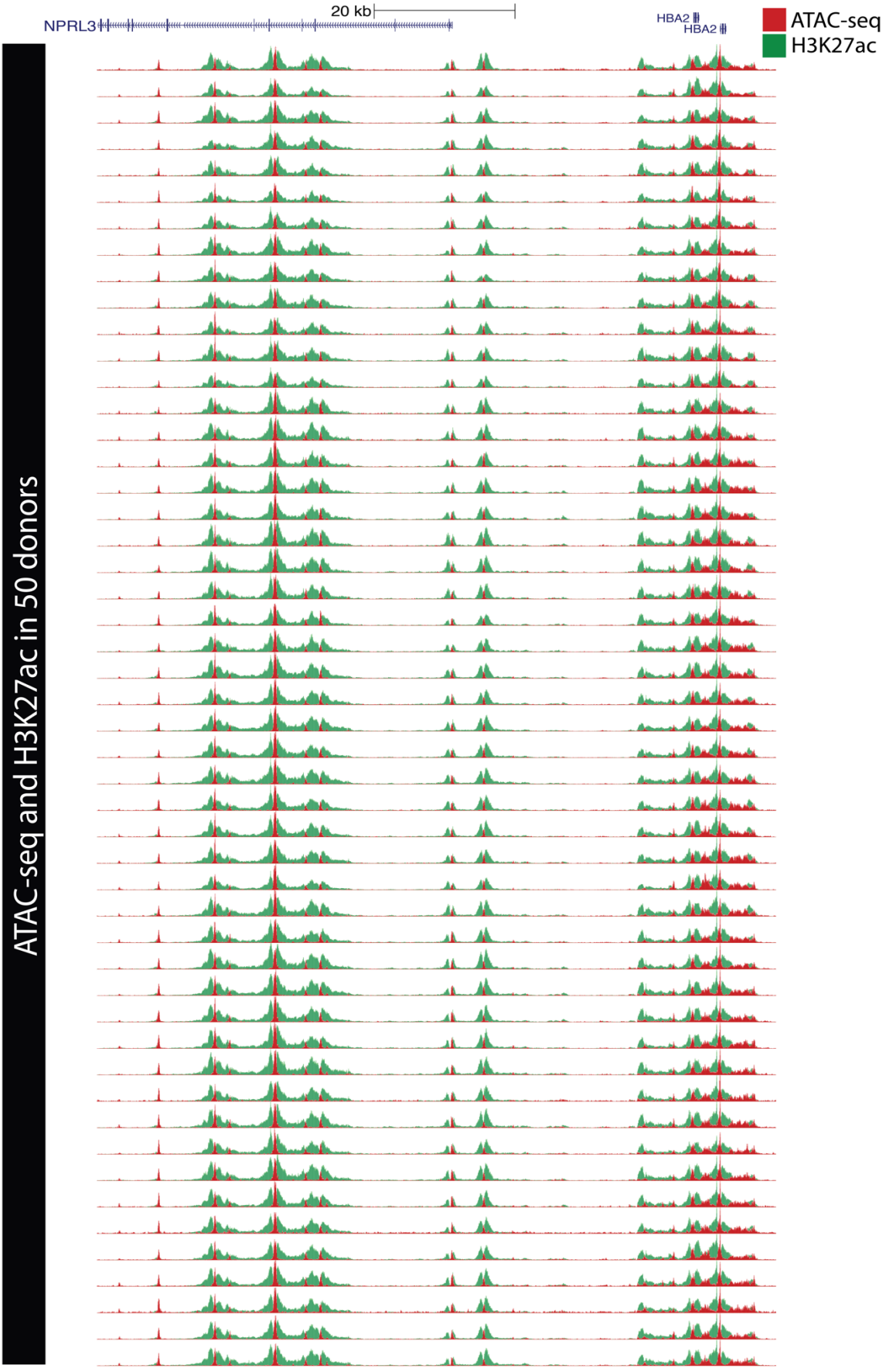

20 kb
NPRL3
HBA2
HBA2
ATAC-seq
H3K27ac
ATAC-seq and H3K27ac in 50 donors

Figure 2.5: Conservation of the regulatory elements at the *HBA* locus. ATAC (red) and H3K27ac (green) overlay from 50 individuals shows that the open chromatin and activity profiles across this population show no variation across all present regulatory elements.

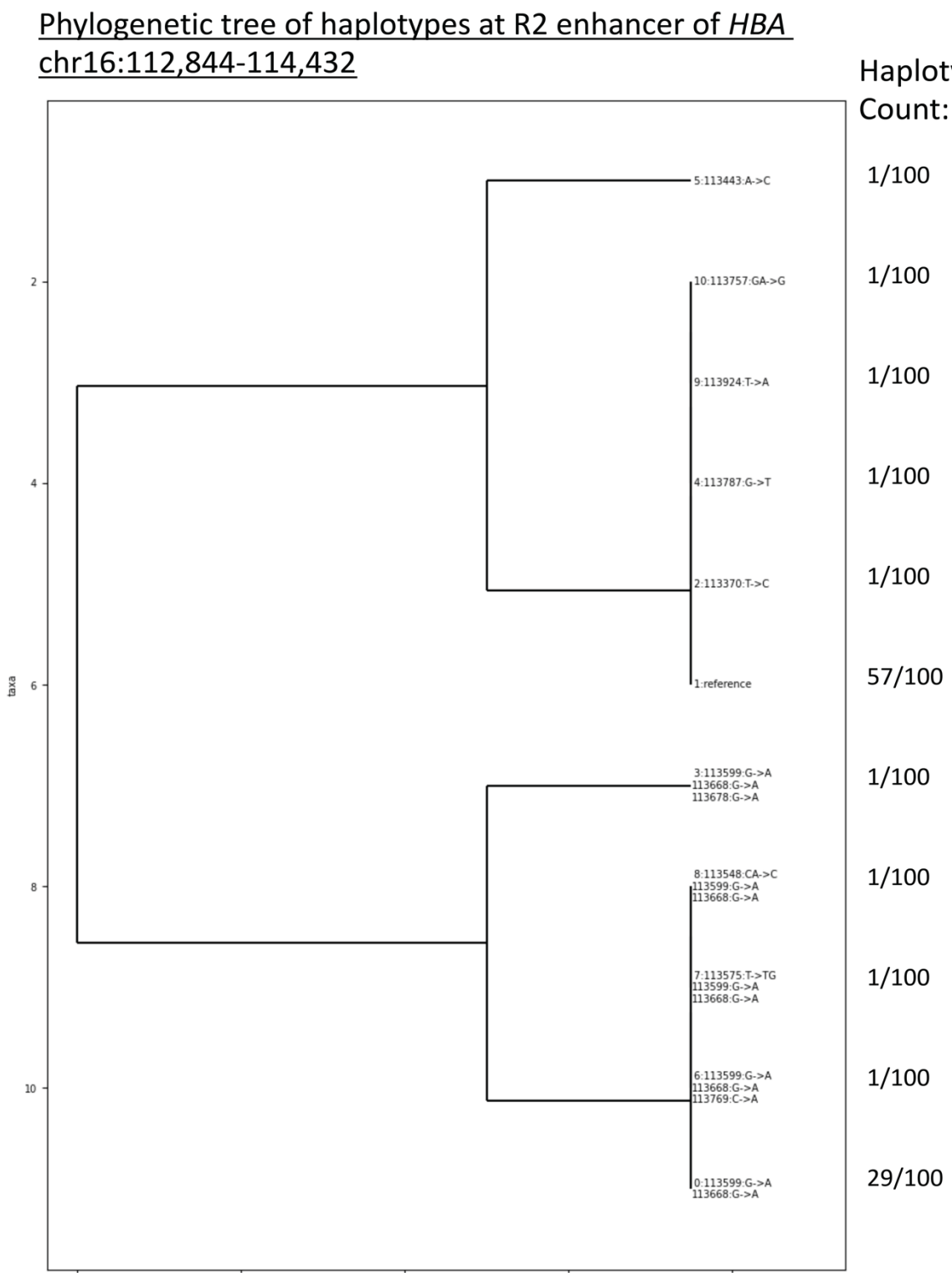


Figure 2.6: Dendrogram of haplotypes present at the R2 enhancer of *HBA.* Tree depicts the variety of haplotypes present in 50 individuals at R2. Total of 96 confident haplotypes of a possible 100 form 11 distinct haplotypes, the most common of which matches the reference. All other counts for each haplotype are included. All the R2 peaks at these haplotypes have strong activity and no allelic skew.

A novel peak is unique to a single donor and represents a true gain of function variant

In contrast to the highly stable right hand distribution of Figure 2.4, the regions in the left-hand distribution represent novel elements. These should be both convincingly active across a broad range of epigenetic assays, but highly restricted in the haplotype structure on which they are active to fulfil the definition of gain of function.

As an exemplar of such regions, we have chosen one of the most extreme cases in the dataset, which is associated with only one haplotype in the 100 haplotypes sampled. A gain of activity can be seen in the ATAC-seq track only for donor 1. The activity of this peak is further confirmed by more wide-spread enrichment for H3K27ac only in the same donor around the open chromatin sites (Figure 2.7). The region lies within the 3rd intron of the *PTPRK* gene, which encodes for a receptor type tyrosine-protein phosphatase (Figure 2.8). Though widely expressed in a variety of cell types and tissues, *PTPRK* it is not expressed in late erythroid cells when the region in question is active (Figure 2.8) (locus TAD structure shown in supplementary Figure 2). ChIP-seq for H3K4me1 and H3K4me3 was available for donor 1 in addition to H3K27ac and CTCF. These data indicated that the epigenetic profile of the region matched that of an enhancer and further confirmed the peak as a gain of function in this individual only (Figure 2.8).

Using the combined haplotype-aware allelic skew analysis described previously in this chapter, this showed that the gain in activity was only associated with one allele of this donor (Figure 2.9) and that this allele was unique to this individual in our population. Within the body of the new peak there is a difference of a single base pair at position chr6:128256094 (hg38). With this change of an A>G relative to the other haplotypes, the G allele is associated with activity. However, by manual inspection of the genome sequencing reads of donor 1, this variant is confirmed to be absent from the current release of dbSNP[151] (Supplementary Table 1). Chr6:128256094 A>G is also not associated with any human trait in the GWAS catalogue, however this is confounded by its absence from dbSNP. It is therefore unlikely to have been part of existing GWAS designs, which are largely based on these data.

PhyloP analysis [152] of conservation shows that the *de novo* peak is in an area depleted of conservation, although multiple sites of non-coding conservation exist within the same intron of

PTPRK, with the nearest being 2.4kb away. However, none of these non-coding conserved regions are associated with open chromatin peaks in erythroid cells (Figure 2.7).

Taken together this data strongly suggests that the substitution of an A residue for a G residue at this position makes a binding motif for a TF active in erythroid cells. However, it is known that a single TF binding site is unlikely to be sufficient for activity in mammalian cells. Therefore, the new site must work alongside other pre-existing TF binding motifs already present in the region. Considering the lack of conservation, it would suggest that these motifs existed cryptically and by chance. In this light, it is highly interesting that when the ATAC-seq data from all 50 donors is aligned, as in Figure 2.6, a weak increase of ATAC-seq signal can be seen across some of the donors at exactly the position of the gained open chromatin site; however, despite this accessibility, a H3K27ac signal is consistently absent. While well below the level detectable by peak calling methods, this could be suggestive of low-level priming by pre-exiting motifs, which is then driven to full activity in the presence of a new motif formed by the G allele.

At present there is no obvious candidate for the motif that is formed and hence the TF bound. The new sequence does not correspond to any of the binding motifs of the well know regulators of erythroid biology (GATA1/2, NFE2, KLF1, TAL1, etc). Similarly, analysis of the 15 bp sequence around the variant using tools such as footprintDB[153] that aggregates multiple motif databases, did not indicate any convincing candidates that change with the inclusion of the variant. However, despite the erythroid system and its TFs being well studied and defined, our total knowledge of the TFs active in the system is highly incomplete. We will therefore need to employ other methods to confirm this hypothesis and mechanisms (Chapter 3).

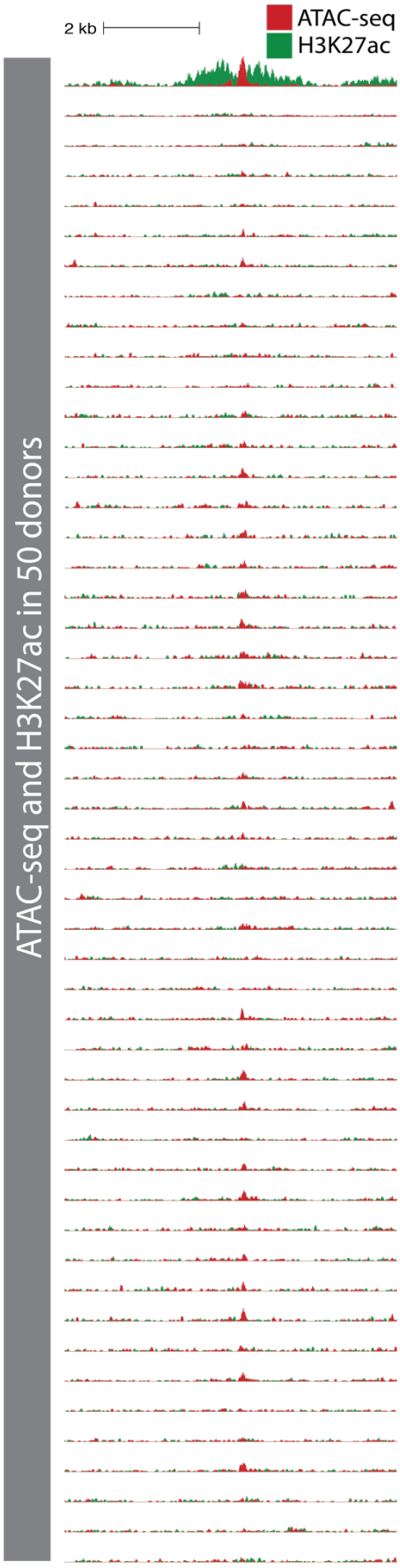

2 kb
ATAC-seq
H3K27ac
ATAC-seq and H3K27ac in 50 donors

Figure 2.7: A gain of function variant found in *PTPRK* is only present in one donor. Overlayed ATAC (red) and H3K27ac (green) from 50 individuals shows that a peak is only present in a single donor. This peak is both open and marked with H3K27ac. A weak ATAC signal can be seen in other individuals, but never with corresponding H3K27ac marks.

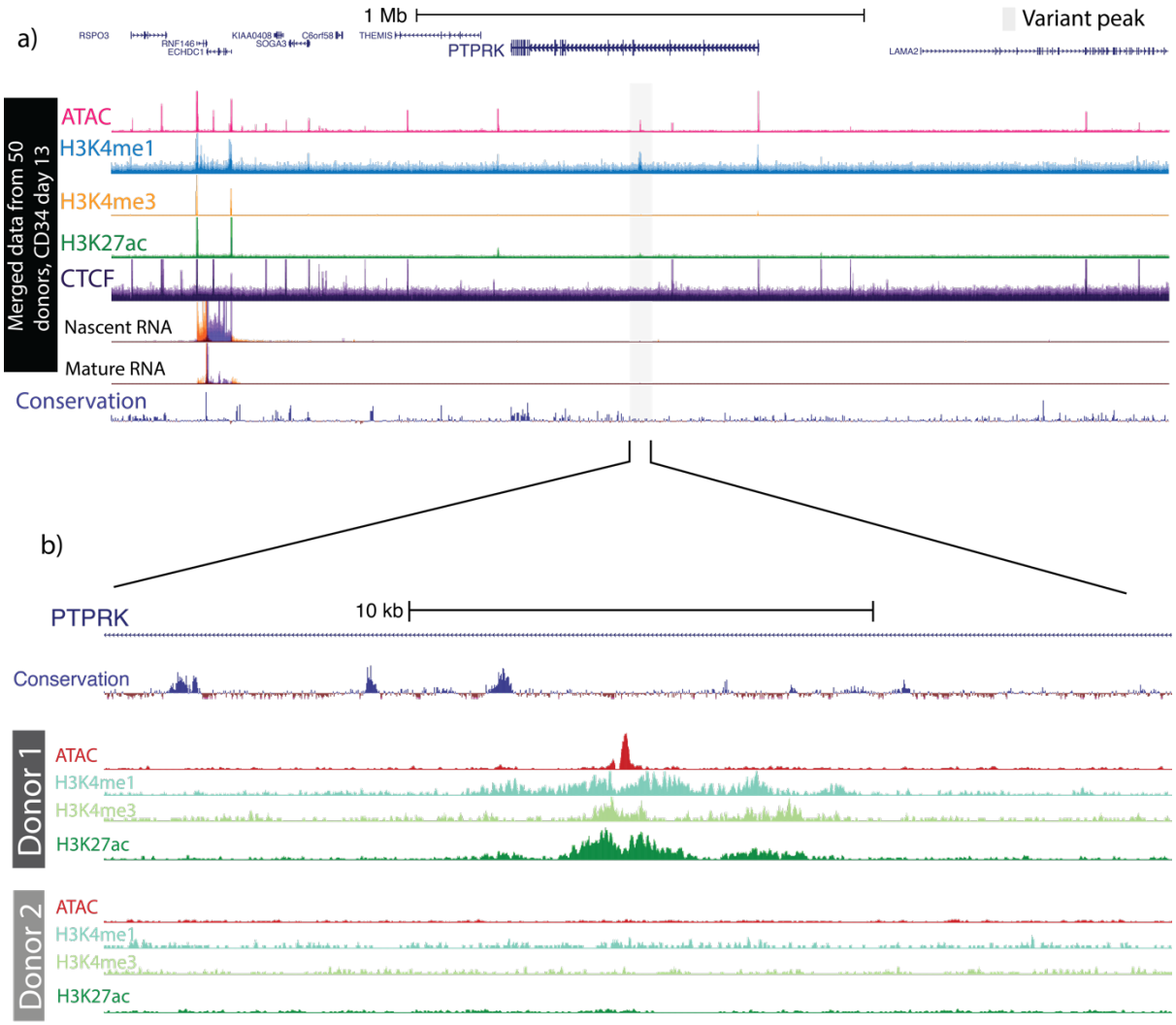
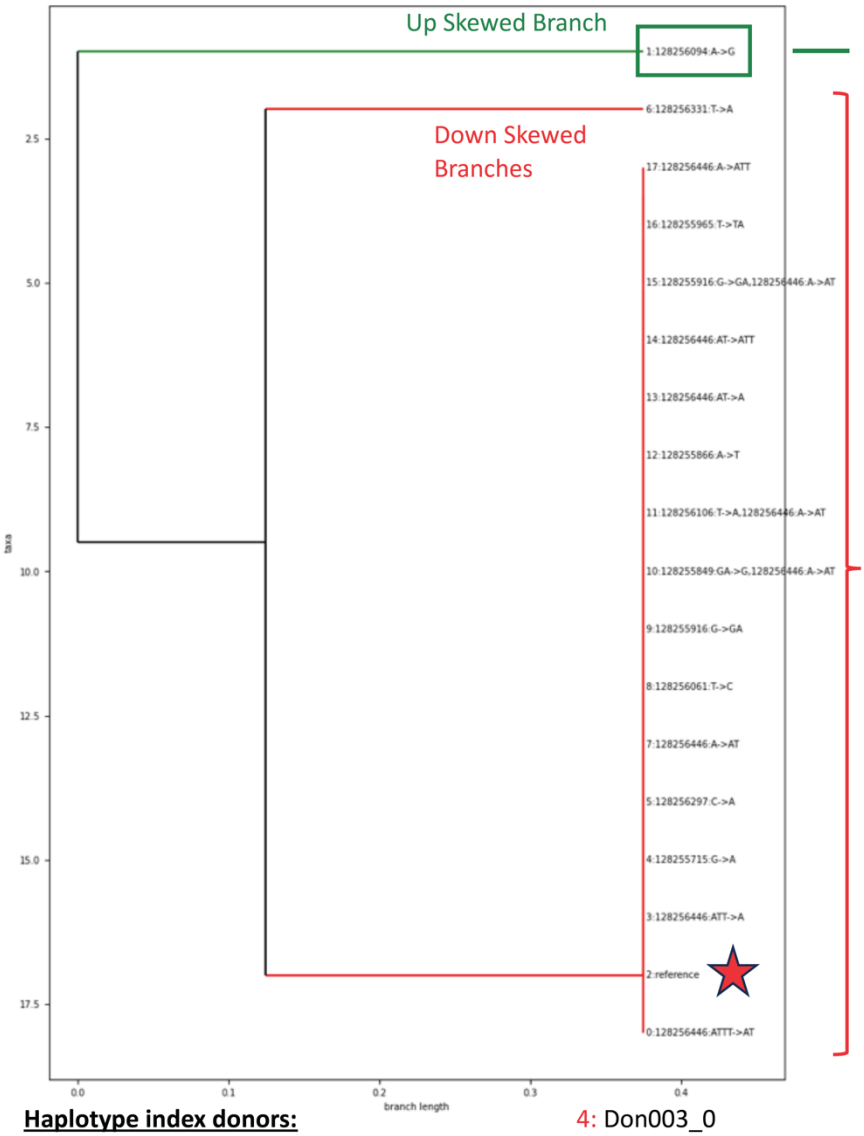


Figure 2.8: Genomic landscape and histone modifications around a gain of function peak in an intron of *PTPRK*. a) Epigenetic profile at the *PTPRK* locus. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Nascent RNA (PolyA-) and Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). Grey highlight depicts the variable gain of function peak. Because the pink ATAC track is an amalgamation of the 50 donors, the novel peak found in a single donor is visible in this track b) Zoom in on novel peak with epigenetic data from two donors, one has the peak present, the other has no peak. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). ATAC (red), H3K4me1 (light blue), H3K4me3 (green) and H3K27ac (dark blue). Peak and all associated marks are only present in one donor.

Phylogenetic tree of haplotypes at gain of function peak at *PTPRK* chr6:128255641-128256494



Haplotype Count:

1/100
2/100
1/100
1/100
2/100
1/100
1/100
1/100
1/100
2/100
1/100
1/100
7/100
1/100
1/100
1/100
42/100
1/100

Haplotype index donors:

1: Don001_1

0: Don001_0

2: Don002_0, Don003_1, Don004_1, Don006_0_1, Don007_1, Don009_0_1, Don010_0_1, Don011_0_1, Don13_0_1, Don014_0_1, Don015_0_1, Don016_0, Don017_1, Don018_0_1, Don019_0, Don020_1, Don021_0_1, Don022_0, Don023_0, Don025_0_1, Don026_1, Don028_1, Don029_0_1, Don030_0, Don031_1, Don032_0_1, Don033 0, Don035 0, Don035 1, Don036 0, Don037_0_1, Don038_0_1, Don039_0, Don040_0_1, Don041_0_1, Don042_1, Don044_0_1, Don045_0_1, Don046_0_1, Don047_0, Don049_1, Don050_0

3: Don002_1

4: Don003_0

5: Don004_0

6: Don005_0, Don005_1

7: Don007_0, Don019_1, Don031_0, Don039_1, Don047_1, Don048_0_1, Don050_1

8: Don016 1

9: Don017_0

10: Don020_0

11: Don022 1

12: Don023_1

13: Don026_0, Don036_1

14: Don028 0

15: Don030_1, Don049_0

16: Don033_1

17: Don042_0

Figure 2.9: Dendrogram of the haplotypes at a gain of function variant at *PTPRK.* Tree depicts the variety of haplotypes present in 50 individuals at the peak depicted in Figures 2.7 and 2.8. Total of 68 confident haplotype calls of a possible 100 form 18 distinct haplotypes. Red branches depict down skewed branches with no ATAC activity. Single green branch depicts the single up skewed haplotype that exhibits open chromatin and activity. Index list of which donors and which alleles belong to each haplotype. Red star marks the reference, which is the most common haplotype. All other counts for each haplotype included.

An enhancer proximal to SCARB2 shows variation in presence and activity across 50 individuals

Many loci across many individuals contain the variant peaks found in the middle distribution of Figure 2.4. When considering the emergence and selection of these variable peaks, several possible interpretations to explain their frequency exist. The variation may be linked to a highly common LoF allele or alleles, where the frequency is common enough that homozygous inheritance is also common. The alternative is that they arise as a GoF allele and through selection become common in a population. For these variants the activity is associated with a specific common allele whilst inactivity is associated with a normal, varied haplotype distribution.

The locus of *SCARB2* shows an active ATAC peak that is present in several but not all the donors (Figure 2.10). Similarly to the previous example, the presence of open chromatin correlates with the active H3K27ac histone mark, showing that accessibility and activity are linked. Amongst the individuals, varying degrees of peak strength can also be observed. When accessible, the peak is marked with the enhancer histone modifications H3K4me1 and H3K27ac, but these marks are reduced when the region is less open (Figure 2.11b). With the donors shown in Figure 2.11, the strong reduction in accessibility is apparent, though a small ATAC peak and some histone modifications can still be seen, suggesting some weak activity is still present. The enhancer lies between the promoter of *SCARB2*, which shows active transcription, and *FAM47E,* which is inactive. Upstream of *SCARB2* is also the active gene *NUP54. SCARB2* encodes scavenger receptor class B member 2, which in turn is involved in the production of a lysosomal membrane protein. NUP54 is Nucleoporin 54, a subunit of the nuclear pore of the nuclear envelope (locus TAD structure shown in supplementary Figure 3).

By haplotype analysis, it can be seen that the haplotypes for this region are complex, with several non-adjacent branches showing positive skew (Figure 2.12). As a complex dendrogram, it is therefore

difficult to determine whether the presence of this peak is due to a loss or gain of function. One possible explanation is that it represents a loss of function, where the top branch retains the function, and then in the lower ones some variants have damaged it, whilst others are silent and have retained the activity. There are two SNPs found within the main body of this variant peak: rs11097297 and rs76504716 (Supplementary Table 1). rs11097297 (Chr4:76233712 C > T) is more likely to be causal as its frequency amongst the individuals varies (homozygous C/C in 34/50 donors, heterozygous C/T in 14/50 donors, homozygous variant in 2/50 donors) and matches the activity. In contrast, rs76504716 (Chr4:76233804 T > G) is much more common (Homozygous T/T in 46/50 donors, heterozygous in 4/50 donors) and its presence does not correlate with activity. When assessing the primary three donors, donor 2 has the largest ATAC peak, with donors 1 and 3 showing much smaller peaks. rs11097297 was shown to be homozygous T/T for donor 2 and homozygous C/C for donors 1 and 3, showing that the SNP was associated with an increase in activity. It can therefore be assumed that rs11097297 is the causal SNP. When considering the conservation over this region, there is very little within the local vicinity of the variant peak (Figure 2.11). Conservation can be seen at the introns of the neighbouring *SCARB2* and *NUP54* genes, but not over the enhancer peaks.

By assessing potential binding factors near rs11097297, the base change lies 4bp away from an identified RUNX2 motif. Though not associated with erythropoiesis, the related family member RUNX1 is critical in erythroid differentiation and shares a common binding site. When assessing donors with the T/T genotype (Figure 2.11b), though reduced, the chromatin is still open, and histone modifications signify activity. This suggests the SNP drastically reduces the accessibility but does not fully abolish the region's ability to open, bind factors and to be active. As it is merely adjacent to this motif, perhaps this explains the lack of complete abolition. I will functionally follow up on this element in the subsequent chapter.

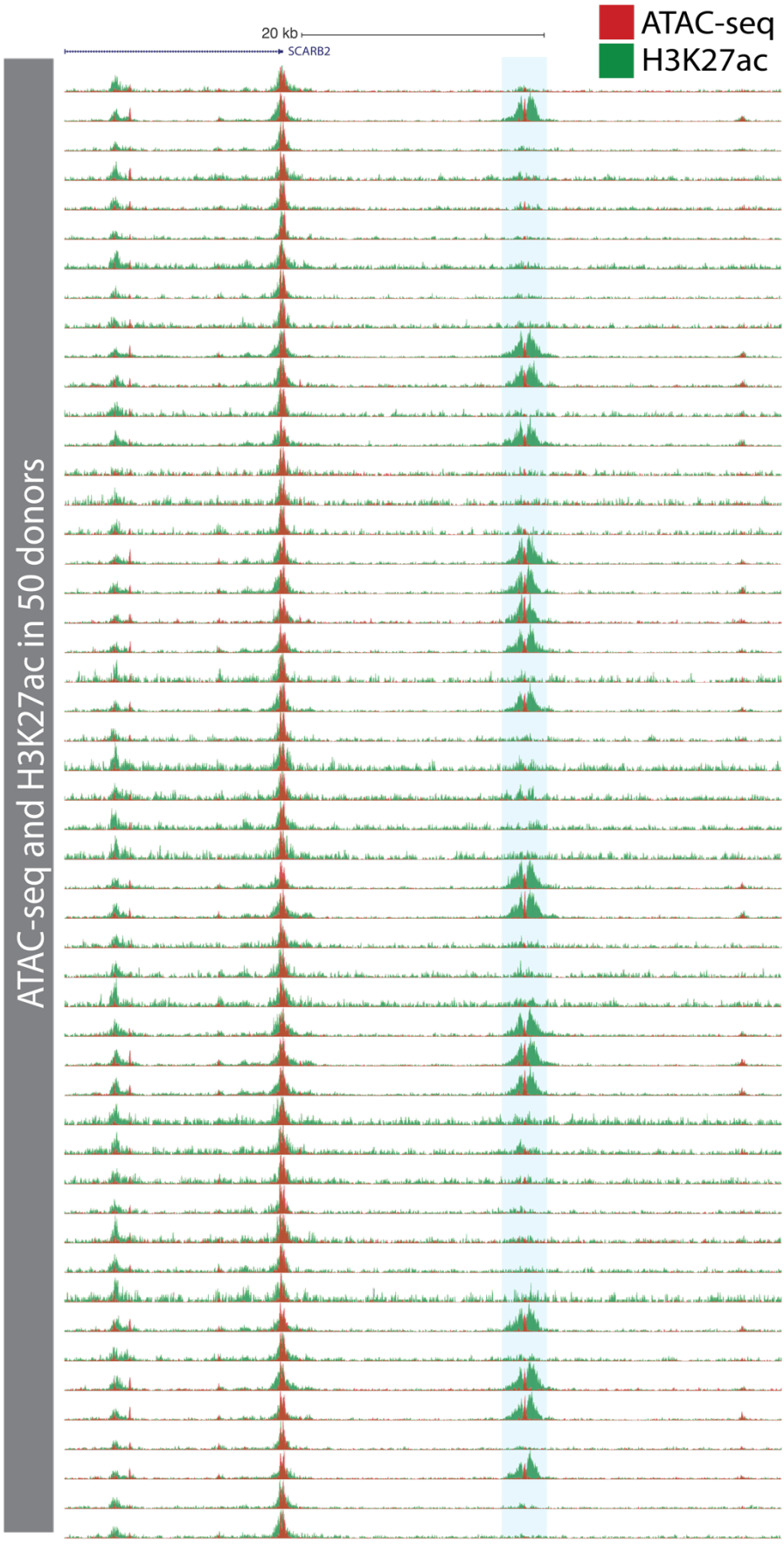

20 kb
SCARB2
ATAC-seq
H3K27ac
ATAC-seq and H3K27ac in 50 donors

Figure 2.10: Variation at the *SCARB2* locus. Overlayed ATAC (red) and H3K27ac (green) from 50 individuals shows that the peak corresponding to the *SCARB2* promoter is conserved, whereas the downstream enhancer, highlighted in blue, shows variation across the individuals. For each individual with an open chromatin peak, H3K27ac marks are also present.

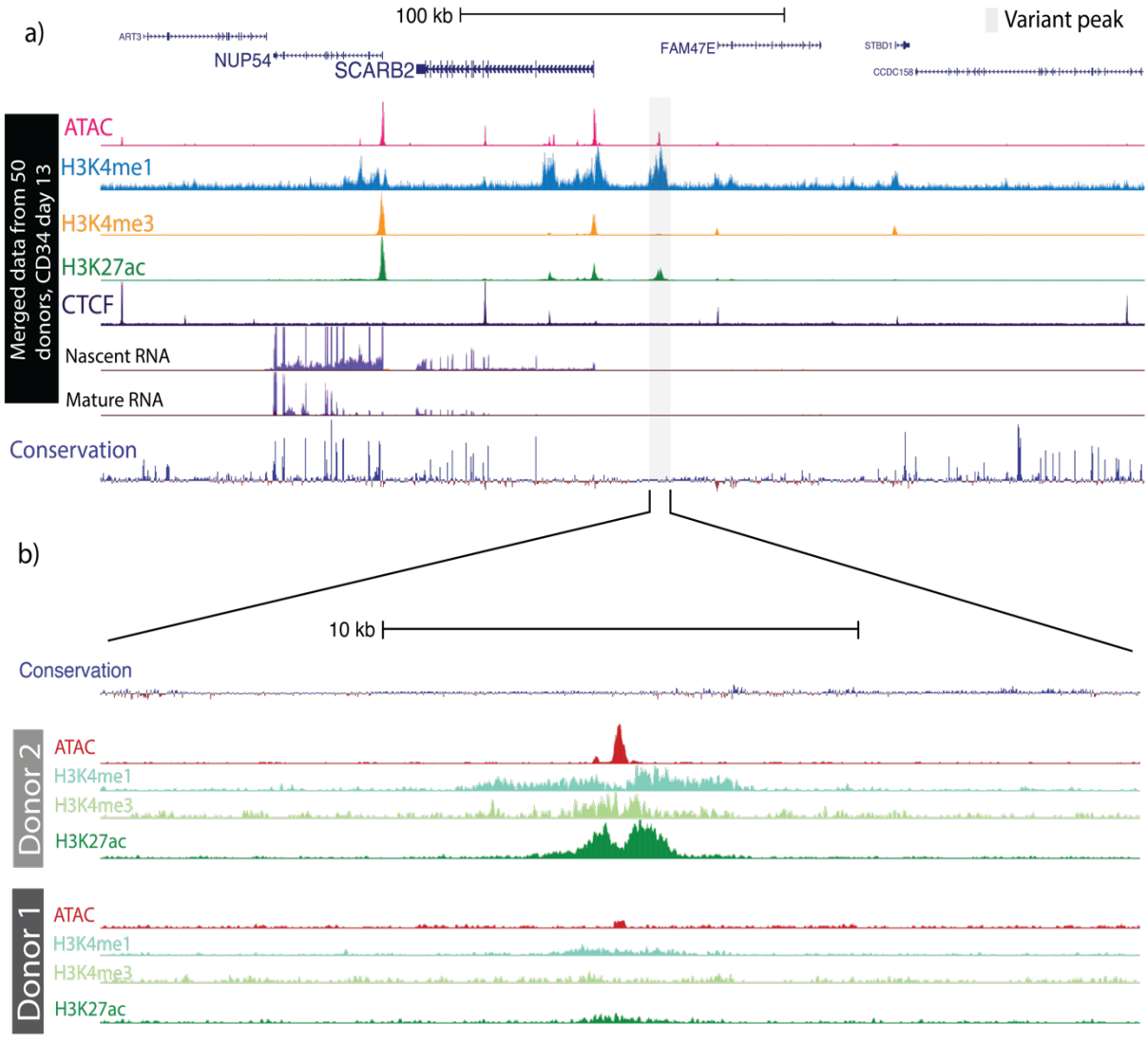


Figure 2.11: Genomic landscape and histone modifications around a variable enhancer downstream of *SCARB2*. a) Epigenetic profile at the *SCARB2* locus. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP-seq from the 5 available sets (Table 2.1). Nascent RNA (PolyA-) and Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). Grey highlight depicts the variable enhancer peak. b) Zoom in on novel peak with epigenetic data from two donors, one has the peak present, the other has a much weaker signal. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). ATAC (red), H3K4me1 (light blue), H3K4me3 (green) and H3K27ac (dark blue). Peak and all associated marks are stronger in one donor.

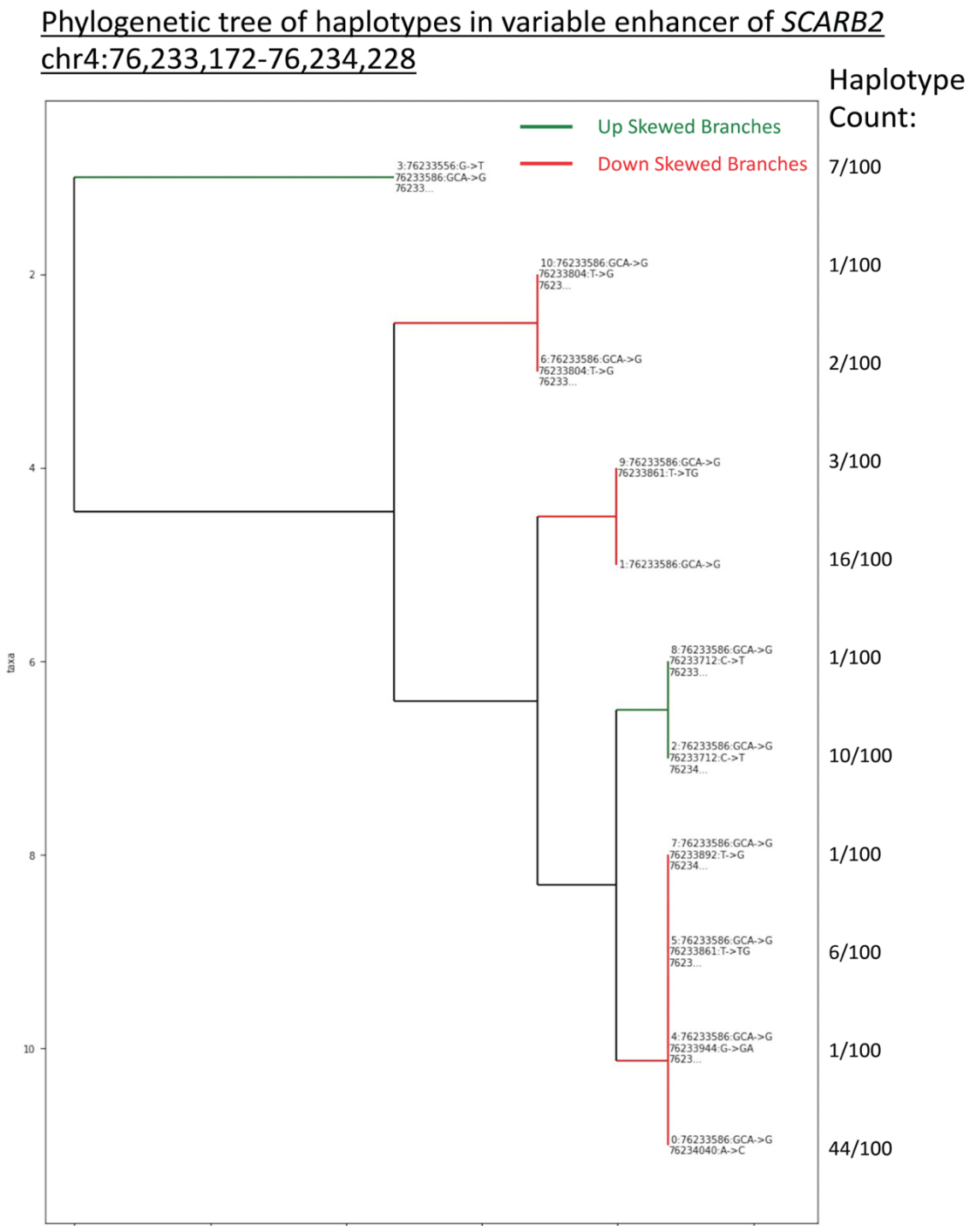


Figure 2.12: Dendrogram of the haplotypes at a variable enhancer peak downstream of *SCARB2*. Tree depicts the variety of haplotypes present in 50 individuals at the peak depicted in figures 2.10 and 2.11. Total of 92 confident haplotype calls of a possible 100 haplotypes form 11 distinct haplotypes. Red branches depict down skewed branches with no ATAC activity. Green branches depict up skewed haplotypes that exhibit open chromatin and activity. Counts for each haplotype included. Complex inheritance pattern of the haplotypes makes defining this peak as a loss or gain of function difficult.

The variable genomic landscape proximal to PPP1R3B exhibits both gain and loss of function enhancers

Within our analysis we found an interesting example of a highly complex haplotype that affects two different elements within the same regulatory landscape. In the amalgamated data set the *PPP1R3B* locus shows a simple landscape with a promoter and two downstream enhancer elements (Figure 2.14) (locus TAD structure shown in supplementary Figure 4). However, our analysis suggests that of the two enhancers working in this locus, the first peak (peak 1) can be subjected to a LoF variant whilst the second peak (peak 2) is created by a potential GoF variant. Additionally, this locus has a known haplotype linked to an erythroid associated GWAS.

SNP rs3748136 in this haplotype resides in an erythroid open chromatin peak (peak 2) in donors 1 and 2, but not donor 3. The SNP rs3748136 is associated with a reduced reticulocyte count [72], and based on chromatin marks (from donor 1 and 2) sits in an enhancer upstream of the *PPP1R3B* promoter (Figure 2.14). When looked at across the full 50 individuals, this locus demonstrated a highly complex interplay between the loss of peak 1 and the gain of peak 2 (Figure 2.13). Peak 1 is also associated with enhancer chromatin marks, and is situated more proximal to the *PPP1R3B* promoter (Figure 2.14). In contrast to conserved examples, the occurrence of both enhancer peaks varies widely across the 50 individuals. Allelic skew and haplotype analysis shows that three main classes of haplotype exist in this enhancer landscape.

The first class represents a range of haplotypes where peak 1 is active and peak 2 is inactive on the same allele (Figure 2.15). In these haplotypes the active peak 1 is associated with the reference sequence for rs330946 (chr8:9164667, A > C). On these haplotypes peak 2 is reference for two linked SNPs rs3748137 (chr8:9172593, G > C) and rs3748136 (chr8: 9172650, G > A), a haplotype which is associated with a lack of activity (Supplementary Table 1). Two haplotypes then branch from here, where one haplotype has two active enhancers (Haplotype A) whilst the other has no active enhancers (Haplotype B) (Figure 2.15). In haplotype A peak 1 still lacks the rs330946 C allele and so remains active while peak 2 acquires both rs3748137 (C allele) and rs3748136 (A allele) and gains activity, so the locus now has two active enhancer elements (Figure 2.15). This haplotype was therefore subjected to the

GoF variant but not the LoF variant. In haplotype B peak 1 gains the rs330946 C allele and therefore loses activity while peak 2 remains as reference for rs3748137 (G Allele) and rs3748136 (G allele) and so still shows no activity. This haplotype was therefore subjected to the LoF variant but not the GoF variant. The dendrograms of the haplotypes at both the peaks are presented in Figures 2.16 and 2.17 and reflect the complexity of this locus. Both trees show a clear distinction between the skewed haplotypes which associated with the active peaks and the SNPs described. The complexity is due to several haplotypes containing these active SNPs and exhibiting further variation within the haplotype blocks.

Peak 1 appears poorly conserved suggesting this is a rapidly evolving regulatory landscape. Though peak 2 is adjacent to a block of sequence conservation it does not overlap with the peak itself and so appears unrelated to function (Figure 2.14). Interestingly, the gain of activity in the peak 2 haplotype is associated with the simultaneous creation of a GATA1 binding site (rs3748137, C allele) and an NFE2 binding sites (rs3748137 A allele) 57 bp apart. Both GATA1 and NFE2 are critical TFs at this stage of erythroid development and are known to be essential for upregulation of the globin genes active at this time point. This would be a credible explanation for the formation of an active element in this haplotype. However, no clear interpretable motif was identified as lost in peak 1 in the presence of the rs330946 C allele.

This locus highlights the incredible regulatory variability and interplay between regulatory variants. While not formally linked to the expression of the *PPP1R3B* gene, which encodes the protein phosphatase 1 regulatory subunit 3B, the locus has been linked to multiple GWAS traits including glycogen synthesis and Type 2 Diabetes onset [72, 154-156]. Due to the interesting patterns of variation and links to human traits this locus was prioritised for follow up in the subsequent chapters.

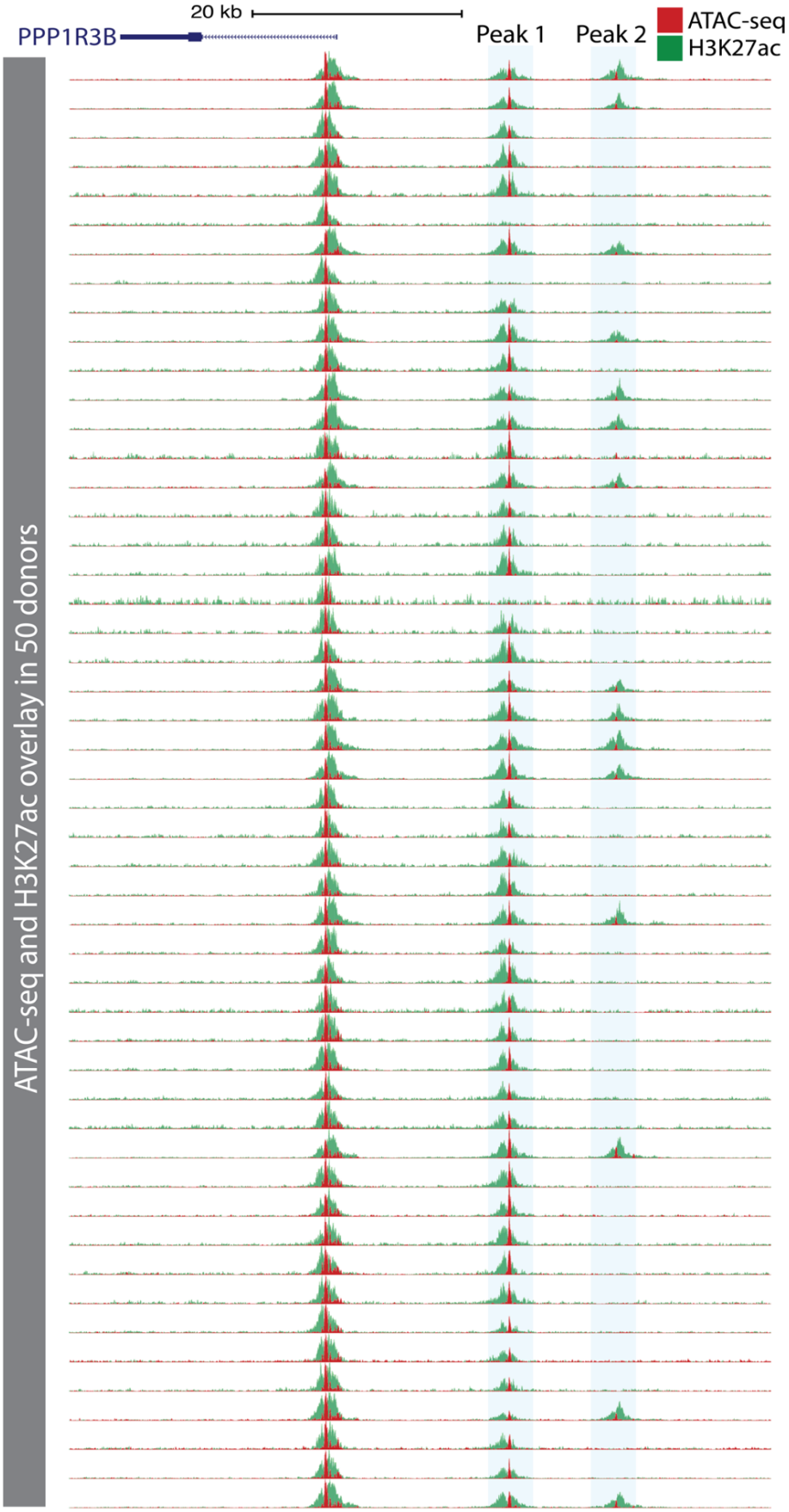

20 kb
PPP1R3B
Peak 1
Peak 2
ATAC-seq
H3K27ac
ATAC-seq and H3K27ac overlay in 50 donors

Figure 2.13: Variation at the *PPP1R3B* locus. ATAC from 50 individuals shows that the left hand peak corresponding to the *PPP1R3B* promoter is conserved, whereas the two downstream enhancers, highlighted in blue, both show variation across the individuals.

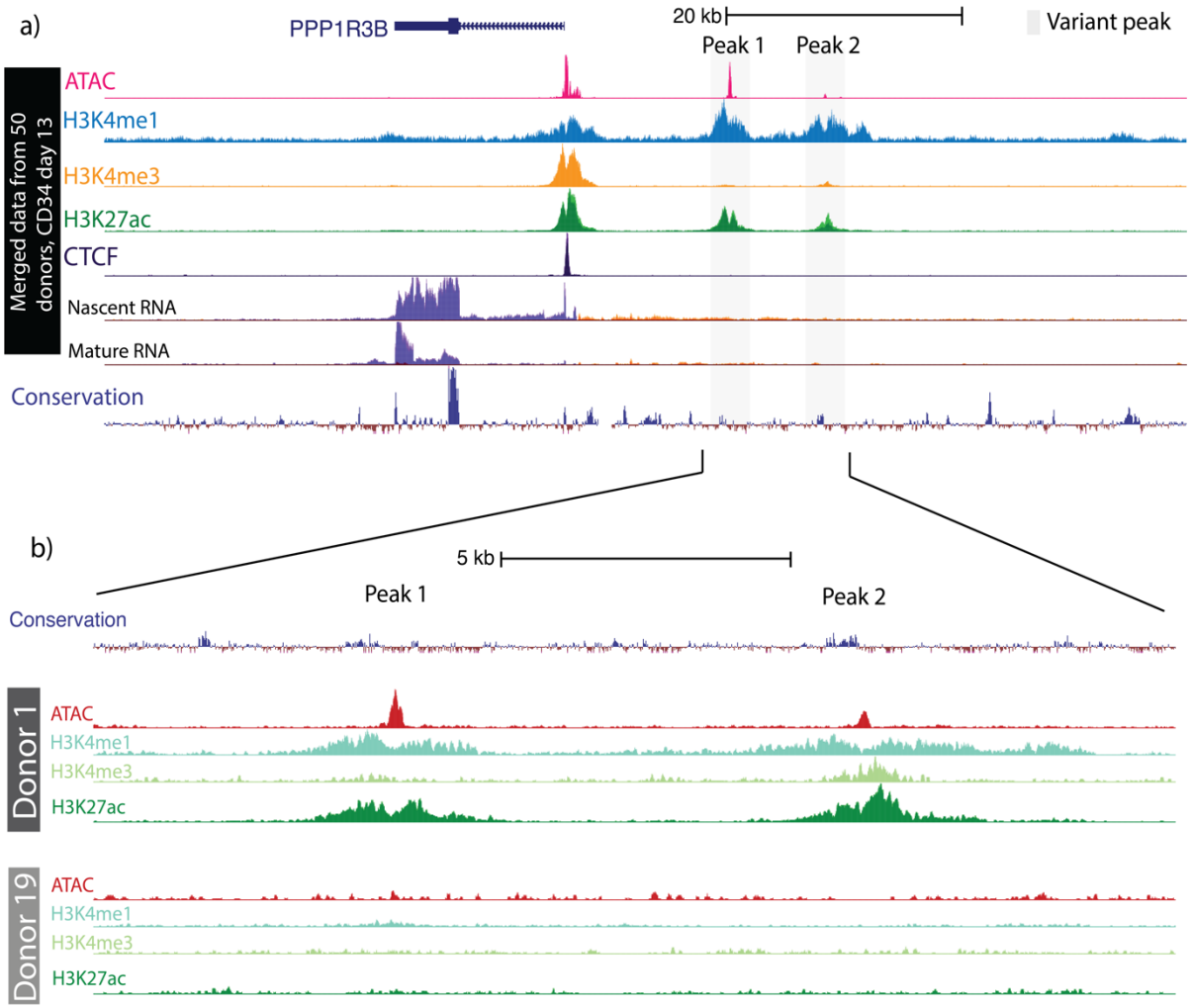


Figure 2.14: Genomic landscape and histone modifications around two variable enhancers downstream of *PPP1R3B*. a) Epigenetic profile at the *PPP1R3B* locus. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP-seq from the 5 available sets (Table 2.1). Nascent RNA (PolyA-) and Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). Grey highlight depicts the two enhancer peaks (peak 1 and peak 2). b) Zoom in on novel peak with epigenetic data from two donors, one has both peaks present, the other has a neither peaks. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). ATAC (red), H3K4me1 (light blue), H3K4me3 (green) and H3K27ac (dark blue). Peak and all associated marks appear together in donors.

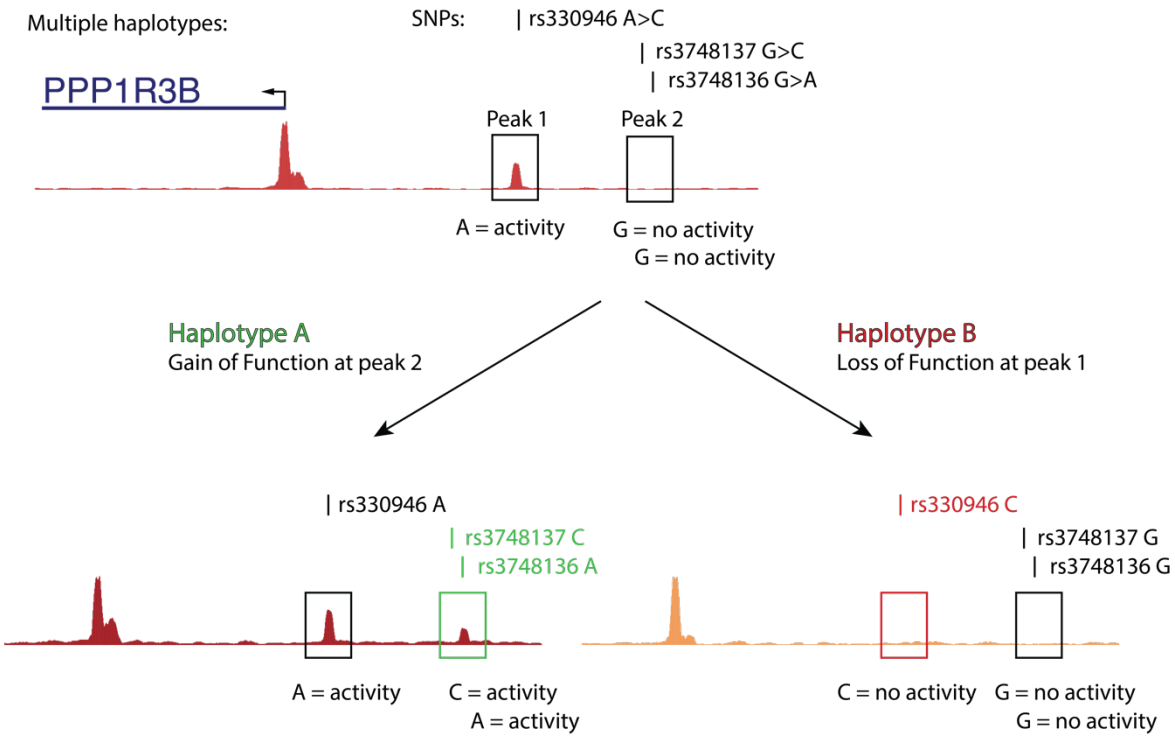


Figure 2.15: The enhancers of *PPP1R3B* show complex inheritance across multiple haplotypes. Multiple haplotypes show the presence of the *PPP1R3B* promoter and an enhancer at peak 1. Peak is reference A for the SNP rs330946. In these haplotypes peak 2 remains closed and is reference G and G for the SNPs rs3748137 and rs37481386 respectively. Haplotype A depicts regions that have been subjected to the two SNP changes in peak 2, with rs3748137 and rs37481386 changing from G > C and G > A respectively. These act as a gain of function and so these haplotypes have both active enhancers. Haplotype B depicts regions that have been subjected to the SNP change in peak 1, with rs330946 changing from A > C. This acts as a loss of function and so these haplotypes have no active enhancers.

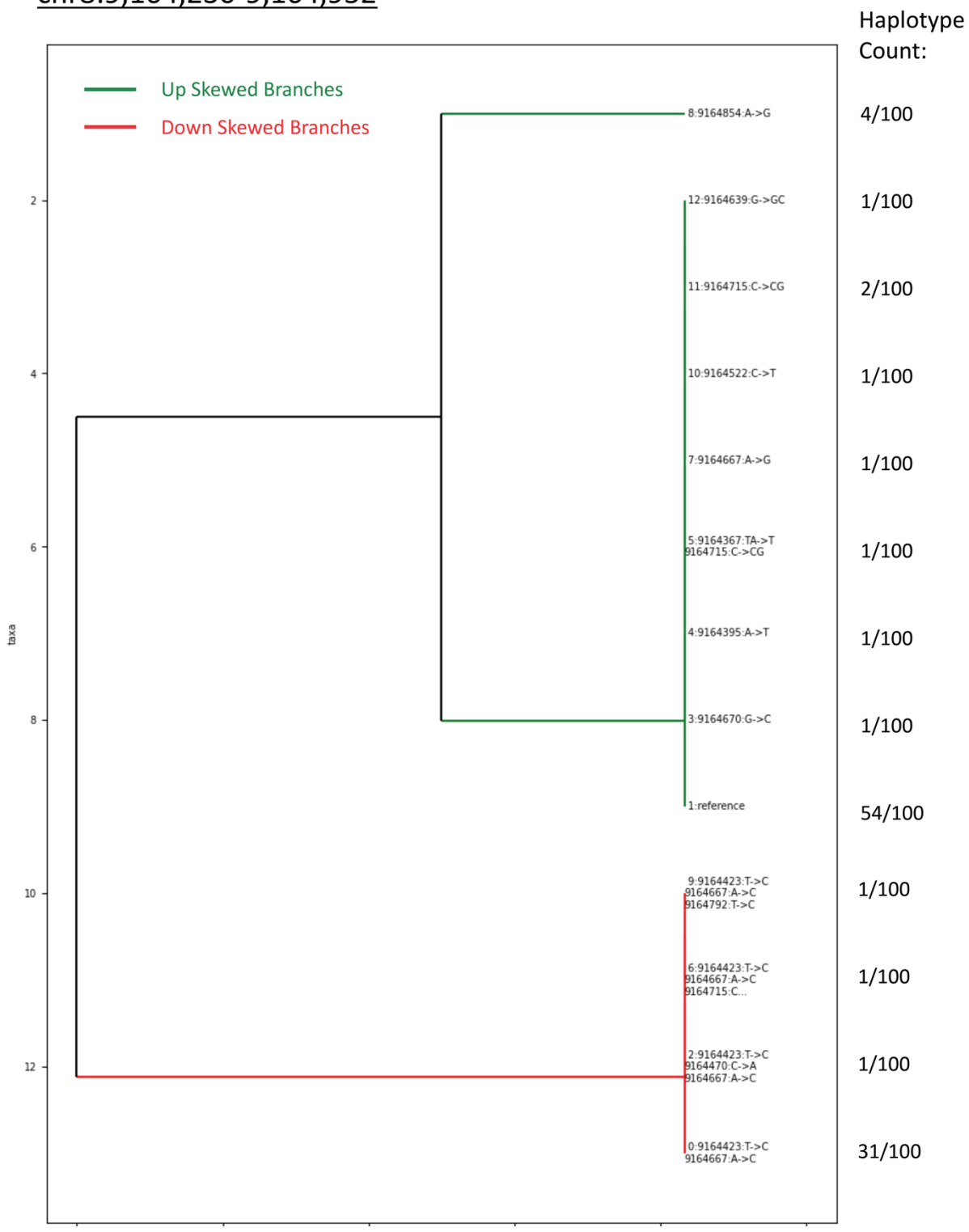

Phylogenetic tree of haplotypes at LoF peak 1 at PPP1R3B chr8:9,164,230-9,164,952
Haplotype Count:
Up Skewed Branches
Down Skewed Branches
taxa
2
4
6
8
10
12
8:9164854:A->G
12:9164639:G->GC
11:9164715:C->CG
10:9164522:C->T
7:9164667:A->G
5:9164367:TA->T
9164715:C->CG
4:9164395:A->T
3:9164670:G->C
1:reference
9:9164423:T->C
9164667:A->C
9164792:T->C
6:9164423:T->C
9164667:A->C
9164715:C...
2:9164423:T->C
9164470:C->A
9164667:A->C
0:9164423:T->C
9164667:A->C
4/100
1/100
2/100
1/100
1/100
1/100
1/100
1/100
54/100
1/100
1/100
1/100
31/100

Figure 2.16: Dendrogram of the haplotypes at the loss of function peak downstream of *PPP1R3B.* Tree depicts the variety of haplotypes present in 50 individuals at peak 1 depicted in figures 2.13, 2.14 and 2.15. Total of 100 confident haplotype calls of a possible 100 haplotypes form 13 distinct haplotypes. Red branches depict down skewed branches with no ATAC activity. Green branches depict up skewed haplotypes that exhibit open chromatin and activity. Counts for each haplotype included.

Phylogenetic tree of haplotypes at GoF peak 2 at *PPP1R3B* chr8:9,172,375-9,172,880

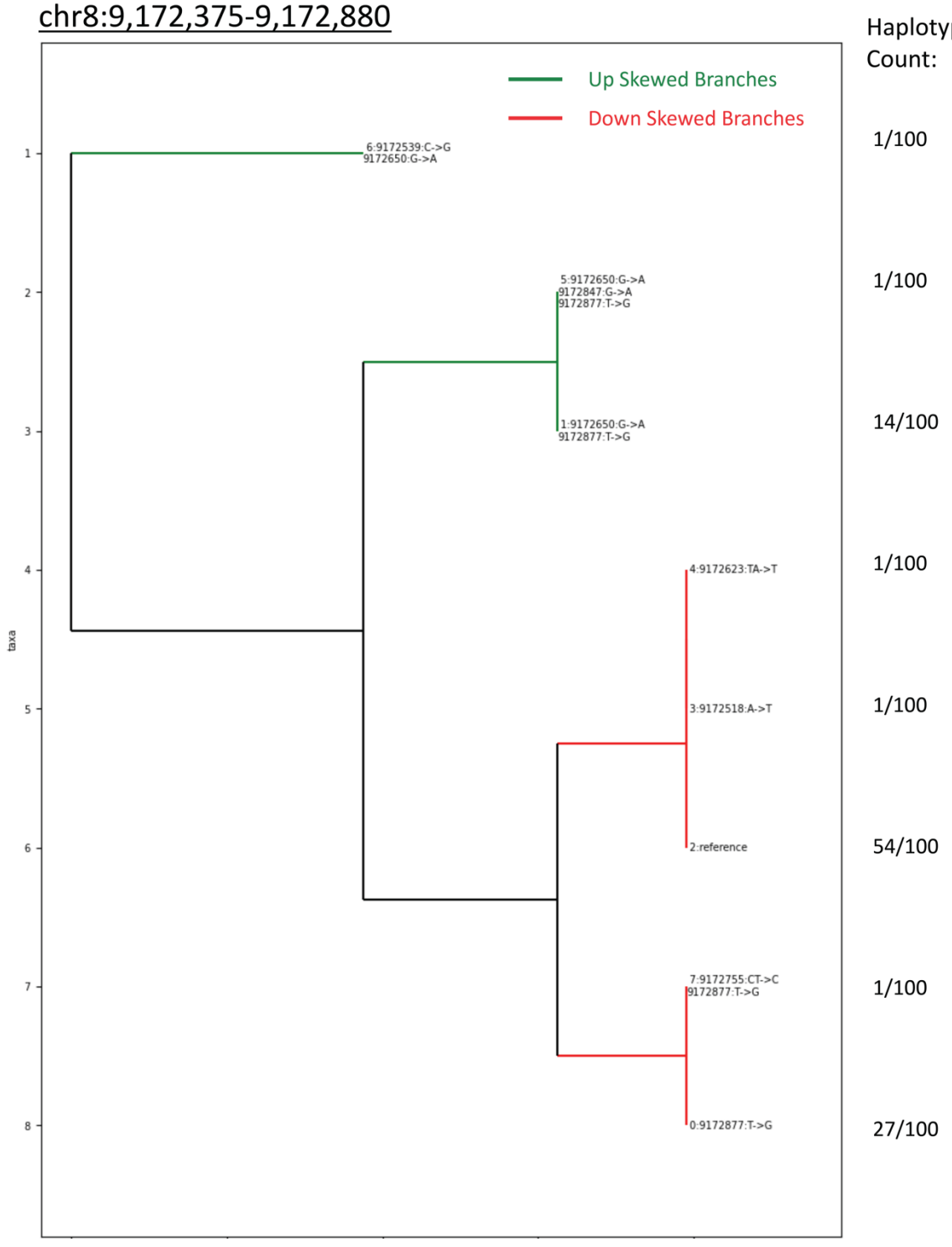


Figure 2.17: Dendrogram of the haplotypes at the gain of function peak downstream of *PPP1R3B.* Tree depicts the variety of haplotypes present in 50 individuals at peak 2 depicted in figures 2.13, 2.14 and 2.15. Total of 100 confident haplotype calls of a possible 100 haplotypes form 8 distinct haplotypes. Red branches depict down skewed branches with no ATAC activity. Green branches depict up skewed haplotypes that exhibit open chromatin and activity. Counts for each haplotype included.

Variable regions can exhibit multiple peaks appearing simultaneously in certain individuals

Another important observation that this analysis highlights it that of multiple peaks arising or being lost simultaneously in the same landscape. Figure 2.18 shows four variable peaks within the 4th intron of *CLNK* on chromosome 4, which when present also show H3K27ac activity. Across the individuals, the landscape will either show all of these peaks simultaneously, or none at all. Histone modifications and CTCF-ChIP seq show that from left to right the four peaks present are likely enhancers with a CTCF site separating them (Figure 2.19). The right most peak (peak 4) also shows promoter associated H3K4me3 marks, but as there are no active transcripts from this region, this region is not behaving as an active promoter in this cell type. Interestingly, for individuals missing the central ATAC peak (peak 3) that overlaps the CTCF site, CTCF binding is still present even when the peak appears closed. This CTCF site is bound in all but one of the donors, even when the ATAC peak is absent. This therefore suggests that even though the peak is closed, CTCF is still able to bind. *CLNK* encodes Cytokine Dependent Hematopoietic Cell Linker or Mast Cell Immunoreceptor Signal Transducer (MIST), which is involved in the regulation of immunoreceptor signalling, and is off in erythroid cells (Figure 2.19). However, the upstream *ZNF518B* Zinc finger protein gene is expressed, showing that the region overall is active (locus TAD structure shown in supplementary Figure 5).

Of the donors with these peaks, 9/10 share a common haplotype, and all are heterozygous at this region. There are SNPs present in three of the peaks (two in peak 2, one in peak 3, three in peak 4), and so determining causality proves difficult. This also raises an important question: are these regions opening independently or because of the local context? For these landscapes there are two possible causes: either a variant in a single active region is recruiting certain factors that once in the local vicinity, are also able to open other regions; or each individual peak has a unique SNP that is causing the region to open. It has been shown via genomic editing that introducing active elements into previously inactive regions can affect the local landscape and induce activity [22]. By recruiting strong influential transcription factors to a genomic region, this can induce the opening of surrounding

regions. Determining the exact causality of this type of landscape proves difficult, and furthermore the causes are not mutually exclusive. For this region in particular, all the peaks reside on a common haplotype shared by the donors that display the open sites, and the two largest peaks both contain SNPs. PhyloP conservation shows little conservation in this region, except at the downstream *CLNK* intron, and a small amount over the central CTCF site. This means it is difficult to determine the exact effect of the SNPs and the active relationships between the peaks. However, this example does highlight the importance of considering the surrounding genomic context when assessing differential peaks or variants.

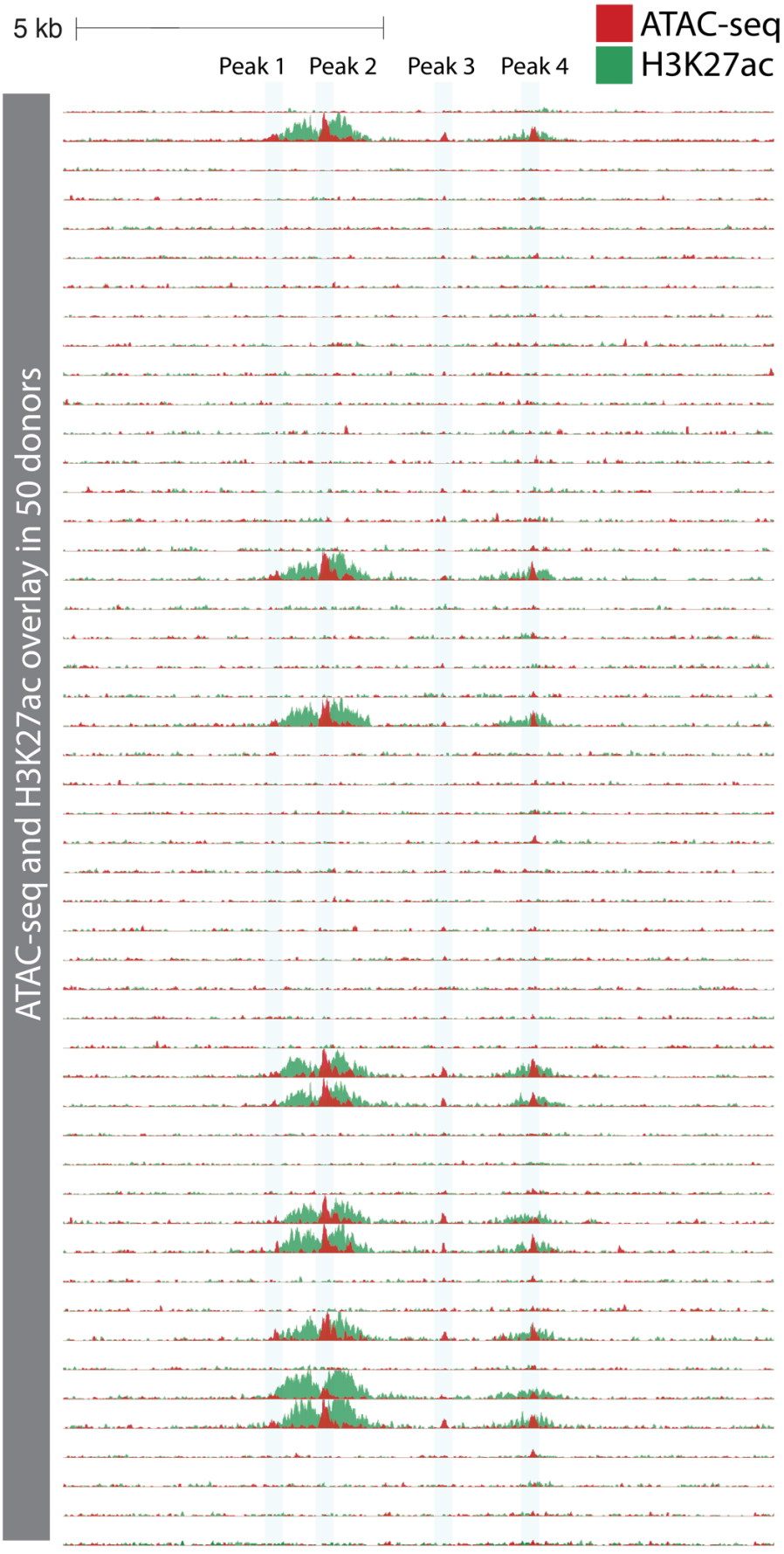

5 kb
ATAC-seq
H3K27ac
Peak 1
Peak 2
Peak 3
Peak 4
ATAC-seq and H3K27ac overlay in 50 donors

Figure 2.18: Multiple peaks arising simultaneously within the same haplotype. Overlayed ATAC (red) and H3K27ac (green) from 50 individuals shows that a region within an intron of *CLNK* shows variation across the individuals. Donors show either all or none of these peaks, and there is correlation between accessibility and activity.

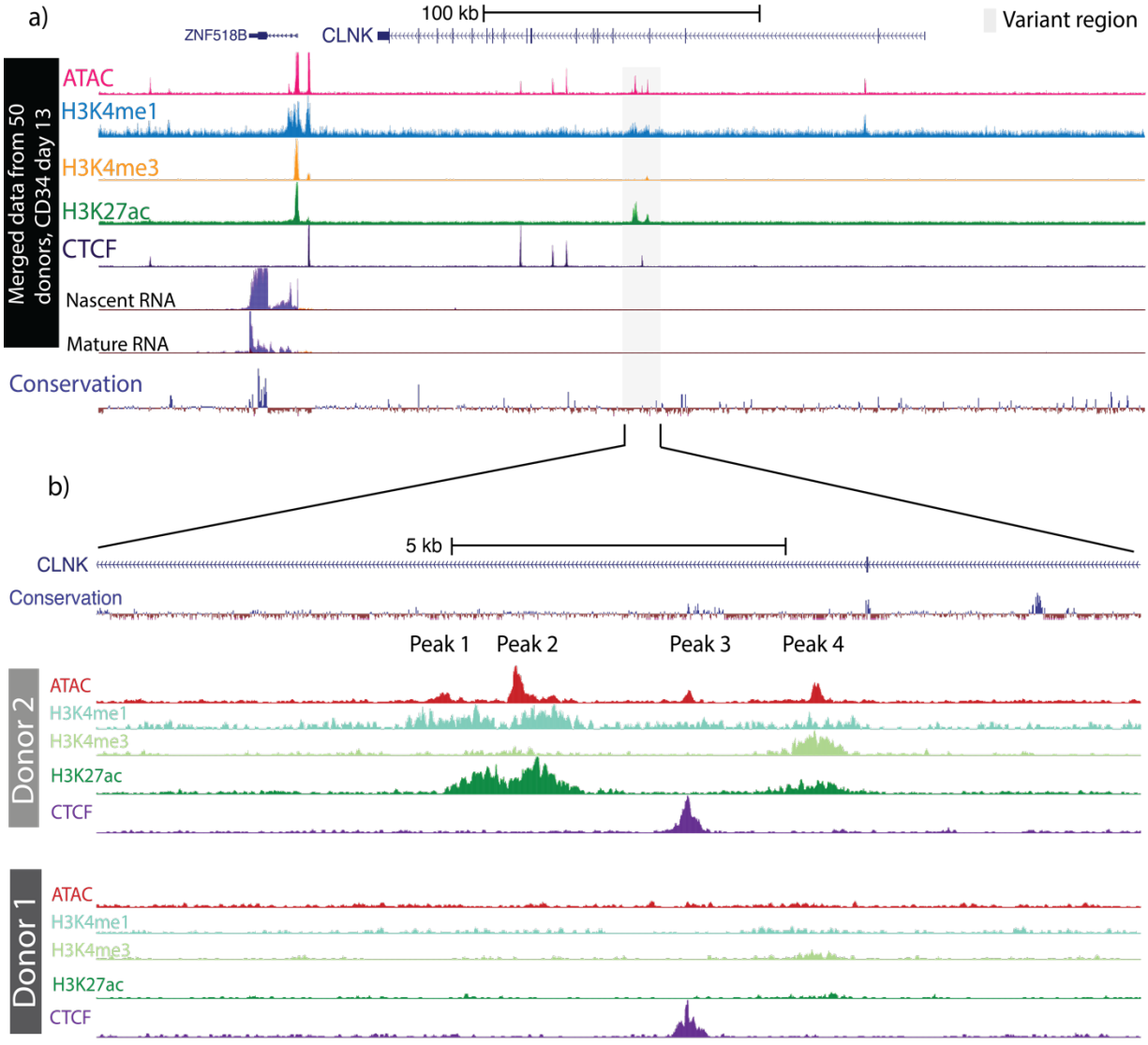
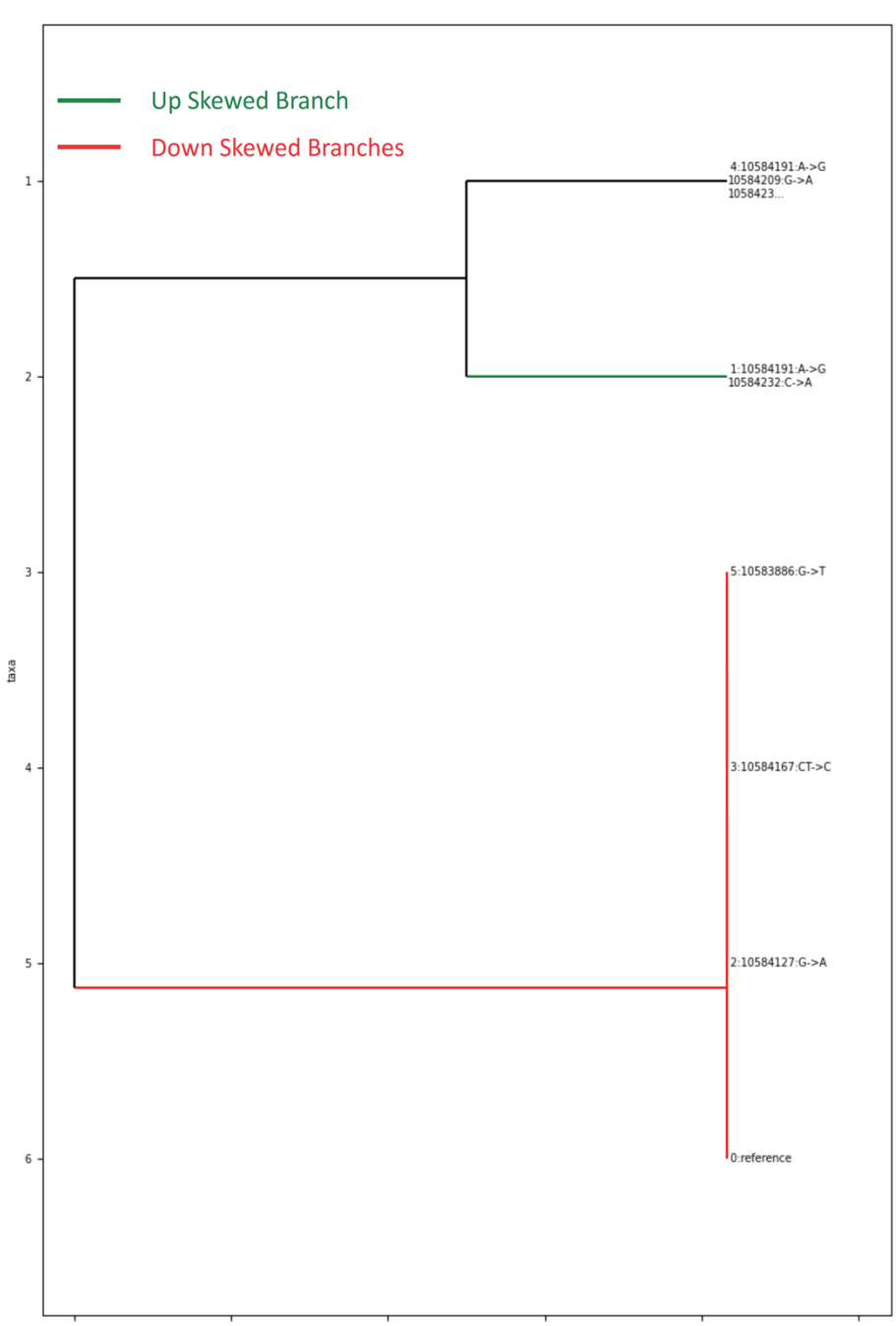


Figure 2.19: Genomic landscape and histone modifications around variable peaks in an intron of *CLNK* occur simultaneously in donors. a) Epigenetic profile at the *CLNK* locus. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Nascent RNA (PolyA-) and Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). Grey highlight depicts the variable regions of peaks. b) Zoom in on the variable peaks with epigenetic data from two donors, one has all peaks present, the other has no ATAC or histone peaks. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). ATAC (red), H3K4me1 (light blue), H3K4me3 (green), H3K27ac (dark blue) and CTCF (purple track). CTCF ChIP seq shows CTCF binding at peak 3 in both donors, even in the absence of an ATAC peak.

Phylogenetic tree of haplotypes in variable peak 4 of *CLNK* peaks chr4:10,584,012-10,584,320



Haplotype Count:

4/100

10/100

1/100

1/100

1/100

81/100

Figure 2.20: Dendrogram of the haplotypes at the fourth peak of the simultaneously arising peaks at *CLNK.* Tree depicts the variety of haplotypes present in 50 individuals at peak 2 depicted in figures 2.18 and 2.19. Total of 98 confident haplotype calls of a possible 100 haplotypes form 6 distinct haplotypes. Red branches depict down skewed branches with no ATAC activity. Single green branch depicts up skewed haplotype that exhibits open chromatin and activity. Counts for each haplotype included.

Variants can also affect CTCF sites with reduced open chromatin reflecting a reduction of CTCF binding

In contrast to the previous examples that show variants affecting enhancers, examples of variation affecting other regulatory elements can also be found. As described in the introduction, CTCF has a central role in gene regulation as a boundary element capable of blocking the interactions between enhancers and genes [14]. It therefore follows that changes in CTCF binding could change enhancer promoter interactions and hence influence gene expression.

As CTCF binding is dictated by sequence motifs, polymorphisms in these motifs can cause variability in binding across individuals (Figure 2.21). Generally, using ATAC-seq to detect changes in open chromatin associated with CTCF binding is challenging as the peaks are usually weaker than those seen at other regulatory elements. However, all the individuals in our biobank also have ChIP-seq data for CTCF from our erythroid cells. When comparing ATAC-seq to the CTCF ChIP-seq at the same region, it is clear to see that a lack of open chromatin also correlates with a lack of CTCF binding. As is typical of CTCF sites they lack other histone modifications at their binding sites, and at this site no marks are present (Figure 2.22b). This clear example of polymorphic binding is found in the 32nd intron of *MYH15*, which encodes Myosin Heavy Chain 15. Myosin is more commonly expressed in muscle tissue, and the lack of RNA transcripts from the erythroid samples show that this gene is inactive in this cell type. Downstream genes, such as *CIP2A*, are being actively transcribed in the erythroid cells, and so the region overall is not inactive (locus TAD structure shown in supplementary Figure 6). This example proves interesting as while CTCF sites are often bound in multiple cell types, and show much less tissue specificity than promoters and enhancers, they can affect the expression of genes over large distances and in a cell type specific manner.

When assessing links between haplotypes and CTCF binding, rs936266 (Chr3:108406767 C > T) is found in several donors (homozygous C/C in 5/50 individuals, heterozygous C/T in 22/50 donors, homozygous T/T in 23/50) with homozygous C/C correlating with no peak (Supplementary Table 1). Because of this association it can be assumed that this is the causal SNP. Furthermore, a previous study has also shown rs936266 to be associated with altered CTCF binding [157]. The haplotypes at this region are varied, with two distinct branches of up and down skewed haplotypes (Figure 2.23). This makes assigning this region as a loss or gain of function ambiguous. The conservation around the peak is strong at the *MYH15* introns, and there is a weak signal adjacent to the peak. Using Homer motif analysis [158] a CTCF binding site can be seen under rs936266.

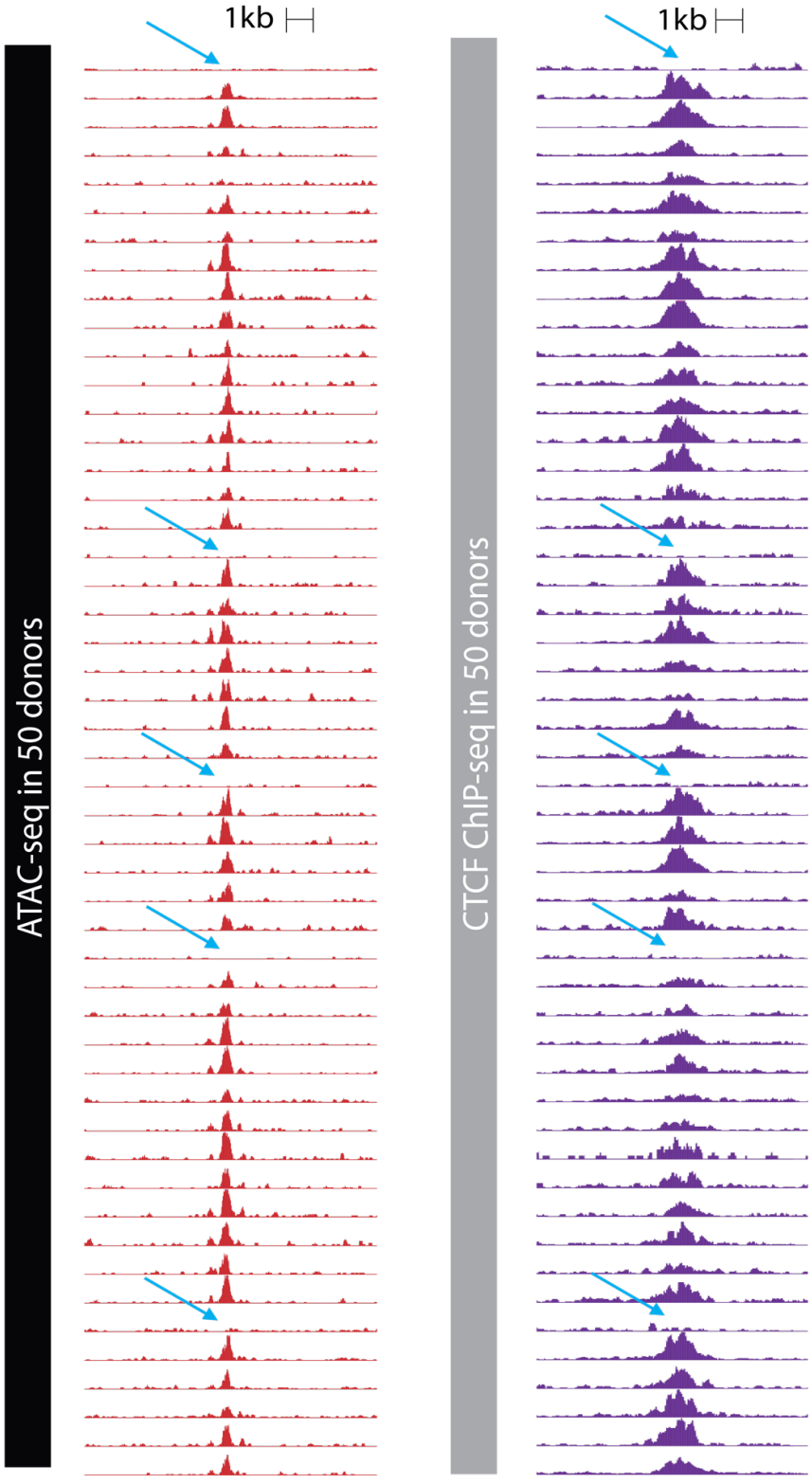


Figure 2.21: Variation at a CTCF site. ATAC and CTCF ChIP-seq from 50 individuals at the same genomic location. Blue arrows indicate individuals who at this region show closed chromatin which in turn also shows a lack of CTCF binding.

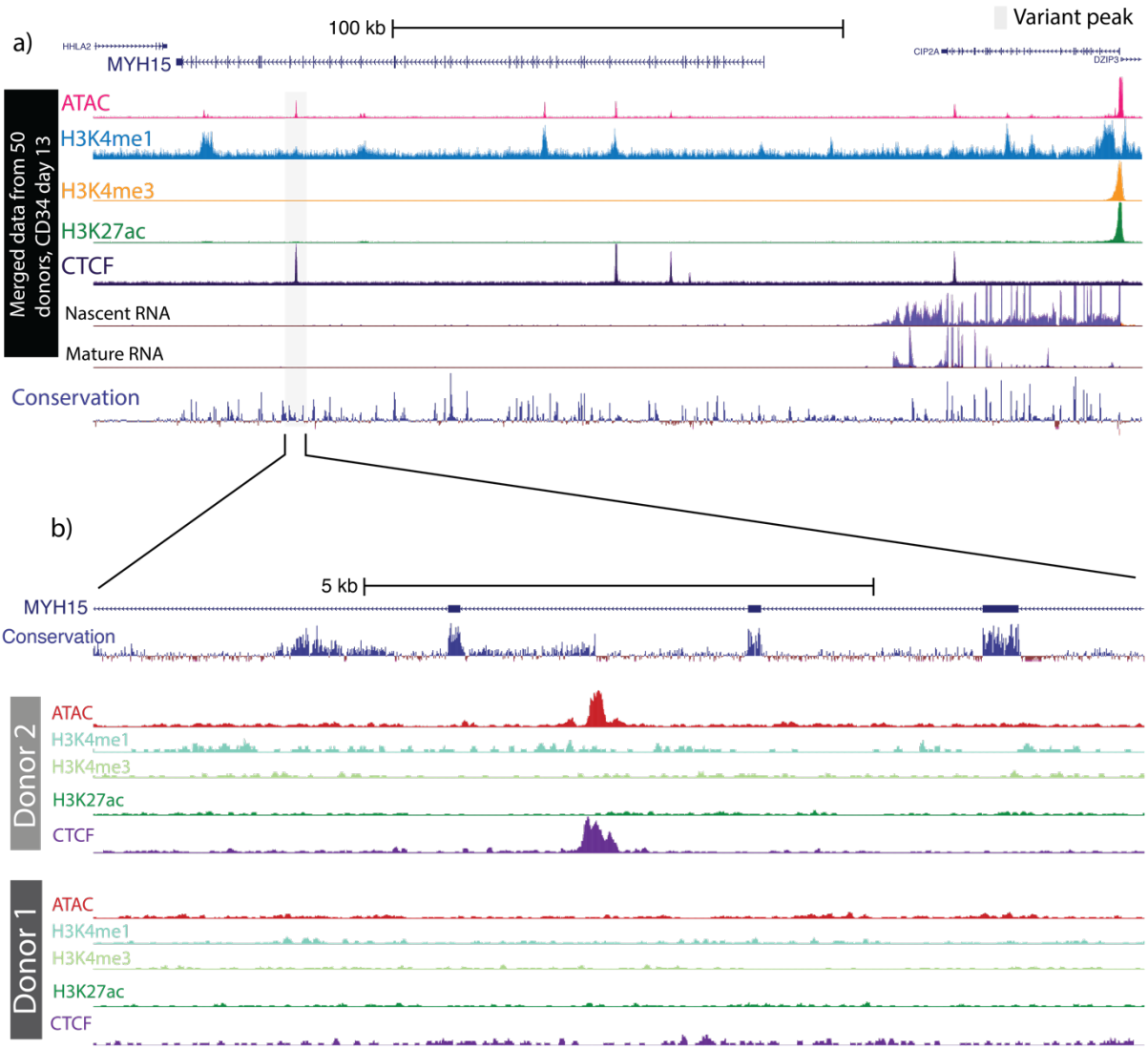


Figure 2.22: Genomic landscape and histone modifications around a variable CTCF site in an intron of *MYH15*. a) Epigenetic profile at the *MYH15* locus. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Nascent RNA (PolyA-) and Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). Grey highlight depicts the variable CTCF peak. b) Zoom in on the variable peak with epigenetic data from two donors, one has ATAC and CTCF ChIP peaks present, the other has no ATAC or CTCF peaks. CTCF peak is not associated with any other histone modifications. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). ATAC (red), H3K4me1 (light blue), H3K4me3 (green), H3K27ac (dark blue) and CTCF (purple track).

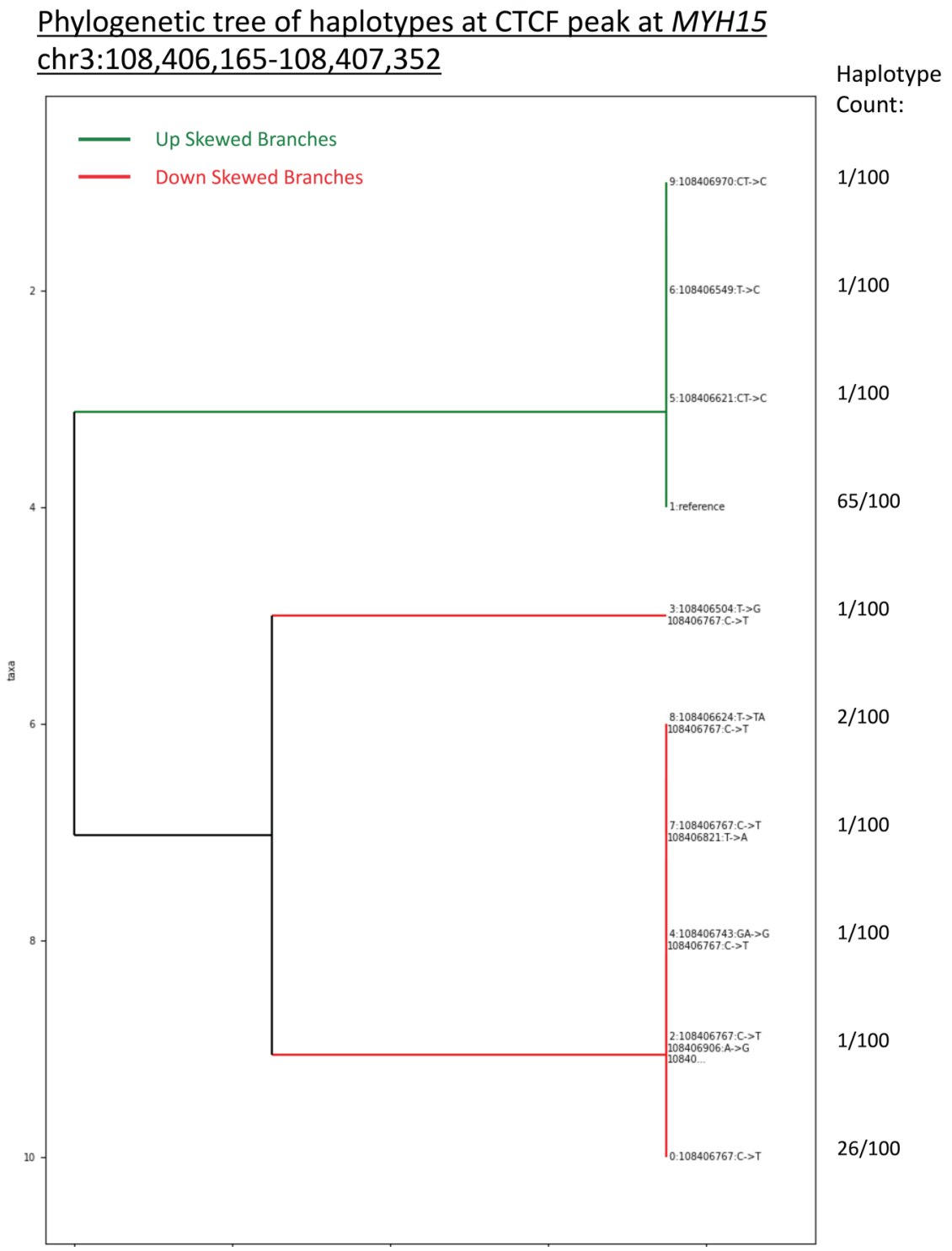


Figure 2.23: Dendrogram of the haplotypes at the variable CTCF peak at *MYH15*. Tree depicts the variety of haplotypes present in 50 individuals at a variable CTCF binding site depicted in figures 2.21 and 2.22. Total of 100 confident haplotype calls of a possible 100 haplotypes form 10 distinct haplotypes. Red branches depict down skewed branches with no ATAC activity. Green branches depict up skewed haplotypes that exhibit open chromatin and activity. Counts for each haplotype included.

A deletion variant relative to the reference genome creates a promoter-like region associated with active transcription

Due to the matched mature and nascent RNA-seq data within our dataset we can also detect variation that causes the generation of novel non-coding transcripts. An ATAC peak present in multiple donors also shows active transcription marks and acts as a highly polymorphic active promoter-like element that produces nascent and stable transcripts (Figures 2.24). The extent of the novel transcript is extremely large, in excess of 80kb in length and so considerably larger that RNAs associated with enhancer elements (eRNA), which are typically around 150-500bp in length [159]. The peak is found in a generally inactive region, with the closest gene being the inactive *FERMT1*, 140kb away. Genes 250kb further upstream show active transcription (Figure 2.25) (locus TAD structure shown in supplementary Figure 7). H3K4me3 marks at the peak suggest promoter activity, and corroborate the peak as an active, transcribing promoter. However, it is known that enhancer elements can also act as promoter-like elements driving stable non-coding transcription [160, 161] so it cannot be assumed that this element lacks enhancer activity.

This peak is unique to a single haplotype shared by the individuals exhibiting the novel peak and transcripts (Figure 2.26). At this site the variant rs71182124 is a 6 bp deletion (GCAGACC > G) relative to the reference sequence (Supplementary Table 1). There are several peaks of conservation within the region, but none over the peak. Conservation also does not appear to overlap with the introns of the stable transcript. It is presumed that the deletion of 6 bps of DNA sequence creates a *de novo* binding site for a TF active in these cells, which cooperates with other potential TF binding site to form an active element. However, motif analysis of the resultant sequence did not convincingly highlight any known erythroid TF motifs.

In summary, transcriptional activity and open chromatin is only associated with a single haplotype containing the 6bp deletion, while all other haplotypes are inactive. The region lacks conservation and together with its unique haplotype specificity this strongly suggests a gain of function mechanism. While at present it is unclear whether this is a *de novo* promoter or an element with regulatory activity or even if the transcript produced has any cellular effect, it shows the highly polymorphic and complex variation found in normal individuals. Also as this variant is relatively rare (MAF of 0.14) it also

highlights the need for deeper sampling of the population using such epigenomic and transcriptomic approaches to understand normal human variation.

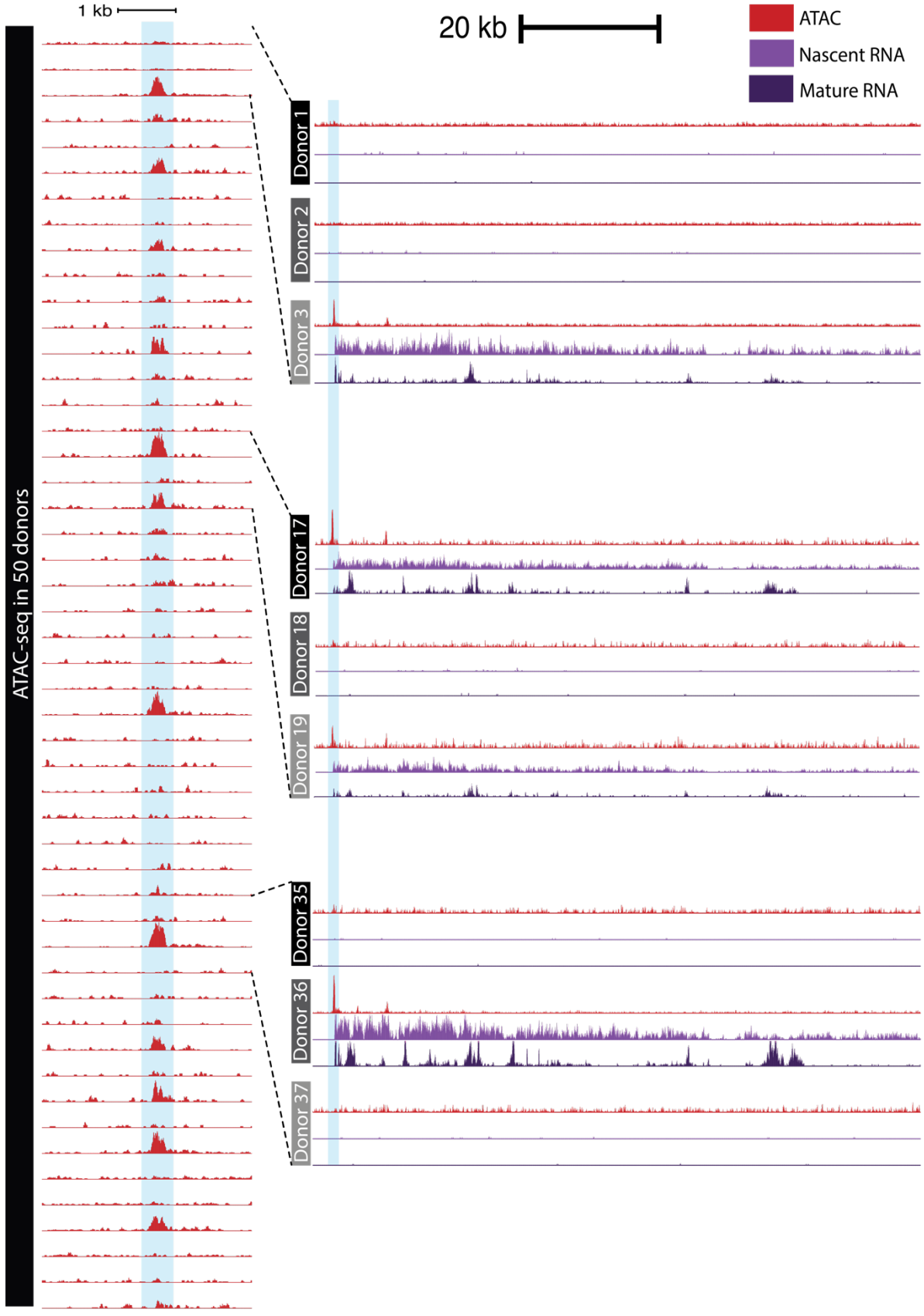

1 kb
20 kb
ATAC
Nascent RNA
Mature RNA
ATAC-seq in 50 donors
Donor 1
Donor 2
Donor 3
Donor 17
Donor 18
Donor 19
Donor 35
Donor 36
Donor 37

Figure 2.24: Variation at a site downstream of *FERMT1* is associated with novel active transcription. Left hand tracks depict ATAC (red) from 50 individuals at the same genomic location upstream of *FERMT1* and blue highlight shows variation amongst the individuals. Blue highlight in right hand tracks depicts the same differential ATAC peak along with a selection of donors and their nascent RNA levels (light purple) and mature RNA levels (dark purple). Examples of both active and inactive regions are shown, to highlight the correlation between the open chromatin and the active transcription.

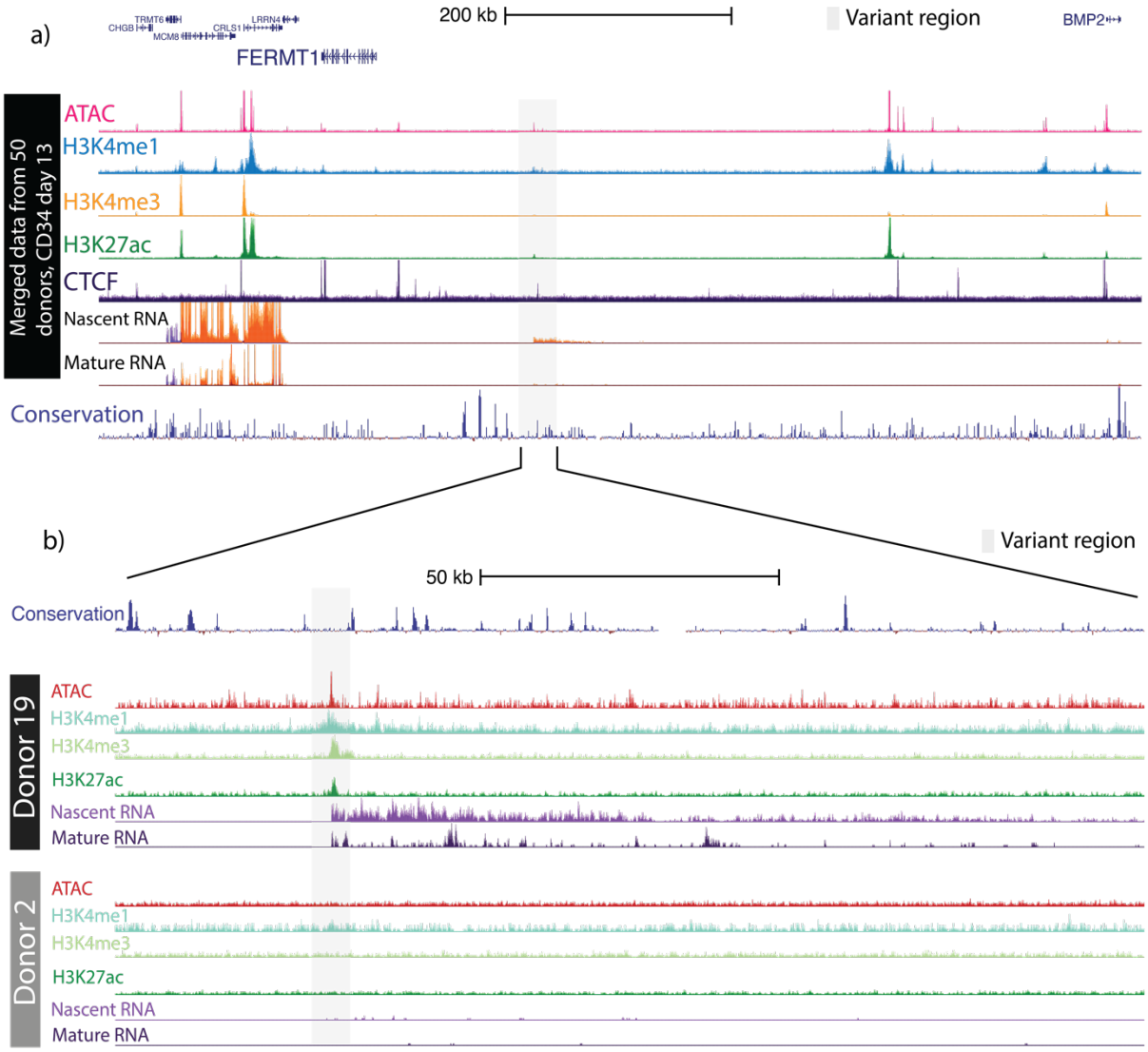


Figure 2.25: Genomic landscape and histone modifications around an active ATAC peak associated with novel transcription. a) Epigenetic profile at the *FERMT1* locus. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Nascent RNA (PolyA-) and Mature RNA (PolyA+) combined from 50 donors, with orange signal

denoting forward transcription and purple signal denoting reverse transcription. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). Grey highlight depicts the variable peak associated with the novel transcription. b) Zoom in on the variable peak with epigenetic data from two donors, one has ATAC peaks associated with the production of nascent and mature RNA, the other has no ATAC or RNA activity. Conservation track of PhyloP data (dark blue for conserved, red for less conserved that median). ATAC (red), H3K4me1 (light blue), H3K4me3 (green) and H3K27ac (dark blue), Nascent RNA (PolyA-) (purple) and Mature RNA (PolyA+) (dark purple). Donor with open chromatin site shows histone modifications and active transcription, other donor shows no peak and no active transcription.

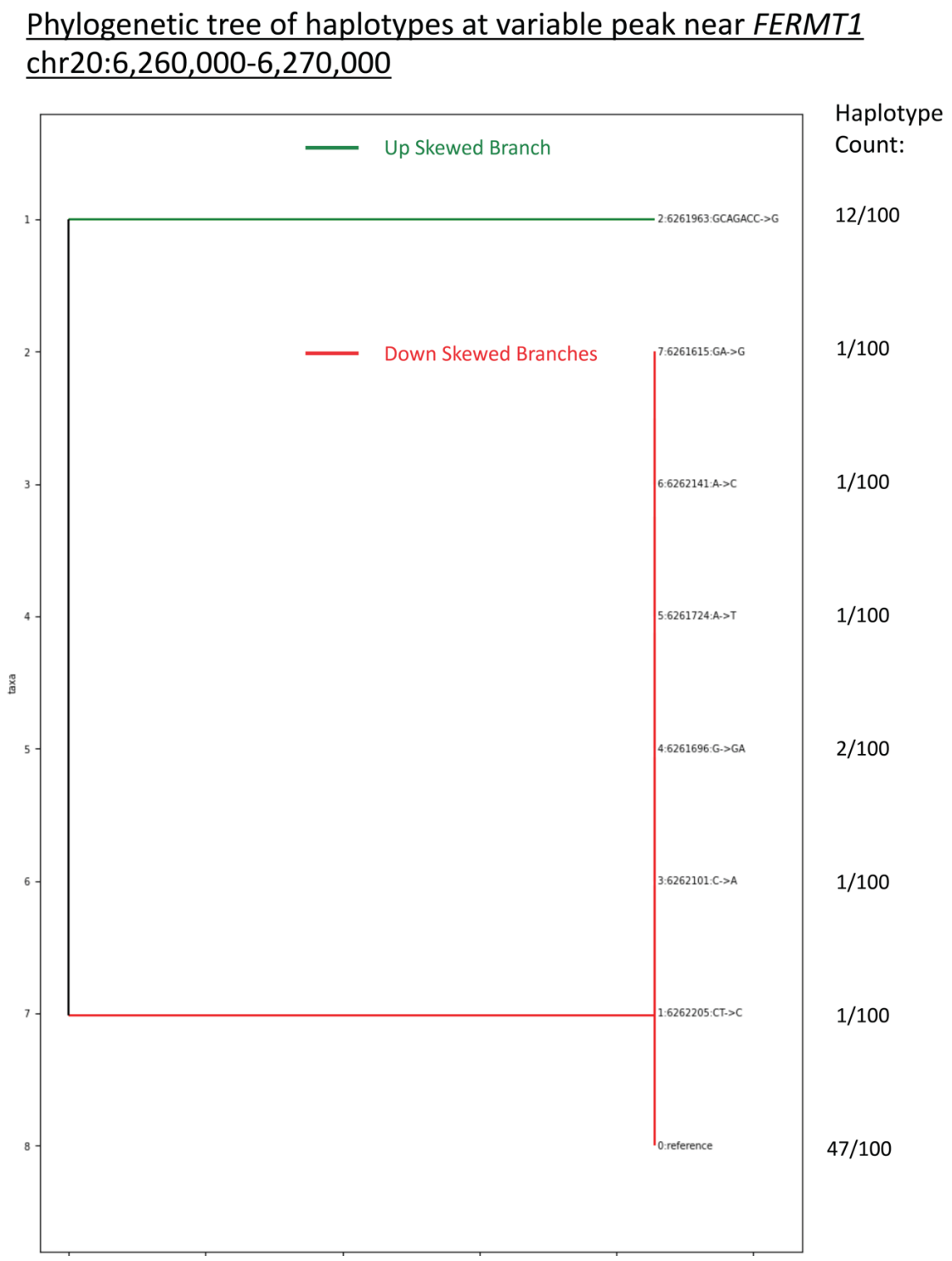


Figure 2.26: Dendrogram of the haplotypes at a site of novel transcription downstream of *FERMT1*. Tree depicts the variety of haplotypes present in 50 individuals at a site of novel transcription depicted in figures 2.24 and 2.25. Total of 66 confident haplotype calls of a possible 100 haplotypes form 8 distinct haplotypes. Red branches depict down skewed branches with no ATAC activity. Single green branch depicts up skewed haplotype that exhibits open chromatin and activity. Counts for each haplotype included.

Discussion

The work in this chapter represents an initial exploration of a very large multidimensional dataset, comprised of phased WGS, with matched genomics and transcriptomics. To fully describe and analyse such a complex dataset will be a long-term collaborative project for the Hughes lab. My goal in this chapter was to start to interpret the types of variation so far found in terms of likely biological mechanism. This exploration will therefore guide ongoing analysis in the lab but has also provided real examples of variation that I can further analyse in the next chapters.

Gains and losses of regulatory activity in the human population

Many examples of SNPs damaging TF binding sites and leading to a loss of function have been documented [11, 43, 137], and we found more such events in this chapter. However, fewer examples of gain of functions currently exist, likely due to ascertainment basis and how likely causal regulatory SNPs are prioritised. To some degree definition of loss or gain of function has been dependent on the reference sequence from which the relative gain or loss is determined.

The reference sequence is an amalgamation of several individuals, compiled during the completion of the human genome project [2] and though it has worked well as a proxy for the human genome, it does not represent a true baseline for any population. Therefore, when considering whether variants are true losses or gains, it is important to consider the genomic context from which they have arisen. In this work we have addressed this by assessing the regulatory landscape of a sizable sampling of the population, to gain some understanding of the haplotype structures and the variance associated with regulatory activity.

A loss of function can be described as damage to a pre-existing, common element that can be seen to reduce the activity of the element. These elements are likely to be more functionally conserved, and are therefore found unaltered in their activity in many haplotypes. In contrast, a gain of function can be described as the opposite, an alteration or creation of an element linked to a specific haplotype. Both of these types of variation may be widespread within a population, either as a result of selection

or genetic drift, but their true frequency across cell types and links to human genetics is still to be uncovered. However, our initial analysis in this chapter suggests that gain of function across a population may be considerably more common than is currently thought (Figure 2.4).

Here I have presented an exploration of a human population to find examples of these cryptic variants. These variants by definition are more elusive, as they may only be present in limited individuals, and therefore their link with activity will only become apparent with deeper sampling. This lies at the heart of why they may be underappreciated due to ascertainment bias. It is common to take a GWAS catalogue and intersect it with epigenetic data as a means to select likely causative variants. However, with these cryptic gains of function, as the variant lies outside of open chromatin in most haplotypes, it will be excluded as an unlikely cause when compared to a generic genomics reference set.

There are two potential methods to discover these cryptic GoFs: deep sampling of a population as used here or machine learning based prediction of function from sequence. Machine learning approaches rely on pre-existing genomic and epigenetic data to predict cell type specific regulatory activity from sequence. As these predictions are sequence based they have shown the potential to predict changes in function when the appropriate variants are included in the sequence [31, 90]. These approaches can predict both damaging and enhancing changes, but such approaches are in their infancy and so require *in vivo* or *in vitro* validation. However, they do represent a promising approach to direct the investigation of potential GoF mechanisms, when combined with recall by genotype or genome editing approaches. However, the analysis in this chapter makes a strong case for the analysis of regulatory variation in large cohorts in the same cell types and lays a ground work for the combinations of data types required to detect and classify types of variation.

Genomic mechanisms affected by variation

Once found, there are a variety of ways in which a sequence change can alter an REs activity. For the gain of function variants described in this chapter, it is therefore important to determine what effect they are having on the local genomic landscape. We can hypothesise that a GoF can be having one of several effects: it can arise and create an open, active region that interacts with local REs and influences gene expression; it can appear and by recruiting certain transcription factors it activates

other regions, which in turn have an effect on gene expression; it can appear and influence the local landscape, by opening either one or several peaks, but without affecting gene expression. To discern these effects, further downstream assays in the appropriate cell type would need to be performed.

In this chapter, I have looked for open chromatin variation amongst ATAC profiles of a population to try and find functional variants. One such example found in an intron of *PTPRK* was only present in a single haplotype and is therefore a novel GoF. However, as open chromatin is a general marker of activity it is also important to utilise broader data sets and to consider the impact of variants on all RE classes. For example, a poorly considered mechanism was highlighted at the *MYH15* locus where variation was linked to polymorphic binding of the structural boundary protein CTCF. As previously mentioned, variation in the binding of CTCF, through its ability to form regulatory domains and block enhancer-promoter interactions has the potential to cause widespread changes in gene expression [21, 162]. This has instigated a more general analysis of the polymorphic binding of CTCF and its links with erythroid genetics within the group, which remains on-going. I have also prioritised this for allelic 3C analysis in the subsequent chapter to investigate the effect on the regulatory interactions of the locus.

I also showed the creation of either a highly transcriptionally active enhancer element or a *de novo* promoter element at the *FERMT1* locus. I will also follow up this example using 3C analysis to determine if this element forms regulatory interactions with surrounding promoters. Additionally, this has also instigated a project in the group leveraging the nascent transcriptomics data to search for further polymorphic sources of transcription across our biobank of individuals. At present no tools exist to quantify changes in nascent transcription in the absence of gene annotation and so this will require the development of novel analysis approaches.

Determining the functional downstream consequences of these mechanisms is likely to be complex. Unlike enhancer elements CTCF elements do not typically form strong 3D interactions with the genes they influence, so the target gene can be more complex to determine. Also, as CTCF elements tend to be less tissue specific their loss or gain may have effects in multiple cell types and on different genes in the locus. *De novo* promoter elements may interfere with the correct regulation of the locus they reside in, or the transcript produced may have a functional effect and so require extensive functional

follow up [146]. However, these complexities aside it is important that the links between variation, genome function, output and human genetics be thoroughly catalogued to direct research in the future.

In the following chapter I will optimise the novel MCC technology for erythroid cells and adapt it to allele specific analysis. I will follow up the exemplar loci described in this chapter to add base pair and allele-specific regulatory interactions to the compendium of genomics data presented in this chapter to further investigate the likely mechanistic effects of regulatory variation described in this chapter.

Acknowledgements

For this chapter I extracted CD34 cells from the two additional donors, cultured these cells and performed downstream ATAC and ChIP assays. I computationally looked for differential ATAC peaks across the donors, and worked with computational members of the Hughes group for subsequent peak exploration. I worked alongside Edwards Sanders for the profiling of each locus.

Caroline Scott provided extensive advice and support as well as reagents for all CD34 cultures used throughout this project, and helped with cell culture during restricted lab time. Data from the biobank individuals was generated by myself and other alumni members of the Hughes lab prior to the start of this DPhil: Damien Downes, Stephanie Carpenter (nee Hill), Priscilla Hirschfeld.

The bioinformatic elements of this chapter were a collaborative effort with several members of the Hughes group. Edward Sanders generated all allelic skew data sets, dendrograms and the meta-analysis of variant distribution. Robert Beagrie provided python lessons. Lance Hentges provided help with LanceOtron peak calling as well as python lessons and support for the differential ATAC peak analysis. Martin Sergeant provided help with data visualisation and filtering in the MLV browser. Jeanne Inchauspe generated the combined ATAC and H3K27ac analysis for differential peak analysis. Jennifer Herrmann helped with motif searches.

Chapter 3 - Using MCC to determine the mechanisms of loss and gain of function variants in late erythroblasts

Introduction

Gene expression levels determine the fate, form and function of a cell. The transcription of genes is controlled by regulatory elements, which can be in close proximity to the gene (promoters) or distal in the linear sequence (enhancers). Distal regulatory elements are thought to come into close physical contact to induce gene expression [7, 163-166]. Therefore, determining the structural conformation of the genome and the precise interactions between genomic regions is important in understanding the genomic elements that control gene expression. Furthermore, the variants such as those found in the previous chapter can change the activity of regulatory elements and alter the structure of chromatin.

Since the first Chromosome Conformation Capture (3C) method was published in 2002 [92], many new proximity ligation assays have been developed from these principles to uncover genome architecture [93, 99, 167]. Originally used on plasmids the foundations of formaldehyde fixation, DNA digestion and ligation has become common in discerning regulatory interactions [168, 169]. Over the last two decades technical advances combined with the falling cost of sequencing have generated deeper and more informative interaction maps.

In this chapter, MCC was used to interrogate regulatory genome structure at high resolution and specificity [99, 100]. This is an oligonucleotide capture based 3C method, based in part on the Capture-C technology [6], that utilises the largely non-sequence specific cutting of MNase along with other modifications of the protocol to generate higher resolution interaction profiles than is possible with restriction enzyme-based methods. Moreover, because of the use of MNase, probes can be designed to anywhere in the genome, enabling easier generation of allele specific interaction profiles.

Due to its use of MNase the positioning of capture probes is not restrained in the same way as with the restriction enzyme based methods. This means that capture probes can be positioned optimally within the open chromatin site of the element under question to precisely detect interactions. This also means probes can lie directly over a variant of interest within these elements. This means that all MCC interaction profiles can typically be split allelically when performed on heterozygous samples.

Allelic split has been described previously with other 3C methods using heterozygous sequence variants to generate separate contact profiles from each allele [170]. However, these allele-specific 3C experiments have typically been very limited due to the requirement for the heterozygous variant to lie close to the restriction site being used. Interaction profiles are bioinformatically split based on the detected genotype within the captured and sequenced fragment. Therefore, a variant is required to be within this region to determine the allele of origin. However, the design of probes in restriction enzyme based approaches is much more inflexible, and sites are very infrequently near enough to the element or the variant of interest. At present this approach has been employed sporadically at best.

However, the ability to allelically split interactions is extremely experimentally beneficial. Comparing the activity of one chromosome against the other in the same sample essentially removes intrasample and experimental variability and noise. This means that changes in interactions can be reproducibly detected with much greater sensitivity and interpretability.

Performance of MCC

Like other 3C proximity ligation assays, MCC is performed on a population of formaldehyde fixed cells. Formaldehyde forms covalently bonded links between DNA and the associated nucleosomes and proteins in chromatin. This fixes chromatin structure and allows it to be maintained during the subsequent steps in the protocol. Following fixation, cells are permeabilised with digitonin instead of harsher conditions and detergents typically used in other 3C protocols. This enables MNase and other enzymes access to nuclear chromatin whilst minimising the disruption of nuclear structure.

The fixed DNA in the chromatin is then fragmented in a largely sequence independent manner with a titrated volume of MNase to ensure optimum cutting: too little will result in large fragments that yield

limited, less specific data, whilst too much will over digest the DNA. The prevention of over-digestion in MCC is critical to avoid trimming the DNA to such a degree that it inhibits the subsequent ligation reactions in the protocol.

Once digested, samples are cooled and EGTA is added to chelate the Ca^{+} ions essential for MNase activity, thus stopping the MNase reaction. Next, DNA fragments are ligated with ligase to form concatemers that reflect the 3D shape. As DNA is organised around nucleosomes, MNase concentration is titrated to achieve an optimal fragment size of ~180bp. This ensures that the DNA around the nucleosomes have a linker tail that can ligate to proximal fragments. Since an MNase cut leaves a single base overhang, this stage also relies on the 3' to 5'exonuclease activity of Klenow to trim the overhang, providing a blunt end for ligation. This ensures that any fragments can ligate. This differs from restriction enzyme digestion, where complementary sticky ends are generated. Once ligated, proteinase K digestion is used to degrade the proteins and the formaldehyde linkers, leaving a DNA library with a linear sequence that now reflects proximity in the 3D nucleus (Chapter 1 Figure 1.5). This library is then purified using standard Qiagen column purification.

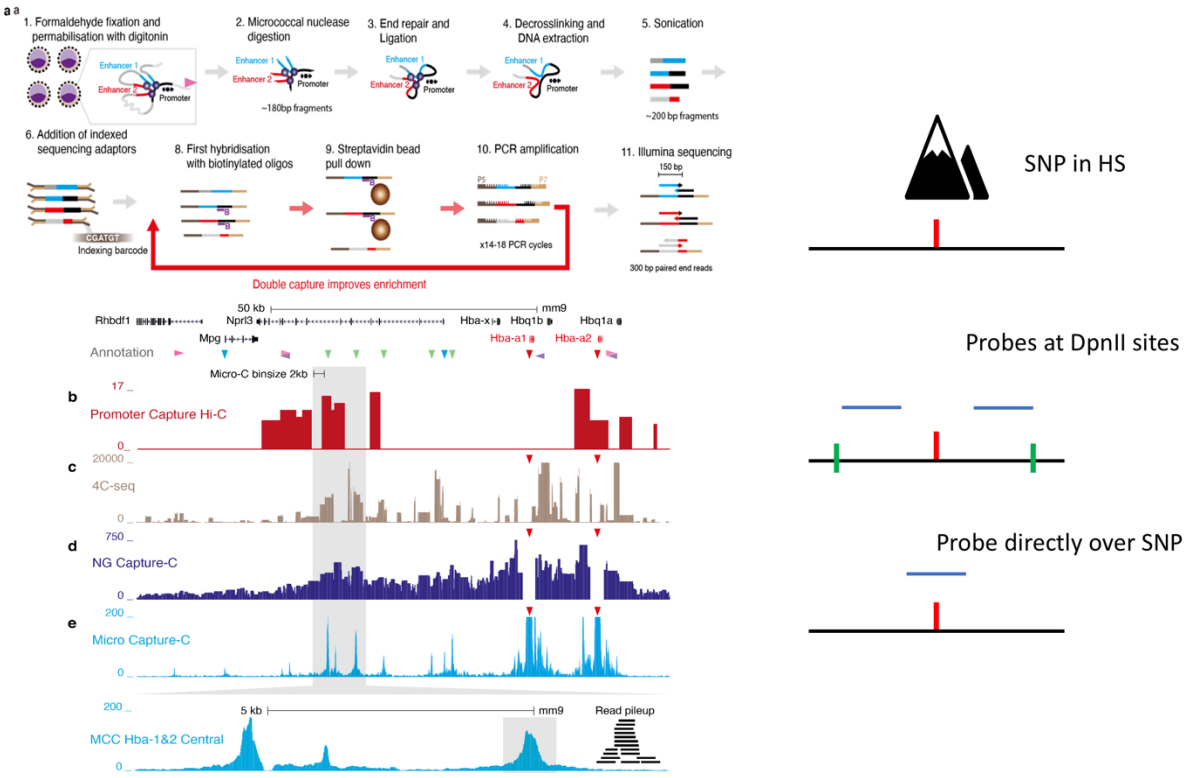


Figure 3.1: MCC overview. a) Overview of the MCC process: Cells are formaldehyde fixed and digested with MNase, ligated, decrosslinked, sonicated, adapters are added and double capture enriches for regions of interest. Compared to other techniques such as b) Promoter Capture Hi-C, c) 4C-seq and d) NG Capture-C, e) MCC produces much higher resolution interaction profiles that precisely overlap with REs. f) When studying variants and designing capture probes, MCC allows for the design of probes directly over SNVs or SNPs. This is because of the MNase which is contrary to 3C that rely of restriction enzymes, at which cut sites determine where probes must be designed to. This means you can design over specific regions or SNPs of interest.

Sonication of libraries generates uniformly sized fragments that are conducive for the flashing of reads after sequencing to completely reconstruct the fragment; reads can be flashed when sequencing data from each read end overlaps in the centre of the read, allowing for more confident base calling [171]. This then enables the subsequent confident calling of ligation junctions, with base pair resolution, within the junction fragments. To achieve this sonication settings are optimised to generate libraries that average 200bp fragments and the reaction is cleaned with Ampure beads. Libraries are then indexed using the NEBultra kit, which end repairs, 5' phosphorylates and adds a dA overhang for the addition of an adaptor sequence. The adapter is ligated as a hairpin structure containing a central uracil, which is then excised. Libraries can then be PCR amplified with library specific index primers. Indexed libraries can then be pooled. In common with the original Capture-C technology, this means that biological replicates or libraries from other samples can be enriched in the same downstream steps. This will minimise technical noise and variation and increase throughput. This pooled indexed library is then captured with biotinylated oligonucleotide probes designed to hybridise to the sequences of interest. For this step, libraries are denatured and hybridised to biotinylated probes; these probes will bind to specific regions of interest. Binding with streptavidin beads will then pull down and enrich for regions of interest, and as with Capture-C, performing a double capture further improves enrichment of signal relative to background. After amplification captured libraries can then be sequenced on Illumina platforms and analysed. As genomic structure is dynamic, at the moment of fixation each individual cell within a population will have a unique interaction profile. By aggregating the interactions detected from each captured fragment across a population of cells an overall representative interaction profile is generated.

The use of MNase as the enzymatic fragmentation agent has a further advantage in the analysis of regulatory variants. It has been long known that MNase preferentially cuts naked DNA in chromatin,

and its preferential cutting of nucleosomal linker regions has been used historically to assay nucleosome positioning [172]. Also similar to DNase I it's cutting is inhibited by the binding of TFs and proteins such as CTCF [173]. Therefore, with deep sequencing MNase can also generate footprinting data analogous to DNase I footprints, that assess the bound state of DNA by factors in chromatin. In the context of MCC data this is performed by compiling the ends of mapped digestion sites in the ligated fragment, this shows precisely where the DNA was cut and therefore where no protein was bound. Due to the capturing enrichment of probe sites in the libraries, these data are especially deep and informative over the MCC capture site. Furthermore, at heterozygous sites this can show differential binding of transcription factors directly over variants. In alleles that also demonstrate differential interactions, it can be hypothesised that the change in transcription factor binding is directly affecting the interactions between regulatory elements.

Overall aims of the chapter

In this chapter I will:

1. Optimise the MCC protocol for use in primary human erythroid cells and generate libraries from the donors introduced in the previous chapters.
2. Assess the effects of the variants introduced in the previous chapter to demonstrate their effect on genome interactions.
3. Utilise allele specific MCC analysis to show the effect of variation across separate alleles within the same cell.
4. Generate MNase footprinting tracks to determine TF binding footprints and their haplotype specific variation.

By using MCC I have assessed the interactions of differential genomic regions including those found in the previous chapter. As these variants were found in our erythroid differentiation system, I have used the same cell type and stage in this chapter to ensure I am capturing from the defined open regions. I have generated libraries from several of the biobank donors to demonstrate a variants ability to alter interactions, both between individuals and between heterozygous alleles. Of the LoFs and

GoFs shown previously, some show interaction profiles with neighbouring REs, whilst others appear to have no contacts. This demonstrates the importance of incorporating interaction data when deciphering the effects of variation on gene regulatory networks and gene expression.

Results

Initial optimisations

Prior to starting this project, MCC had only been performed successfully in mouse cells and human cell lines. Due to the novelty of the protocol the successful production of informative libraries was known to be highly sporadic, especially when changing cell type, even for the Davies group who developed the protocol. Therefore, a critical aspect of my DPhil was to help optimise the protocol, in collaboration with the Davies Lab, and adapt the protocol for the generation of MCC libraries from primary erythroid cells.

Due to the techniques requirement of large cell numbers and the complexity and expense of our erythroid differentiation system, I chose to do the initial optimisations on Human Umbilical Cord Blood-Derived Erythroid Progenitor (HUDEP) cells. These are an immortalised human erythroid cell line that are used as a model for the cells generated by the CD34 differentiation system and have many properties in common with the primary cells [106]. I then moved on to primary CD34 derived cell samples from a generic blood cone sample (blood transfusion service) once the protocol and conditions were generating consistent results. I used the cone samples for initial optimisation due to logistical concerns.

During these initial experiments, my main focuses were ascertaining the MNase concentrations for optimal digestions, testing different ligation conditions and improving the overall yield of libraries. The first conclusion from these tests was the importance of getting an optimal digestion, that balanced almost complete chromatin fragmentation with the retention of sufficient inter-nucleosome DNA linker length, as this proved crucial for subsequent ligation efficiency and yield. To titrate the MNase concentration, $1x10^7$ fixed cells were divided into three reactions and each digested with a different concentration of MNase, then processed and library size assessed. This therefore required several

libraries of large cell numbers. Figure 3.2 shows exemplary tapestation reports of a well digested and ligated library, in which the digested sample shows standard nucleosomal digestion banding. The smallest peak being ~180bp is indicative of the 143bp of DNA wrapped around a nucleosome plus some intact linker DNA also. This linker DNA enables more efficient ligation to other nucleosome bound DNA, and a successfully ligated sample is demonstrated by a shifted profile, with a reduction in the amount of DNA in the smallest peak, and more yield in the di-, tri- and larger nucleosomal bands. These tapestation profiles are indicative of a successful library preparation, and are used as a quality control throughout this chapter.

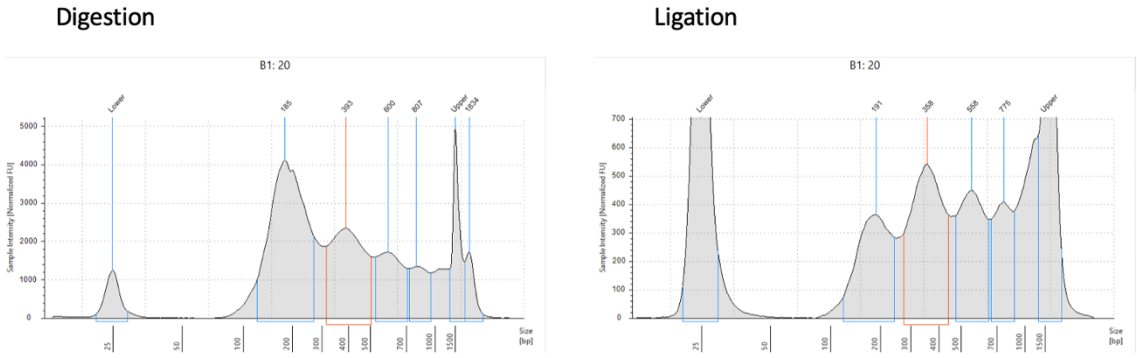


Figure 3.2: Tapestation profiles showing DNA patterns of a well digested and ligated MCC library.
Tapestation profiles show quantitative fragment size distribution of a DNA library. This library was digested with 20 units of MNase, and each profile has a lower and upper marker used to determine fragment size. The lefthand digested profile shows the highest peak at 185bp, with further stereotypical di- and tri-nucleosomal banding at 393bp and 600bp. These larger fragment fractions as less concentrated that the 185bp peak. The righthand profile shows the ligation of the lefthand digested sample. Nucleosomal banding is still present, but the biggest peak is now at 358bp, showing that the DNA has ligated successfully and the average fragment size has increased.

The first ligation conditions tested incorporated the inclusion of a wash after the Klenow reaction, and also changing the temperature of the ligation reaction. Prior to ligation, Klenow's exonuclease activity removes the dA overhang to allow for blunt end ligation, and is optimal at 37°C. Like many other enzymes, ligase works optimally at 37°C, however, to encourage close range ligations of *cis* interactions, ligation reaction temperatures can be lowered (≥4°C) and the ligation time extended. Lowering the temperature will reduce the overall energy within the system, reducing molecular movement and keeping fragment ends together longer for the reaction to occur. Usually lower

temperatures also help to stabilise sticky ends, which is not necessary here. There is therefore a balancing act to get the correct ligation conditions for the reaction.

These first optimisations failed as the DNA continued to digest prior to EGTA addition and washes. This can be seen as the failed ligations have smaller fragments than the digestions (Figure 3.3).

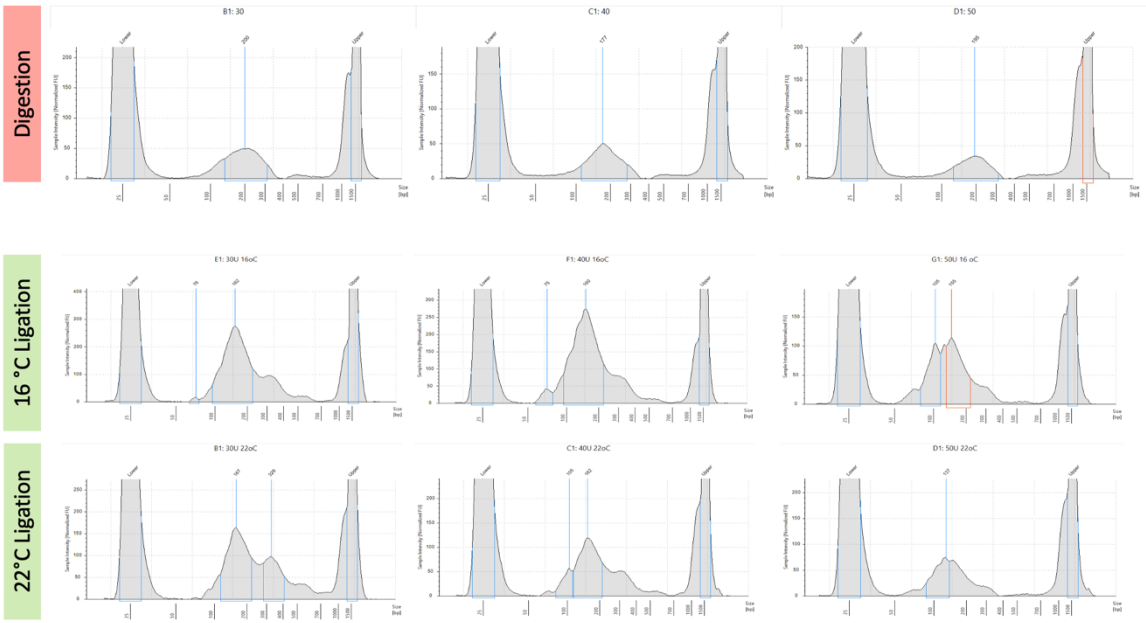


Figure 3.3: Initial experiments showed digestion but failed ligations.
Tapestation profiles showing digestions (red row) with varying amounts of MNase, 30, 40 and 50 units of MNase corresponded to peaks of 200bp, 177bp and 195bp respectively. Ligations (green rows) at 16°C (upper) and 22°C (lower) show a failed ligation reaction with insufficient shift towards larger peaks. Despite some large fragments being present, the largest peaks remain and show more digestion prior to the reaction with an smaller average fragment size, indicating that the digestion reaction was not stopped.

MCC uses programmable benchtop thermomixers, with the standard ligation reaction at 37°C for 2 hours, and then 16°C for 8 hours. This stage includes a 37°C incubation for the Klenow polymerase activity, but as this exonuclease role is only required until the ligation has happened, we wanted to see if it was over digesting the linkers before the ligase had a chance to work. I therefore tried adding a wash step to remove the Klenow before adding the ligase. There was still over digestion in the libraries and so the ligations failed. This experiment was repeated with the additional change of ligation reaction temperatures (16°C and 22°C). Though minimal ligation was seen in these reactions,

and neither of the samples produced workable libraries, both ligation profiles were comparable, suggesting ligation temperature was not an issue (Figure 3.4).

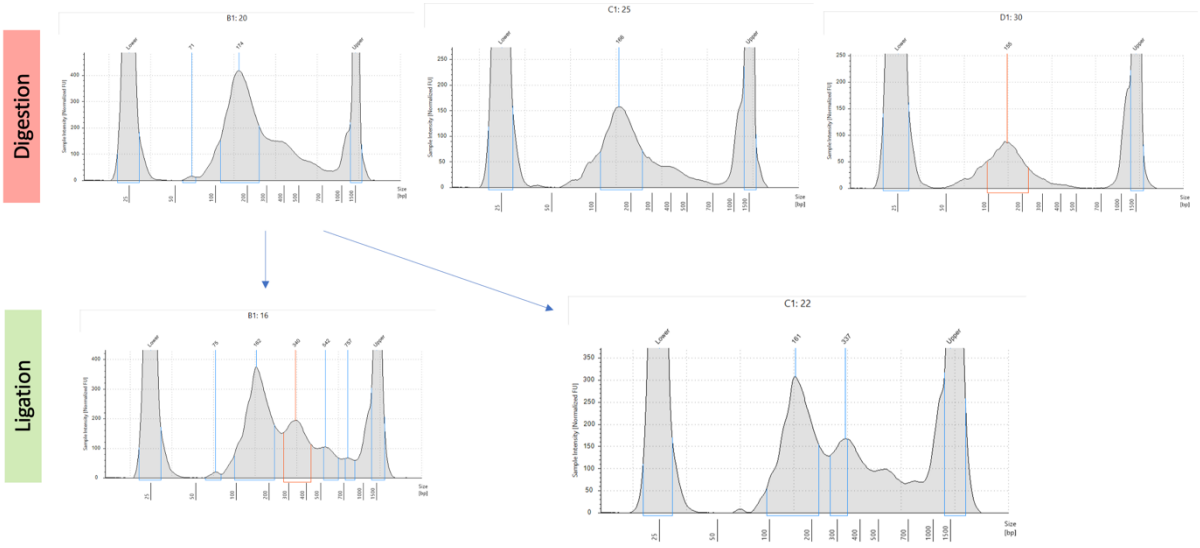


Figure 3.4: Bad digestions lead to bad ligations, regardless of temperature.
Tapestation profiles for a series of digestions (red row) with 20, 25 and 30 units of MNase. Digestion with 20 units was an acceptable digestion, if slightly over digested, 25 and 30 were not carried forward. Ligations (green row) at 16°C (left) and 22°C (right) show some ligation, but still suboptimal, with evident more digestion.

The next condition I tested was the speed and frequency at which libraries were shaken during the digestion. It had been anecdotally suggested that using a different thermomixer could result in failed libraries. As the temperature should be a consistent factor across machines, I therefore wanted to alter the shaking to see if this would have an effect. The protocol used horizontal shaking at 500rpm during the MNase digestion and 550rpm for the ligation. I incubated my libraries with two further conditions: intermittent 500rpm for 15 seconds every 10 minutes and no shaking. The point of shaking is to keep the solution homogenous so that MNase is not static and preferentially cutting the DNA within some nuclei or localised areas and not others. I kept post spin supernatants to see whether there was DNA leaking from the cells (Table 3.1). Furthermore, from the tapestation analysis it could be seen that material was being lost in the shaking samples (Figure 3.5). The supernatant DNA

concentrations did not show any difference between the conditions, but the yields at both the digested and ligated stages did show a reduction.

In conclusion, the changes to the reagents and temperatures used during the ligation reaction had little to no effect on a successful library generation. Instead, after changing the shaking conditions I was able to generate high yield, well ligated libraries. Though this shows the ligation conditions were effective, it is unclear whether combining the new shaking conditions with other reagent or temperature changes would also yield better results. However, since the technique was now producing good quality libraries, I continued with my experiments.

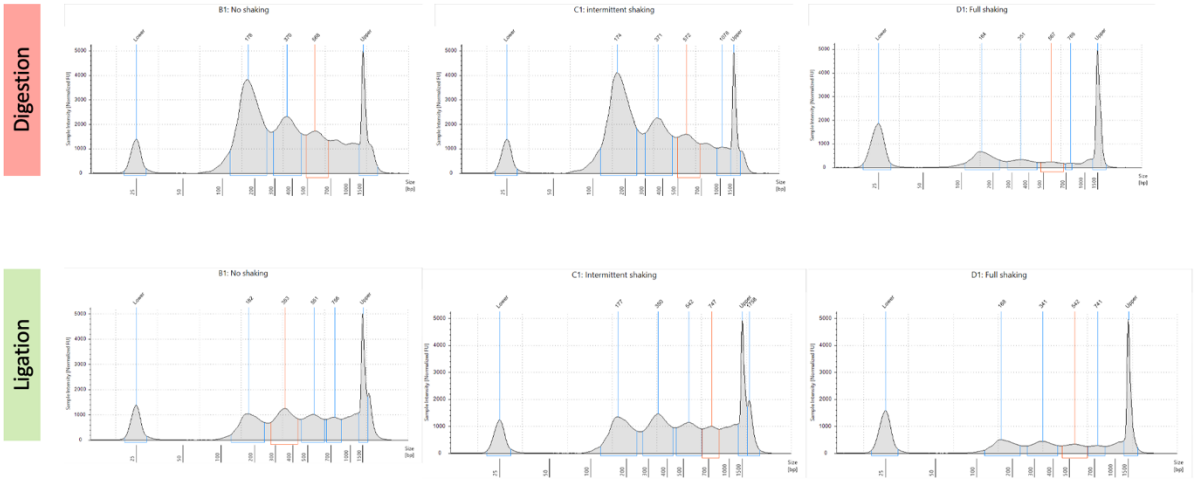


Figure 3.5: Shaking MCC libraries during digestion caused loss of material.
Tapestations of digestions (upper red row) and corresponding ligations (lower green row) with varying shaking conditions. Left was not shaken, middle was shaken intermittently, right was shaken constantly. Loss of DNA yield can be seen from the shaken sample.

Stage	Condition	Concentratio ng/ul	volume ul	DNA ug
Digested control	No shaking	104	20	2.1
	Intermittent	111	20	2.2
	Full shaking	45.5	20	0.9
Digested control supernatant	No shaking	20.8	20	0.4
	Intermittent	18.5	20	0.4
	Full shaking	19.8	20	0.4
Ligated sample	No shaking	59.2	150	8.9
	Intermittent	58.5	150	8.8
	Full shaking	49.2	150	7.4
Post ligation supernatant	No shaking	9.49	20	0.2
	Intermittent	7.24	20	0.1
	Full shaking	10.4	20	0.2

Table 3.1: DNA yields of supernatants and samples from shaking experiment.

Once the protocol was working on the HUDEP cells I moved onto erythroblast samples derived by CD34 cell differentiation. I successfully made three biological replicate libraries from late erythroblasts derived from a blood cone sample (CD34 differentiation to day 10) (Figure 3.6). These were then indexed with two indices per library to increase complexity in the final pool; PCR duplicates are removed during analysis and so by increasing the amount of indices assigned to a library, more real interactions are kept.

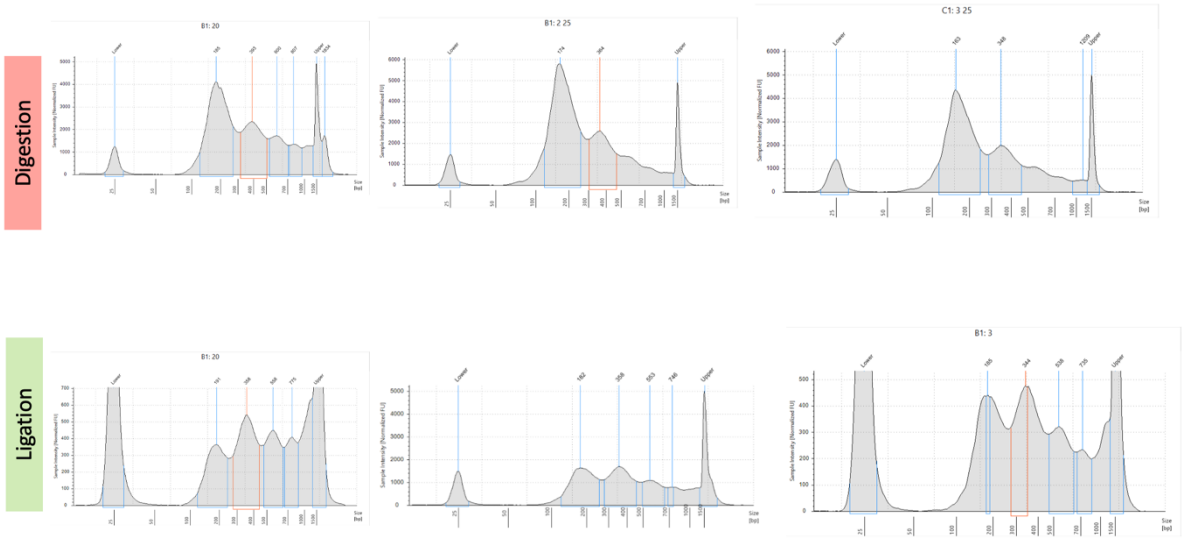


Figure 3.6: Successful MCC libraries of late erythroblasts derived from a blood cone sample.
Tapestations showing three replicates of digestions (upper red row) and ligations (lower green row) of successful MCC libraries from a CD34 cone sample.

Designing MCC probes

To test the quality of the erythroblast MCC libraries generated, I designed a panel of probes for use in a capture (MCC probe set 1). This first set of probes were designed to a range of regions, including previously known variant regions, differential ATAC regions described in Chapter 2, and loci with active erythroid genes with known enhancer interactions as controls (Methods Table 6.3). For this design, I manually selected 120bp sequences in the centre of peaks of interest and ran BLAT to determine how unique the sequence was, and whether there was likely to be any off-target capture. Additionally, as BLAT is prefiltered to remove low complexity regions the sequences were also scored for low complexity repeat regions.

Unique probes are essential as they enrich for regions of interest, and if there is duplicity they will pull down similar, non-specific regions, which may swamp real signals during sequencing and data analysis. Due to the flexibility of the probe design in MCC and the fact that they can be positioned directly over functional elements, this is less of a problem in MCC probe design than in other 3C methods as regulatory regions are generally less repetitive. However, even a single rogue probe can completely swamp the signal in a capture based 3C experiment, so it is still critical to check. As several regions were derived from the differential peak analysis, they were in fact repeat regions with high similarity scores, and therefore excluded from the design.

Interaction profiles at the HBA locus confirms library quality

As a highly characterised locus with well-defined REs and interactions, the *HBA* locus is ideal for checking the quality of a capture experiment and the libraries used. Figure 3.7 shows the interactions to the *HBA* promoters, with the expected signals from all four enhancers and CTCF sites. Since the two *HBA* genes are duplicated, the probes will enrich for signal from both, and this is shown as two capture points in the figure and the profile represents their combined interaction profiles.

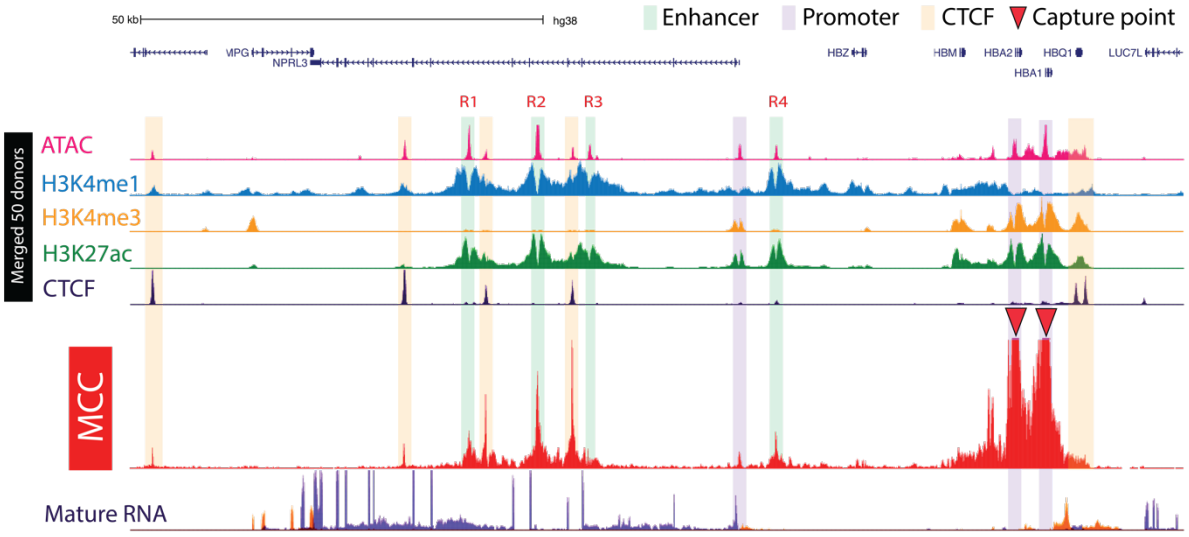


Figure 3.7: MCC interactions from cone sample at *HBA* promoter showing expected profile.
Genomic landscape at *HBA* with MCC profile from the promoters shows expected interaction profile between REs. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Due to the duplication of the alpha genes, the aligned reads are split between the two promoters.

Using MCC data to determine the genetics of an uncharacterised donor

This initial set of probes were designed based on differential ATAC peaks found in the primary three donors. As this first MCC sample was derived from an undefined and uncharacterised blood donor, I initially did not have any genetic data or SNP information for their genotype. However, due to the sequence enrichment in the MCC protocol, deep sequencing data from over the probe sites from this individual was obtained. Therefore sequence variation in this individual could be determined from the BAM files of the MCC experiment itself. This could either be done bioinformatically with samtools [174] mpileup or manually by direct visualisation of the BAM files in the UCSC browser (Supplementary Figures 8 and 9). BAM files are binary alignment files of individual reads mapped to the best matched regions in the reference genome. As such, when reads contain bases that do not fully match the reference this will be apparent in the BAMs. When visualised on platforms such as UCSC, the variants in the sequenced material will be displayed as contrasting to the reference.

Using a more computational approach, samtools mpileup can annotate all the base calls at a certain region and generate a file of this information. When a specific coordinate is then searched for within this file, the likely genotype or potential variation can be determined. Consequently, I was able to computationally screen all the capture sites for likely variation and quickly manually confirm the existence of the variants. Both homozygous and heterozygous variants were determined in a subset of capture site reads, and so this provided me with a list of regions to attempt to bioinformatically split these sites allelically for the first time.

Utilising heterozygosity within capture regions to generate allele specific MCC profiles

Because SNPs can alter regulatory elements, heterozygous alleles can be compared to determine the effect of specific SNPs. By scanning through the capture sites, I was able to find a list of regions that were heterozygous in this individual. Although the majority of these SNPs had no discernible differences in interactions between the two alleles and so were likely neutral, two regions had pronounced changes.

Allele specific tracks utilise a heterozygous SNP within the region captured by a probe. Due to the MCC probe design process, probes can intentionally be designed over regions with known SNPs. Standard 3C data from diploid cells, including the MCC track shown in Figure 3.7, depict the interactions coming from both alleles. As these techniques are population-based assays, each read will have originated from an individual allele from an individual cell, and so the amassed data will therefore represent how common these interactions are across the whole population at the time of fixation. However, it has previously been shown using the NG Capture-C technologies that 3D genome interactions occur predominantly in *cis* and that interactions between two different alleles occur at the same background rate as the rest of the genome. Hence interactions between elements on different alleles are almost indetectable compared to the frequency of interactions on the same allele [6, 175]. Therefore, if the sequences over the capture regions can be assigned to a specific allele using the presence of a distinguishing variant, the interaction profiles derived from the ligation partners of these fragments in the 3C library will represent the interaction profile of that specific allele.

Figure 3.8 represents how this is achieved. Because every read in the standard profile will have a certain amount of the probe sequence within it, a proportion of the sequenced reads will contain the SNP. For a standard profile, this slight deviation is not enough to influence the alignment, and thus all the reads from both alleles make it into the final data. However, bioinformatically reads can be separated using the known SNP present in the capture site, creating two separate profiles, for the reference and alternate alleles. These two new data sets will be distinguished by the base present at the SNP location, and therefore which allele the reads originated from.

SNPs may affect the activity of a regulatory element by altering transcription factor binding dynamics. This, in turn, may alter the interaction frequency between two or more regulatory elements in a locus, ultimately changing gene activity and transcriptional output. If a SNP alters the interaction of an enhancer with a promoter, this could be detected in MCC data as a change in peak strength. Where a probe is designed across a heterozygous SNP, the reads from each allele can be separately analysed, allowing direct measurement of the effect of that SNP on interaction frequencies (Figure3.8).

Allele specific 3C has been achieved before with other methods but comes with restrictions [170]. Due to the digestion of these assays relying on restriction enzymes, capture probes for these techniques are designed to the edges of these resultant fragments. This means that a SNP must be in this precise region in order to be used for allelic skew analysis. MCC probes however can be designed anywhere in the genome, and so are not limited in this way.

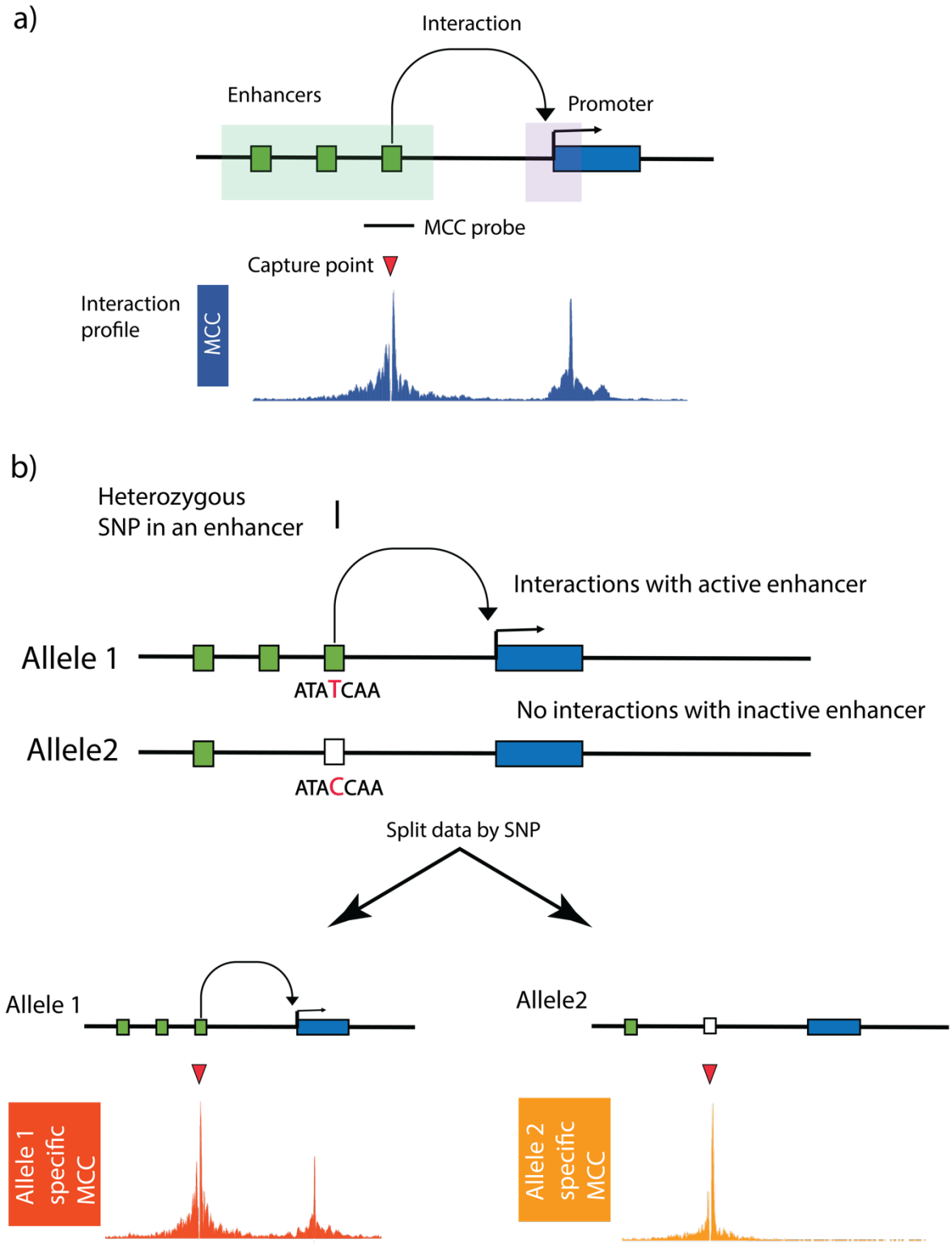


Figure 3.8: Schematic of allelic MCC track generation.
a) Standard genomic assays are visualised as a single chromosome. For example, a standard MCC track at an active region where an enhancer is interacting with a promoter. A probe designed to the enhancer (red arrow)

will pull down these interactions, and these can be displayed as a UCSC track. b) If a heterozygous SNP is present, for this example in an enhancer, and alters the activity, each allele will be behaving differently, one with promoter interactions and one without. This would be visualised as a single track as in a). However, bioinformatically splitting this data by the base information at the SNP during analysis generates distinct tracks, each reflecting the activity at that allele. Here, the first allele was contributing all the interaction data whereas the second allele shows no interactions, indicating that this SNP alters the function of the enhancer.

Allele specific MCC profiles from a SCARB2 enhancer show different interactions with surrounding elements

Interestingly, our cone derived donor could be shown to be informative for variants described in the previous chapter. *SCARB2* was shown to have a variable enhancer across the 50 donors and was shown in Figure 2.9. Capturing from this enhancer upstream of *SCARB2* that contains a heterozygous SNP allowed for the generation of allele specific MCC tracks. The probe was designed over the enhancer harbouring the SNP rs11097297, with the captured donor being a heterozygous C/T. As seen in the standard MCC track, the enhancer interacts with the *SCARB2* promoter, another enhancer, and two flanking CTCF sites. However, when splitting the reads from this track into the two separate alleles, much stronger interactions can be seen from the allele with the alternate T variant. Mapping was checked by assessing the amount of reads in each track, with C and T alleles having 1,055,612 and 537,167 reads respectively (Figure 3.9). Though these numbers differ, both have substantial amounts of data and so the lack of interactions from the C allele are not merely an artifact of low data coverage. The exact reasoning behind these differences has also been observed in other data generated outside of this project, and the exact reasoning behind it is currently under discussion by the Davies lab. Several possible biophysical explanations have been suggested, including the openness of the site allowing easier access for MNase, or the fragment sizes of libraries causing biases during capture and mapping.

As previously demonstrated in the 50 donors, this SNP had a likely influence on activity based on ATAC and H3K27ac. Using individuals with genotype C/C, C/T, T/T at rs11097297 these showed reproducible differences in the open chromatin and the activity at these regions (Figure 3.9). In these individuals as the genotype progressed from T/T, to C/T, to C/C, this correlated with a continuous reduction in both hypersensitivity and activity. This in turn correlated with a reduction in the interaction profiles as seen in the MCC.

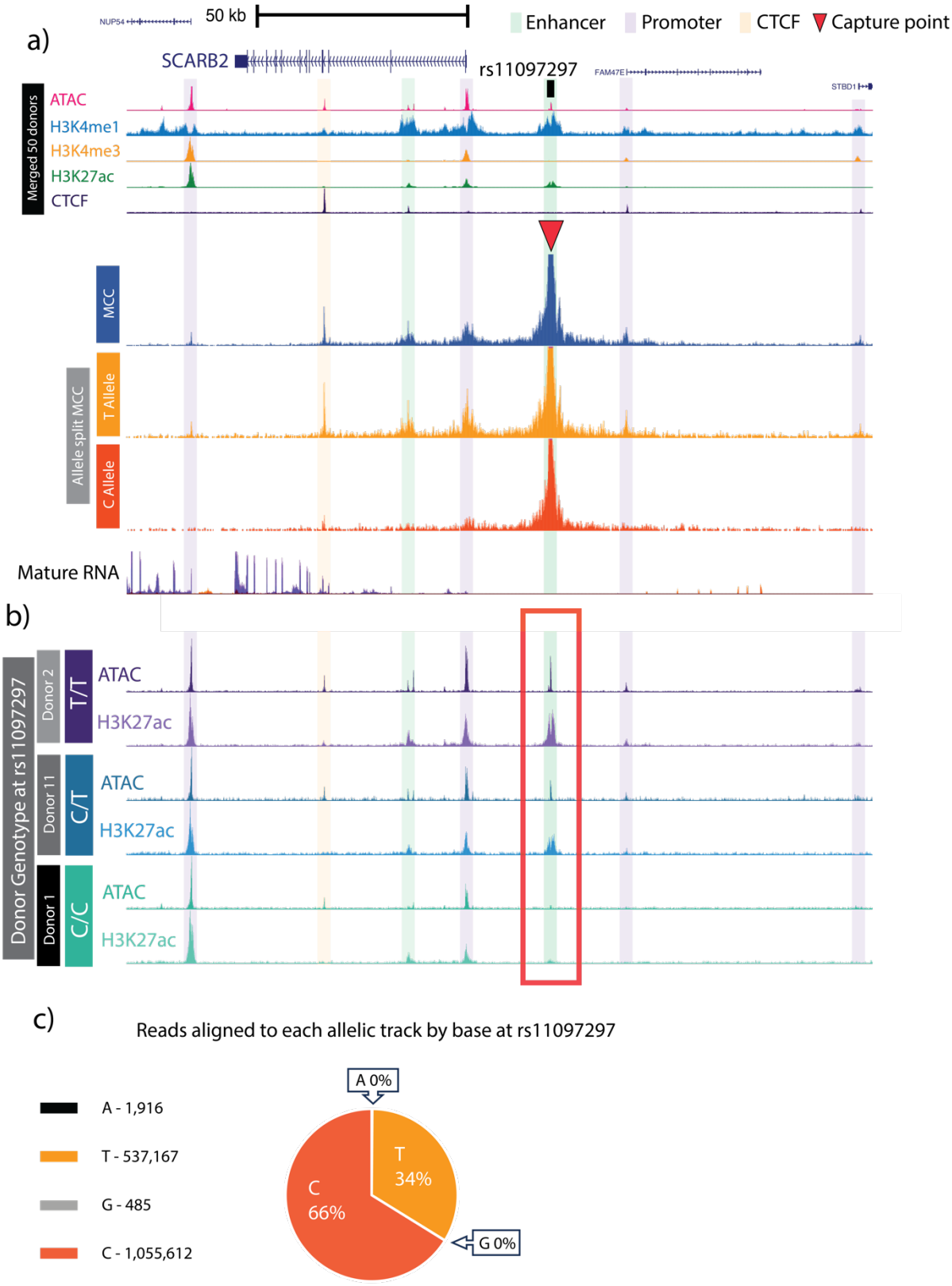

a)
50 kb
Enhancer
Promoter
CTCF
Capture point
SCARB2
rs11097297
Merged 50 donors
ATAC
H3K4me1
H3K4me3
H3K27ac
CTCF
MCC
Allele split MCC
T Allele
C Allele
Mature RNA
b)
Donor Genotype at rs11097297
Donor 2
T/T
Donor 11
C/T
Donor 1
C/C
ATAC
H3K27ac
c)
Reads aligned to each allelic track by base at rs11097297
A - 1,916
T - 537,167
G - 485
C - 1,055,612
A 0%
T 34%
C 66%
G 0%

Figure 3.9: Example of allelic split at *SCARB2* enhancer, with rs11097297.
a) Genomic landscape at *SCARB2*. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Dark blue MCC track depicts standard MCC from an enhancer (red arrow) housing rs11097297, interactions can be seen to surrounding promoters, including the upstream *SCARB2* promoter, also another enhancer and a CTCF site. Yellow and orange MCC track depict allele specific MCC from the T and C allele respectively. Much stronger interactions can be seen in the yellow T track. b) Genomic landscape can also be assessed by finding individuals with certain genotypes. Each donor depicted shows ATAC (upper, darker track) and H3K27ac (lower, lighter track). Genotypes at rs11097297: Donor 2, T/T, purple tracks; Donor 11, C/T, blue tracks; Donor 11, C/C, green tracks. ATAC accessibility and activity at the affected enhancer highlighted with red box can be seen to reduce in the C/T heterozygous individual, and even more so in the C/C individual. c) Number of reads assigned to each allelic track during allelic split analysis. The lack of interaction signal from the C allele is not a result of reduced reads.

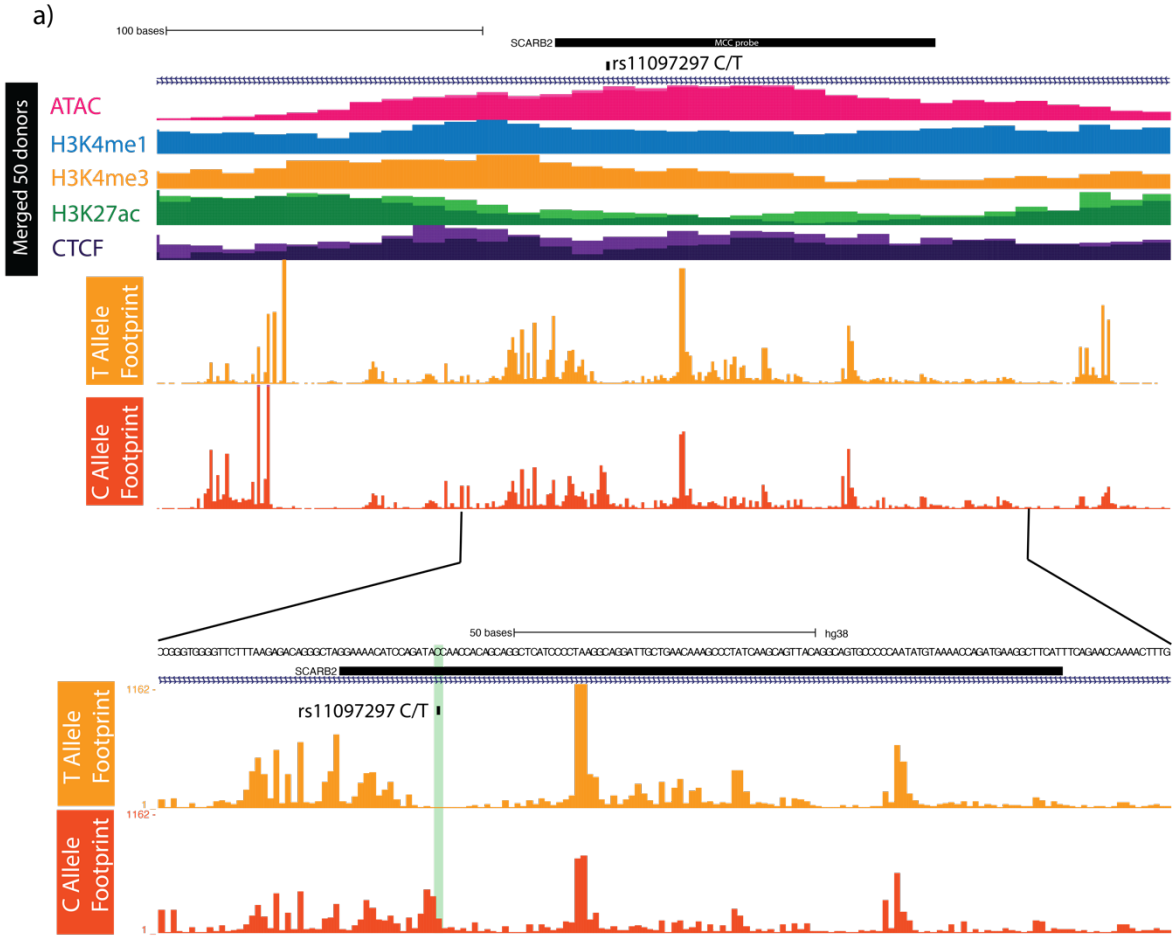
Combined footprint and motif analysis to support the causality of the variant.

To further confirm the causality of the variant, footprinting data from the MCC can be used to see where potential transcription factors are bound and altered in the presence of the variant (Figure 3.10a). Even though both the ATAC and the MCC show that activity is reduced in the C allele, it is still accessible to a lesser degree. This suggests that some factors are still binding and causing a weak open chromatin signal. MCC footprinting can be used in combination with other approaches that are also designed to support causality of a variant. Sasquatch is a computational tool that uses hexamer sequence based DNase footprint data from multiple cell types to predict the effects of noncoding variants on transcription factor binding in specific cell types [32]. It is then able to predict how a single base pair change within a hexamer sequence will affect the transcription factor binding. Interestingly, in this case despite the clear change in MCC footprinting data, the sasquatch prediction shows only a weak change, with a slight increase in footprint prediction at the T allele that exhibits stronger binding (Figure 3.10b). However, since both alleles are open and have activity, a complete loss of binding is not to be expected. From MCC footprinting data, it can be seen that the alternate allele with the T and the reference C have different footprint shapes, with more binding seen in the T allele (Figure 3.10a).

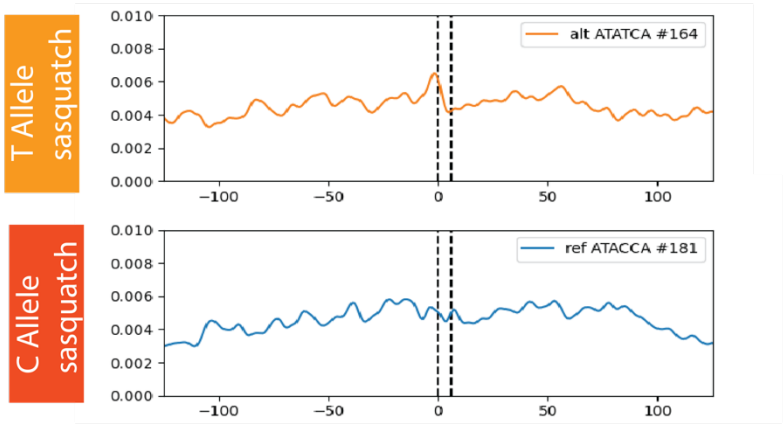
Furthermore, motif analysis can be performed with software such as HOMER [158]. By providing the reference and altered sequence HOMER analysis can indicate which underlying motifs may be being affected by the variant. For this variant, the site beneath was predicted to be a binding site for the

transcription factor MYB (Figure 3.10c). However, the MYB motif was correlated with the ref C that in turn correlated with the decreased interactions. MYB is thought to act as a transcriptional activator, and therefore the prediction would be that this would increase interactions on the C allele. The observation that the C allele decreases interactions indicates that a different TF, potentially one which has not been assessed with ChIP and so therefore the consensus motif is not contained in HOMER, is binding. This therefore highlights that bioinformatic knowledge of motifs is incomplete and predicting transcription factor binding is difficult. This also means downstream prediction of SNP effect is equally challenging. However, the use of predictive methods such as sasquatch for TF binding or HT-SELEX for binding site preferences are useful tools in determining this likely causality of SNPs [176, 177]. Combined with the fact that ChIP is not possible for every TF, this shows the power of footprinting as a proxy for all TF binding.

a)


100 bases
SCARB2
MCC probe
rs11097297 C/T
Merged 50 donors
ATAC
H3K4me1
H3K4me3
H3K27ac
CTCF
T Allele Footprint
C Allele Footprint
50 bases
hg38
SCARB2
rs11097297 C/T
1162
T Allele Footprint
C Allele Footprint

b)


T Allele sasquatch
alt ATATCA #164
C Allele sasquatch
ref ATACCA #181
0.010
0.008
0.006
0.004
0.002
0.000
−100
−50
0
50
100

c)

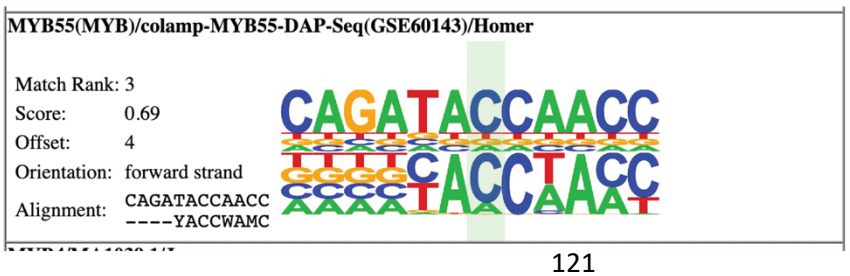

MYB55(MYB)/colamp-MYB55-DAP-Seq(GSE60143)/Homer
Match Rank: 3
Score: 0.69
Offset: 4
Orientation: forward strand
Alignment: CAGATACCAACC
----YACCWAMC

Figure 3.10: Foot-printing data show rs11097297 disrupts protein binding.
a) Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Black bar shows location of the MCC capture probe over the site that contains rs11097297. Yellow track shows foot-printing from the T allele and orange track shows foot-printing from the C allele. In upper tracks foot-prints are similar across the two alleles showing both are accessible. When zoomed in, depicted by black lines and shown in lower section, SNP at green highlight can be seen to be disrupting binding. Yellow allele that shows the binding is also associated with stronger interactions. b) Sasquatch analysis shows footprint prediction when either a T (top track) or C (bottom track) is present in the sequence. Trace depicts a footprint prediction, with a higher signal predicting a lack of cutting and a lower signal predicting a bound factor protecting the DNA c) Motif analysis by HOMER suggests the site is a MYB binding site, and is disrupted by the SNP (N.B. this analysis contradicts the actual binding shown by the foot-printing).

Allele specific analysis of transcription to confirm effect of variation on interacting promoters

At this stage in the analysis, we have shown that the variant not only affects open chromatin and associated TF footprints but also shows differential interactions with a specific gene promoter. This now forms a clear hypothesis to be tested, that the disrupted enhancer is having an allele specific effect on the local promoter and subsequent gene expression. The MCC has shown that these elements are interacting on one allele but not the other, and this decrease in enhancer activity and interaction with the promoter would be predicted to decrease allele-specific gene expression.

This hypothesis can be explored using the matched expression data from the 50 individuals. As well as the ATAC and H3K27ac from the 50 donors, we also have existing allele specific RNA-seq data, from both nascent and mature transcripts. These data are intrinsically allelic because SNPs within introns and exons can typically be used to accurately distinguish from which allele a transcript has originated. From these data, I looked at the genotypes of all the donors and looked at the corresponding RNA outputs of *SCARB2* (Figure 3.11) (Supplementary Tables 2 and 3). From these data it can be seen that (excluding some outliers) homozygotes have matching expression levels, with homozygous C/C conferring to low expression and the T/T being associated with high expression for both alleles. Conversely the het C/T alleles have different expression levels from each allele. The outliers may result from polygenic factors affecting those donors or experimental noise. This shows that the disrupted enhancer is also associated with a lower gene output.

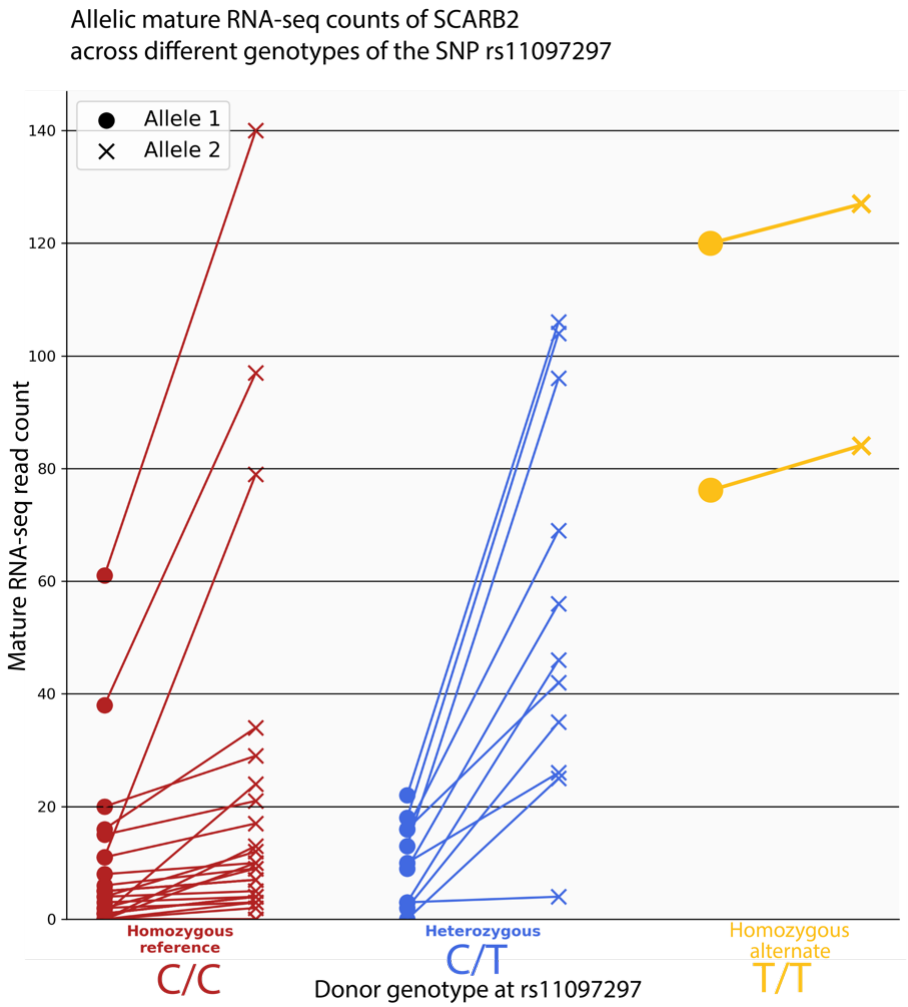


Figure 3.11: *SCARB2* mature RNA allelic counts, split by rs11097297 genotype.
Raw mature RNA-seq read counts of *SCARB2* at 50 individuals. Leftmost, red points from donors who are homozygous C/C; middle, blue points from donors who are heterozygous C/T; rightmost, yellow points from donors who are homozygous T/T. Dots and crosses depict the different alleles of each individual, with the lower expression always as a dot on the left.

Allele specific MCC profiles show a variant disrupting the binding of CTCF and subsequent interactions

The next example demonstrates an allele specific CTCF site. When considered in the context of regulatory networks and deciphering SNP mechanisms, determining the overall effect of losing a CTCF site is challenging. Broadly speaking, with an enhancer and a promoter, an interaction between them will increase the expressional output [7]. When using 3C to show that in the presence of a variant

these interactions are reduced, it can therefore be hypothesised that this will subsequently decrease gene activity. CTCF's structural role in the genome has been shown to play a key role in the formation of regulatory interactions and in blocking interactions between enhancers and promoters. However, CTCF is largely non-cell-type specific, and so the effects of altering its activity may have repercussions in multiple cell types across multiple genes. Studies have shown that CTCF depletion has a profound effect on local architecture by reducing intradomain interactions and also increasing interdomain interactions. However, a change in architecture does not always lead to a change in expression [178, 179]. Taking all of these factors into account, deciphering and predicting the global impact of altering a single CTCF site across multiple cell types is challenging.

The CTCF site shown in this example lies within an intron of *MYH15*, with CTCF binding varying across the donors as shown in Figure 2.8. The donor sample was reference C and alternate T at rs936266. From the standard MCC track, the site can be seen to be interacting with several other CTCF sites up and downstream of the capture site (Figure 3.12a). By splitting the tracks into the separate alleles, it can be seen that all the interactions to the other CTCFs were coming from the reference C allele. Similarly to the previous example, though not equal in read counts, each allele had enough data to show that it was a real allelic influence and not just a lack of data from one allele (1,055,612 reads for the C allele and 537,167 reads at the T allele).

When looking at CTCF ChIP-seq data from donors with different genotypes, a clear depletion of CTCF binding can be seen (Figure 3.12b). For homozygous C/C there is strong binding, with this being reduced in the heterozygous C/T and undetectable in the homozygous T/T donor. When assessed with Sasquatch, the alternate T variant is associated with a damaged footprint and therefore less binding. This is reflected in the real footprinting profiles from the MCC data (Figure 3.13). A clear pattern of binding can be seen on the C allele with the bound CTCF, whereas the T allele with no binding shows reduced hypersensitivity and footprinting. As there is no open chromatin at this altered site, this suggests there is no binding, the chromatin is not opening, and so MNase could not cut the site frequently to form a footprint.

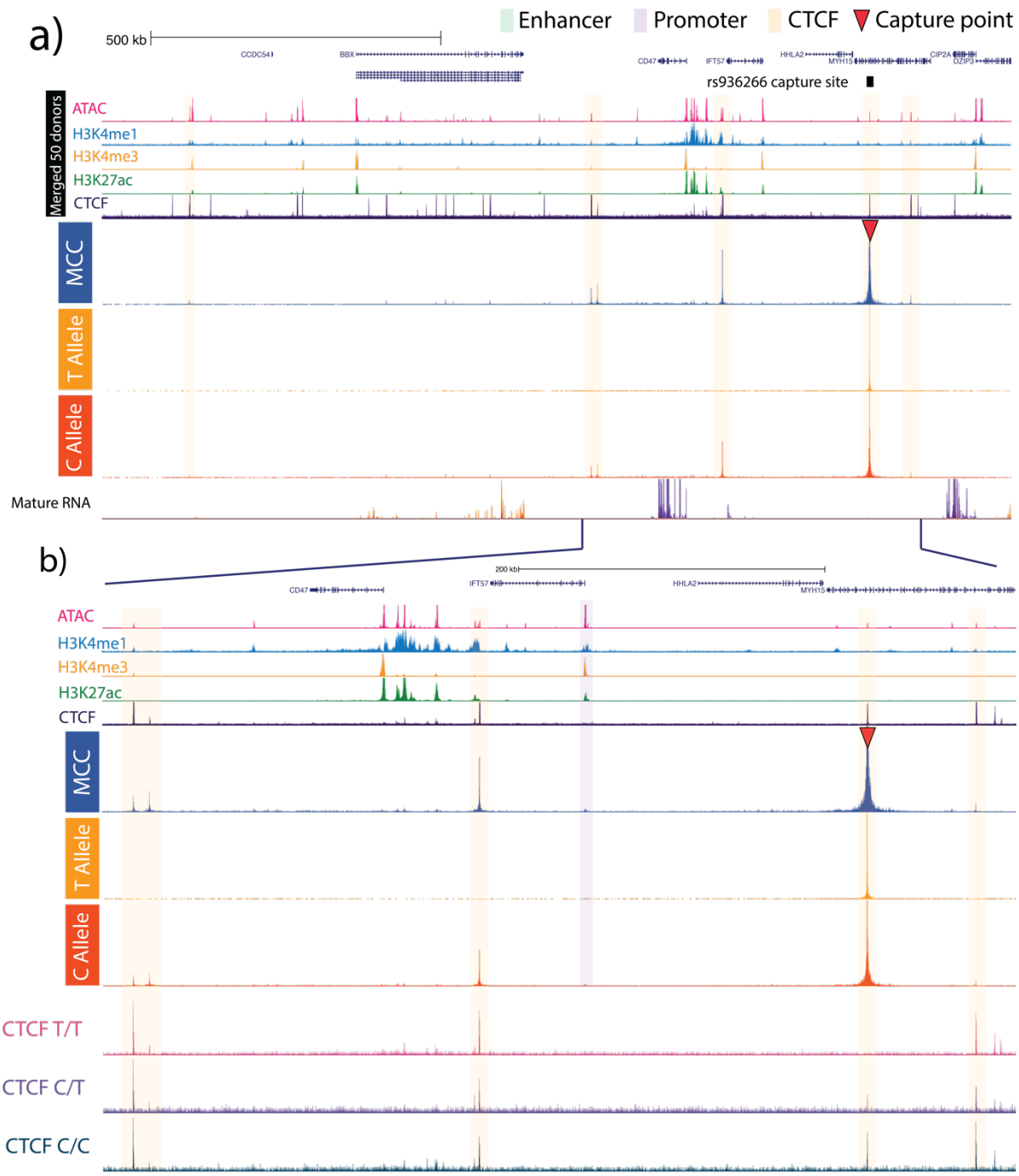


Figure 3.12: Example of allelic split at a CTCF site, with rs936266, in *MYH15 intron*.
a) Genomic landscape at *MYH15*. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Dark blue MCC track depicts standard MCC from the CTCF site (red arrow) housing rs936266, interactions can be seen to an adjacent promote and several other surrounding CTCF sites. Yellow and orange

MCC track depict allele specific MCC from the T and C allele respectively. Much stronger interactions can be seen in the orange C track. b) Zoomed tracks of region shown in a) with the same tracks from 50 donors, plus the standard and allele MCC. Genomic landscape can also be assessed by finding individuals with certain genotypes. Each track depicted shows CTCF ChIP-seq. Genotypes at rs936266: Upper pink track, T/T; middle purple tracks, C/T; lower blue track, C/C. CTCF binding at this site (red box) is absent in the T/T individual, present in the C/T individual, and stronger still in the T/T individual.

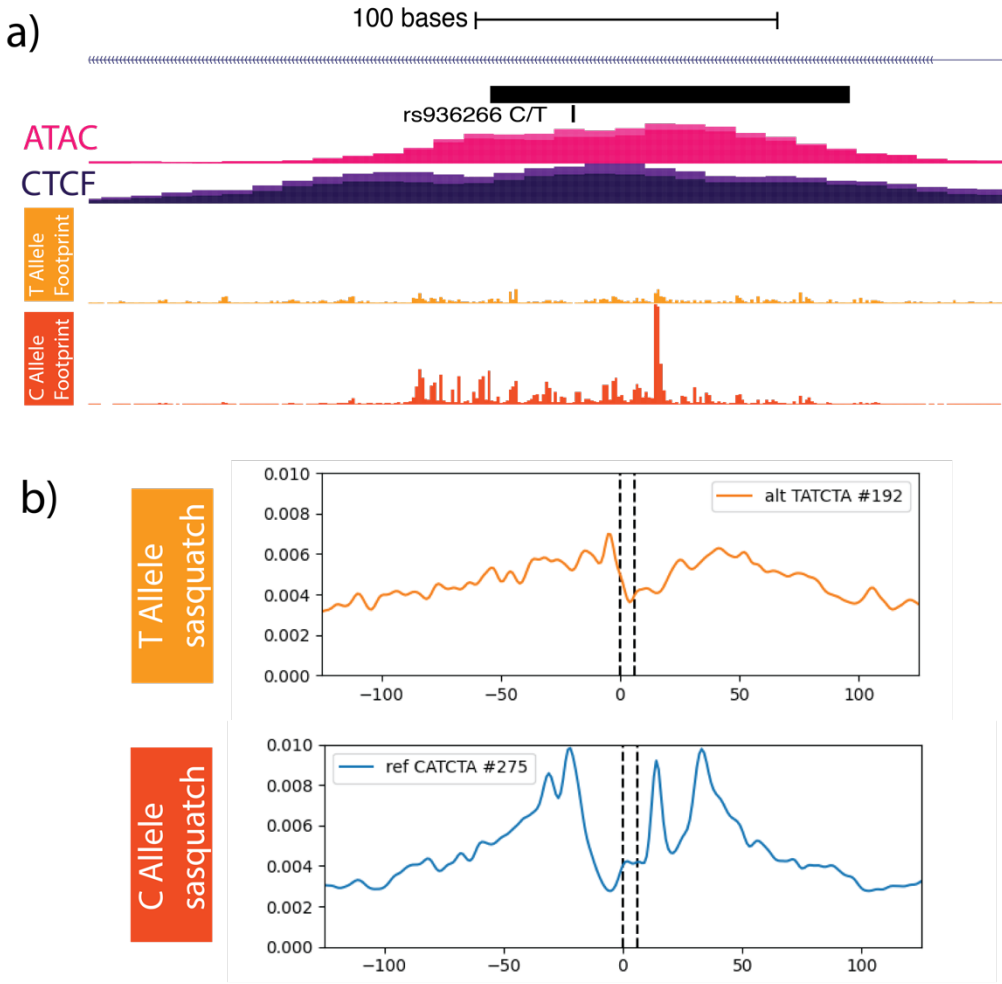


Figure 3.13: Footprint and sasquatch data showing damage at rs936266 CTCF *MYH15* site.
a) Black box shows capture probe location, with rs936266 SNP location used for allelic split. Track heights auto scaled to data view, Track heights auto scaled to data view, ATAC (pink track) and CTCF (purple track) ChIP-seq combined data from 50 data sets show location of open chromatin site and CTCF binding. Yellow and orange tracks show footprints at T and C alleles respectively. The higher signal at the C allele shows more accessibility and more binding. b) Sasquatch analysis shows footprint prediction when either a T (top track) or C (bottom track) is present in the sequence. Trace depicts a footprint prediction, with a higher signal predicting a lack of cutting and a lower signal predicting a bound factor protecting the DNA. These match the real data. Upper track shows weak footprint, which associates with no CTCF binding. Lower profile shows the C allele with a prediction of strong binding, which associates with CTCF binding.

Designing and validating MCC probes to variant regions

With the MCC protocol set up and a pilot probe set validated, I designed another set of capture probes based on analysis of the regulatory variation found in the 50 donors (MCC probe set 2). This design process utilised a modified version of the Capsequm tool used for Capture-C capture probes [6]. The next set of probes was comprised of: regions highlighted by the differential ATAC and H3K27ac peaks from the analysis in Chapter 2; extra variants of interest; promoters of skewed genes; and controls (Table 6.4). The skewed gene promoters were selected by assessing the allelic RNA-seq from the 50 donors. Genes that showed strong skew in their expression were included to see which enhancers were controlling them, to see if their skewed expression was due to variation in distal regulatory elements. These regions were not explored further in this thesis as certain genes inherently show skew through multiple mechanisms such as imprinting and will be subject to further analysis in the future [115, 116, 180].

A variant in the R2 enhancer of HBA is a neutral SNP that does not affect interactions

Probes were validated with the existing erythroid libraries. Figure 3.14 shows the MCC profiles from R2, the strongest enhancer of the *HBA* locus, and displays the expected interactions for the domain. This probe was designed over the SNP rs11248850, which is a heterozygous A/G in the cone donor, and allele specific tracks were generated. This SNP was chosen as a control for the allelic MCC analysis as it was shown to be unskewed in open chromatin signal in the donors. It was therefore proposed to be neutral in terms of effect on enhancer function. In confirmation of this the allelic tracks split based on rs11248850 showed no discernible differences in their interactions, and the allele specific tracks both show mirrored interactions to all the expected REs and *HBA* promoters. This provides an example of a neutral SNP that has no effect on interactions [52], providing an internal control for the influence of mappability or alignment of allele specific tracks. This neutrality is further reflected in the footprinting data, with a strong TF binding site being seen equally in both alleles, over the site of the SNP. No difference in footprints is detected in Sasquatch analysis (Figure 3.14).

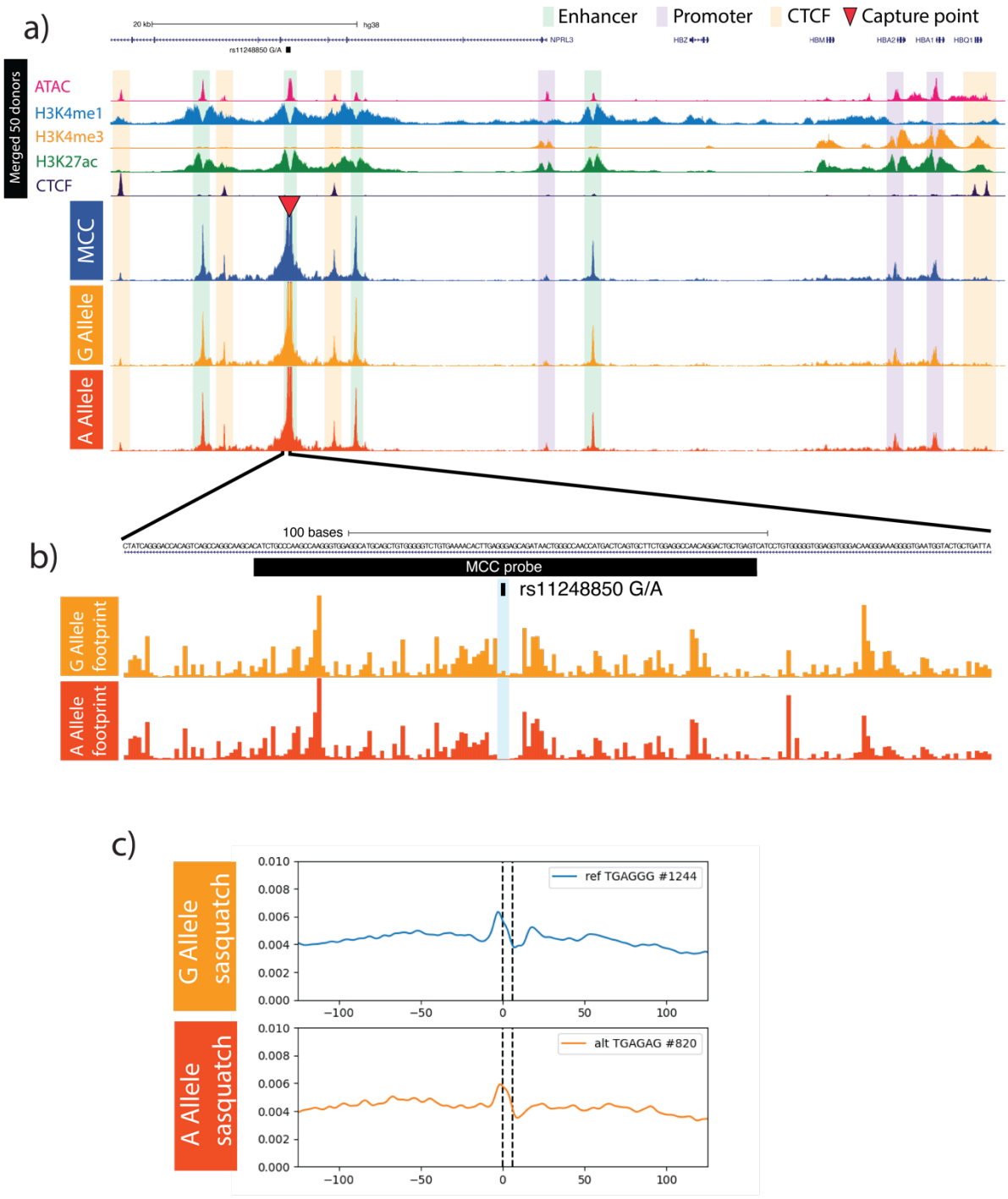


Figure 3.14: Validating the second set of probes and demonstrating that a neutral SNP has no effect on allele specific tracks.

Genomic landscape at *HBA* locus. Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Blue MCC track shows standard MCC from the enhancer R2 (red arrow)

and shows strong interactions with the other enhancers as well as interactions with the alpha promoters. Allelic data split with rs11248850, G allele in yellow and A allele in orange both match each other and standard blue track, showing SNP has no effect on interactions. b) Lower tracks show zoom in of R2 enhancer, probe and SNP location (blue highlight), and unaltered footprinting between the G allele (yellow) and A allele (orange). c) Sasquatch analysis shows footprint prediction when either a G (top track) or A (bottom track) is present in the sequence. This prediction shows no change between the alleles.

Generating MCC libraries from characterised biobank donors

The previous examples were captured from an uncharacterised individual. To properly utilise the data from the biobank donors, more MCC libraries were generated from characterised individuals. By using and incorporating the existing epigenetic data from these individuals, more in-depth insights can be made from the MCC data. Despite previously generating successful libraries, these library preparations were again experiencing low yields which impeded progress. Prior to successfully generating these libraries, more optimisation and troubleshooting was required. The shaking conditions of the digestion stages were eventually assessed and changed again, and full troubleshooting experiments are described in the methods chapter. Once the protocol was working reproducibly again, I was then able to generate libraries from four donors (donors 2, 3, 54 and 55).

Generating MCC libraries from samples with a low cell input

CD34s from Donor 16 were differentiated for another project in the lab, and so I was able to get a $3x10^7$ aliquot of cells for MCC. However, since in the previous library generations I had needed several rounds of titrations and troubleshooting to get successful libraries, these had used multiple $1x10^7$ aliquots. For this low cell sample I therefore titrated the MNase concentration on a smaller aliquot of cells to determine the MNase concentration required for the libraries. Due to slight differences in fixation conditions, time and exposure to formaldehyde, exact cell numbers and digitonin permeabilisation efficiency, the MNase concentration required for optimal digestion will vary across samples. By reducing all the cell numbers and reaction volumes by 90%, and by using the whole sample for MNase titration, I was able to deduce the approximate MNase concentration required on the real sample. This proved successful for both donors 16, and donor 1, of which I also had a limited number of cells available.

Exploring interaction profiles at sites of variation in characterised individuals

With my next probe set validated (MCC probe set 2) and the libraries from my characterised individuals generated, I performed another capture on all the libraries: with 82 probes across six donors generating 492 unique viewpoints, and additional allelic split tracks of any probe site over a heterozygous variant. Due to the large number of individuals and viewpoints generated in this experiment it is difficult to present all of the data and analysis generated in this chapter. Broadly, a range of landscapes were observed, from regions that showed no discernible interactions, to complex landscapes that encompassed long range interactions (~1 Mb) between multiple RE classes. I have therefore selected and presented several interesting examples that further explore the variants highlighted in Chapter 2, and demonstrate their local genomic landscape to show the utility of these data.

At present no pipeline for meta-analysis of these data sets and their interactions exists and so instead represents an ongoing effort for the Hughes and Davies labs. One such solution being explored is the repurposing of the inhouse Capture-See software to efficiently visualise the capture profiles from multiple individuals across multiple loci [175]. Once visualised, the next step in quantifying interactions is the implementation of peak calling software. As MCC represents a new data type with a specific and distinctive peak landscape, work with LanceOtron[149] is currently underway to determine the best way to call these peaks. Once called, packages such as ChromHMM[181] or Regulamentary (unpublished, Simone Riva) can be used to determine the likely REs interacting with the capture site. Combined, these tools will make the qualitative and quantitative assessment of MCC profiles much more high throughput.

To start, library quality was checked and verified at the *HBA* locus (Figure 3.15). For this region, each donor's MCC profile can be seen to reflect the interactions expected and seen in Figure 3.7. Additionally, despite the low quality ATAC data for donor 54 (Chapter 2, Figure 2.2), the MCC data from this individual shows the same consistent and reproducible high-resolution interactions with the known elements. This emphasises the interpretability of MCC data even in the absence of matched open chromatin data. As is common in MCC, dependent on the probe location, interactions between

the classes of REs are often the strongest, such as enhancers with other enhancers. This is especially apparent with the strong CTCF-CTCF interactions, reflecting their structural, pair wise or hub interacting properties. These enhancer-enhancer and promoter-promoter interactions can be seen in Figure 3.15, depending on the viewpoint. With previous techniques the interaction landscape has been well established, especially the links between the alpha promoters and its four regulatory enhancers (R1-4). With MCC, when capturing from R2, while strong interactions with the promoters can be seen, interactions with the other *HBA* enhancers (R1, 3 and 4) are much stronger.

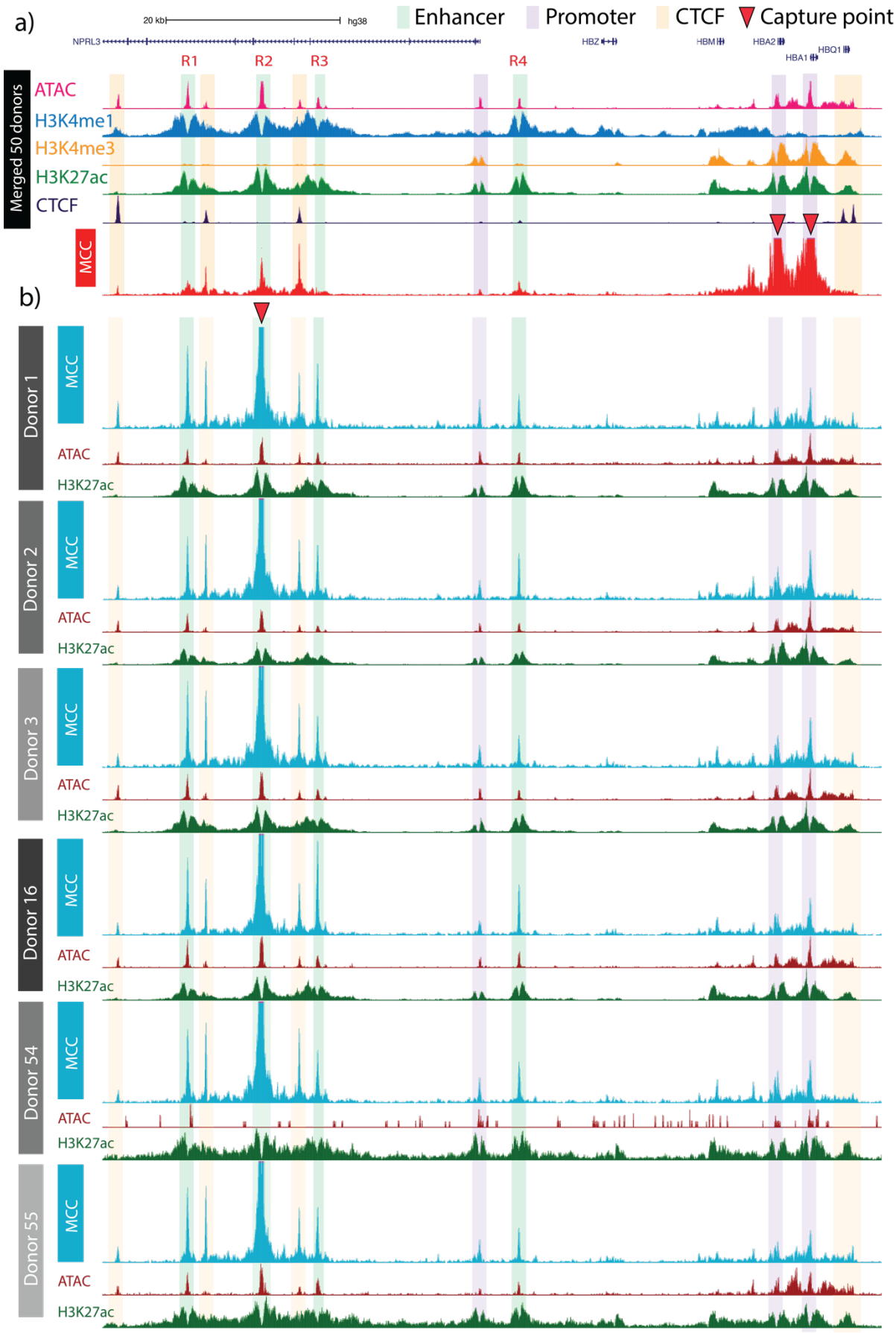

a)
20 kb
hg38
Enhancer
Promoter
CTCF
Capture point
NPRL3
HBZ
HBM
HBA2
HBQ1
HBA1
R1
R2
R3
R4
Merged 50 donors
ATAC
H3K4me1
H3K4me3
H3K27ac
CTCF
MCC
b)
Donor 1
Donor 2
Donor 3
Donor 16
Donor 54
Donor 55
MCC
ATAC
H3K27ac

Figure 3.15: Validation of six MCC libraries at the *HBA* locus.
a) Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Red MCC track of cone sample captured from the promoters (Figure 3.7) shows interactions with enhancers and other REs. b) MCC tracks (blue) of donors 1, 2, 3, 16, 54 and 55 from the R2 enhancer (red arrow) show expected interactions to the other enhancers and the alpha promoters. Each donor set also shown with the individuals ATAC (red) and H3K27ac (green) data. N.B. ATAC from donor 54 was bad quality, but the other assays show the expected patterns.

MCC from a novel gain of function shows no interactions to surrounding landscape

The unique gain of function peak within the *PTPRK* intron as shown in Chapter 2 had an open ATAC peak with active H3K27ac marks (Figure 2.7). This peak is unique to a single haplotype only seen on one allele of donor 1. Here it can be seen that despite this activity, when looking at the interaction profiles in both donor 1 and the donors without a peak, there is no gain in local interactions with the gain in activity. In the local area there is an upstream enhancer with weak H3K27ac marks and two downstream CTCF sites, none of which appear to be interacting with the active peak. This suggests that this GoF variant is indeed open and active, but is currently not interacting with any other regulatory elements and therefore not contributing to any regulatory networks. This is still an interesting and informative result even though the element seems not to interact with other elements when active. In fact, this is a logical predicted outcome of a novel gain of function. As they are generated by the per chance creation of TF binding sites in a locale with other cryptic binding sites, they are not under any selection for function. This means they can appear in places in the genome where they cannot interact with existing regulatory processes and hence would be deemed neutral.

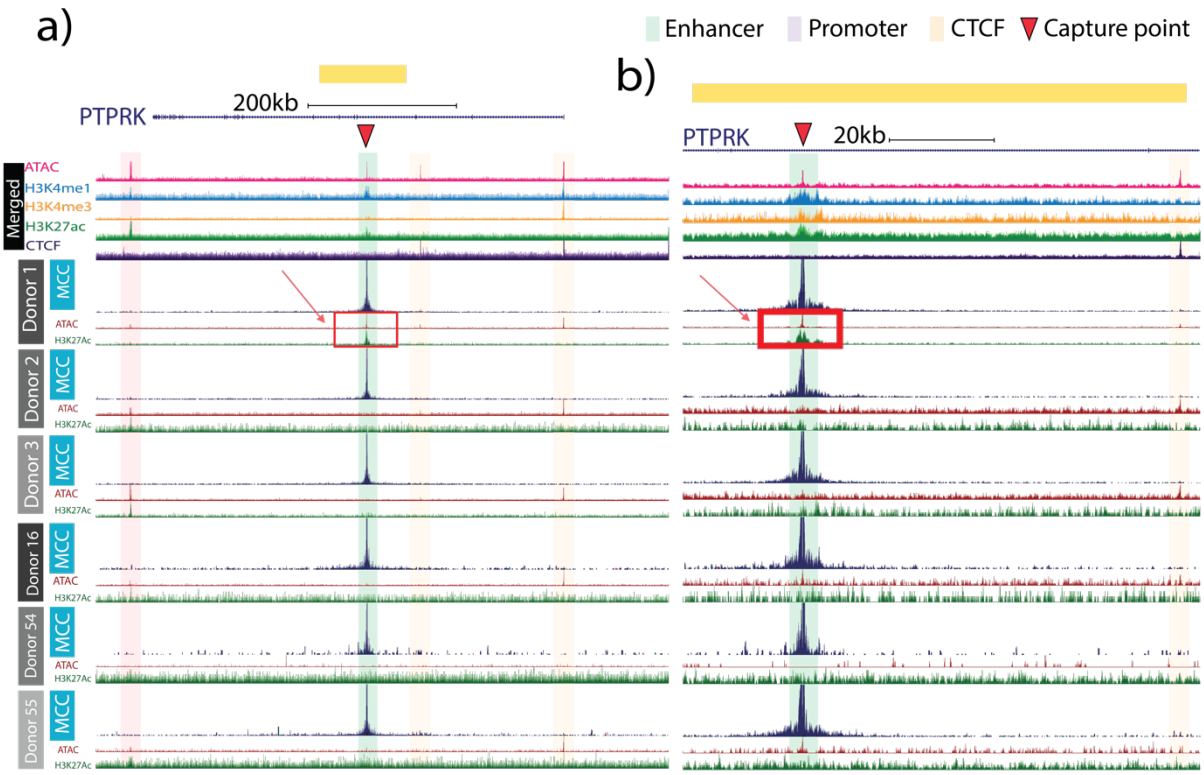


Figure 3.16: A Gain of Function variant unique to donor 1 shows no interactions.
a) Region within a *PTPRK* intron containing a gain of function variant only present on one allele of donor 1. Red box marked by red arrow depicts the open chromatin and activity open present in donor 1. Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. MCC tracks (blue) of donors 1, 2, 3, 16, 54 and 55 from the peak (red arrow) show no interactions to surrounding region. Only active peak in donor 1 also shows no interactions. b) Zoom in at the same region, area depicted by yellow box.

MCC from co-appearing open chromatin peaks show weak local interactions

The next example shows the peak within the *CLNK* intron, as shown in Figure 2.18. In this region multiple peaks can be seen arising simultaneously in individuals. Although open and active, when capturing from one of these peaks, there is not a clear interaction profile with the surrounding elements. Weak interactions can be seen to the most adjacent peaks, as well as a CTCF site further upstream. Though active in both H3K27ac and showing interaction in a limited local region, these peaks do not appear to be interacting with any expressed genes, and are therefore unlikely to be influencing a regulatory network in this cell type. However, as this data is representative of a certain

time point within a differentiation, these elements have the potential to be acting in different cell types or stages.

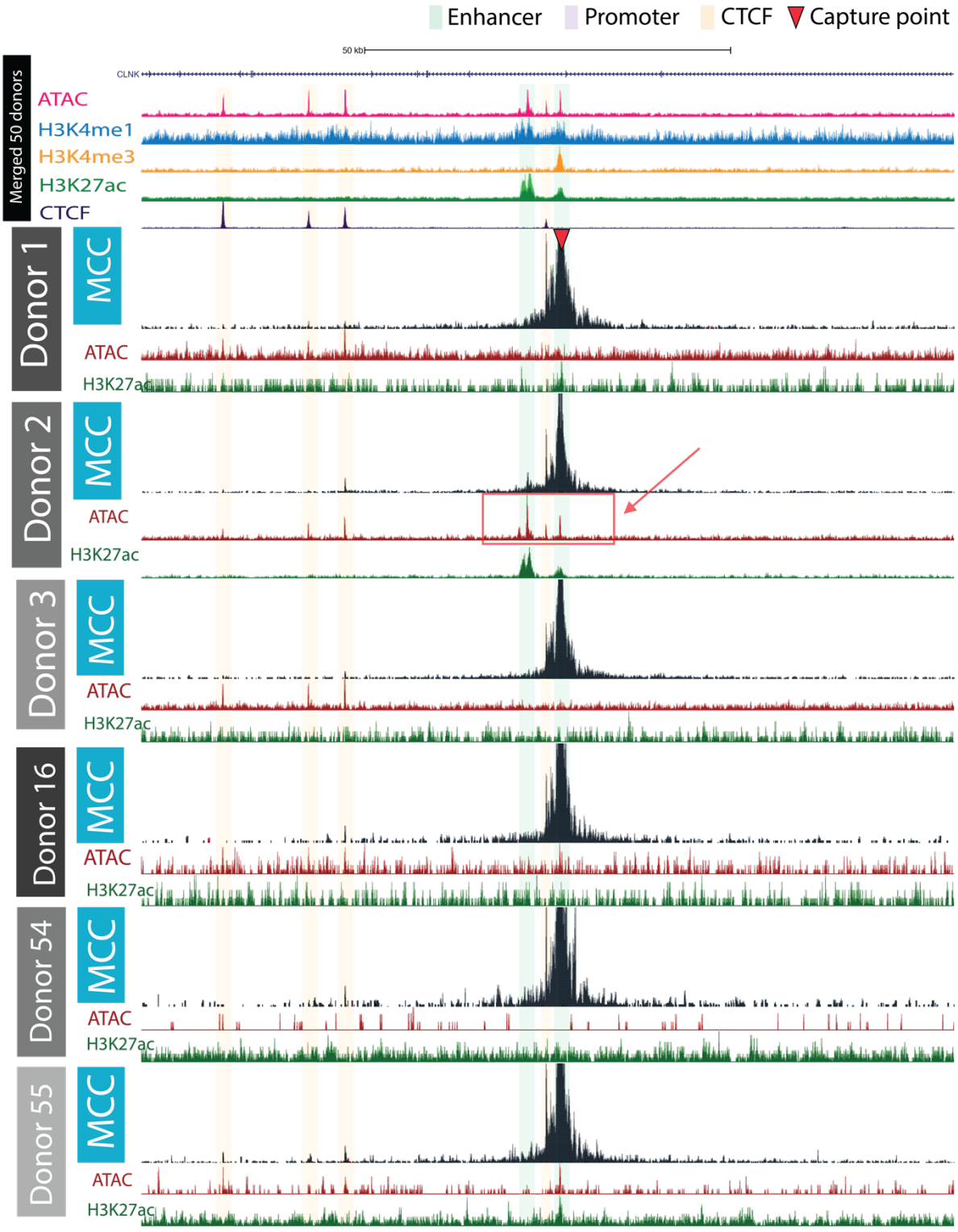


Figure 3.17: Interactions at *CLNK* site of multiple novel peaks.
Region within a *CLNK* intron containing multiple peaks in donor 2. Red box marked by red arrow depicts the open chromatin and activity open present in donor 2. Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. MCC tracks (blue) of donors 1, 3, 16, 54 and 55 from the peak (red arrow) show no interactions to surrounding region. Only active peak in donor 2 shows interactions with the other proximal open regions.

MCC from a site of novel transcription shows weak local interactions

The peak downstream of *FERMT1* associated with the novel transcript from Figure 2.24 is similar to the *CLNK* example in that several peaks can be seen in the same local vicinity in several donors. Despite the clear presence of transcription from this peak that associates only with the donors with the open regions, there appears to be minimal interactions to the local landscape. Two downstream ATAC peaks may show some weak interactions. As promoters are sufficient at driving transcription without the interaction of enhancers, this still proves an interesting example [182]. As a rare and therefore likely new promoter, it would take time for a local enhancer to arise and contribute to the regulatory network. Furthermore, for an enhancer to arise, interact with the promoter and persist within a population, positive selection pressure would be required.

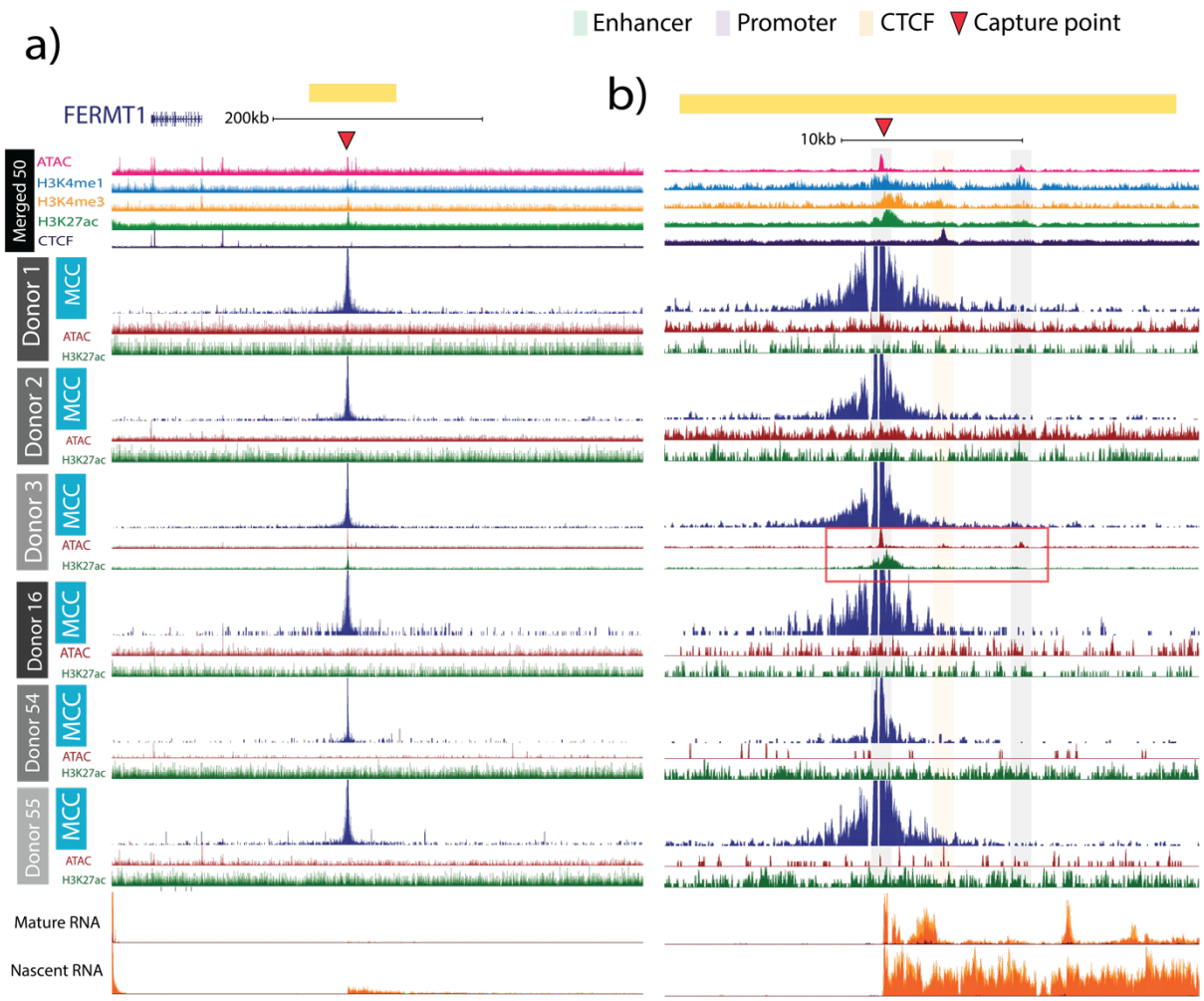


Figure 3.18: No interactions from active peak, despite active transcription.
a) Region downstream of *FERMT1* containing an open region that produces transcripts in donor 3 (and other donors not shown) Red box depicts the open chromatin and activity present in donor 3. Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. MCC tracks (blue) of donors 1, 2, 16, 54 and 55 from the peak (red arrow) show no interactions to surrounding region. b) Zoom in at the same region, area depicted by yellow box. Only active peak in donor 3 also shows no interactions.

MCC from the PPP1R3B enhancers show different interactions relative to genotype

As shown in Figure 2.13, the regulatory landscape of the *PPP1R3B* gene is relatively simple, with a promoter being controlled by two downstream enhancers. However, the pattern of open chromatin across this region showed that the presence of these two enhancers varied across the donor

population (Table 3.2). What seems like a simple locus becomes much more complex when considering the effects of SNPs within these regions as both enhancers are associated with SNPs. Rs330946 is located in the enhancer more proximal to the promoter (peak 1 Figure 2.13) and the presence of an alternate C acts as a loss of function. In contrast, rs3748136 and rs3748137 are in the more distal enhancer (peak 2 Figure 2.13) and the presence of an alternate A and C act as a gain of function. Rs3748136 is also a GWAS hit associated with a reduced reticulocyte count, and so relevant to the erythroid cell model. For simplicity in figures, tables and text only rs3748136 will be referred to, as the SNPs within this enhancer are linked.

Genotype at SNP				Donors	Figure
rs330946	ATAC peak?	rs3748136	ATAC peak?		
C/C	No	G/G	No	Donor 55	Figure 3.19
A/C	Yes	G/G	No	Donors 3, 16	Figure 3.20
A/C	Yes	G/A	Yes	Donors 1, 2, 54	Figure 3.21

Table 3.2: Genotypes and activity at rs330946 and rs3748136 in six donors.
Figure number for MCC tracks of each individual at both SNPs.

There are a variety of open chromatin patterns amongst the captured individuals, and each behaves differently. In total, these donors exhibit three different combinations of SNPs, each displaying expected interactions to the promoter relative to the presence or absence of open chromatin and activity.

Figure 3.19 shows one donor, with a genotype that is homozygous C/C and G/G at rs330956 and rs3748136 respectively. At both these sites the ATAC shows no open chromatin, and this lack of activity is reflected in the MCC tracks from both enhancer regions as there are no interactions to the promoter. Missing genotype data for the individuals without WGS was acquired using mpileup of the BAM files (Supplementary Figure 8)

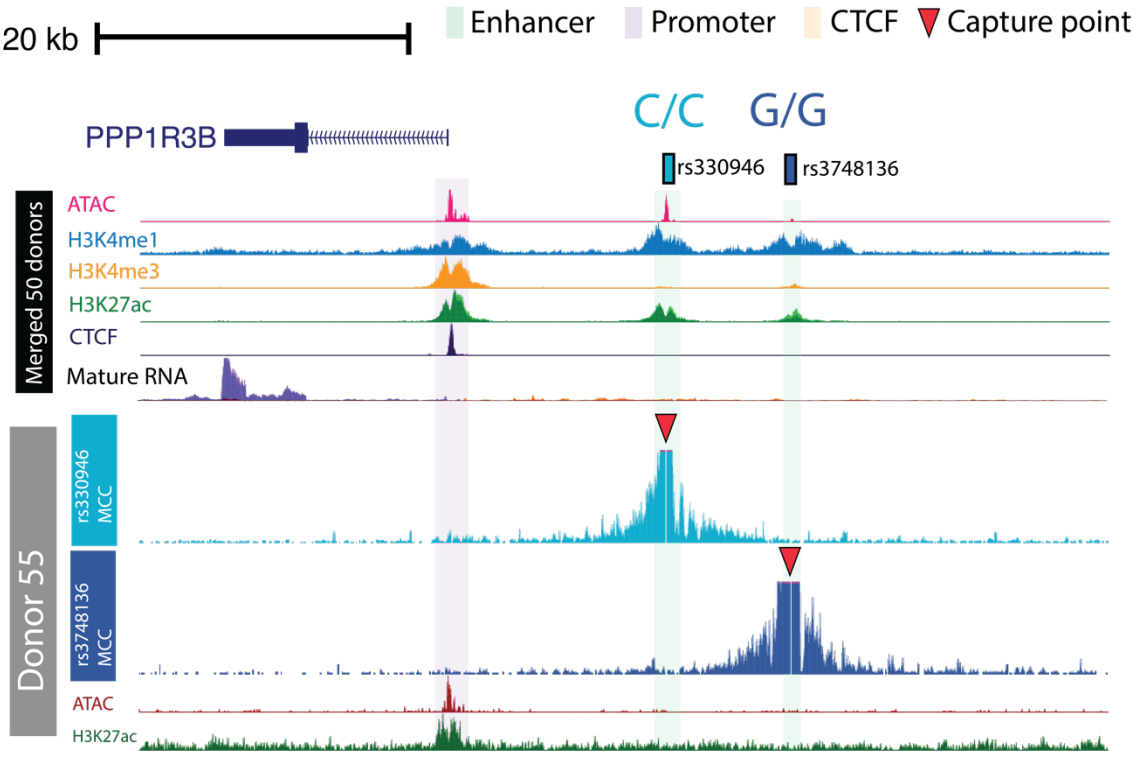


Figure 3.19: Interactions at *PPP1R3B* from a C/C rs330946 genotype, and a G/G rs3748136 genotype.
Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Blue MCC tracks from one individual, light blue from rs330946 and dark blue from rs3748136 (red arrows), red track ATAC and green track H3K27ac show no open chromatin, activity or interactions.

Figure 3.20 shows two donors, who are both heterozygous A/C and homozygous G/G at rs330956 and rs3748136 respectively. When compared to donor 55, these individuals do have the first enhancer, which is associated with the A allele. In the MCC, this active enhancer can be seen to be interacting with the body of the promoter, as well as having a peak at the CTCF site adjacent to the promoter. There are no interactions with the downstream, inactive enhancer, which in turn also has no interactions from it.

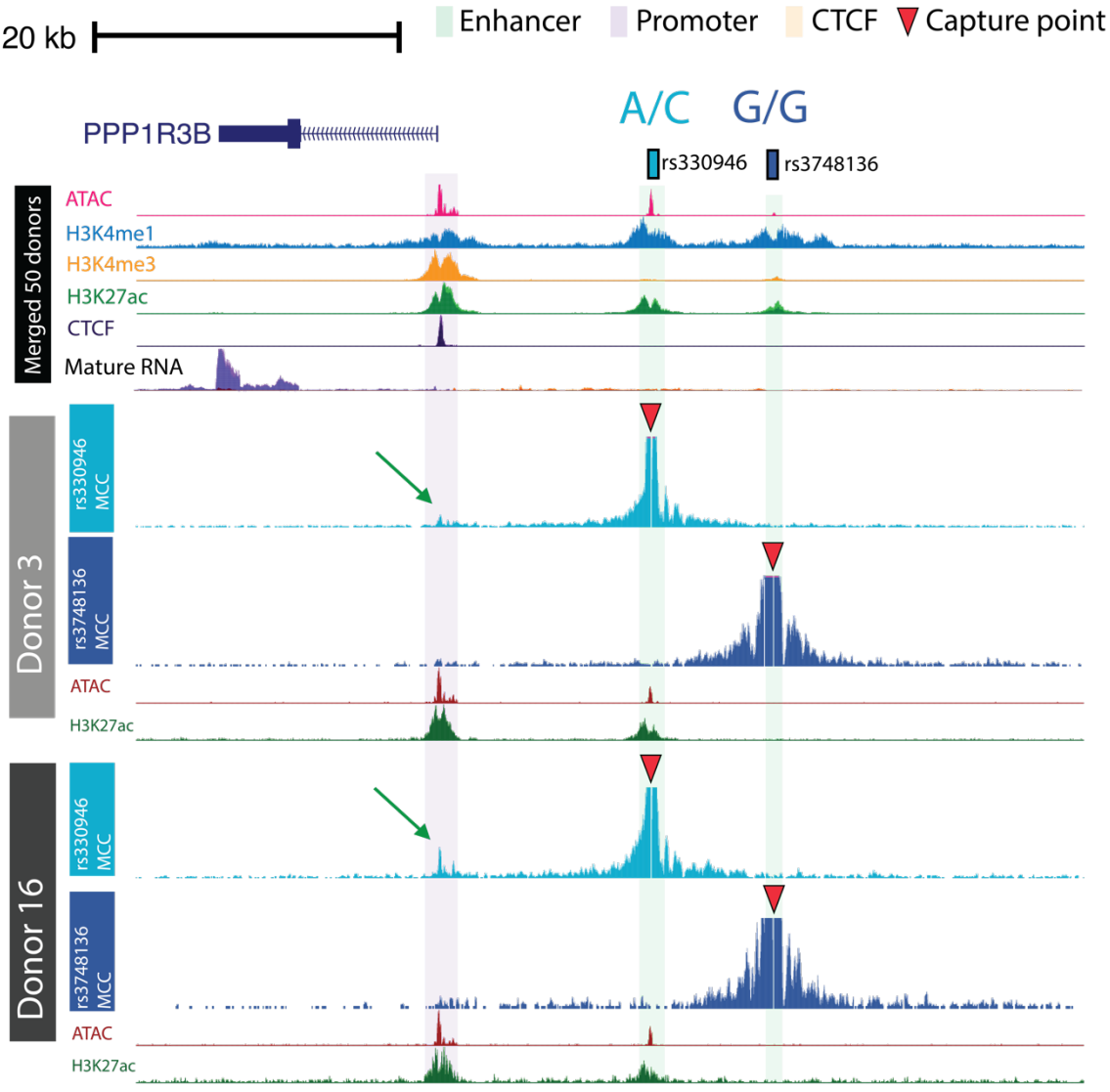


Figure 3.20: Interactions at *PPP1R3B* from an A/C rs330946 genotype, and a G/G rs3748136 genotype. Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5

available sets (Table 2.1). Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks from two individuals, with blue MCC tracks (light blue from rs330946 and dark blue from rs3748136) (red arrows), red track ATAC and green track H3K27ac. Both individuals show open chromatin and activity at the leftmost enhancer, plus light blue MCC shows interactions to the *PPP1R3B* promoter (green arrows). Both individuals show no open chromatin, activity or interactions at the rightmost enhancer.

Figure 3.21 shows data from three donors who are heterozygous A/C and G/A at rs330956 and rs3748136 respectively. The ATAC reads from rs3748136 have been shown to be allelically skewed to the A allele (Supplementary Figure 10). From the MCC profiles, both enhancers can be seen to be interacting with the promoter. Donor 54 also seems to have interactions between the enhancers, but this is not seen in the other two donors.

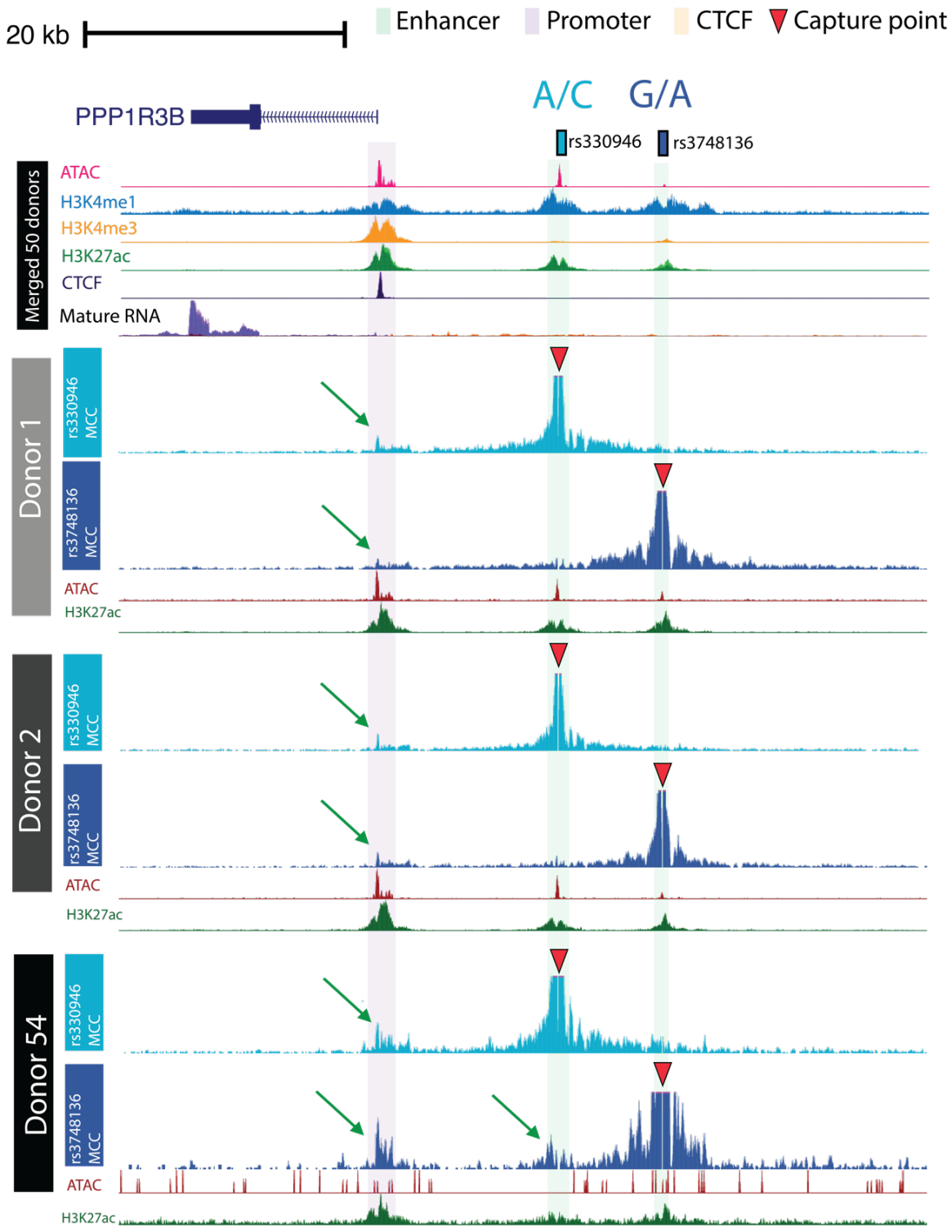


Figure 3.21: Interactions at *PPP1R3B* from an A/C rs330946 genotype, and a G/A rs3748136 genotype. Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward

transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks from three individuals, with blue MCC tracks (light blue from rs330946 and dark blue from rs3748136) (red arrows), red track ATAC and green track H3K27ac. All show open chromatin and activity at both enhancers, plus MCC shows interactions to the *PPP1R3B* promoter (green arrows).

Allelic MCC tracks from the PPP1R3B enhancers show minimal interactions between enhancers

Allelic MCC tracks from a single donor exhibiting heterozygous SNPs in both enhancers show that the enhancers are not interacting (Figure 3.22). Phased genomes allow us to determine which SNPs and therefore which open chromatin sites reside on the same allele (Supplementary Figure 12). This donor has both *PPP1R3B* haplotypes depicted in the previous chapter (Chapter 2 Figure 2.15), with allele 1 (rs220946 C, rs3748137 G, rs3748136 G) exhibiting the LoF and therefore having no peaks, and allele 2 (rs220946 A, rs3748137 C, rs3748136 A) exhibiting the GoF and therefore having both enhancer peaks. The allelic interactions from this locus clearly show that only the active enhancers are interacting with the *PPP1R3B* promoter. However, in contrast to other loci, the enhancers are surprisingly showing minimal interactions to each other, despite being on the same allele. This is highly interesting, as usually like RE classes will interact strongly (Figure 3.15). The exact reasoning for this is currently unclear, but may be due to the presence of noncompatible TFs.

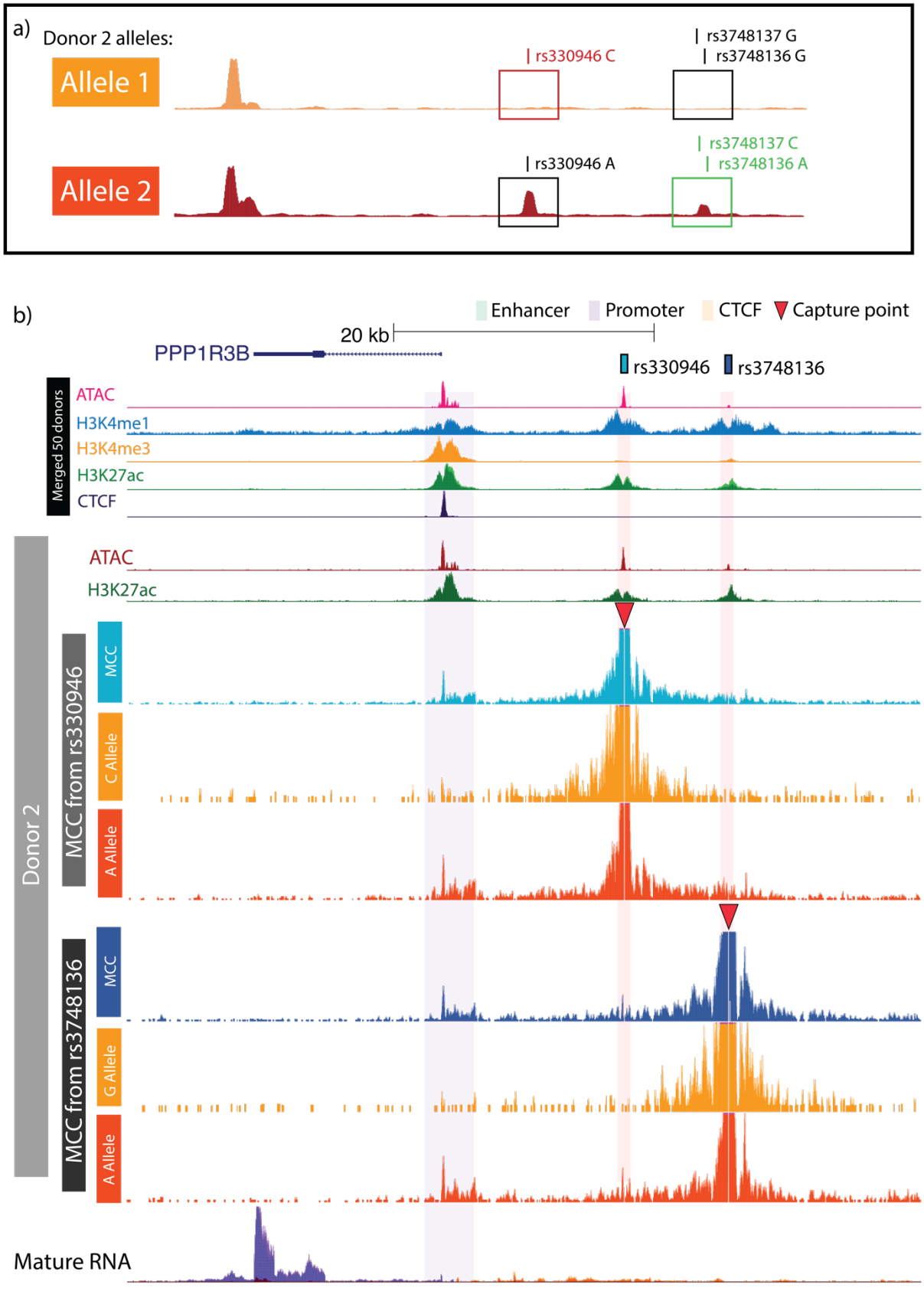
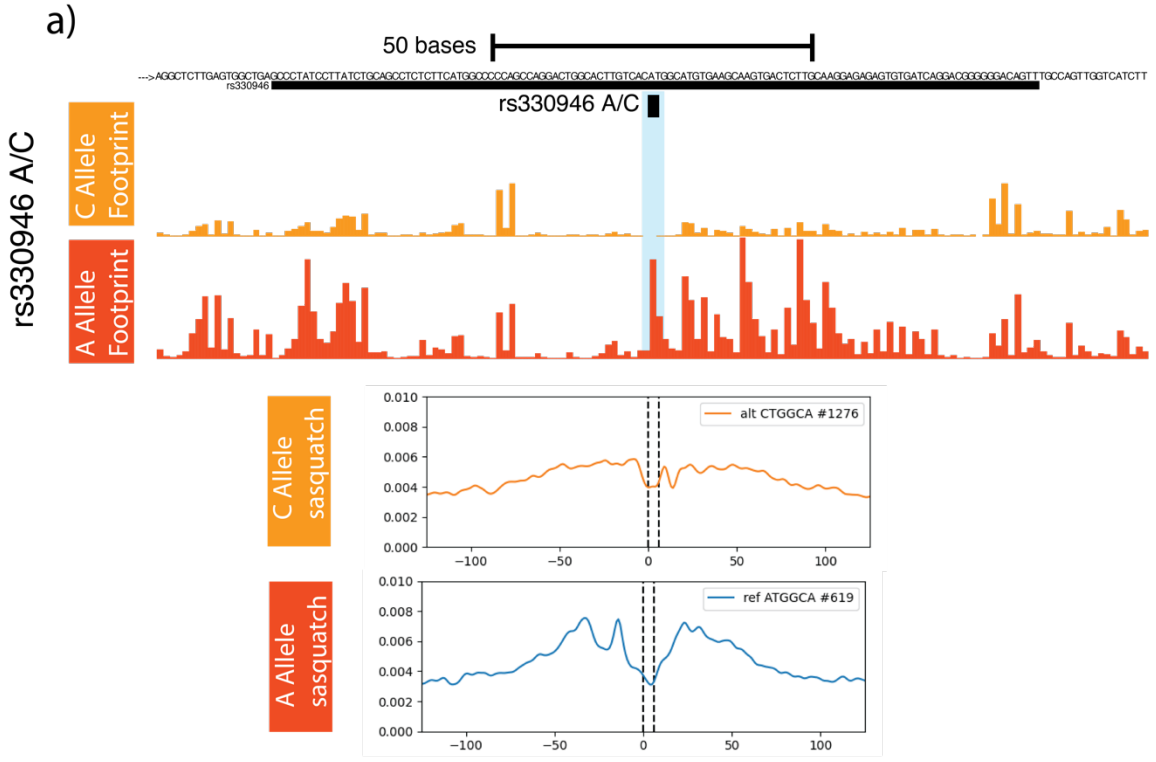

a)
Donor 2 alleles:
Allele 1
Allele 2
rs330946 C
rs3748137 G
rs3748136 G
rs330946 A
rs3748137 C
rs3748136 A
b)
Enhancer
Promoter
CTCF
Capture point
20 kb
PPP1R3B
rs330946
rs3748136
Merged 50 donors
ATAC
H3K4me1
H3K4me3
H3K27ac
CTCF
Donor 2
ATAC
H3K27ac
MCC from rs330946
MCC
C Allele
A Allele
MCC from rs3748136
MCC
G Allele
A Allele
Mature RNA

Figure 3.22: Allelic MCC at *PPP1R3B* from an A/C rs330946 genotype, and a G/A rs3748136 genotype.
a) graphical depiction of the alleles present in donor 2 with the genetics and associated peaks. Allele 1 has the LoF (red box) and so no enhancers. Allele 2 has the GoF (green box) and so two enhancers. Phased genetics confirmed in Supplementary Figure 12. b) Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks show the ATAC (red track) and H3K27ac (green track) for this donor and show open chromatin and activity at both enhancers. Following MCC tracks are from rs330946 (upper) and rs3748136 (lower) (red arrow), with blue tracks showing standard MCC, and yellow and orange tracks showing the allele split tracks from these data. In both cases, all the interactions can be seen to be coming from the lower orange track and not the yellow track.

When comparing the allelic footprints at these SNPs, the peak at rs330946 demonstrates how the lack of an open chromatin peak can influence these data. The differences seen in the figure suggest that the depletion of signal, or the lack of clear strong footprints, are a result of the inactive allele being closed and therefore inaccessible to the MNase. The C allele shows minimal cutting across the length of the probe region, reflecting the closed structure, whereas the A allele with the increased interactions shows clear footprinting. This therefore suggests that the A variant is contributing to a motif that enables the opening and therefore activity of this site. In turn, this means that the alteration to the C allele is disrupting binding and causing the loss of function variant.

When compared to the footprints at rs3748136, both profiles show a relatively comparable distribution of cut sites. The precise point at which the SNP resides in the A allele does seem to reduce the footprint and therefore the binding of a factor, and correlates with this allele's reduced interaction profile. It can therefore be assumed that the alternate A allele is creating a TF binding site which increases the activity and is acting as a gain of function variant.


a)
50 bases
rs330946 A/C
rs330946 A/C
C Allele Footprint
A Allele Footprint
C Allele sasquatch
A Allele sasquatch
alt CTGGCA #1276
ref ATGGCA #619

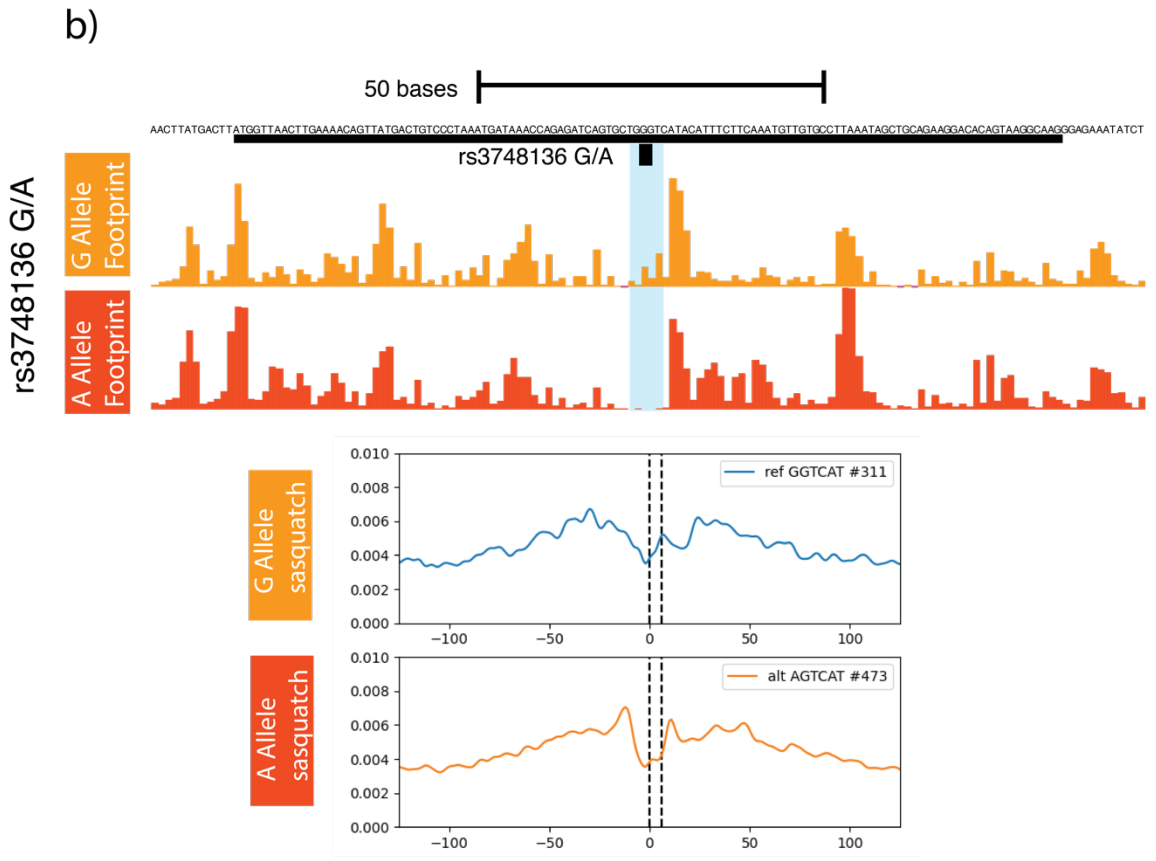

b)
50 bases
rs3748136 G/A
rs3748136 G/A
G Allele Footprint
A Allele Footprint
G Allele sasquatch
A Allele sasquatch
ref GGTCAT #311
alt AGTCAT #473

Figure 3.23: MCC foot-printing data at the *PPP1R3B* locus.
Foot-printing and sasquatch predictions at a) rs330946 and b) rs3748136. In both examples, the orange tracks represent the alleles with interactions. a) Foot-printing is generally stronger for the A allele, showing that this region is more accessible than the other allele. This matches a strong prediction in the sasquatch plot. b) Foot-printing is largely similar across the alleles, showing that both are accessible. Blue highlight shows the SNP location, with a binding site in the orange A track, but not the yellow G track. This matches a strong prediction in the sasquatch plot.

Exploring the interaction profiles at allelic skewed open chromatin sites in characterised individuals

My next set of probes were designed to ATAC peaks that exhibit allelic skew in our erythroid samples (as described in Chapter 1) (MCC probe set 3). In particular, these probes were designed to regions that were shown to be skewed in the four donors in which I manufactured MCC libraries. The majority of probes were also selected as the peaks contained a SNP that was a known GWAS hit, including several studies that found associations with erythroid development phenotypes. With 91 probes and 3 individuals this amounted to 273 viewpoints, with many intentionally chosen for their heterozygous variants to increase that number. From an initial screen I then selected 19 probes from which 28 allelic skew viewpoints were obtained. As in the previous section, I have again selected several examples to demonstrate the effect of these variants, and in particular the behaviour of individual alleles. The *HBA* locus was again used as a control to check the capture efficiency (Supplementary Figure 11). Though the quality of the libraries was prior checked, validation of each new experiment ensures quality capturing, sequencing and analysis.

MCC profiles from regions of open chromatin skew also show interaction skew

The following examples were all designed to peaks in which a donor exhibited allelic skew in the ATAC profiles of late erythroblast cells, matching that of the MCC libraries. For these individuals, if heterozygous variants were present, allelic MCC tracks were also generated.

Figure 3.24 is a capture from rs148015593, a heterozygous SNP found in all three donors captured here. The underlying ATAC peak though small is associated with enhancer marks, and from the MCC tracks can be seen interacting with a downstream enhancer and the upstream promoter of the actively transcribed *MLXIP*. As all three donors are heterozygous and all the allelic specific tracks match, this shows the reproducibility of this technique and analysis across individuals.

Skew analysis in the ATAC-seq from donor 3 showed that the T allele was more active and open than the C allele (Peak on T allele 23 reads, peak on C allele 5 reads). However, when assessing the allelic footprints, a clear increase in cutting can be seen directly at the variant site, and a stronger footprint is seen in the C allele (Figure 3.25a). This contradicts the sasquatch analysis that predicts a strong

footprint in the T allele (Figure 3.25b). By visual inspection it is clear that the T variant associates with a GATA motif, a prominent TF in erythroid cells. Combined, these data do not all align, with the footprinting contradicting the activity seen in the ATAC, the interactions seen in the MCC and the predicted GATA binding site from the Sasquatch prediction. This example demonstrates the technical consistency between the donors, but that certain variants remain challenging to fully understand.

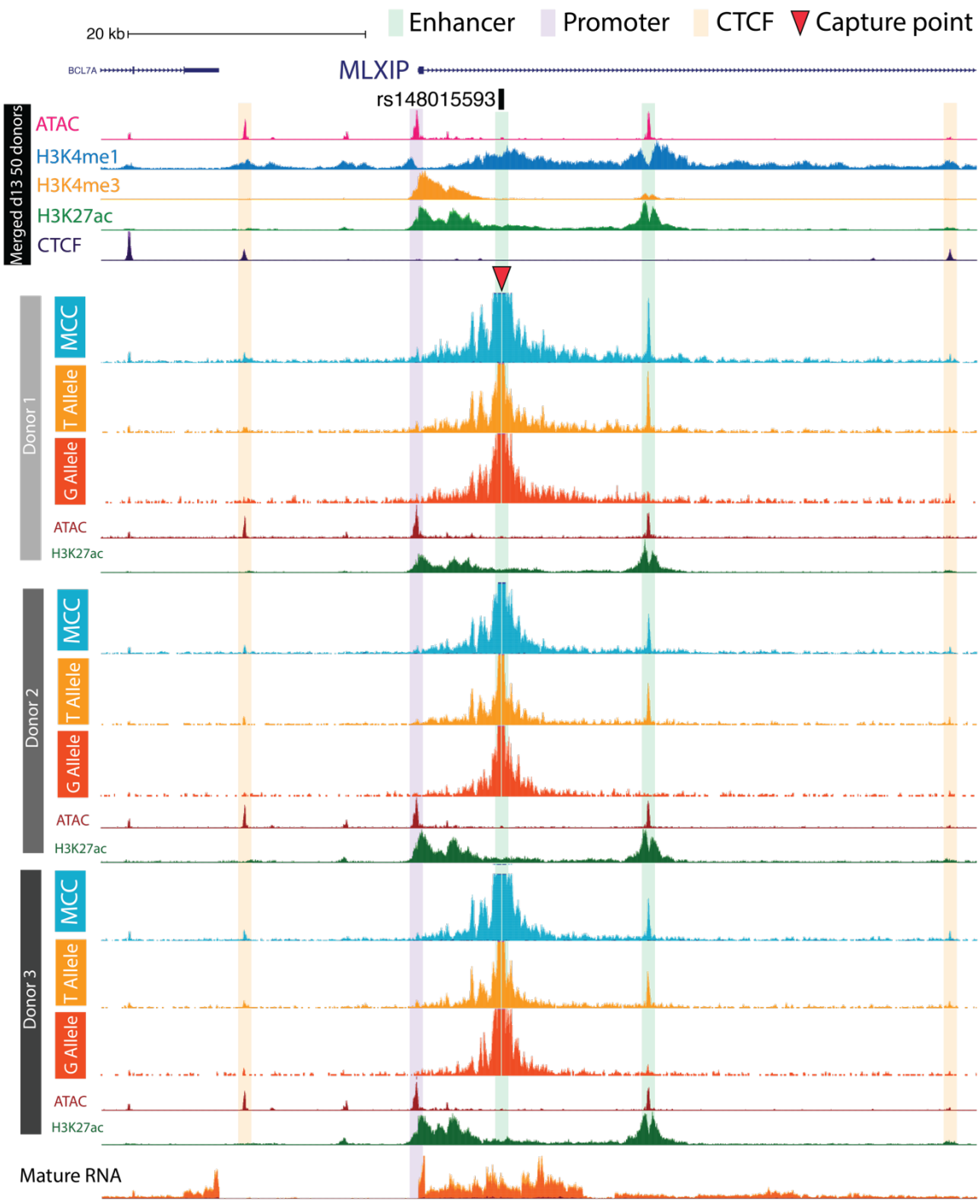
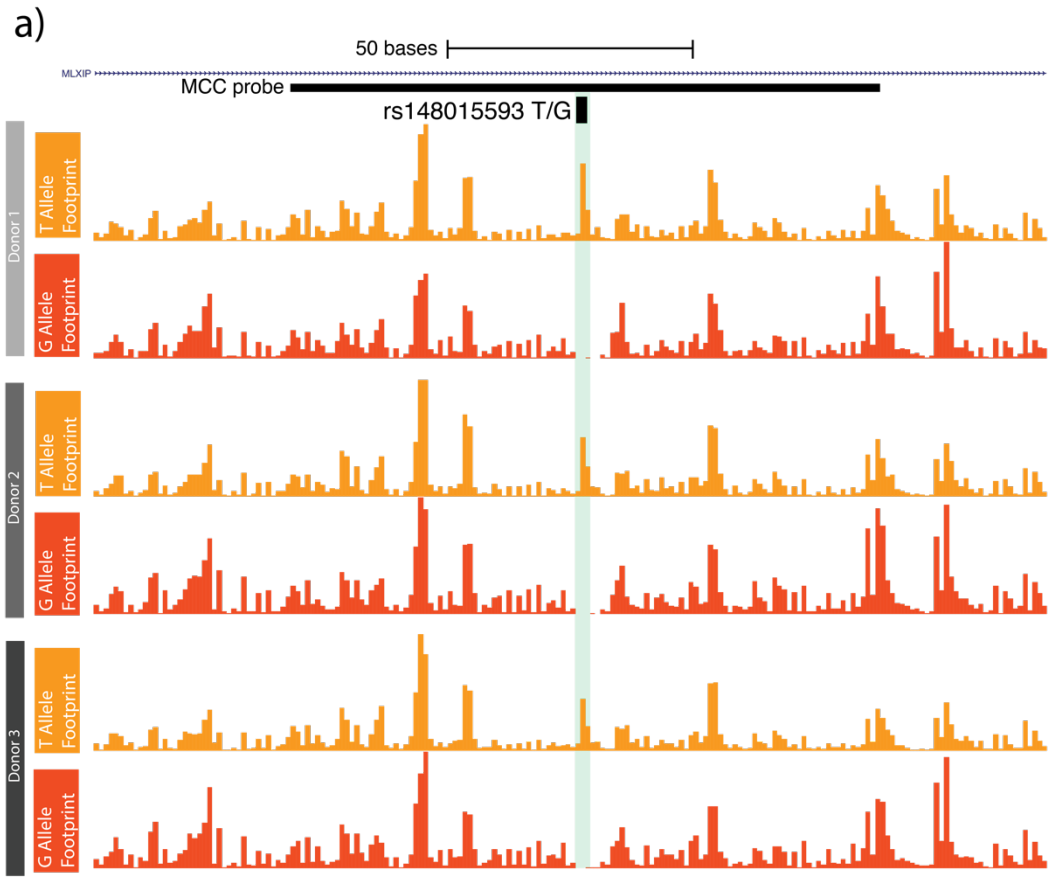


Figure 3.24: Allelic MCC from rs148015593 shows consistency across individuals.
Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks from three individuals, with blue standard MCC

tracks captured from rs148015593 (red arrows), yellow allelic track from T allele and orange allelic track from the G allele, red track ATAC and green track H3K27ac. All donors show the same allelic split pattern across the locus and interactions, with more interactions from the T allele.



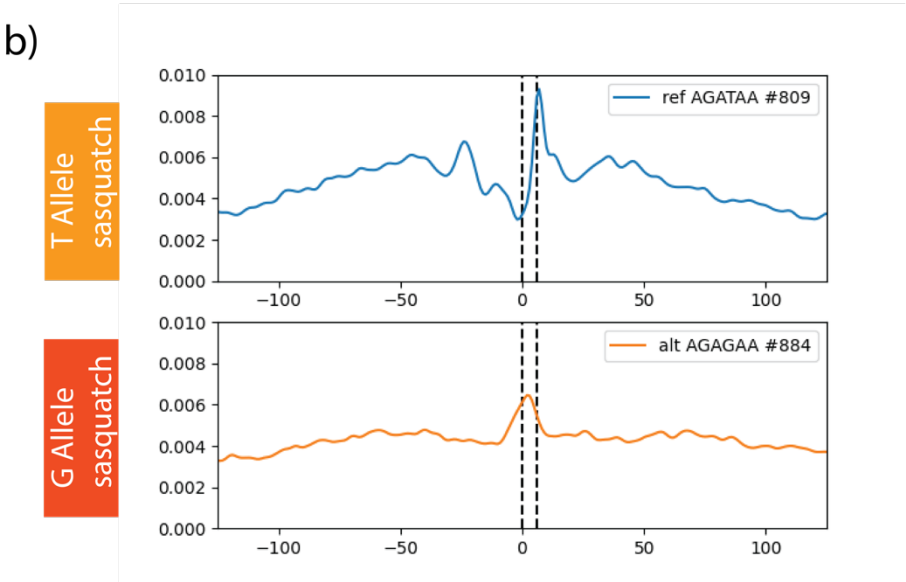


Figure 3.25: Allelic MCC footprinting from rs148015593.
a) MCC probe used for capture in black with rs148015593 SNP used for allelic split, location in green highlight. Yellow and orange tracks show footprints at T and C alleles respectively. FP in G allele is not consistent with the reduced interaction in G allele in Figure 3.24. All tracks consistent across all the donors. b) Sasquatch analysis shows footprint prediction when either a T (top track) or G (bottom track) is present in the sequence. Trace depicts a footprint prediction, with a higher signal predicting a lack of cutting and a lower signal predicting a bound factor protecting the DNA. Upper T track shows a strong footprint, likely a GATA site. Lower profile shows the G allele with no predicted FP.

MCC profiles from skewed CTCF sites show strong interaction skew

The next example demonstrates allele specificity of interactions across CTCF sites. As in Figure 3.12, these CTCF tracks show very strong, precise, far reaching interactions to neighbouring CTCF sites over 70kb away. This is to be expected from these structural proteins, and is indicative of TAD structures and DNA loops, and also corroborates the existence of larger scale rosette structures [26]. As in previous examples, the allele tracks show the complete loss of interactions between variant tracks, and ChIP shows a depletion at the capture site. Donor 1 is homozygous C/C with no interactions, donor 2 is homozygous G/G with all the interactions, donor 3 is heterozygous C/G with the allelic split. The causality of the variant is confirmed by loss of MNase footprinting and further supported by a predicted loss of footprint in Sasquatch analysis (Figure 3.27).

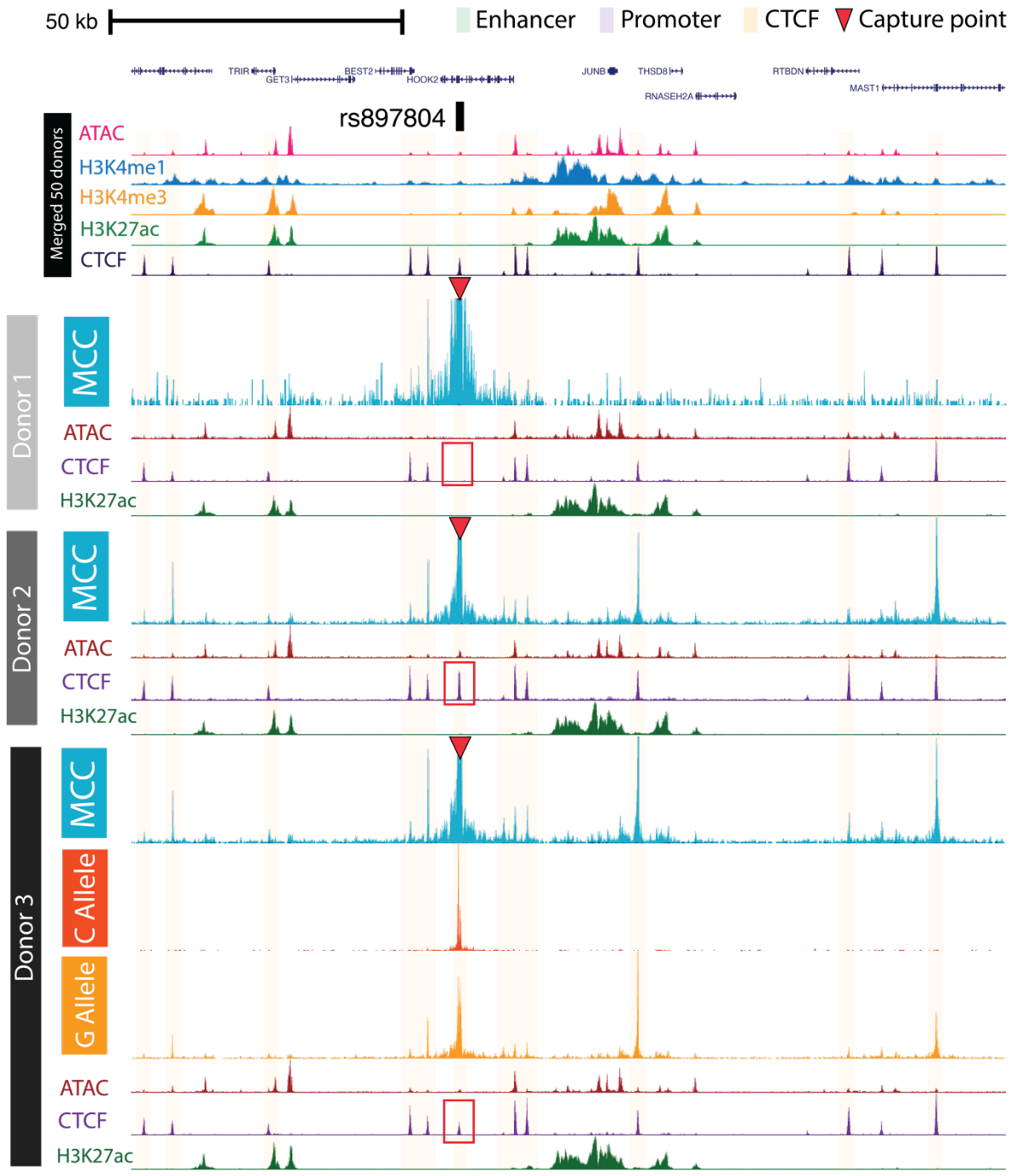


Figure 3.26: Allelic MCC tracks from rs897804 show differing CTCF interactions.
Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks from three individuals, with blue standard MCC tracks captured from rs897804 (red arrows), orange allelic track from the C allele and yellow allelic track from G allele, red track ATAC, purple track CTCF ChIP-seq and green track H3K27ac. Red box highlights the change in CTCF binding across the individuals,

which in turn reflects the interactions shown in the blue MCC tracks, with CTCF binding correlating with interactions. When split in the heterozygous individual, the interactions are coming from the G allele.

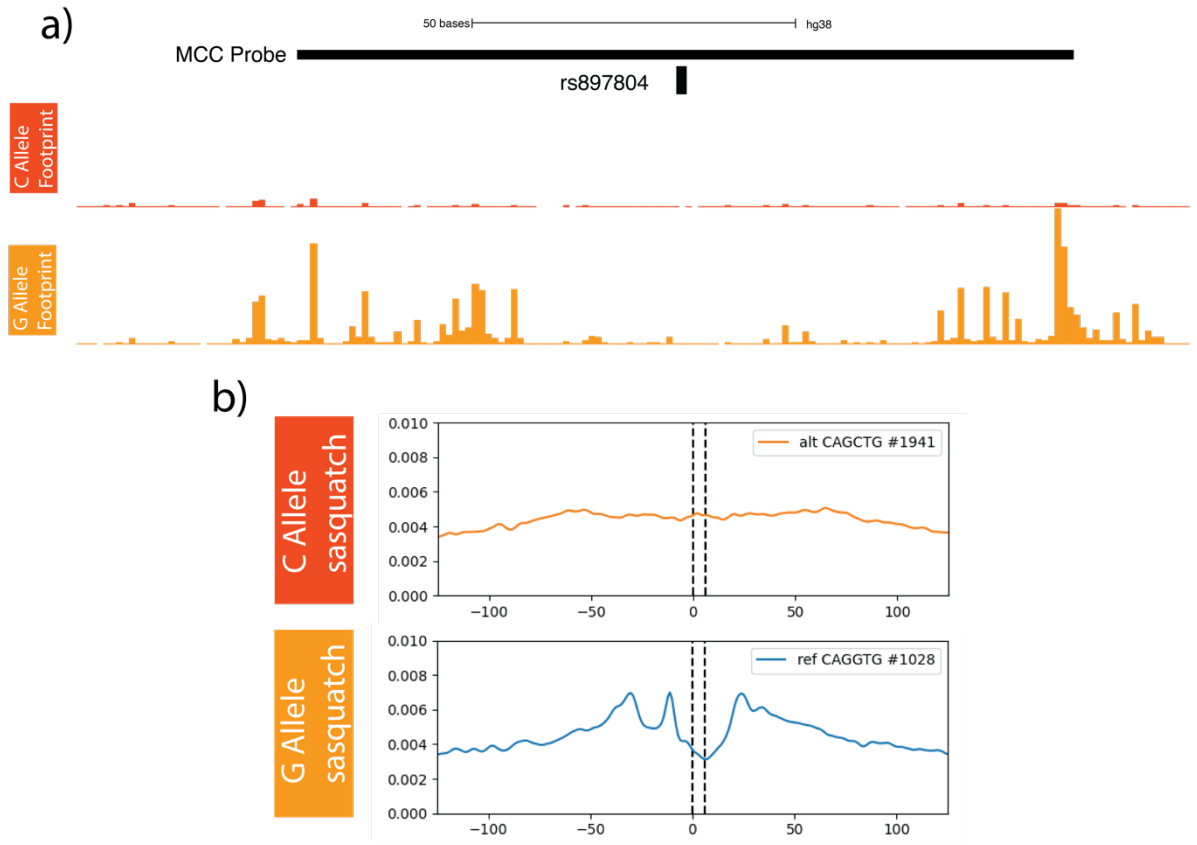


Figure 3.27: Allelic MCC footprint at rs897804.
a) MCC probe used for capture in black with rs897804 SNP used for allelic split. Orange and yellow tracks show footprints at C and G alleles respectively. Orange C allele shows no footprint, consistent with a lack of activity. Yellow G allele shows accessibility and TF binding. b) Sasquatch analysis shows footprint prediction when either a C (top track) or G (bottom track) is present in the sequence. Trace depicts a footprint prediction, with a higher signal predicting a lack of cutting and a lower signal predicting a bound factor protecting the DNA. Upper C track shows no footprint. Lower profile shows the G allele with strong FP prediction, matching with the real interactions shown in the MCC.

MCC profiles from a skewed enhancer containing a GWAS variant show differential interactions with JAK2 promoter

Rs10758656 resides within an enhancer and is present as a heterozygous in donor 2. This SNP is associated with several erythroid traits, including mean corpuscular haemoglobin [72]. An A>G change

destroys a binding site for the key erythroid TF GATA1 and leads to strong skew in open chromatin as well as loss of binding footprints in MCC footprints and Sasquatch analysis (Figure 3.29). Donors 1 and 3 are both homozygous A at this SNP, and both show strong interactions to the surrounding landscape. Donor 2 however is a heterozygous A/G, and the allele split tracks from this heterozygous donor show that all the interactions are coming from the A allele that associates with open chromatin. This active enhancer can be seen to be interacting with other regulatory elements including promoters, CTCF sites and other enhancers (Figure 3.28). Importantly, one of these genes is Janus-kinase 2 (*JAK2*) which has been shown to have a role in haematopoiesis and is therefore involved in the development of erythroblasts [183]. The interaction of this enhancer to the surrounding promoters suggests it plays a role in the regulatory network of this gene, which in turn could be having an effect on erythroid development. This hypothesis was previously confirmed in the Hughes group using HDR CRISPR editing to introduce the variant into the HUDEP erythroid cell model and showed that the presence of the variant was associated with decreased expression of *JAK2 [184]*. This experiment therefore shows for the first time, in high resolution, the change in regulatory interactions that underlie that change in expression. Future analysis using the phased genomes to determine the allelic RNA output as described at *SCARB2* will also be used on this locus to determine whether the SNP is having an influence of gene expression.

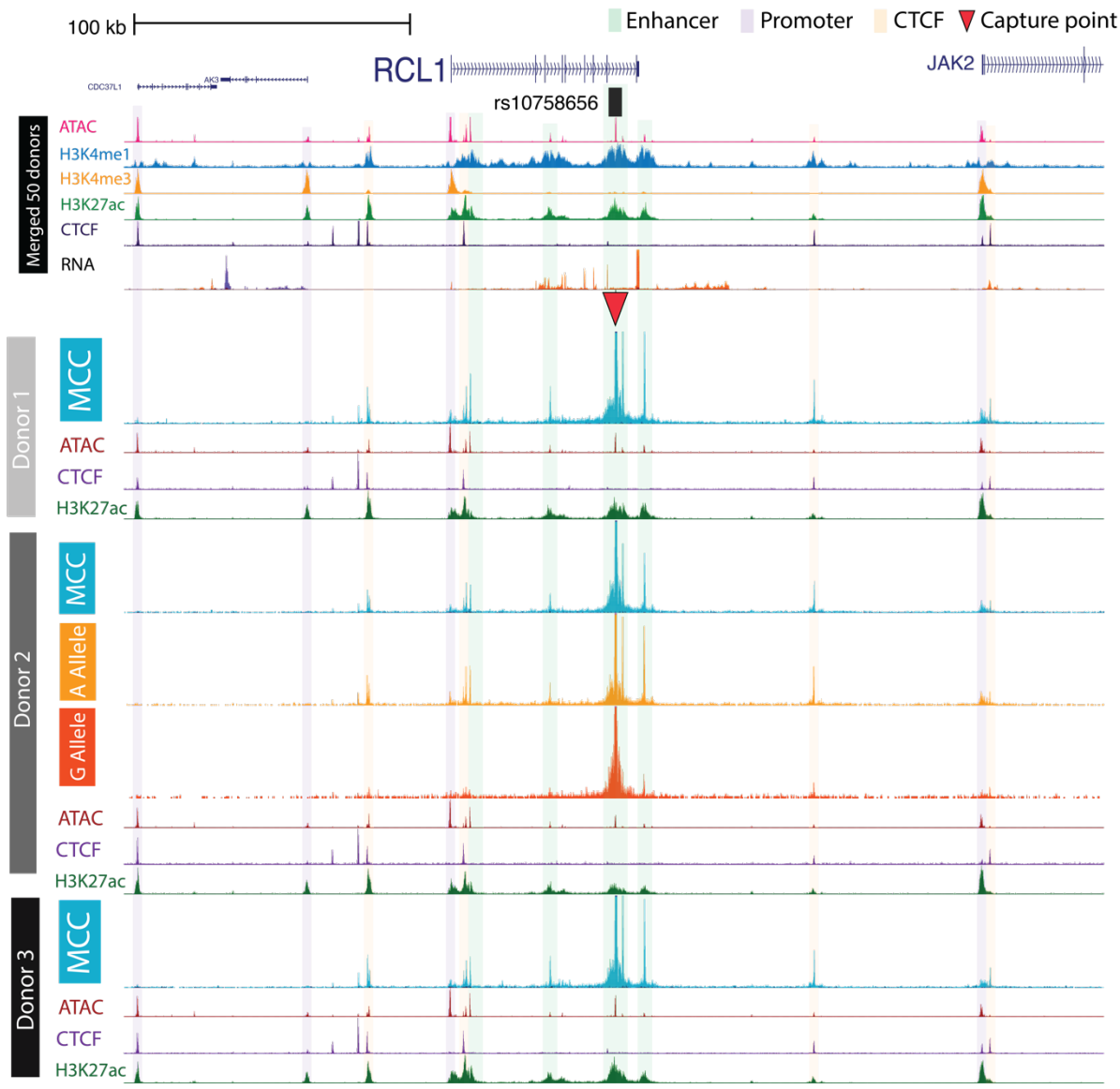


Figure 3.28: Allelic MCC tracks from rs10758656 show differing enhancer interactions to *JAK2*.
Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks from three individuals, with blue standard MCC tracks captured from rs10758656 (red arrow), yellow allelic track from the A allele and orange allelic track from G allele, red track ATAC, purple track CTCF ChIP-seq and green track H3K27ac. From the heterozygous data, more interactions can be seen from the yellow A allele, but some interactions can also be seen from the orange G allele.

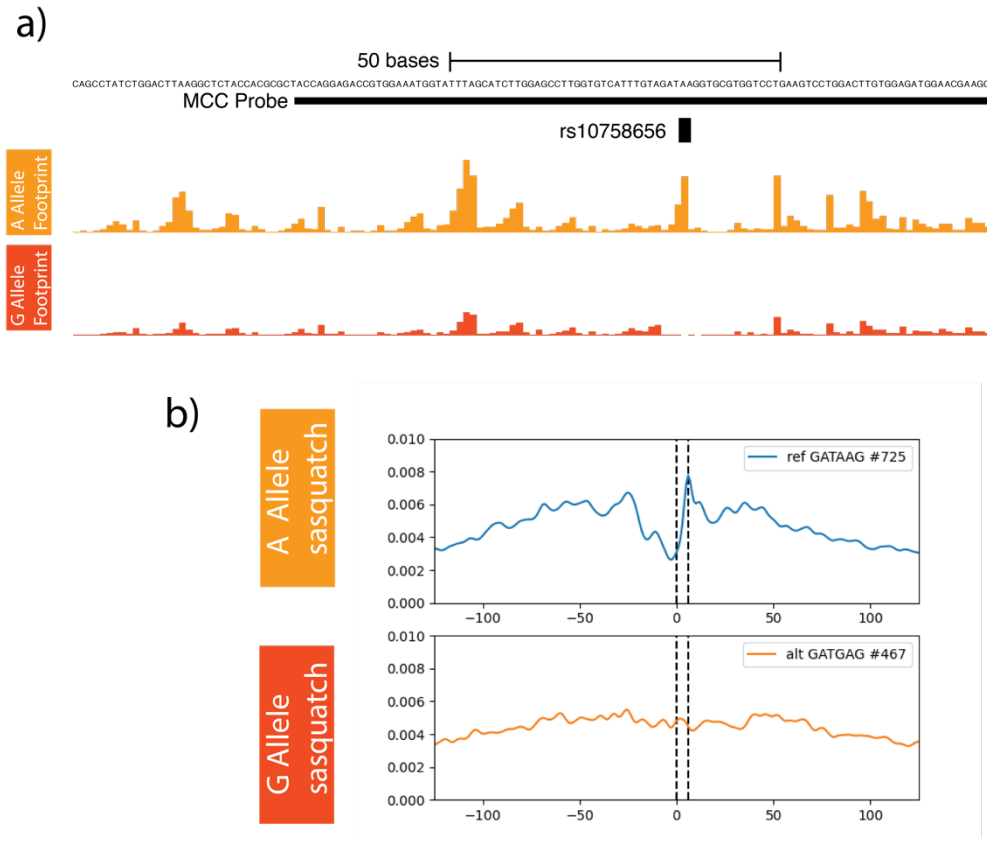


Figure 3.29: Allelic MCC FP from rs10758656.
A allele is more accessible and correlates with more binding, shown in the sasquatch prediction. a) MCC probe used for capture in black with rs10758656 SNP used for allelic split. Yellow and orange tracks show footprints at A and G alleles respectively. Yellow A allele shows accessibility and TF binding. Orange G allele shows no footprint, consistent with a lack of activity. b) Sasquatch analysis shows footprint prediction when either an A (top track) or G (bottom track) is present in the sequence. Trace depicts a footprint prediction, with a higher signal predicting a lack of cutting and a lower signal predicting a bound factor protecting the DNA. Upper track shows the A allele with strong FP prediction, matching with the real interactions shown in the MCC. Lower G profile shows no footprint.

MCC profiles from an enhancer containing the erythroid GWAS variant rs696281 show strong allelic skew in a highly variable region

Rs696281 is associated with several GWAS and erythroid traits including altered mean corpuscular haemoglobin [70, 72]. The functionality of the element in which this SNP resides is therefore likely to be active during erythropoiesis. Across the donors 1, 2 and 3, it is present as homozygous T/T, heterozygous T/G and homozygous G/G respectively, providing a useful example to look at the

effected interactions in both homozygosity and heterozygosity. ATAC-seq from these individuals shows a strong open chromatin site associated with the homozygous T individual, and an almost total loss in the individual with homozygous G. The SNP itself resides in an enhancer, and in Figure 3.30 can be seen to interact in MCC with many other elements of the local landscape, including upstream and downstream enhancers at distances of 420kb and 540kb respectively. These long-range chromatin interactions reveal this locus to be a large complex regulatory domain.

Though the open chromatin sites differ across the individuals, the TF footprint profiles are relatively comparable (Figure 3.1). A slightly larger FP can be seen on the T allele, as well as stronger footprinting which is supported by the stronger open chromatin peak. This increase in hypersensitivity and TF binding associates with the interactions. Sasquatch analysis predicts the SNP to have a very weak effect, although the G allele has a slighter stronger FP. This contradicts the real FP data, which combined with a lack of strong motifs at this site, further suggests the effect is not that strong.

When zooming into the more proximal interactions (Figure 3.32), more contacts with the local enhancers can be seen, as well as strong interactions with the promoter of the active *BZW2* and *ANKMY2*. These encode Basic Leucine Zipper And W2 Domains 2 and Ankyrin Repeat And MYND Domain Containing 2 respectively. In particular, in between the promoter and captured enhancer is another cluster of contacted enhancers. These show a pattern reminiscent of the *CLNK* variant peaks (Figures 2.15 and 3.17) in that the activity of these peaks appears to be associated with the activity of the variable peak that contains the variant. These interactions are flanked by a pair of CTCF sites, and although interactions are found outside of this region, the majority are confined to within these boundaries. When split allelically, despite some weak interactions to the promoter region in the C allele, the vast majority of the contacts are being formed by the T allele (Figure 3.30 and 3.32).

To further investigate the apparent functional link between the variant containing peak and the other local polymorphic peaks their co-variation in the 50 individuals can be investigated. Figure 3.33 shows the open chromatin landscape of this region across the individuals, with the rs696281 T/T, T/C and C/C individuals grouped and coloured together. Despite the distinct landscapes that matched up with their genotype in donors 1, 2 and 3, this pattern was not reflected in the other individuals. Instead, several patterns of variation with multiple combinations of peaks and no peaks appearing both at the

rs696281 site and in the local landscape are seen. Furthermore, the pattern of this landscape seems to be unrelated to rs696281. Additionally, the presence or absence of the upstream enhancer peaks did not correlate with the presence or absence of the rs696281 peak. This locus further exemplifies how complex the genetics of such regions can be, and how different the landscapes of individuals are.

In these examples, the interactions and therefore regulatory networks of the donors captured from them can be confidently called. However, predicting the interactions present in the rest of the donors at this locus could prove challenging. Instead, for each individual a full dissection of the genome sequence, the epigenetic landscape and the interactions seen would be needed to fully understand the regulatory network. This locus represents a highly complex region. This therefore highlights the importance of the genome context in which a SNP resides, not just between cell types, but also individuals and also their distinct combinations of alleles inherited at the same locus. Potential variation affecting the activity of these other variable peaks is now under investigation.

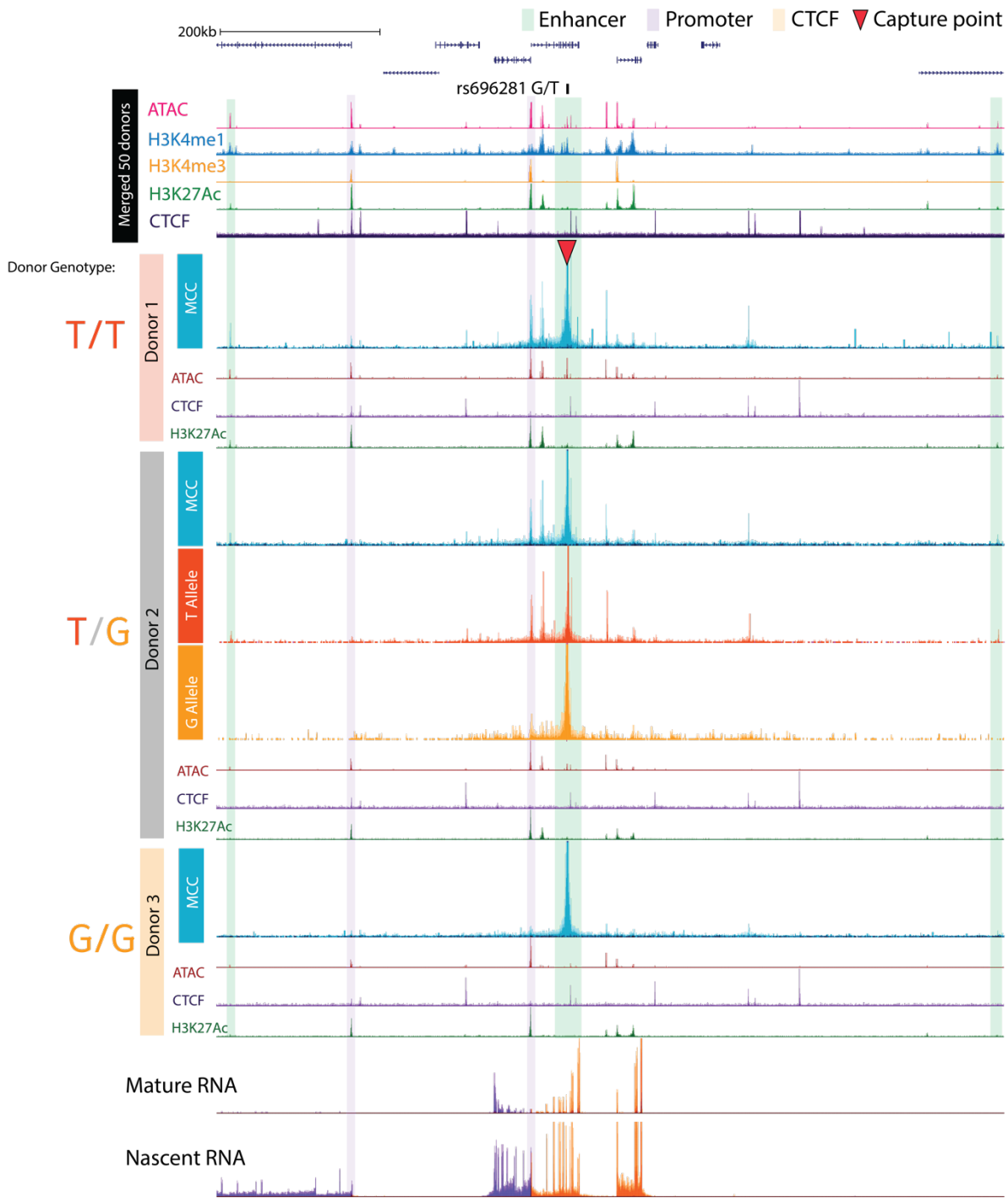


Figure 3.30: Allelic MCC tracks from rs696281 show differing enhancer interactions.
Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Nascent RNA (PolyA-) and Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks from three individuals, with

blue standard MCC tracks captured from rs696281 (red arrow), orange allelic track from the T allele and yellow allelic track from G allele, red track ATAC, purple track CTCF ChIP-seq and green track H3K27ac. From the heterozygous data, more interactions can be seen from the orange T allele.



Figure 3.31: Allelic MCC FP from rs696281.
T allele is more accessible and correlates with more binding, with a subtle change in the sasquatch prediction. a) MCC probe used for capture in black with rs10758656 SNP used for allelic split. Orange and yellow tracks show footprints at T and G alleles respectively. Both the tracks show accessibility and TF binding. b) Sasquatch analysis shows footprint prediction when either an T (top track) or G (bottom track) is present in the sequence. Trace depicts a footprint prediction, with a higher signal predicting a lack of cutting and a lower signal predicting a bound factor protecting the DNA. A slightly stronger FP prediction can be seen in the G allele.

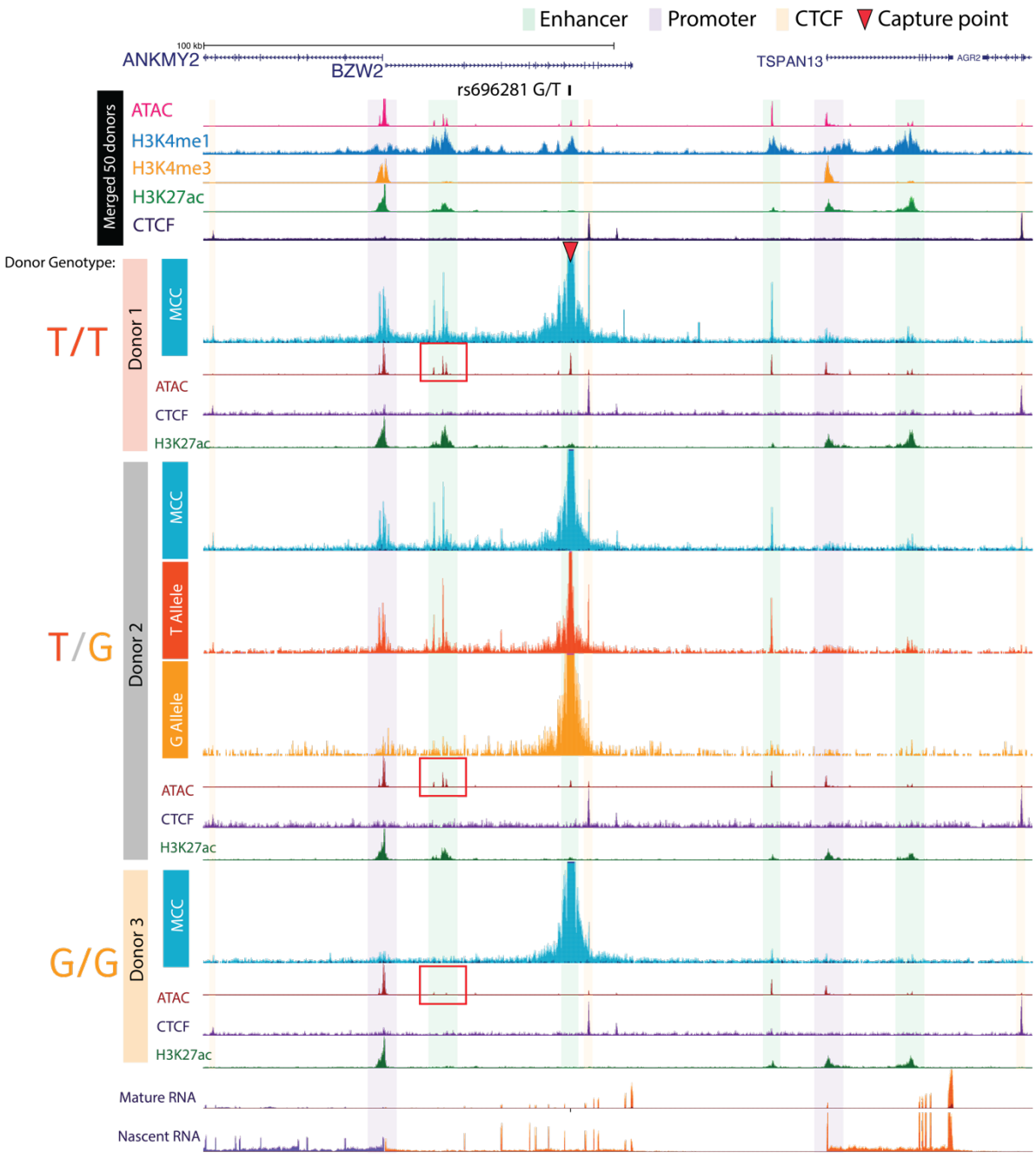


Figure 3.32: Close up of allelic MCC tracks from rs696281 show differing enhancer interactions.
Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Nascent RNA (PolyA-) and Mature RNA (PolyA+) combined from 50 donors, with orange signal denoting forward transcription and purple signal denoting reverse transcription. Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks from three individuals, with blue standard MCC tracks captured from rs696281 (red arrow), orange allelic track from the T allele and yellow allelic track from G allele, red track ATAC, purple track CTCF ChIP-seq and green track H3K27ac. From the heterozygous data, more interactions can be seen from the orange T allele. Red box indicates enhancers that are only appear simultaneously with the peak captured from.

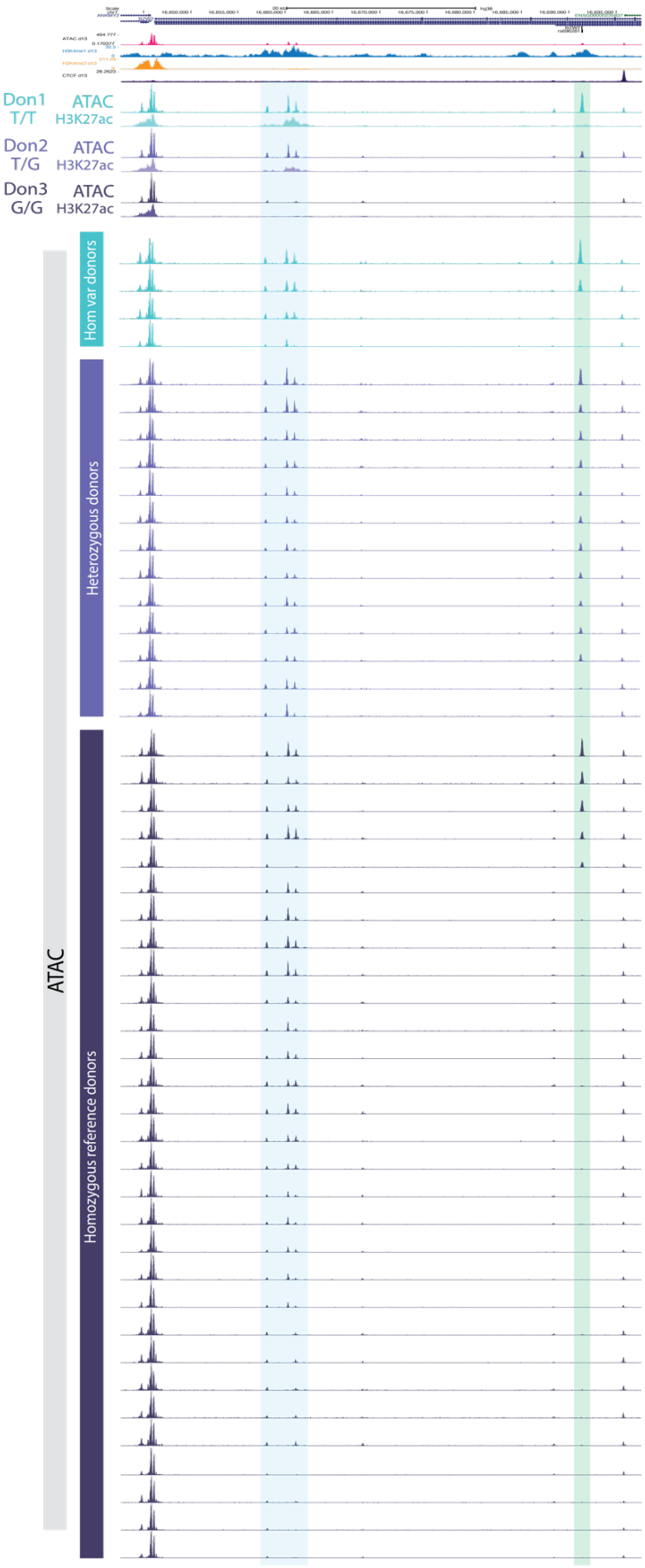

Don1
T/T
ATAC
H3K27ac
Don2
T/G
ATAC
H3K27ac
Don3
G/G
ATAC
H3K27ac
Hom var donors
Heterozygous donors
Homozygous reference donors
ATAC

Figure 3.33: Variation at the *BZW2* locus shows landscape pattern is independent of SNP.
Variation at the region shown in figure 3.29. Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green highlights depict enhancers with rs696281, from which genotype distinction is made. Blue highlight depicts further variable region Next tracks depict the ATAC (darker) and H3K27 (lighter) from individuals with a T/T (blue), T/C (violet), and C/C (purple) genotype at rs696281. Following tracks are the remaining 47 donors colour coded similarly: T/T (blue), T/C (violet), and C/C (purple) genotype at rs696281. Despite what seemed to be a fairly straight forward pattern demonstrated in the initial three donors, this is not reflected across the whole population, with multiple different patterns that do not match the genotype at rs696281.

Discussion

In this chapter I have set up MCC in primary human erythroid cells and presented high resolution interaction data. From these tracks I have generated allele specific interaction profiles and footprinting. By capturing from regions of interest found in Chapter 2, I have demonstrated methodologies for finding and functionally validating human variants by uncovering their interaction profiles. Uncovering cell specific regulatory networks is a crucial step in functionally validating variant effects, and MCC provides precise interaction profiles of unprecedented resolution.

Optimising and implementing MCC in primary human cells

My first objective was to set up the MCC protocol in primary human cells, which required optimisation and adaption of the existing protocol. The importance of limiting excessive MNase digestion after one hour was highlighted by the reduced efficiency of the ligation in over digested libraries. Losing the linker sequence massively reduced the chances for fragment ends to contact and ligate, and so effectively stopping residual MNase activity was crucial. Furthermore, careful treatment of cells during the shaking of digestion and ligation reactions reduced DNA loss and low library yields. Cell numbers remain a limiting factor for MCC, as a degree of MNase titration is required for every library preparation. For my samples that had lower cell numbers to work with, I was able to do an investigative scaled down titration in order to successfully make libraries without the need for multiple sample aliquots. Work by the Davies lab is ongoing to further lower the cell number requirements.

My second main objective was to implement allelic MCC to determine the effect of variants on regulatory interactions in a highly controlled manner. Determining the allele from which a read and therefore interaction originates allowed for the sensitive detection of changes associated with variants, especially between enhancers and promoters. Enhancers interacting with genes including *SCARB2*, *PPP1R3B*, *JAK2* and *BZW2* have all been shown to contain SNPs that have enabled allele specific MCC. More importantly, from all of these examples a clear difference can be seen between the alleles. Additionally, allele specific footprinting has been used to show a variants ability to disrupt TF and structural protein binding. This is especially clear in the examples of variants affecting the binding of CTCF. CTCF is instrumental in the formation of chromatin loops and therefore shows strong,

precise interactions with neighbouring CTCF sites, which are lost when a site is disrupted. Defining these interactions between different classes of REs is a vital step in identifying the target genes of variant affected regulatory elements. Moreover, these allele specific MCC examples demonstrate the importance of assessing each allele individually when functionally validating a variants impact on gene expression.

Importantly, as well as showing examples of interactions changing in the presence of variants, I have also shown the effects of a neutral SNP. Rs11248850 within the R2 enhancer of *HBA* was shown to have no effect on the local interactions, nor did it change the footprint or TF binding at this site . This demonstrates a silent base pair change, and highlights that not all variants in an active element will affect interactions and acts as a control for the allelic analysis of MCC data. As an additional control, capturing from multiple donors allowed for the comparison of profiles at sites where genetics differed and therefore comparable profiles were generated. In several examples where the three captured donors exhibited the full range of homozygous alternate, homozygous reference and heterozygous genotypes of a SNP, the effects of these interactions were often clear. This further highlights the power of having several donors from which to capture from.

The ability of MCC to produce TF footprint profiles (also allelically) has been used to demonstrate variants disrupting TF binding. These effects can also be predicted with Sasquatch analysis to aid in the deciphering of variant mechanisms. Many of the examples shown here, though not all, have shown the ability of Sasquatch to accurately predict these changes. These approaches are especially important when considering the limitations of using motif analysis and ChIP-seq. Thousands of potential motifs exist, and so predicting them or showing TF binding to them via ChIP is costly in both time and resources. Though it does not prove exactly which TF is bound or disrupted by a SNP, footprinting is a powerful tool that can be used to prove that a change in binding has occurred. Due to the extremely deep data that is captured and sequenced over an MCC probe site, it means the data generated from this assay is highly informative.

MCC can be used to decipher the impact of GWAS variants

One of the ultimate goals of using MCC and its allelic capabilities is to functionally link SNPs of interest in enhancers to the genes they control. This capability has been highlighted by capturing from the SNP rs10758656 that lies in an enhancer upstream of *JAK2*. This SNP and enhancer have previously been shown by Capture-C and genome editing to interact with *JAK2* and be associated with a reduced *JAK2* gene output. The use of MCC here provides a higher resolution interaction profile to further confirm this and to show the utility of MCC data at a characterised locus. Additionally, the allelic tracks generated further prove that the A allele associated with expression has more interactions than those from the G allele that associated with lowered expression. This example shows that MCC could be used on further loci and SNPs of interest to decipher the functional mechanisms of these variants.

This approach can be combined with the allele skew analysis demonstrated in the previous chapter for the prioritisation and then interaction determination of heterozygous variants. Though considerable computational development is needed to deploy at scale, I have piloted an approach that combines these allele specific epigenetics with allele specific MCC. Additionally, the implementation of phased allele specific RNA expression, in combination with these data, can be used to determine further the regulatory effect of variation on genome interactions and subsequent gene expression.

Several CTCF site disruptions have been exhibited and validated in this chapter. However, using these inferences to discern their full effects on regulatory networks remains challenging. Disrupting CTCF binding can have a profound effect on genome architecture and domain structures, and MCC can show the effects at the boundaries of these domains. However, it will not fully decipher how the other elements within these regions are interacting, or more importantly the downstream effects of this structural disruption. As shown in previous studies, altering boundaries and domains can lead to ectopic promoter-enhancer interactions that cause disease [185]. From these observations, it could be hypothesised that damaging a CTCF site could result in such ectopic interactions. It would therefore be logical to do further captures from the other REs around a disrupted CTCF site to see if or how the interactions change. This can be used in combination with phased transcriptomics to also assess gene

expression changes of genes on the allele with the variant CTCF sites. This would then provide more information that could be used to decipher the regulatory networks of the region.

The work presented in this thesis so far has demonstrated just how much variation there can be at loci in a human population, and that with MCC we can determine the landscape at these regions. However, it has also highlighted how complex the human genome can be within a population, with certain loci having multiple haplotypes coinciding. In the example of rs696291 at *BZW2*, polygenic alleles make fully deciphering the region challenging, especially when considering the amount of variation between individuals. These examples and those of the GoFs show that each locus within an individual can be behaving in a unique and complex way. Because of this, a thorough and systematic approach, as well as an appreciation of this potential complexity and variation, is needed to functionally validate variants of interest.

In summary, by capturing from sites containing heterozygous variants, I have generated allele specific interaction profiles from each allele in our human erythroid cell model. From heterozygous individuals, we can see differential interactions from each allele, from within the same cell population. Moreover, by combining these insights with ATAC and RNA data, I have shown that these differential interactions are correlated with reduced hypersensitivity and mRNA expression. Combined, MCC is able to unambiguously link variants in open chromatin regions to other regulatory elements and the genes they control, allowing us to infer the activity imparted by their presence and interpret GWAS hits.

Having performed these analyses in late erythroblasts I next wanted to assess the genomic landscape in earlier cell populations. In the following chapter I will use a scMultiome system to explore open chromatin and gene expression profiles across earlier stages of erythropoiesis. This can be used to further find variant regions of interest, from which we can use 3C techniques such as MCC to elucidate interaction profiles.

Acknowledgements

For this chapter I extracted and cultured all CD34 cells from the biobank donors. I performed every step of the MCC protocol including: the fixation of cells, library generation, probe design, capture, sequencing, analysis and data visualisation.

Early MCC optimisations were performed on cells cultured and fixed by Mohsin Badat. Rosa Stolper provided cell culture support and MCC fixation during restricted lab time. Jennifer Herrmann, Tom Wilson and Nicholas Denny provided help and support with MCC probe design. Tim Rostron of the sequencing facility supported the sequencing runs. Joseph Hamley consolidated scripts for allelic MCC splitting analysis. Matt Baxter provided technical support during MCC analysis. Analysis by Edward Sanders and Jeanne Inchauspe described in Chapter 2 was further used for MCC probe design. Martin Sergeant helped with sasquatch plot generation. Tom Bolton helped with the visualisation of the allelic RNA-seq data.

Chapter 4 - Assessing the changing genomic landscape of erythropoiesis

Introduction

The crucial role of regulated gene expression for accurate cell form and function has been well defined [12, 186]. The fine-tuned spatiotemporal expression of genes via regulatory elements ensures that the right proteins and transcription factors are present when and where needed. During cell differentiation, pluripotent and multipotent stem cells rely on a unique cocktail of factors to progress down a lineage and become a terminally differentiated cell [127]. During this process, the genomic landscape is dynamically changing, with active regions opening and closing, concomitant with the formation of regulatory interactions and changes in gene expression. Cellular differentiation generally proceeds from early multipotent cell lineages to more restricted and more terminally differentiated cells [187]. This progression means that both the quantitative output of differentiation, as well as the form and function of the cells produced further down a lineage can be altered by earlier changes in the biology of preceding cell stages. While the functional genomics analysis of GWAS signals is often focused on the more easily attainable later stages in differentiation, this does not exclude the possibility that many GWAS signals may affect early stages of differentiation.

Therefore, for a more complete dissection of a set of GWAS traits it is necessary to consider all stages of differentiation prior to terminal differentiation. It is therefore also essential to know which SNPs intersect with the changing landscape of active elements across the complete arc of a differentiation. These SNPs could potentially be causal in a trait, as SNPs can alter elements that work at any precise time in differentiation.

Traditionally, generating a complete dataset across multiple stages of differentiation has been a challenging task, both due to the need to purify each population prior to analysis and the often-limiting numbers of cells available [188, 189]. Much of this challenge has now been met with modern

single cell technologies making earlier differentiation and developmental cellular niches available for analysis [190-192].

The 10X chromium single cell multiome platform can simultaneously assay the open chromatin and gene expression of a heterogenous population of single cells. Utilising a unique barcoding system, DNA and RNA from the same cells are indexed and sequenced. After the data is generated, downstream analysis using either 10X or in house software identifies different cellular populations as maps of clustered cells that reflect the similarities between the constituent cells. In samples with distinct cell types, such as PBMCs, these cell types will cluster distinctly and with separation. However, within a differentiating population, the gradual changing of genomic landscapes makes this more challenging. Analysed cells will instead show a trajectory of cellular differentiation rather than distinct clusters. Additionally, using the combined expression profiles and epigenetic landscapes of these cells adds extra information for the clustering analysis. This combination provides insights into the expression of marker genes at each differentiation stage, allowing for the more precise mapping of defined populations within the trajectory. Moreover, by overlaying GWAS data, we can assess SNP enrichment within the open chromatin sites of identified populations of differentiation stages.

Using scMultiome to map erythroid differentiation

The scMultiome workflow starts with the lysis of a population of cells to a single nuclei suspension. These nuclei are then counted and incubated in a transposition mix where mutated Tn5 transposases enter the nuclei to simultaneously cut and add adapter sequences to the DNA preferentially at open chromatin regions. These fragments are retained within the nuclear membrane. Following transposition, nuclei are incorporated into GEMs along with a library master mix. GEMs are formed from oil and contain a uniquely barcoded gel bead, coated with: a poly(dT) sequence that generates a cDNA from poly-adenylated mRNA and produces a gene expression (GEX) library; and a spacer sequence that enables barcode attachment to transposed DNA fragments to generate an ATAC library (Figure 4.1).

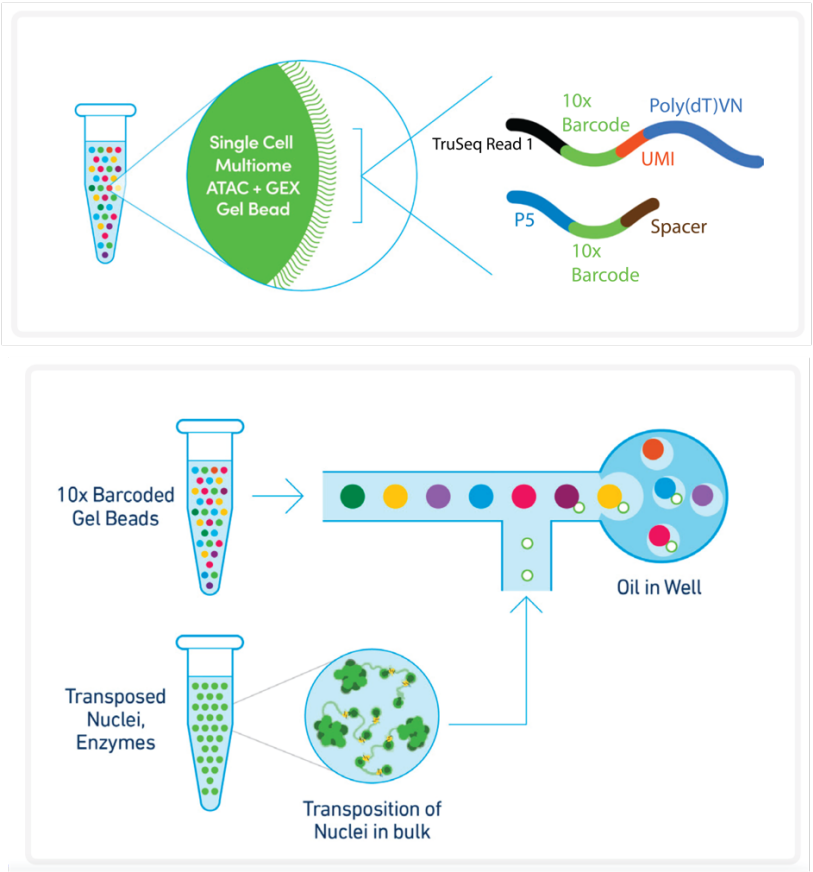


Figure 4.1: scMultiome uses GEMs and unique barcodes to label tagmented DNA and RNA from single cells. scMultiome ATAC + GEX Gel Beads have a poly(dT) sequence for generation of barcoded, full-length cDNA from poly-adenylated mRNA to assess gene expression (GEX) library, a Spacer sequence that barcodes attached transposed DNA fragments for ATAC library. GEMs are made of barcoded Gel Beads, transposed nuclei, a Master Mix, and Partitioning Oil. Single nuclei resolution is achieved by limiting dilution, with the majority (~90- 99%) of generated GEMs containing no nuclei, and the rest containing a single nucleus. (adapted from 10x document CG000338 Rev B)

Unlike many other cell models that rely on assaying terminally differentiated cells, the erythroid cells used in this chapter were harvested at different time points through differentiation to assess the stages of a differentiation arc. To be confident that a population of cells have progressed down a certain lineage into the expected cell types, several checks can be done. These include assessing a cells morphology with microscopy or FACS, its genomic landscape with ATAC or its transcriptome with RNA-seq. However, even when sampling different time points the cell mixture of a population will represent multiple differentiation stages, and bulk ATAC-seq therefore would superimpose the population as a whole and cannot detect cellular heterogeneity within the sample. ScMultiome can be used to

determine both the composition and the exact genomic landscape and expression patterns of the individual cell types within a population. This means that if any cells are progressing at different speeds down the erythroid lineage, which may be more likely at later time points assayed, this diversity in cell stage will be captured and resolved.

By taking samples from multiple time points throughout the erythroid differentiation and performing scMultiome, we aim to capture every cell stage and therefore map the gradual changing of open chromatin profiles. From this we aim to identify SNPs that may have their effect at an early time point of the differentiation. A cells morphology and subsequent function is set up and defined in early stem cell stages and is dependent on the function of preceding differentiation stages. If a stage crucial to development is disrupted in an early cell stage, this will have downstream effects on terminally differentiated cells. To this end, cellular phenotypes seen in later stages of differentiation may have their root cause in any number of preceding stages.

Overall aims of the chapter

Therefore, in this chapter my aims are:

1. To set up a single cell-based methodology to generate epigenetic data from across all possible stages of erythroid differentiation from donors with well sequenced genomes, to complement the analysis done using bulk methods in late erythroid cells shown in previous results chapters.
2. To use these data to perform statistical enrichment and machine learning analysis to see if earlier stages are also enriched for intersection between open chromatin sites and SNPs associated with erythroid traits.
3. To use the heterozygous variant calls from the genomes of the selected donors to see if any allelic skew in open chromatin can be detected in the earlier differentiation stages in scATAC-seq data.

To study SNP enrichment over erythropoiesis we used our CD34 differentiation system as described previously [127]. Isolated circulating CD34+ cells represent an early haematopoietic stem cell

population, and by using a standardised protocol with stem cell and growth factors, these can be differentiated to fully mature red blood cells. As this technique goes through three key stages: proliferation, differentiation and enucleation, that are dictated by the factors added at certain days, the relative staging of this process is well characterised. We therefore decided to use three time points: day 0, day 5 and day 10 to capture as much of the differentiation arc as possible. Day 0 would represent the earliest CD34+ population, isolated stem cells that have yet to expand or differentiate. Analysis of this population would also be an indication of exactly which cell populations are being selected for by the MACS columns and would give an indication of the heterogeneity of this population. Day 5 cells would represent a mid-point to the well characterised day 10 samples used in results chapter 2. Day 5 is at the tail end of the expansion phase, and so although some differentiation should be expected with stem cells progressing down the erythroid lineage, they are thought to be predominantly expanding. As alluded to in previous studies, late stages of differentiation are appearing by day 10, with late erythroid marker genes being highly expressed. The next stage would be enucleation, and so these cells would be unsuitable for any epigenetic analysis.

Because of the nature of the differentiation, whereby cells are likely differentiating at a similar but not identical rates, these time points should contain a variety of cells at different stages of differentiation. Previous bulk ATAC-seq analysis of these time points produced data that shows clear progression along differentiation but loses any heterogeneity in the cellular populations in the sample [188]. The power of single cell techniques used in this chapter will allow us to resolve each cell into its own unique profile, with cells of a similar landscape clustered together, and from these clusters a more precise epigenetic profile of that stage of differentiation is generated.

In this chapter I have generated and assessed scMultiome data of two donors across the erythroid differentiation system. From these data, it is clearly demonstrated that the genomic landscape of a differentiation system is highly dynamic, and that this is reflected in both the open chromatin and expression patterns. This therefore also highlights the importance of considering cell stage when determining the regulatory networks affected by variants.

Results

Set up and optimisation of scMultiome in erythroid cells

As this was the first multiome experiment performed in the group, as well as the first on erythroid differentiation in the institute, and due to the large downstream costs, initial optimisation was necessary. Due to the limiting cell numbers of the early day 0 populations, I performed the first set of optimisations and experiments on day 5 and day 10 expanded and differentiated samples, from donors 2 and 3 due to their local availability. PBMCs were also included during optimisations and as a control cell type. Recommended guidelines for preparation of PBMCs are supplied by 10X, and these PBMC data went on to form part of other research projects within the group.

The first step of the multiome process is to accurately count and then lyse the cells to get a suspension of single nuclei. Lysis was done using a combination of detergents: digitonin, tween and IGEPAL. To achieve optimal lysis, equating to ~100% of cells lysed without diminishing nuclei quality, lysis buffer strength could be altered as well as lysis time. I first opted for a standard 1X concentration of lysis solution, and a 6-minute lysis. Lysis rate was gauged using a cell counter with AO/PI (Acridine Orange/ Propidium Iodide Stain) that can detect intact nuclei; the viability reflecting the nuclei count and therefore lysis. For the first test, fresh PBMCs and CD34 day 5 and day 10 samples were lysed for 2, 4 and 6 minutes. During these optimisations, cell loss was prominent and so final nuclei counts for the CD34 samples were not reliable.

The RNA integrity of nuclei was then checked by generating SMART-seq libraries (performed in WIMM single cell facility). The RNA quality of the day 5 samples was good, and had a large globin peak as expected. The PMBCs were of low quality, likely due to a lack of FACS sorting to remove granulocytes, which are rich in RNases. Once optimised I was then able to perform the real experiment on two individuals and had cell counts and lysis as follows:

Sample	Cell state	Cell conc (cells/ml)	Volume (ml)	Cell count	Viability	nuclei retained
D2 day 5	Pre lysis	4.06x10^5	1	40,600	87%	
	Post lysis	5.2x10^4	0.2	10,400	0%	26%
D3 day 5	Pre lysis	8.49x10^4	1	80,000	67%	
	Post lysis	2.83x10^4	0.2	5,600	0%	7%
D2 day 10	Pre lysis	1.03x10^5	0.8	80,000	100%	
	Post lysis	3.78x10^4	1	37,800	6%	47%
D3 day 10	Pre lysis	9.56x10^4	0.8	77,000	95%	
	Post lysis	5.03x10^4	1	50,300	0%	65%

Table 4.1: Cell counts of optimisations, lysis efficiency and nuclei retention.

Lysis for all the samples was good, however, according to the cell counts the cell retention varied from 7-65% and so it was clear that this stage of isolation and preparation needed further optimisation. However, sufficient cells remained to continue with the experiment, therefore nuclei were passed onto the single cell facility for downstream processing and returned as indexed libraries for sequencing.

Sequencing of the day 5 and day 10 libraries

For multiome experiments, deep sequencing of the samples is required to generate high quality data. However, due to the discrepancies in nuclei yields prior to processing, and with the potential for material loss or PCR duplication during processing, the exact amount of sequencing required needed to be determined empirically. Therefore, to estimate the amount needed to sequence to the correct depth, an initial round of sequencing can be done to quantify the number of unique barcodes in each sample, and therefore determine the number of cells in each sample. This initial sequencing was done in house at low depth, with the ATAC and GEX libraries each being run on a separate high output NextSeq 500 run. This acted as both a library quality control and a determination of required sequencing depth prior to deep sequencing via commercial vendors. From this initial sequencing it was shown that there was considerable variation between the desired and actual number of cells captured, which is a known challenge in all single cell methodologies (Table 4.2).

Sample	nuclei count	nuclei detected	% captured
Don2 day 5	10400	15059	145%
Don 3 day 5	5660	978	17%
Don2 day 10	15000	2663	18%
Don3 day 10	15000	3877	26%

Table 4.2: Cell counts for day 5 and day 10 samples.

Of the four prepared samples, three were below the expected cell count, and one sample was over (Table 4.2). However, previous analysis within the group (Dr Ravza Gur) showed that as few as 40 cells are required to form a distinguishable cluster and 500 hundred cells when aggregated would produce an ATAC-seq track comparable with bulk data (paper in preparation). Therefore, for these samples the degree of heterogeneity of the population and the cell numbers within each will be an important factor when combining and amalgamating ATAC tracks for each cluster. This would also partially determine the ultimate quality of the data. However, as this was the first single cell experiment performed on these time points, the level of heterogeneity was unknown. Therefore, it was decided to go ahead with the deep sequencing of this pilot experiment as cluster composition could not be determined from the low depth sequencing.

The 10X guidance for required sequencing depth for each sample, was:

ATAC sequence depth = 25,000 read pairs per nucleus
GEX sequence depth = 20,000 read pairs per nucleus

Based on this guidance I therefore calculated the total amount of reads required for each sample, and subsequently the ratio of the libraries to be pooled as a sequencing library to acquire the correct sequencing depth.

Further lysis optimisation

In follow up experiments I used the control PBMC cells and unexpanded CD34+ cells derived from the MACS separation from the same donors. Considering the discrepancies with the cell numbers in the

previous attempts, and the low RNA yield from the PBMC SMART-seq quality control, I performed further optimisation steps. For the PBMCs I included a FACS sorting step to remove the granulocytes and neutrophil extracellular traps (NETs). Because of their role as immune cells, granulocytes by function have large quantities of RNases which can degrade a samples RNA prior to GEX library construction [193]. Neutrophils have a defence mechanism of NETs whereby they extrude web-like structures of protein assembled on a scaffold of decondensed chromatin, which when tagmented causes large amounts of background in the ATAC sample [194]. Using FACS sorting with forward and side scatter, both of these cell types, along with the NETs and RNase they produce, can be removed (Methods Figure 6.1). For the lysis tests, on both the CD34+ cells and PBMCs I again optimised the lysis strength and the time, using cell viability as a guide. I also used the 10X optimised low cell number protocol to try and alleviate the problem of cell loss during the washes.

For these optimisations, I started with the blood separation to obtain a fresh PBMC population. Once purified, I took $1x10^7$ cells for FACS sorting as described previously, and the rest were used for CD34+ cell MACS separation. Once separated, CD34s were either rested in media at 37°C prior to lysis testing or multiome library preparation, or used to seed the erythroid differentiation culture.

After multiple rounds of lysis condition optimisations, I arrived at optimal lysis conditions. All cell types were optimally lysed using the recommended 1X concentration of lysis buffer for 5 minutes. Nuclei integrity was good when checked by microscope (Methods Figure 6.2).

Generating PBMC and day 0 libraries

Having optimised the lysis conditions I proceeded with the real experiment. PBMCs were separated from the same donors used for the above day 5 and day 10 erythroid differentiation. A portion of PBMCs were FACS sorted for granulocyte and NET removal prior to library preparation. The remainder of the cells were used for CD34+ cell separation. Cell counts for each sample were determined with AO/PI Luna counter, and then 10,000 cells were targeted for indexing. Indexing was performed as per the 10X protocol with intermediate quality controls shown by bioanalyzer plots in Figure 4.2. The cDNA fractions show a clear but variable in strength signal across the samples, which indicated variation in cell number as seen before. The ATAC bioanalyzer profiles again show variation in strength

of signal, but all show the expected nucleosomal banding standard to tagmentation. For the samples with lower yields, extra rounds of indexing with more PCR cycles were performed during the indexing stages to increase the final concentration for sequencing.

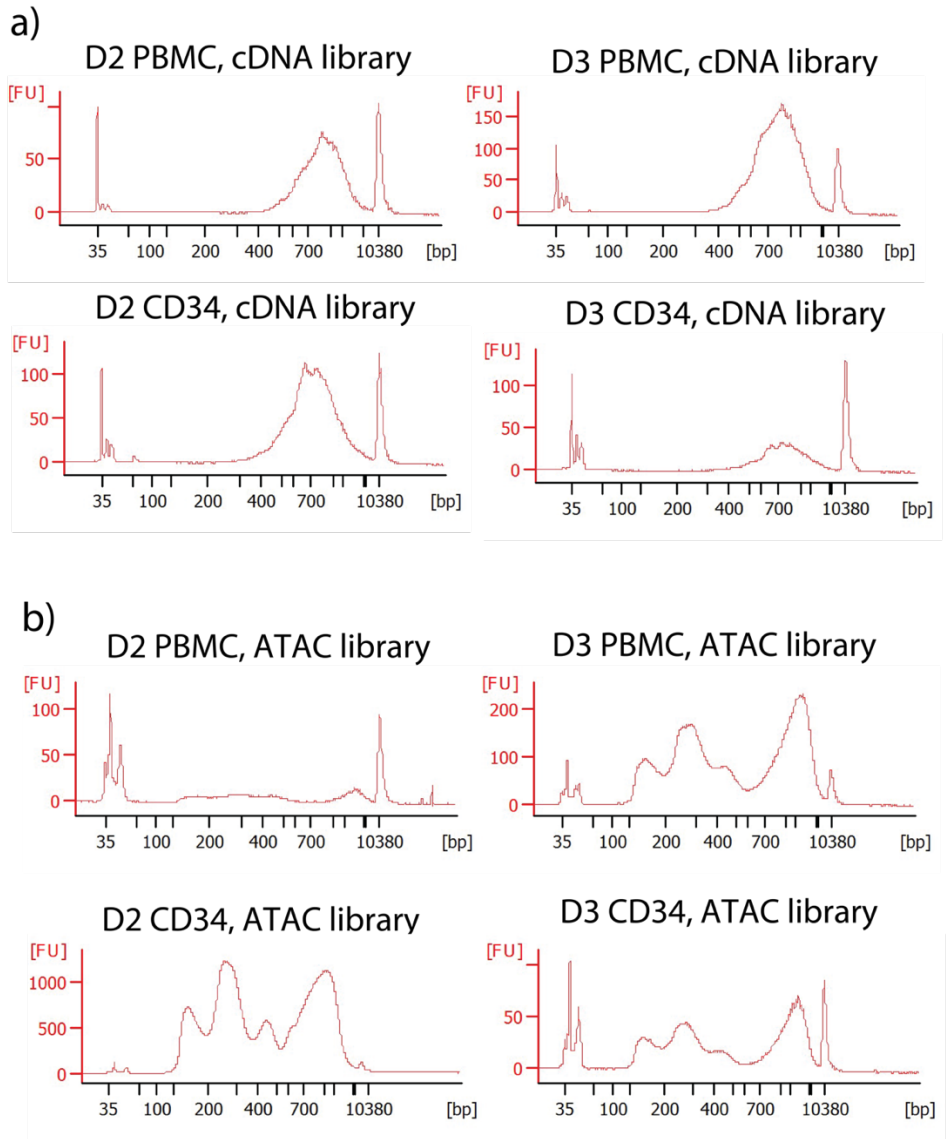


Figure 4.2: Bioanalyser intermediate analysis of multiome libraries.
a) Gene expression libraries at the cDNA stage show a large peak in all the samples. b) ATAC libraries show standard nucleosomal banding.

Sequencing the PBMC and day 0 samples

As with the day 5 and day 10 samples, shallow sequencing of the samples was done to determine true cell numbers captured (Table 4.3). Despite the further optimisations and the changes to the protocol, the cell numbers again did not reflect the number of cells targeted. The actual cell numbers ranged from 216 (Donor 3 CD34 day 0) to 19,927 (Donor 3 PBMC), and so there were still issues with the cell targeting and retention. Using the 10X cell ranger package, data quality and cell numbers can be assessed (Table 4.3) [195]. These real cell numbers per sample were then used to determine exactly how much deep sequencing was required for each sample. For the two ATAC samples with low yield (Donor 2 PBMC and donor 3 CD34 day 0), both replicates of indexed libraries were added to the pool to increase complexity. When combined with the day 5 and day 10 samples, donor 2 had good data coverage across day 0, 5 and 10 with 6,212, 15,059 and 2,663 cells in each fraction respectively.

	Metrics	D2 PBMC	D2 CD34 d0	D3 PBMC	D3 CD34 d0	Target/Ideal
	Cell count	4654	6212	19927	216	10000
	ATAC media HQ frags per cell	1228	378	635	25122	
	GEX median genes per cell	344	366	305	817	
	ATAC library nm	0.4/0.8	17	298.9	1.5/4.7	
	GEX library nm	27.5	22.8	6.9	18.1	
ATAC	Sequenced read pairs	13,474,195	15,103,255	20,324,177	71,570,138	
	Mean raw read pairs per cell	2,895.19	2,431.30	1,019.93	331,343.23	
GEX	Sequenced read pairs	21,692,696	26,502,982	21,645,301	21,402,901	
	Mean raw reads per cell	4,661.09	4,266.42	1,086.23	99,087.50	

Table 4.3: cell counts and quality from the initial sequence of PBMC and day 0 samples.

10X Loupe analysis of erythroid differentiation

By using the 10X Loupe software integrated within the CellRanger package [195], samples can be clustered either by their ATAC or GEX data, and then displayed as UMAP or tSNE plots. Figure 4.2 shows the donor 2 CD34 differentiation clustered by ATAC signal as a UMAP; the d0 sample (CD34+ undifferentiated) clusters away from the later stages of the erythroid culture system (d5 and d10 cells) where these later differentiated cells form a continuum. Overlaying gene expression patterns of erythroid genes on to this ATAC UMAP shows patterns consistent with patterns of change in erythroid

differentiation. Erythroid specific transcription factors such as *GATA1* and *KLF1*[196] can be seen to have consistent active transcription through the mid and late stages (d5 and 10). The late expression of the alpha globin genes (*HBA1* and *2*) and Mitoferrin 1 (*SLC25A37*) can be seen increasing during the differentiation and peak in the latest stages of day 10 (terminal differentiation). Conversely, the early CD34 marker can be seen in the d0 stages, and then is subsequently turned off in later stages as expected. Interestingly, there is no consistent expression pattern of CD34 across the entire CD34 population; this may suggest either contamination of other cell types within the population, or that RNA expression is down regulated but protein levels on the cell surface remain high enough for selection. The latter explanation seems the most likely as the low expressing cells cluster closely or even overlapping with expressing cells, showing they are highly similar and unlikely to be a contaminant.

However, considering our initial aim was to cover the complete arc of erythroid differentiation, this map does not seem to show this as we have two distinct populations in our UMAP. As the data derive from presumed overlapping timepoints starting from the initial CD34+ isolated cells (day 0) we expected to see one continuous connected population. As differentiation proceeds sequentially from the stem cells within the day 0 timepoint, if we had captured all the stages of differentiation within our experimental design, we would expect to see a continuous trajectory of cells spanning day 0, day 5 and day 10. As the day 5 and day 10 timepoints are a continuous arc and day 0 is isolated, this suggests that we are missing some cell stages of differentiation present in the culture between day 0 and day 5. This is surprising as day 0 to day 5 in the culture system is deemed the expansion phase of the culture, thought to be expanding the progenitor cells, with day 7 marking the addition of the differentiation media to the system. This suggests the CD34+ cells from day 0 are actually losing their stemness in the expansion media and beginning to differentiate down the erythroid lineage, despite the presence of SCF. It would therefore be interesting in future experiments to sample a further time point such as day 2.5 from the expansion phase to test this and potentially complete the differentiation arc. However, it is clear from the marker gene analysis (Figure 4.3) that we successfully sample the earliest stem cell population, which appears to be very homogenous in structure although heterogenous in the expression of CD34. Similarly, we have successfully sampled the latter stages of differentiation including terminal differentiation judging by the late erythroid markers (HBA1/2 and SLC25A37).

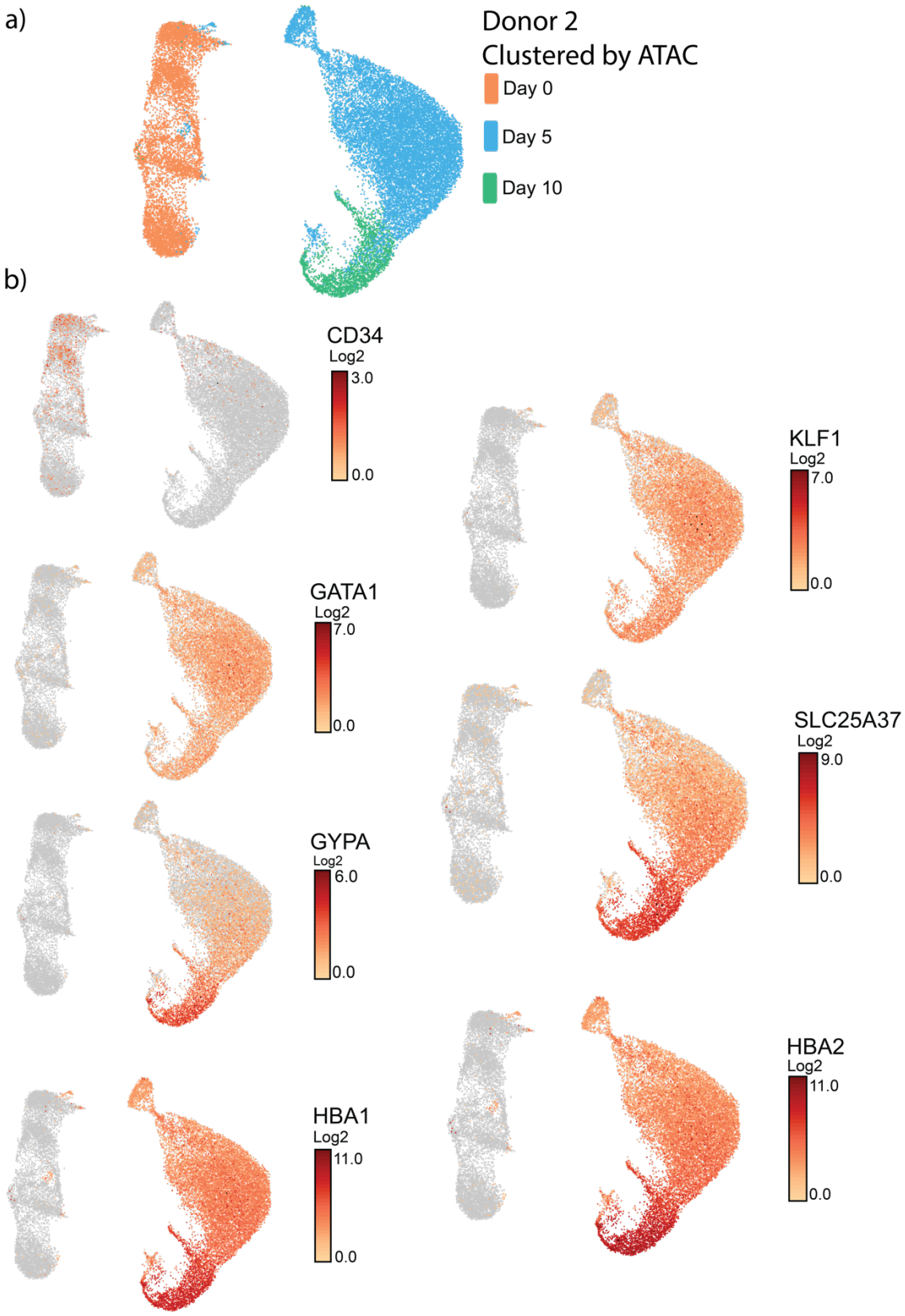

a)
Donor 2
Clustered by ATAC
Day 0
Day 5
Day 10
b)
CD34
Log2
3.0
0.0
KLF1
Log2
7.0
0.0
GATA1
Log2
7.0
0.0
SLC25A37
Log2
9.0
0.0
GYPA
Log2
6.0
0.0
HBA2
Log2
11.0
0.0
HBA1
Log2
11.0
0.0

Figure 4.3: UMAP Loupe analysis of a single donor across erythropoiesis.
UMAPs generated by Loupe software, clustered with ATAC signal. a) Overview of which sample each individual cell/ data point originated from. Day 0 (orange) is seen to form a distinct cluster whilst day 5 (blue) and day 10 (green) form a continuum. b) Gene expression levels of specified genes can be plotted for each data point. Scales vary for all genes, with the darker red indicating higher expression.

Similarly, when visualising the aggregated ATAC data from d0, d5 and d10 as ATAC-seq tracks in the UCSC browser, clear differences can be observed that show progression through erythroid differentiation reproducibly in the two donor cultures. Figure 4.4 shows the ATAC tracks from 2 individuals at the day 0, 5 and 10 time points assayed, over the genomic locus for the erythroid alpha globin marker genes (*HBA*). This locus is one of the characterised loci in terms of its regulation and the position of all the enhancers regions is known. As expected in the CD34 stem cells tracks (d0) the known globin enhancers (R1-R4) appear inactive but there is some suggestion of priming at the major R2 element. The globin enhancers are fully open by day 5, and remain so through to day 10, which is expected considering the fundamental role of the *HBA* genes in late erythropoiesis. It must be noted that due to the stringency of the aligners used by the 10X pipeline and the sequence duplication of the *HBA* genes, reads are not aligned to these duplicated regions, so they appear inactive in this analysis.

In addition to peaks opening, other peaks can be seen to be closing. A peak of unknown function at the 3' end of the *NPRL3* gene (Figure 4.4, red arrow) can be seen to be open in day 0 and day 5, and almost fully closed in day 10. This is confirmed in the merged ATAC from the 50 individuals from the later day 13 culture stage where the peak is also absent (Figure 4.3 top panel) although residual H3K4me1 modification marks remain at this region at d13.

This bulk aggregated analysis confirms the general staging and progression of the samples but ignores the cellular heterogeneity of the data. We therefore turned to a more refined cluster-based analysis to track the differentiation trajectory in our samples.

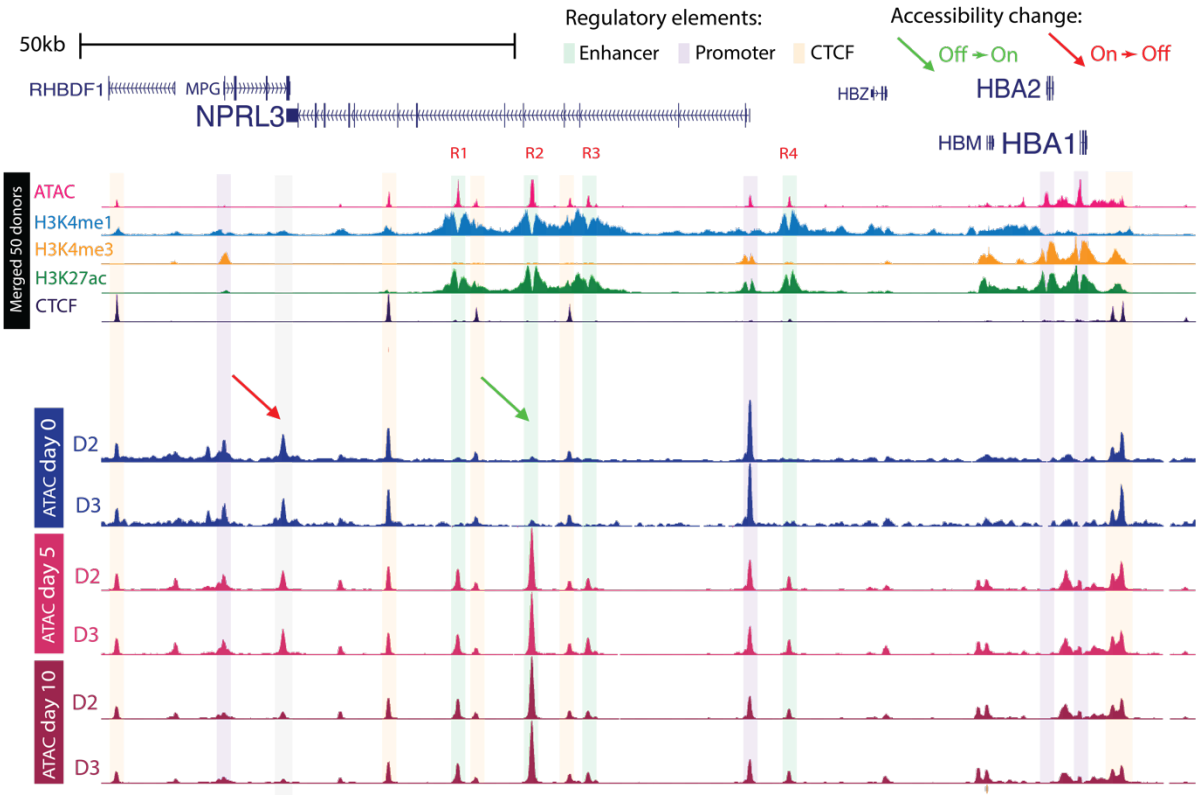


Figure 4.4: Changes in open chromatin can be seen when comparing ATAC from different erythroid stages. CD34 day 13 Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Grey highlight depicts open chromatin region of undetermined activity. Lower tracks depict ATAC for two individuals at day 0 (blue), day 5 (light pink) and day 10 (dark pink). Red and green arrows depict sites of open chromatin that change over differentiation, from off to on and from on to off respectively.

Defining an erythroid differentiation trajectory arc

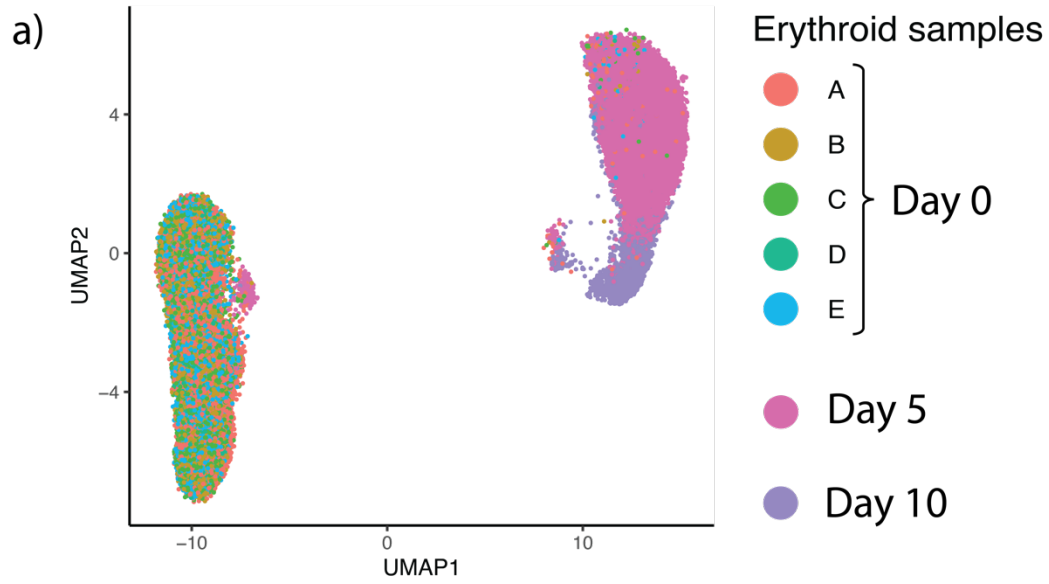
Many new computational approaches have been rapidly developed over the last few years to analyse the highly multidimensional data produced by multiome assays. However, the use of these requires extensive knowledge of R and Python coding languages and I therefore collaborated with Dr Ravza Gur in the Hughes group for this analysis. To generate this new map, CellRanger outputs were reused for further additional analysis using ArchR [197]. ArchR independently regenerates the UMAP analysis I performed with the CellRanger pipeline and these new maps reflected the original Loupe analysis with two distinct populations (Figures 4.3a and 4.5a). In both analyses day 0 clustered away from the day 5 and day 10 samples that in turn formed a continuous arc. This confirmed the initial UMAP

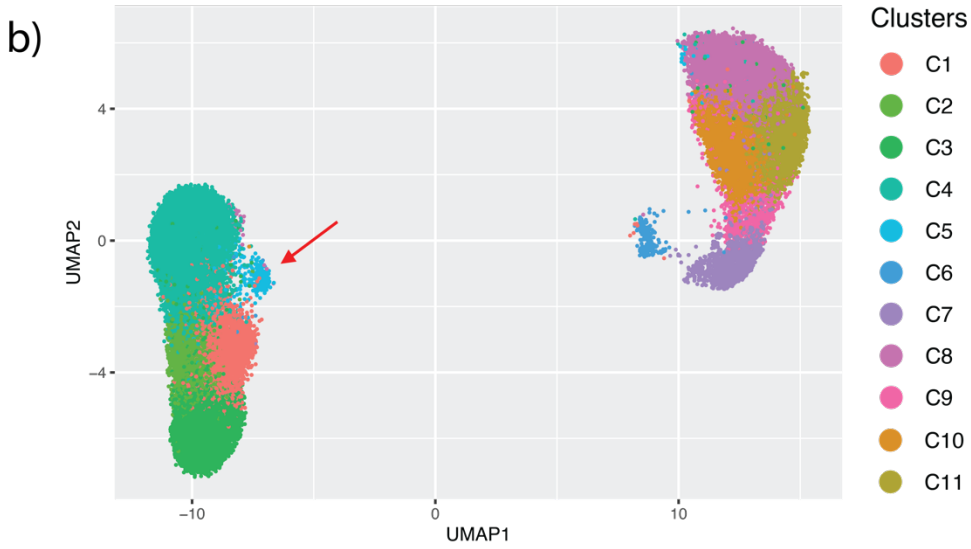
structure was a robust representation of the data and that a proportion of the population are missing from our sampling of the differentiation arc. ArchR was developed to perform advanced quality control and cluster analysis on scATAC-seq data. After removing low quality data ten clusters remained (Figure 4.3b).

However, over clustering of data is a common problem in such analysis and it is typical to manually merge clusters with minimal differences in their underlying data. Several clusters were deemed too similar to retain as independent clusters in this primary analysis. Once surveyed and combined seven clusters in total were carried forward (Table 4.4) (Figure 4.5c). Of these clusters two formed the bulk of the day 0 cluster, four covered the mid to late stages, and the seventh and final cluster reflected cells that were just prior to enucleation. The enucleating cells lay distinct from the bulk of the late erythroid cells as judged by marker gene expression and the fact that they show an extremely decreased number of peaks in the aggregated ATAC-seq data as expected in condensing chromatin prior to nuclear expulsion.

Original clusters		New clusters
C1,2,3	-->	C1
C4	-->	C2
C8	-->	C3
C10,11	-->	C4
C9	-->	C5
C7	-->	C6
C6	-->	C7

Table 4.4: Nomenclature of UMAP clusters.
Original clusters as seen in Figure 4.4b and the corresponding new names for these clusters as seen in Figure 4.4.c.


a)
Erythroid samples
A
B
C
D
E
Day 0
Day 5
Day 10
UMAP2
UMAP1
4
0
−4
−10
0
10


b)
Clusters
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
UMAP2
UMAP1
4
0
−4
−10
0
10

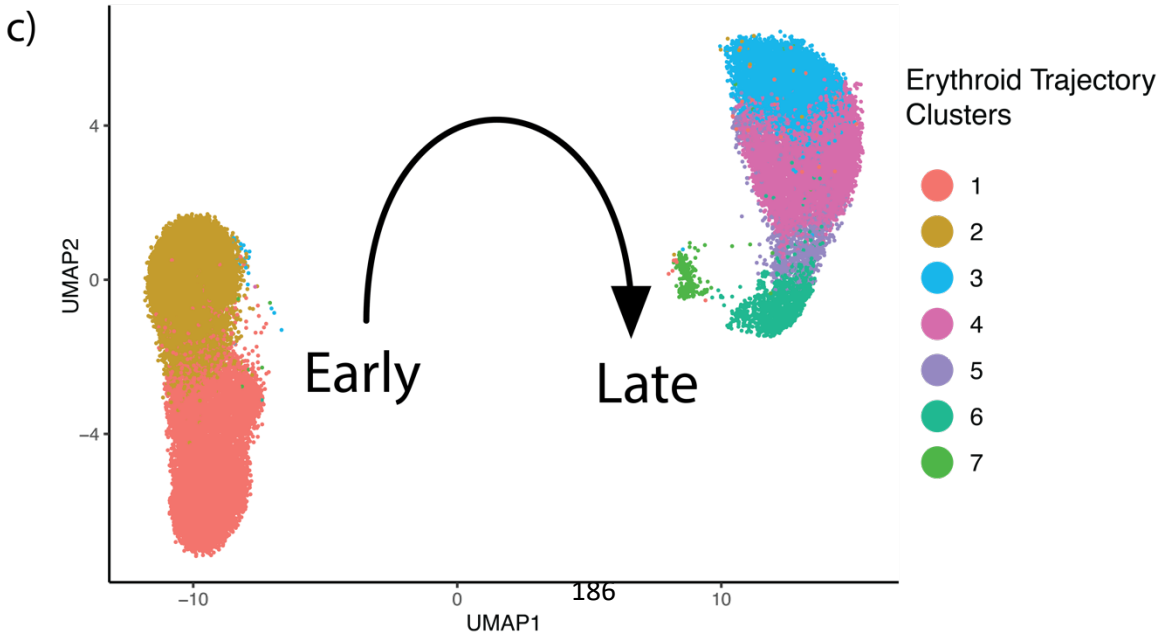

c)
Erythroid Trajectory Clusters
1
2
3
4
5
6
7
Early
Late
UMAP2
UMAP1
4
0
−4
−10
0
10

Figure 4.5: Defining an erythroid trajectory and cell stage clusters.
a) Previously analysed data was further processed using ArchR to generate an erythroid trajectory. Coloured clusters depict which sample each data point came from. Day 0 sample was resequenced and analysed and is shown in red, yellow, greens and blue. Day 5 in pink and day 10 in purple. b) Population clustering formed distinct clusters in each of the samples, with each cluster depicted by a separate colour. Red arrow highlights a population that contained low quality, mostly background data when ATAC was visualised on UCSC and was therefore later removed. c) Clusters with similar ATAC profiles combined to form seven final clusters to define the erythroid trajectory. Early populations on the right, with a clockwise trajectory into the later stages, and ending with the terminally differentiated cells showing chromatin condensation in the final cluster seven (green).

Gene activity of erythroid markers validates the expected trajectory

From these new clusters, as with the Loupe analysis in Figure 4.3, marker gene activity profiles can be overlayed to see the changes across the clusters (Figure 4.6 and 4.7). For these plots, instead of gene expression they represent gene activity, which is a measure of the openness of the chromatin at these genes. Broadly, RNA modality from multiome data is nuclear RNA and contains a more nascent profile that is heavily 3' biased. Nuclear RNA is therefore not a complete enough metric to use to form clusters, and instead chromatin accessibility profiles are used to cluster cells. In both these figures, the clusters are also shown for reference (Figure 4.6a and 4.7a). The activity patterns of these marker genes reflect the expected trajectory expression levels, supporting our working cluster structure for the data.

- Figure 4.6b shows genes that are active at the early stages of the differentiation. CD34 is the marker used to pull down the early stem progenitors, yet is only highly active in a subpopulation of the day 0 sample, and continues to be active through the later stages. The activity is fully gone by the end of the arc. Protein Tyrosine Phosphatase Receptor Type C (PTPRC/CD45) is a common leucocyte antigen, and as an essential regulator of T and B cell activation is expressed in almost all haematopoietic cells except for mature erythrocytes. PTPRC shows strong enrichment for the earliest of the day 0 population. Finally, Nkx2-5 also shows early enrichment, as it regulates the hematopoietic erythroid fate of the early mesoderm precursors [198].

- Figure 4.6c shows activity of the known erythroid transcription factors: GATA1, KLF1, NFE2, TAL1 [196, 199, 200], with all showing strong enrichment in the later stages compared to the early stages. In particular, the activity of KLF1 and NFE2 are shown to increase during the later stages of the differentiation.

- Figure 4.7b shows late stage erythroid genes. The alpha and beta globin genes are highly expressed at the end of the differentiation, with *HBA1* showing increased activity earlier than *HBA2*. Human erythrocyte membrane protein 4.2 (EPB42) [201] is an erythroid specific gene, and has high enrichment in late stage erythropoiesis.

Following on from this clustering analysis, we are now able to aggregate the data to form ATAC-seq tracks with greater cellular granularity, based on clusters rather than just day of culture. When now revisiting known regions such as the alpha globin cluster, the data broadly reflects the patterns shown in the previous sample-based analysis (Figure 4.3), with several peaks opening and closing across differentiation (Figure 4.8).

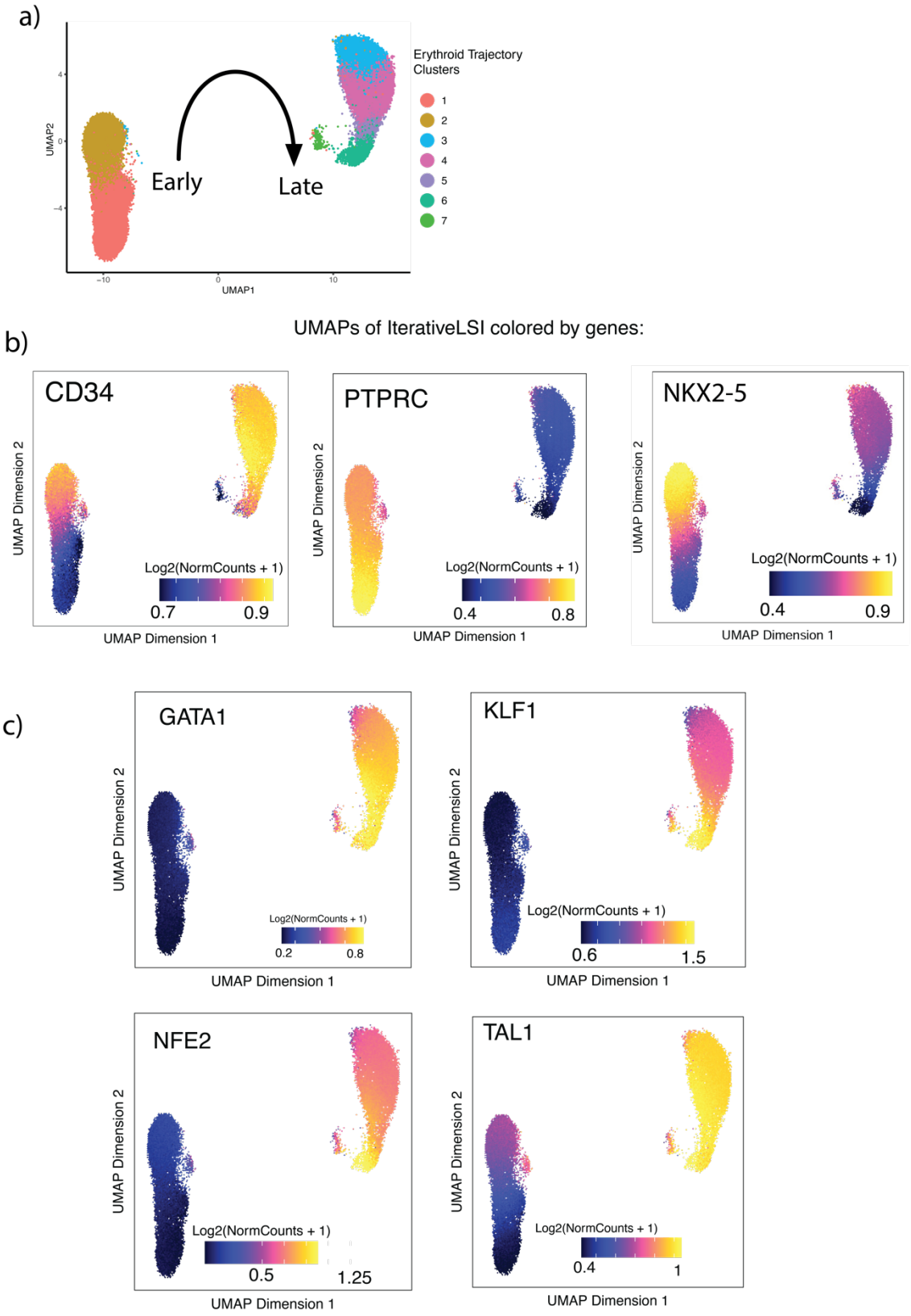

a)
Erythroid Trajectory Clusters
1
2
3
4
5
6
7
Early
Late
UMAP1
UMAP2
b)
UMAPs of IterativeLSI colored by genes:
CD34
PTPRC
NKX2-5
Log2(NormCounts + 1)
0.7
0.9
0.4
0.8
0.4
0.9
UMAP Dimension 1
UMAP Dimension 2
c)
GATA1
KLF1
NFE2
TAL1
Log2(NormCounts + 1)
0.2
0.8
0.6
1.5
0.5
1.25
0.4
1
UMAP Dimension 1
UMAP Dimension 2

Figure 4.6: Gene activity profiles show expected patterns across erythroid differentiation arc.
a) Overview of clusters and staging as seen in Figure 4.5. b) Expression of genes associated with and enriched in early populations. Low enrichment depicted as dark blue, progressing to high enrichment in yellow. c) Expression of erythroid transcription factors show enrichment throughout the later stages of differentiation. Low enrichment depicted as dark blue, progressing to high enrichment in yellow.

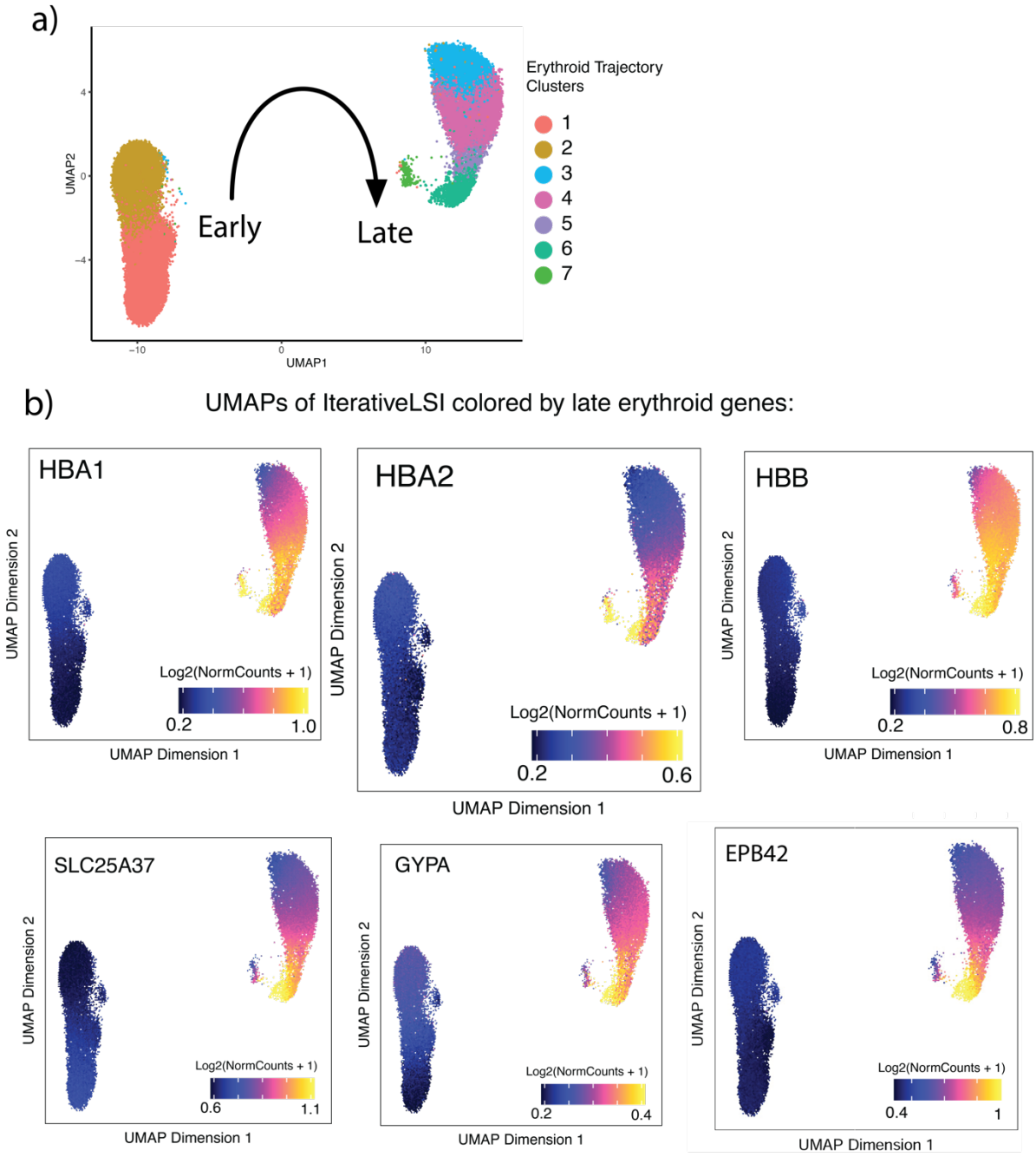


Figure 4.7: Gene activity profiles of late stage proteins show enrichment in later erythroid differentiation stages.

a) Overview of clusters and staging as seen in figure 4.5. b) Expression of genes associated with and enriched in late erythroid populations. Low enrichment depicted as dark blue, progressing to high enrichment in yellow.

Following on from this clustering analysis, we are now able to aggregate the data to form ATAC-seq tracks with greater cellular granularity, based on clusters rather than just day of culture. When now revisiting known regions such as the *HBA* cluster, the data broadly reflects the patterns shown in the previous sample-based analysis (Figure 4.3), with several peaks opening and closing across differentiation (Figure 4.8).

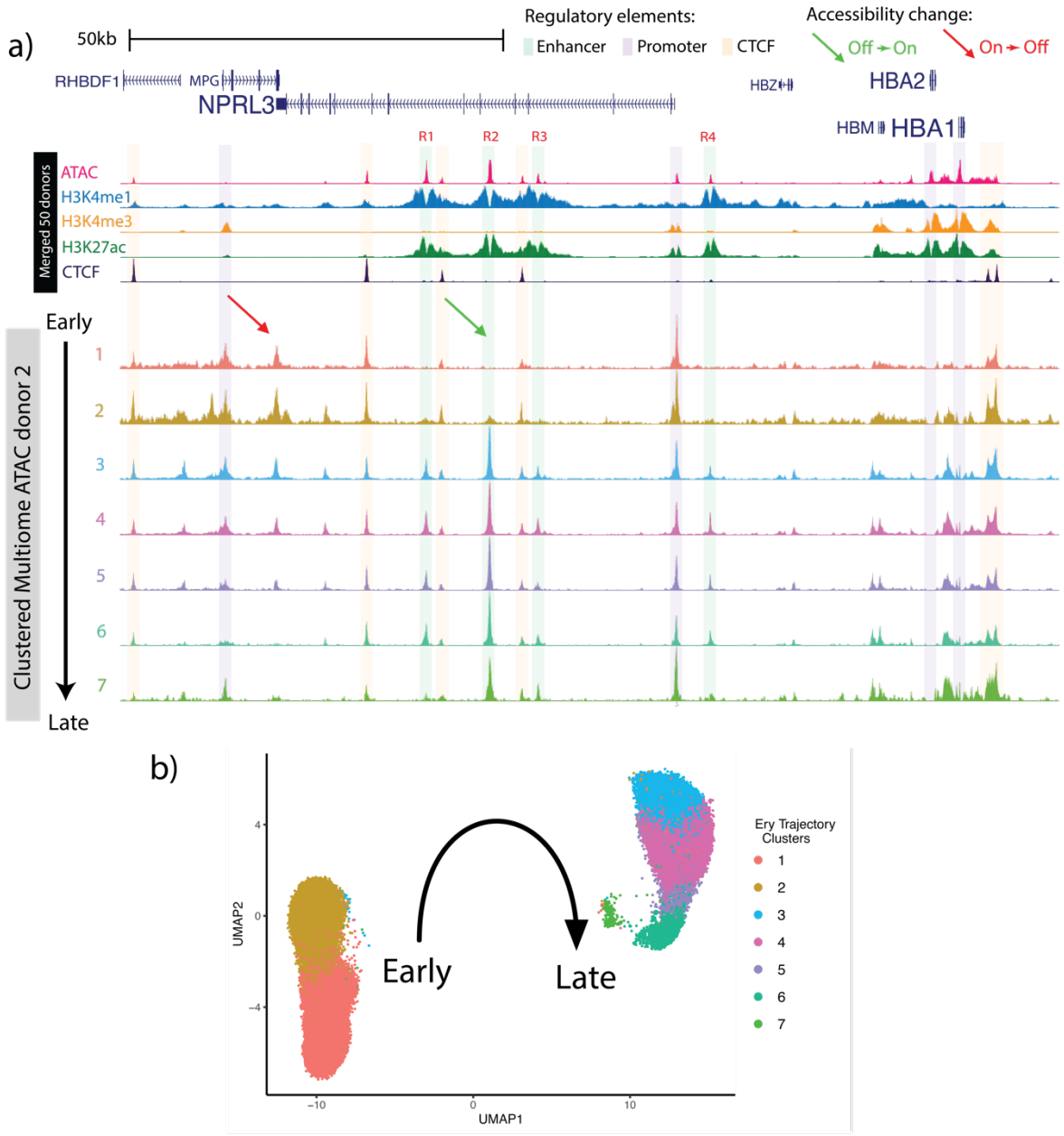


Figure 4.8: Using the trajectory clusters to visualise cell stage specific ATAC profiles at *HBA*.
a) *HBA* locus showing the late stage combined data and multiome ATAC data from across the erythroid differentiation. CD34 day 13 Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks depict ATAC from a single individual, data split by the clusters depicted in figure 4.5c, with corresponding colours, reflecting erythroid progression. Red and green arrows depict sites of open chromatin that change over differentiation, from off to on and from on to off respectively. b) Clusters from figure 4.5c for track colour and cluster number reference. N.B. due to the aligners used by CellRanger, duplication of the *HBA* gene promoters means no reads are aligned here.

Assessing genomic landscape changes at regions of variation

With these data, previous regions of interest can be assessed to see how the open chromatin landscape changes across erythroid differentiation. In particular, we can look at the variant regions identified in previous chapters to see at which stages the open chromatin sites exist. This is important as it will reveal whether a variant intersects and potentially influences a regulatory element at an early stage of differentiation, and not just the later terminal stages. When looking at these regions, changes in accessibility of open chromatin sites (such as those shown at the *HBA* locus in Figure 4.4 and Figure 4.8) can be seen over regions of variation. An obvious limitation is that GoF variants detected in the 50 individuals but absent in the donors on which multiome has been performed will be absent. Due to the complexity of the analysis and time limits of this thesis at present only donor 2 has been fully analysed. Therefore, only cluster split data from this donor will be presented, and data from donor 3 will be represented as aggregated by sample day for reference.

Data in the previous chapters was generated from late erythroblasts harvested at day 10 of the CD34 culture system. By looking at the following regions we can determine if the variants present were specific to certain stages or consistent in their activity through differentiation. From this we can discern which cell stages these genetics are likely to be influencing.

The landscape at *SCARB2* appears to show a simple landscape, with both donors behaving the same, and the enhancers gradually turning on through differentiation (Figure 4.9). In previous chapters this enhancer was shown to vary amongst the individuals and to be linked to *SCARB2* expression. Because of this variation the peak appears stronger in one of the individuals. When comparing the genetics of the causative SNP rs11097297 in these individuals, donor 2 is homozygous T/T and donor 3 is homozygous C/C. From the MCC analysis in the previous chapter, these genotypes were shown to associate with a stronger or weaker interaction score respectively, and this also parallels the strength of the ATAC signals. However, previous bulk ATAC from day 13 shows a much smaller peak in the inactive individuals. Figure 4.10 shows that in the donor 3 day 13 sample, the peak is almost entirely gone. This could be because of two reasons: either the peak is on in the day 5 and day 10 samples, albeit more weakly than in donor 2, and then is off by day 13, or the single cell dataset is technically more able to see weaker peaks. In previous analysis by the group, single cell data has been shown to

be highly sensitive (unpublished, Ravza Gur). Regardless, this comparison of data sets demonstrates the high quality of data generated when combining the ATAC data from single cells. SCARB2 is linked with lysosomes, so the reason as to why it would be turning on in late erythropoiesis is unclear [202]. The activity at the promoter of SCARB2 is consistent throughout the differentiation, but when considering the allelic RNA output (Chapter 3 Figure 3.11) the enhancer is likely increasing the expression of the gene.

The *PPP1R3B* locus also shows a fairly simple increase of activity at the enhancers. As shown in previous chapters, donor 2 has both enhancers, with both here showing a gradual increase in activity. As donor 3 doesn't have the GoF peak 2, it is clearly inactive at all three time points (Figure 3.11a). Both enhancers can be seen to be opening in cluster 2, which would be missed by just using the bulk ATAC. They also seem to be closing when the DNA is condensing (Figure 3.11b).

The variable CTCF peak at *MYH15* shows that even though it varies in presence through the donors (Chapter 2 Figure 2.21) it is open throughout all of the differentiation stages in these donors. In contrast to this consistency, adjacent peaks are opening as the differentiation progresses. This suggests that the CTCF site and the structural boundary domain is set up early in differentiation and remains constant.

When considering more complex regions, the variable co-appearing peaks found within *CLNK* show dynamic opening and closing through differentiation (Figure 2.22). In this region both donors are behaving differently, and peaks are opening and closing across differentiation. This demonstrates the dynamic nature of chromatin landscape changing, combined with the complexity and variability across different donors.

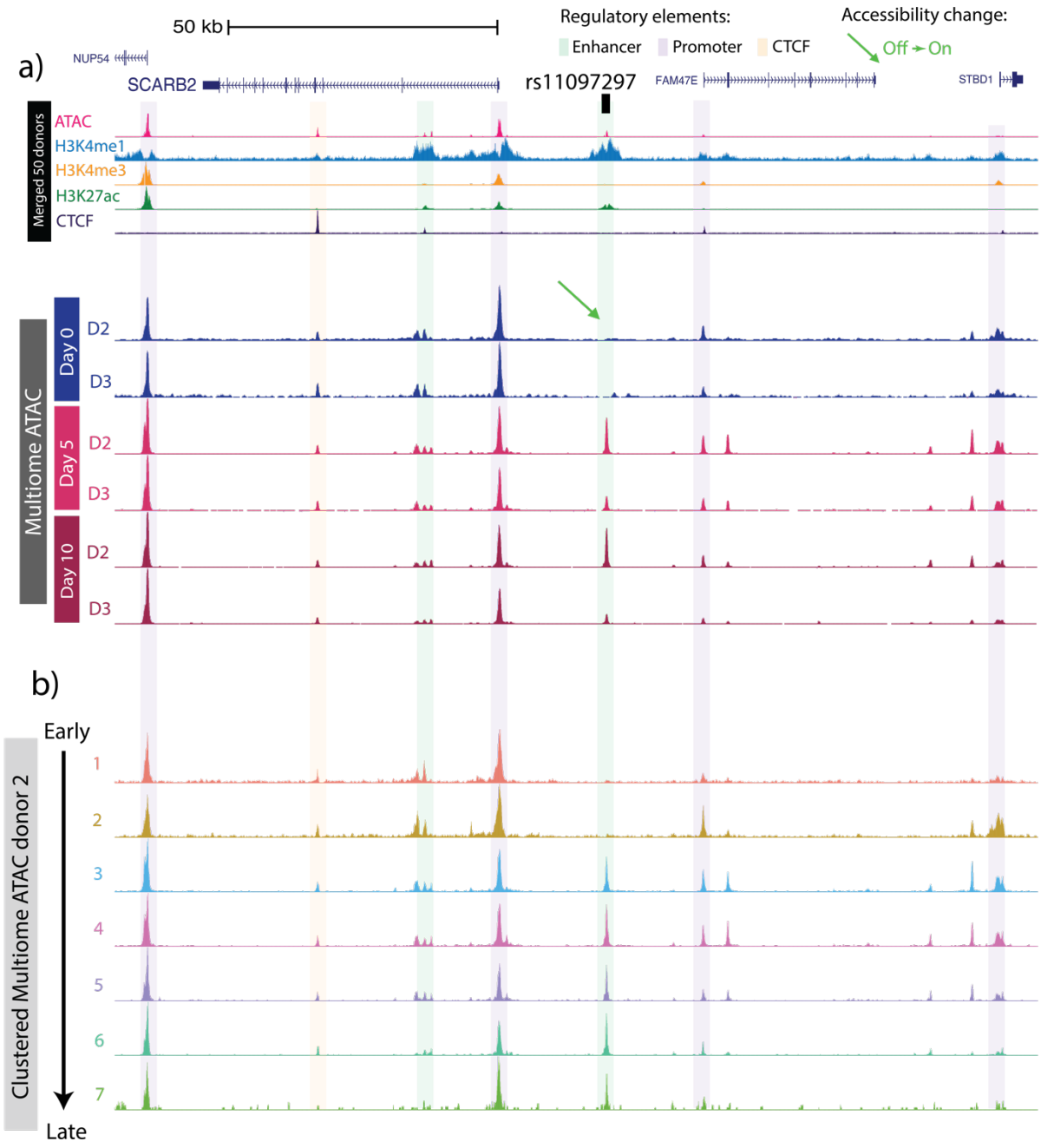


Figure 4.9: ATAC accessibility of a *SCARB2* enhancer increases through erythroid differentiation.
a) CD34 day 13 Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks depict ATAC for two individuals at day 0 (blue), day 5 (light pink) and day 10 (dark pink). Green arrow depicts an enhancer that changes over differentiation, from off to on. b) Same region with tracks depicting ATAC from a single individual, data split by the clusters depicted in figure 4.5c, with corresponding colours, reflecting erythroid progression.

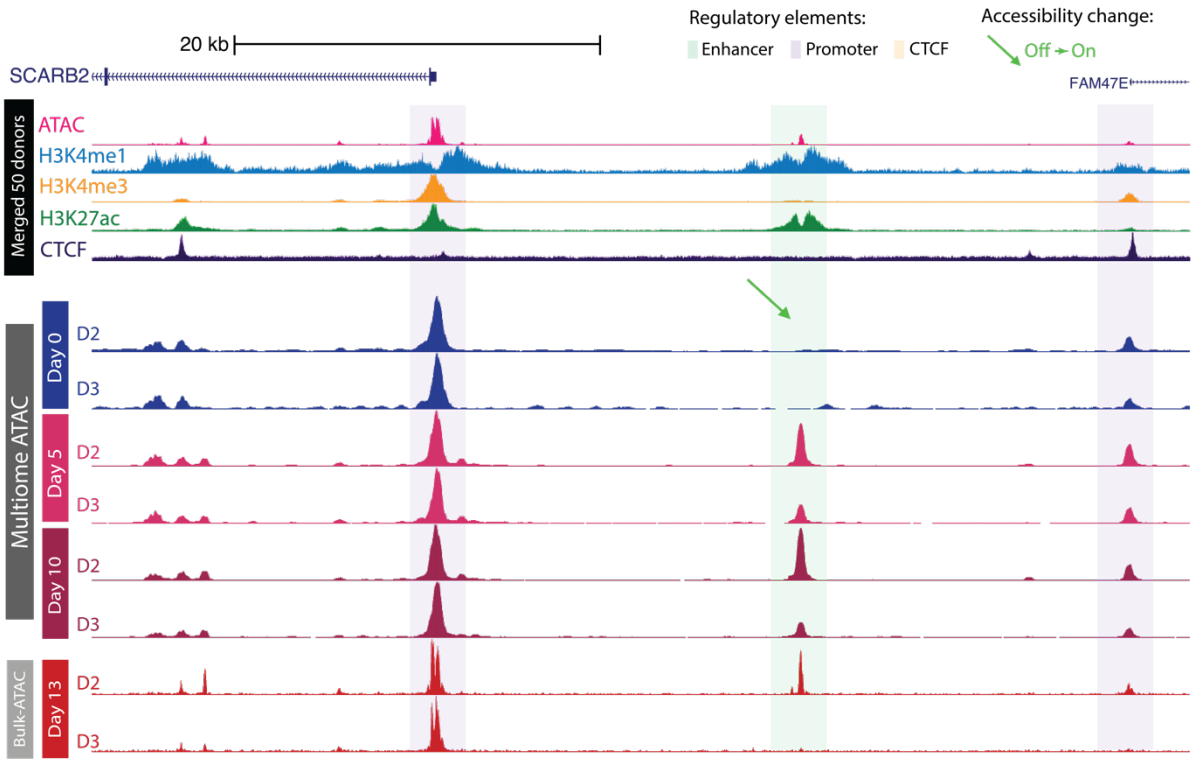


Figure 4.10: Aggregated scATAC and bulk ATAC tracks are comparable.
Tracks are the same as those in Figure 4.9a, zoomed in on the *SCARB2* promoter and enhancer. Additional bulk ATAC data from day 13 show that the data between the two assays is comparable. The difference in signal strength in D3 from the bulk data could either be due to a change in activity at a later stage, or because of increased sensitivity in the scATAC.

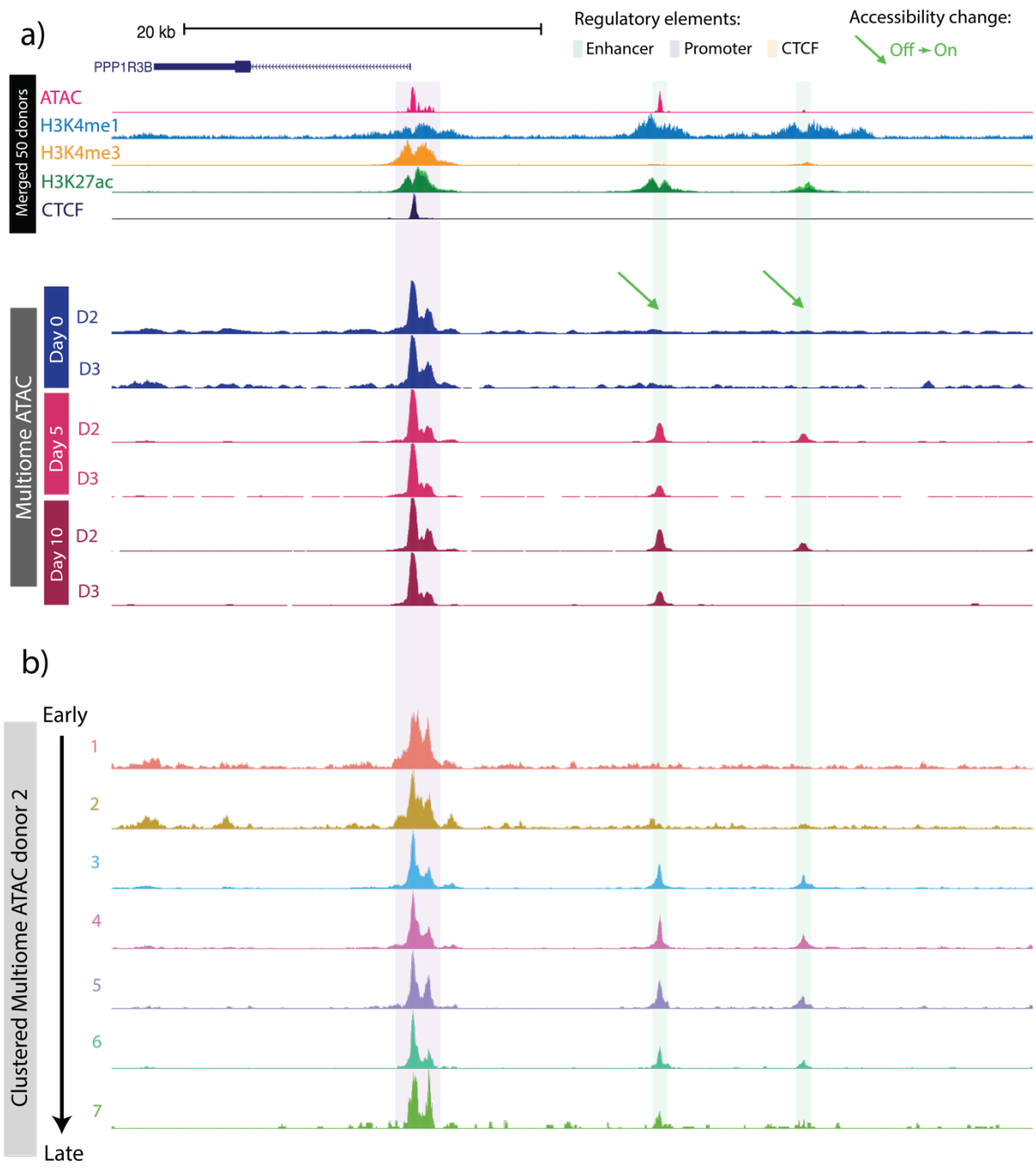


Figure 4.11: ATAC accessibility of the PPP1R3B enhancers increase through erythroid differentiation and differ between donors.

a) CD34 day 13 Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks depict ATAC for two individuals at day 0 (blue), day 5 (light pink) and day 10 (dark pink). Green arrows depict enhancers that changes over differentiation, from off to on. Lefthand enhancer present in both individuals, righthand enhancer only present in one b) Same region with tracks depicting ATAC from a single individual, data split by the clusters depicted in figure 4.5c, with corresponding colours, reflecting erythroid progression. Both enhancers are visible and can be seen to open.

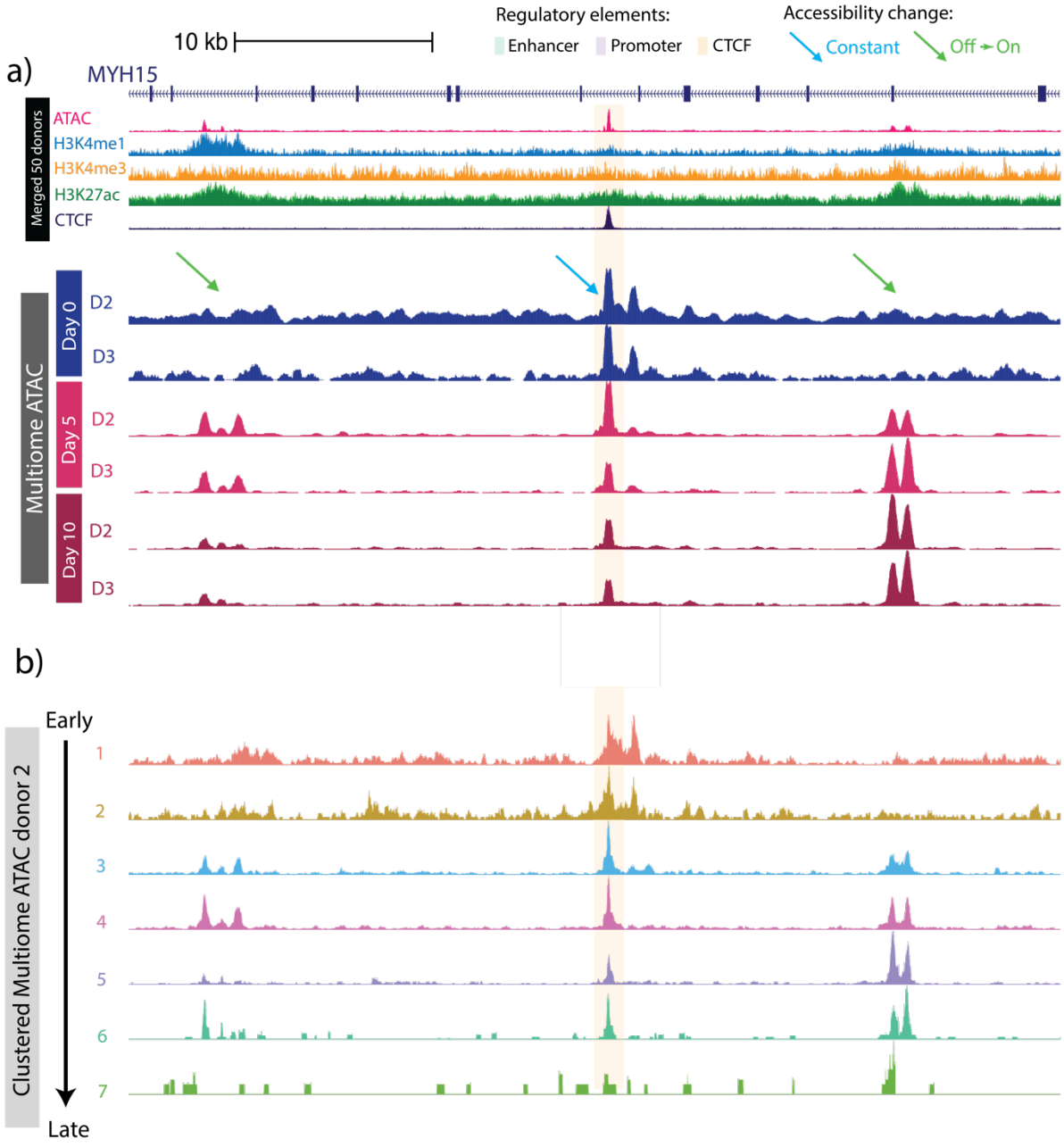


Figure 4.12: ATAC accessibility of a CTCF site remain constantly open through erythroid differentiation.
a) CD34 day 13 Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Lower tracks depict ATAC for two individuals at day 0 (blue), day 5 (light pink) and day 10 (dark pink). Green arrows depict sites that changes over differentiation, from off to on. Blue arrow depicts CTCF site that remains open through differentiation. b) Same region with tracks depicting ATAC from a single individual, data split by the clusters depicted in figure 4.5c, with corresponding colours, reflecting erythroid progression.

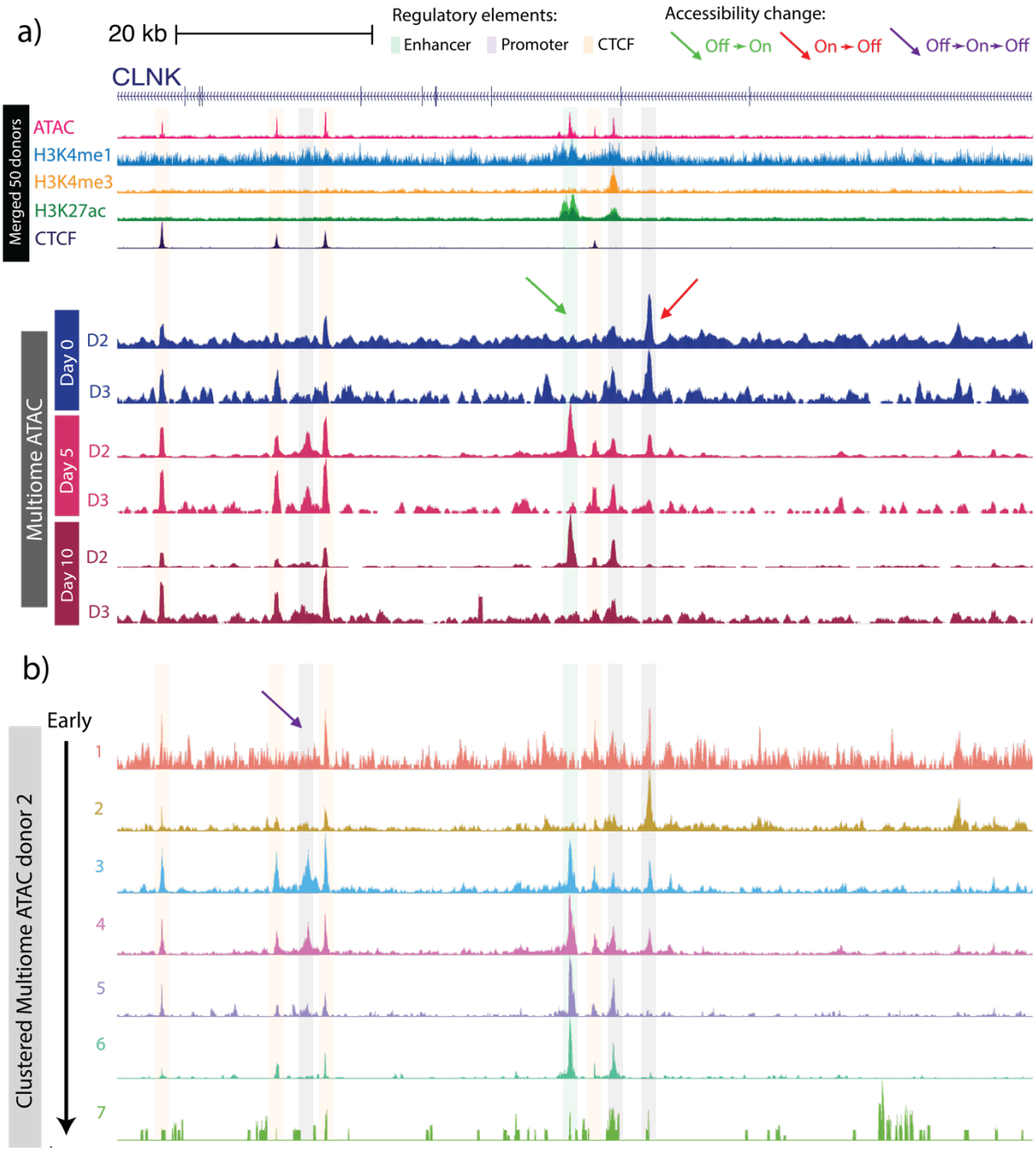


Figure 4.13: ATAC accessibility changes through differentiation are complex over the variable CLNK region. a) CD34 day 13 Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Green, purple and yellow highlights depict enhancers, promoters and CTCF sites respectively. Grey highlight depicts open chromatin region of undetermined activity. Lower tracks depict ATAC for two individuals at day 0 (blue), day 5 (light pink) and day 10 (dark pink). Red and green arrows depict sites of open chromatin that change over differentiation, from off to on and from on to off respectively. b) Same region with tracks depicting ATAC from a single individual, data split by the clusters depicted in figure 4.5c, with corresponding colours, reflecting erythroid progression. Purple arrow depicts complex change that remains goes from closed to open to closed.

GWAS SNP enrichment across erythroid differentiation

As described in the introduction, one of the critical steps in interpreting regulatory GWAS variants is the identification of the relevant cell type or types. It is a reasonable assumption that a large majority of variant haplotypes affect nuclear function by altering the activity of regulatory elements and so should lie in open chromatin regions. In this approach the intersection of the imputed GWAS haplotypes with open chromatin maps is determined over a large population of cell types. This can then find any statistical enrichment for SNPs within a defined cell types open chromatin regions compared to the per chance distribution found in the majority of cells (Figure 4.14). A databank of ~300 cell types derived from both bulk ATAC-seq and scATAC-seq has accumulated in the Hughes group and a statistical analytical platform called "Cell Type Prioritisation Platform" has been built around it [87, 88]. The workflow for this system is visualised in Figure 4.14.

A demonstration of its output is shown below using the combined analysis of scATAC-seq analysis of late erythroid cells (d10 differentiation culture) and PBMCs (Figure 4.15). In this analysis, these distinct cell populations are combined *in silico* and then clustered to form a UMAP. The identified clusters from these data have then been aggregated to form open chromatin datasets, which are then peak called and integrated into the cell type database of the Cell Type Prioritisation Platform as individual cell types. For ease of visualisation the enrichment scores from the platform can be projected back on to the single cell UMAP so enrichment per cluster can be easily seen.

This illustrative example from the platform shows that when erythroid genetics are used as the input the late erythroid cluster is shown to be highly enriched relative to the PBMCs (and all non-erythroid cells in the databank, data not shown). Conversely, when known immune related genetics such as Type 1 Diabetes or Multiple Sclerosis are used the T-cell related clusters and other immune cells are enriched relative to the erythroid cells (and all non-immune cells in the databank, data not shown), though with differences in cluster intensity depending on the genetics used.

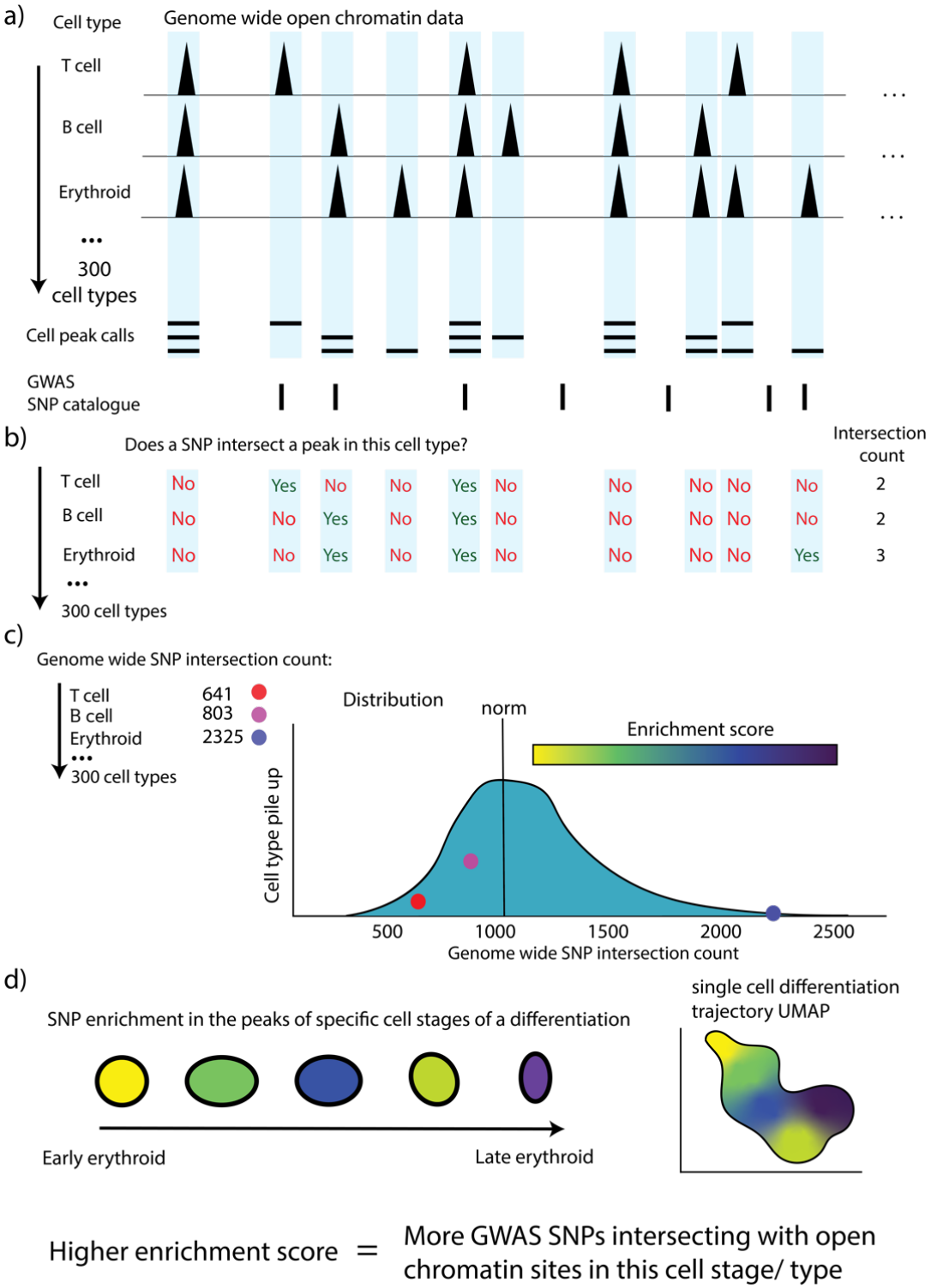


Figure 4.14: Using the Cell Type Prioritisation Platform to find GWAS intersection enrichment in cell populations.

a) Open chromatin profiles for ~300 cell types are peak called and intersection data with a GWAS of interest is determined. b) Intersection counts for each cell type are generated, the higher the number, the more of the

GWAS SNPs are found in open chromatin of that cell type. c) These intersection scores are then plotted and distribution represents the degree of enrichment across all the cell types in the data set. A cell type of interest can then be compared to the other cell types analysed. If this cell type aligns with the central peak or norm of the distribution, this represents the base line enrichment or the per chance enrichment. Cells with a higher intersection score will fall to the right of this peak, and will have therefore be enriched for intersection. The degree of enrichment can be displayed as a heat map. d) By assessing multiple cells within a differentiation system, each cell type will have a unique enrichment score that will reflect the intersection of SNPs. By overlaying these enrichment scores onto UMAPs of clustered single cell data, the differences between clusters can be seen.

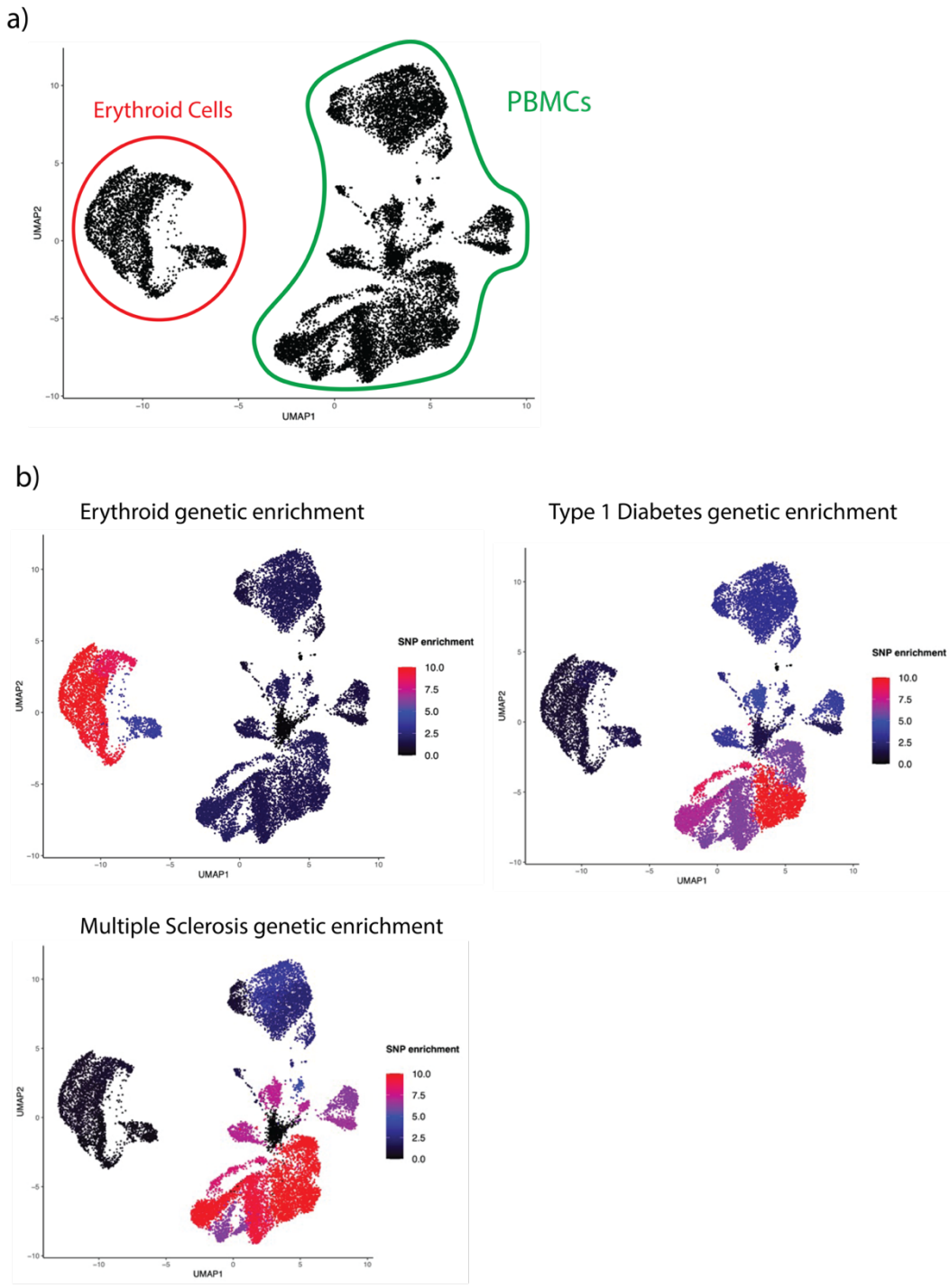


Figure 4.15: Combined erythroid and PBMC UMAPs with overlaid GWAS enrichments.
a) Combined erythroid and PBMC scATAC data ford distinct clusters. b) Cell Type Prioritisation Platform shows the enrichment of certain GWAS. Erythroid genetics are enrichment in the erythroid cells, whilst PBMCs show background levels. Type 1 diabetes and multiple sclerosis genetics are both enriched in the PMBCs but show different enrichment in difference cell type clusters. Both show background levels of enrichment in the erythroid populations.

Therefore, to address the question of whether and to what degree the earlier stages of differentiation may be targets of erythroid regulatory genetics, we integrated the peak calls from our erythroid differentiation arc. The seven final annotated open chromatin clusters from our scATAC-seq data from Figure 4.16a were analysed with three published GWAS datasets for erythroid output parameters (Van der Harst, Astle and Vuckovic; Figure 4.16b, c and d) [70-72]. The SNPs in these studies were all associated with erythroid traits and so enrichment is to be expected within erythroid differentiation (Methods Table 6.12), however the interest here lies in the relative enrichment across the component clusters.

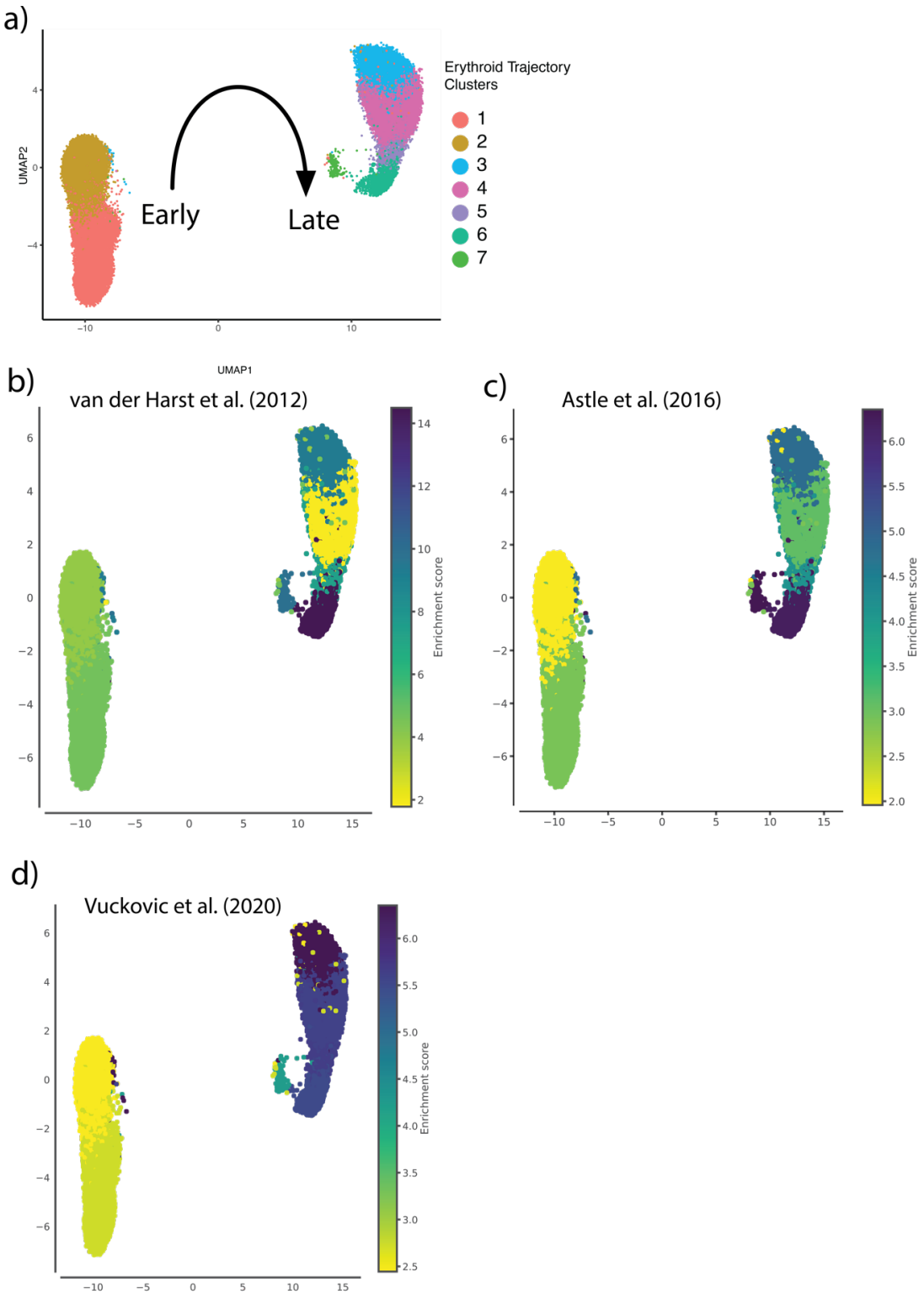


Figure 4.16: Erythroid GWAS SNPs are enriched at all stages across erythropoiesis.
a) Differentiation arc clusters for reference. b-d) Erythroid GWAS catalogues all show enrichment at all stages of erythropoiesis, albeit with different patterns. All show strong enrichment at cluster 6, and enrichment above background across all stages.

While differences in enrichment against cluster structure shows some variation across the three GWAS studies, the main patterns remain relatively stable. These discrepancies are likely due to the different studies sampling different subsets of the genetics and inherent noise. Firstly, while all clusters have some level of enrichment compared to the background, the latter stages of differentiation, and in particular cluster 6, are generally much more enriched (Figure 4.16). Next, both clusters 5 and 3 are also highly enriched, where the intervening cluster 4 appears less enriched (especially in Van der Harst and Astle). The CD34+ population (clusters 1 and 2) in general are more weakly enriched than the latter stages but of the two clusters 1 is relativity more enriched.

As this enrichment is relative to a background signal, all the stages are coloured and therefore have SNPs showing preferential intersection with open chromatin sites in these cells compared to the other cell types in the collection. Though this does not prove causality, it is a strong indication that at least some of these genetics are causal in early differentiation and influencing later phenotypes. However, the main strength in this analysis is to provide guidance as to which cell type or types should be prioritised for functional analysis.

Although we were only able to perform this analysis subsequent to choosing the day 10 differentiation time point and generating all of the epigenetic and transcriptomic data from the 50 individuals in our biobank, it is extremely heartening to find this time point is consistently the most enriched for GWAS intersection in these analyses. Additionally, when assessing donor genotypes it was found that the vast majority of the imputed SNPs found from these studies are present in at least one of the 50 donors (imputed SNPs from GWAS catalogues present in the 50 donors: van der Harst, 1,079/1,095, 1.5%; Astle, 49,834/51,678, 3.6%; Vuckovic, 130,680/135,469, 3.5%). It does however highlight the fact that at least some proportion of these genetics may be acting in these more inaccessible early populations, which poses two interlinked questions. Firstly, is it possible to use allelic skew in scATAC-seq, as we have performed with bulk ATAC-seq, to gain some understanding of whether a variant in an open chromatin peak is likely causal or neutral? Secondly, we have used skew across a population of 50 individuals from different backgrounds to infer causality, but performing scATAC-seq on 50 individuals would be prohibitively expensive; is there a way to prioritise variants that are likely causal at earlier stages of differentiation in order to prioritise individuals who share these genetics? By prioritising individuals with these interesting variants for further scATAC, this will be a more efficient

use of resources, funds and time. In this way heterozygous individuals for these early acting variants could be recalled by genotype specifically for scATAC-seq analysis of their erythroid differentiation arc.

Prioritisation of variants as causal during different differentiation stages using deep neural networks

Deep Neural Networks, as described in the introduction, are increasingly being used for genomic prediction tasks. Due to their relative novelty they still need extensive validation and testing by the field, but they have been shown to successfully predict causality for regulatory variants [184]. Such approaches could therefore be applied to prioritise variants that affect earlier stages of differentiation followed by subsequent recall by genotype and functional testing.

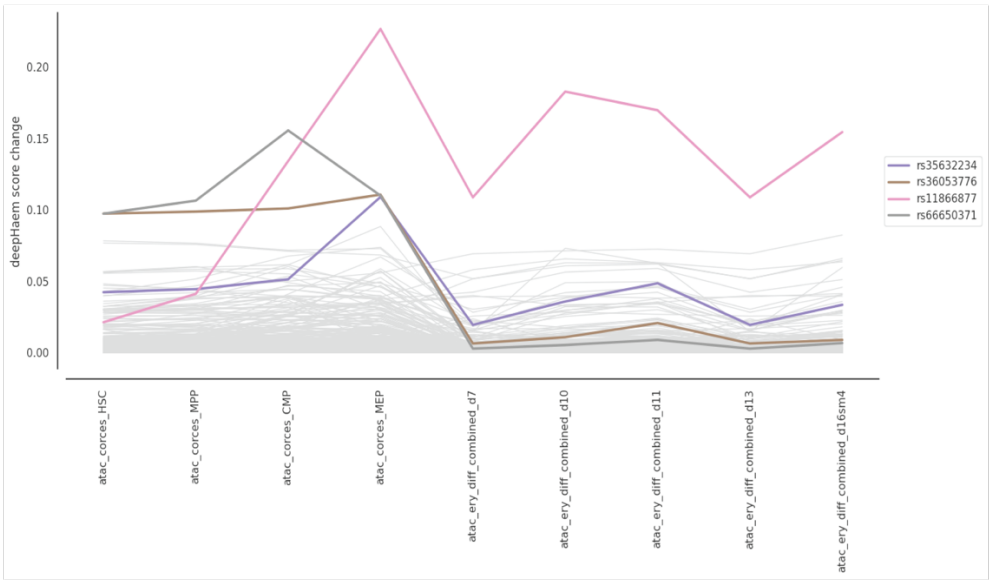
To test the potential for this approach to highlight specific early acting variants we have analysed two comprehensive erythroid datasets (Van der Harst and Astle, Figure 4.16) using the DeepHaem Neural Network from the Hughes group [31]. This approach utilises a deep convolutional neural network to predict chromatin features from DNA sequence. Broadly, it takes a genomic sequence and predicts whether there will be activity at that region and whether it will change in the presence of a variant or variants. Using an input of sequences associated with open chromatin peak calls from a collection of cell types of interest, it can learn which sequence patterns are more likely to be associated with activity in a given cell type. This process is mostly learning the presence and combinatorial patterns of cell type specific transcription factor binding sites. When used in the context of a variant, this software can predict the openness of the chromatin across all of the cell types in its training set. It can also assess how much the predicted openness changes when the variant is included in the input sequence and so assigns a damage score to the variant. Importantly, the scores can be both positive or negative, reflecting a loss or gain of function respectively.

By taking a GWAS catalogue of interest and predicting how SNPs in DNA sequences will be affecting different cell stages, stage specific patterns can emerge. For example, if an open chromatin site is open throughout the entirety of differentiation, the damage score will be consistent across all the stages. If however an open chromatin site is only active in a certain cell stage, the altered score will

be reflected in this stage and not the others. Importantly, these networks learn from the overall patterns of TF binding in the cell type and so do not require genotype specific training data. The DeepHaem network is trained on a large collection of bulk ATAC-seq and now also scATAC-seq datasets. It is combined with publicly available purified stages of erythroid differentiation to supplement the missing stages in our differentiation arc (Multipotent Progenitors (MPP), Common Myeloid Progenitors (CMP) and Megakaryocytic-erythroid Progenitors (MEP)) as well as data from bulk analysis of erythroid differentiation (day 7, 10, 11, 13 and 16). Once our scATAC-seq erythroid differentiation arc is completed the neural network will also be retrained with these data.

Specifically, in this analysis we were looking for variants within these studies that were predicted to alter chromatin accessibility predominantly at early stages of erythroid differentiation (Figure 4.17). Analysis of the van der Harst *et al*. genetics highlighted four such variants: rs11866877, which is precited to damage activity from the MEP stage and subsequent stages of differentiation; rs36053776 and rs66650371 which are predicted to damage activity from the HSC stem cell niche through to the MEP stage; and rs35632234 which is predicted to mostly damage activity at the MEP stage. Astle *et al*. highlighted a further four variants: rs2526354, which like rs36053776 and rs66650371, is predicted to damage activity from the HSC stem cell niche through to the MEP stage; rs11553699 which like rs11866877, is predicted to damage activity from the MEP stage and subsequent stages of differentiation; rs61450169 which is predicted to damage activity predominantly at the CMP and MEP stages; and rs567910619 which interestingly is predicted to have an effect across all stages of differentiation. Unfortunately, our analysed donor is not informative for these variants and so they will be subject to future recall by genotype and single cell epigenetic analysis.

van der Harst et al. (2012)



Astle et al. (2016)

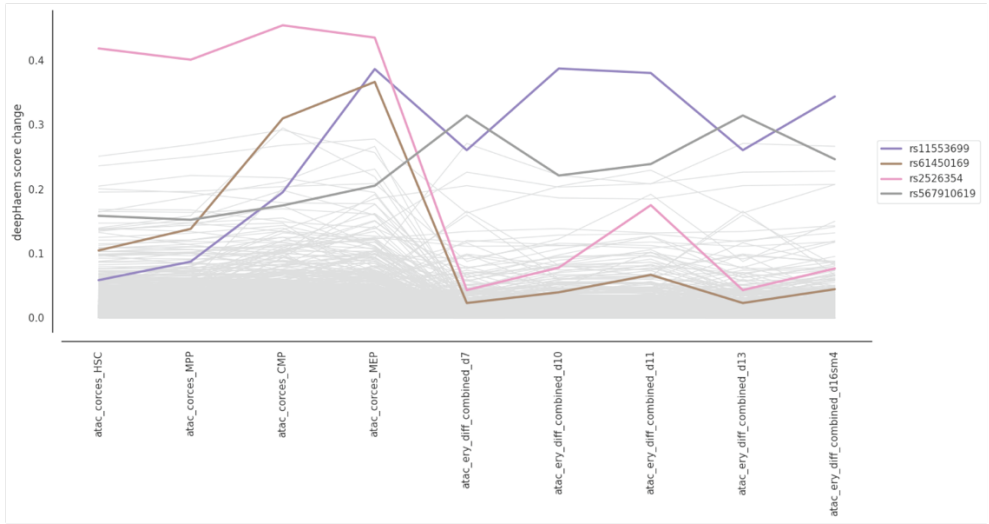


Figure 4.17: DeepHaem variant damage score predictions show changing values across erythropoiesis. DeepHaem score change depicts the effect a variant is predicted to have on a regions accessibility. Each line in the graph depicts a different SNP, with several that exhibit pronounced changes across the cell stages highlighted in colour. GWAS catalogues from the van der Harst [71] and Astle *et al*. [70] publications depicted.

Regions of open chromatin are skewed in early populations and inactive in later stages

While this data from a single donor provides limited power in the analysis of early stage GWAS variants, they did allow us to address a basic question of the technology; can allelic skew be detected in the aggregated single cell clusters from scATAC-seq?

Therefore allelic skew of open chromatin was assessed at the early time points, with a focus on clusters 1 and 2. This enabled the discovery of open chromatin sites that are directly affected by heterozygous variation that has the potential to be causal. The results of this analysis could be filtered to show open chromatin sites that were skewed in early populations but inactive in later stages. Though these results have yet to be fully explored, two examples of such regions present in donor 2 are shown here. In these regions the open chromatin sites are both present only in early populations and show heterozygous skew in their activity (Table 4.5, Figures 4.18 and 4.19). While the SNVs and SNPs at these examples are not presently linked to human genetics, they show that such analysis can be performed in single cell epigenetic data, and this paves the way for large scale analysis of regulatory variation using this technology in the future. The next step of this process will be fully exploring this dataset to find variants for prioritisation that are skewed in early populations.

Gene regions	Coordinates	ATAC read count		Variants/ SNPs present
		Allele 1	Allele 2	
ASB2	chr14:93962492-93963599	17	0	chr14:93963084:G->A = no rsid, too rare
IL16	chr15:81293372-81294718	42	72	chr15:81293722:C->T = rs17875486 chr15:81293903:T->C = rs17875577 chr15:81294587:G->C = rs17875491

Table 4.5: Regions open in early cell stages show allelic skew.

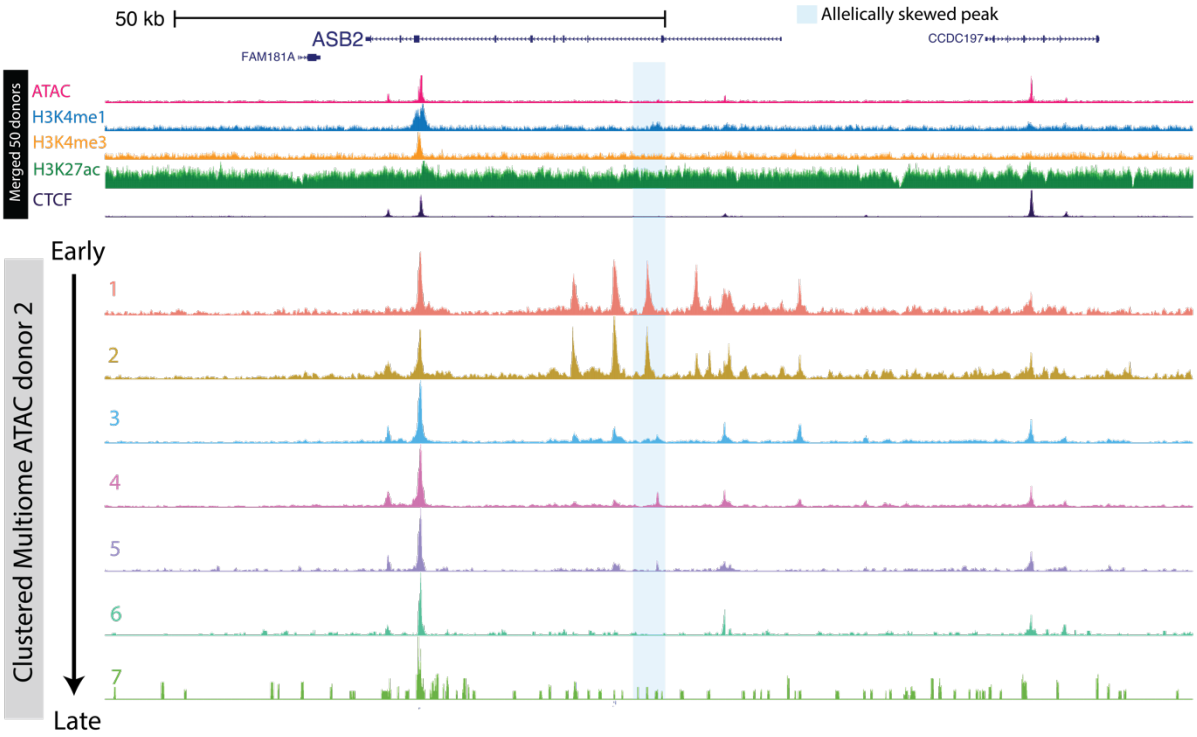


Figure 4.18: ATAC accessibility landscape at *ASB2* shows many peaks turning off simultaneously through erythropoiesis.
CD34 day 13 Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Blue highlight depicts allelically skewed peak caused by a variant in the early cell populations. Lower tracks depict ATAC for a single individual, data split by the clusters depicted in figure 4.5c, with corresponding colours, reflecting erythroid progression. Several adjacent peaks that are open in the early populations turn off as differentiation progresses.

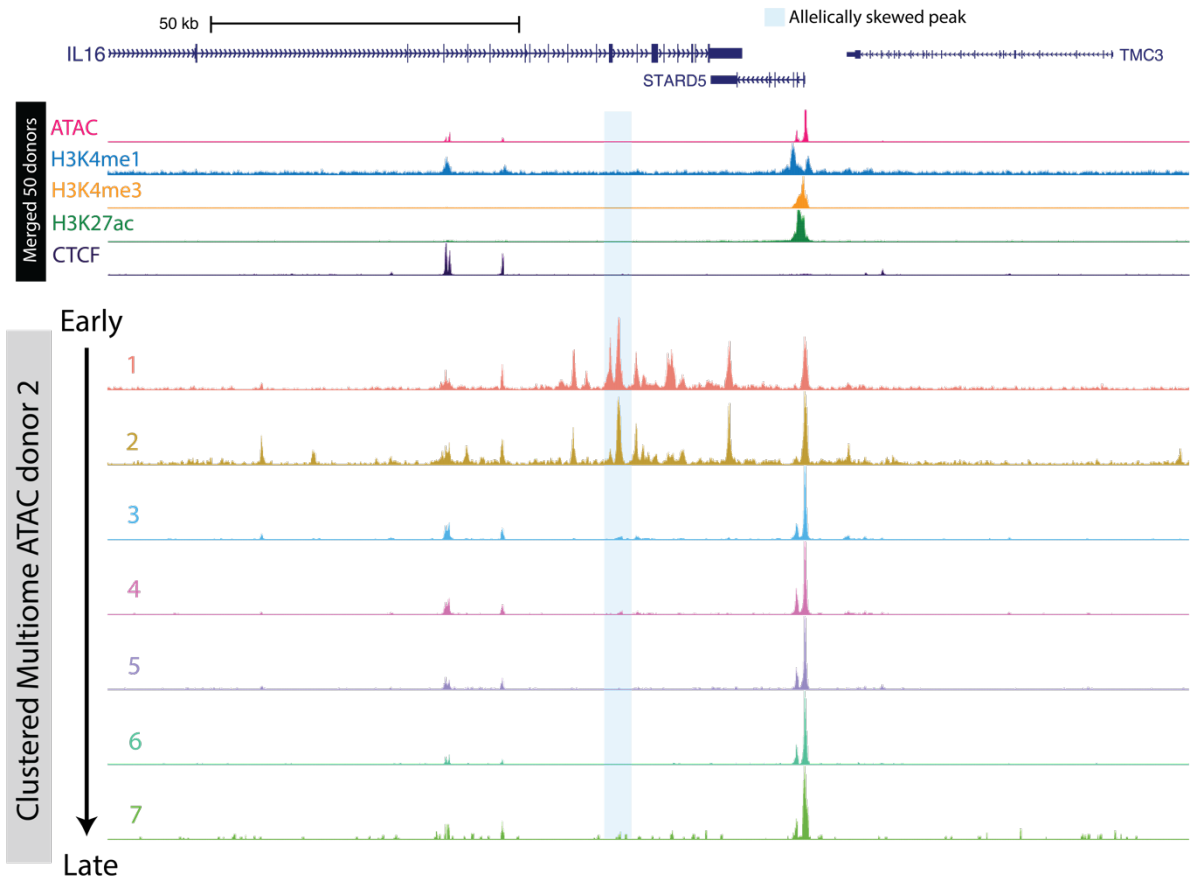


Figure 4.19: ATAC peaks spanning multiple genes show dynamic changing through erythropoiesis.
CD34 day 13 Track heights auto scaled to data view, ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 data sets. H3K4me1 (yellow track) and H3K4me3 (blue track) ChIP seq from the 5 available sets (Table 2.1). Blue highlight depicts allelically skewed peak caused by a variant in the early cell populations. Lower tracks depict ATAC for a single individual, data split by the clusters depicted in figure 4.5c, with corresponding colours, reflecting erythroid progression. Several adjacent peaks that are open in the early populations turn off as differentiation progresses. The lack of early histone modifications makes classification of these peaks difficult.

Discussion

In this chapter I have used the scMultiome system to map the changing regulatory landscape across erythroid differentiation. When performing experiments to functionally validate variants (Chapter 2) and to determine their regulatory interactions (Chapter 3), it is critical to establish the correct cell type and stage in which to do the experiments. In this chapter I have performed preliminary analysis to test the importance of considering the effect of genetics across the full arc of differentiation to understand human traits. With scMultiome data spanning a CD34+ differentiation system, I have shown how both the ATAC signal and expression of marker genes can change profoundly. I have also shown that when aggregated scATAC-seq signal is highly effective at mapping open chromatin sites and allows for allelic skew analysis in relatively inaccessible cell populations.

One of the major technical issues I found whilst using the multiome system was the disparity between the number of cells targeted and the number actually captured. In both attempts, both under and over sampling was observed, despite attempts to improve the accuracy of counting and reduce losses during processing. Whilst possibly reflecting our inexperience with this platform, it does show the requirement of thorough optimisation of the process if the platform is to be scaled up to a large number of samples.

Despite this, I was able to generate data from two separate donors and perform in-depth clustering on one of these donors. Samples were obtained from days 0, 5 and 10 of the CD34 differentiation system. Initial analysis of these data using the 10X Loupe software showed a distinction between the early day 0 samples and the later day 5 and 10, with the latter showing a continuous trajectory. By assessing the gene expression data and overlaying enrichment onto the UMAPs, it could be confirmed using known erythroid specific markers that the trajectory followed an expected erythroid path. When analysed further using ArchR, this gap between the early and later stages remained. It should therefore be concluded that the samples obtained likely did not contain all intermediate stages of erythropoiesis (Chapter 1 Figure 1.6) as expected. With the early stages of the differentiation between days 0 and 7 being defined as the expansion phase, with stem cell factors to encourage proliferation over differentiation, this was surprising. Instead of spanning several intermediate stages, the day 5 samples were much further along the differentiation trajectory than expected. This therefore means

that to fill in this gap and complete the data set, another time point between those of days 0 and 5 should be obtained, analysed, and added to the trajectory.

Despite this missing data from the absent cell stages, I was still able to use one donor for further clustering analysis. From these clusters I was then able to demonstrate the changing landscape of the open chromatin across erythropoiesis. In particular, I looked at the variants explored in previous chapters to show how dynamic and specific these regions are to the latter stages of differentiation. Due to difficulty of generating ChIP-seq data for marks such as H3K27ac in the earlier populations using epigenetic marks to determine the overall activity at these regions remains difficult. However, the changes in peaks demonstrated regions that were either open in early but off in late cells; off in early but on in late; a combination of both or in some instances stable. Notably CTCF seemed to be the most stable across the stages, which reflects its often cell non-specific nature. The enhancer and enhancer-like regions however were shown to be much more dynamic. These changes to the genomic landscape were to be expected, especially the dynamics of the enhancers. The process of erythroid differentiation sees a stem cell going through extreme specialisation, and so the remodelling of the genomic landscape is essential for the regulatory networks to turn on the essential erythroid pathways.

Assessing allelic skew and GWAS intersection in early cell stages

In addition to the previously explored variants, allelic skew in open chromatin sites of the early populations defined further regions potentially affected by variants and confirms the utility of these data for allelic based analysis. In these examples, filtering of allelic skew data found peaks in the early populations that were skewed by a variant and also off in later stages. These examples therefore highlight the importance of considering earlier populations, as they demonstrate regions that are affected by variants, but will not be intersecting open chromatin regions in later cells. Since it is often the terminally differentiated late cells that are assayed for accessibility, these potentially causative variants would be discarded from any SNP interrogation. However, variation in open chromatin does not prove causality and in these examples no phenotype is currently associated with them. By using tools such as deepHaem, the damage imparted by a known variant can be predicted across multiple cell stages to allow for the specific recall of donors for differentiation analysis. If the donor sampled

has heterozygous variants, scMultiome data can be used to assess allelic skew and therefore altered activity in these early populations. When performed across multiple individuals, confidence is gained as to the causal link between the variant and activity, but importantly they also act as validation for the Neural Network tools themselves. This is particularly important at present due to the limited datasets of known causal regulatory variants to act as validation and training sets for these approaches.

Moreover, enrichment analysis using three established GWAS catalogues associated with red blood cell traits demonstrated enrichment across differentiation. In these analyses, strong enrichment could be seen across all studies. Also, enrichment over the baseline was seen across all stages to varying degrees. This highlights that these GWAS SNPs associated with red cell traits are enriched in the open chromatin profiles of all erythroid cell stages, and therefore could be causal at many stages. Despite the general patterns showing some variability from study to study, the stage of differentiation from which we have generated the vast majority of our analysis was consistently confirmed as the most enriched and so most relevant.

Where the data exists, this combination of approaches can be used with other cell systems and associated GWAS to assess whether SNPs are also enriched in the open chromatin of earlier cell stages. More importantly, this could also provide the explanation of the causality of SNPs and the cell stages in which the mechanisms are being affected, which are current unsolved in these genetics.

Determining the mechanisms of early variants

However, as demonstrated in the previous chapter, linking a variant with open chromatin, even when also active, is not enough to prove effect or infer the gene it controls. We have shown that MCC is able to link regulatory elements in late-stage erythroblasts at high resolution and even shows allele specific changes in interactions and TF binding footprints.

However, with the limiting cell numbers typically available from early populations MCC in these stages is currently not possible. While research to lower the required cell input for MCC are currently ongoing in both the Davies and Hughes labs, other potentially appropriate and informative methods exist. Tiled

Capture-C is a 3C method that can generate high resolution “many versus many” interaction profiles similar to Hi-C at targeted regions, from as low as 2,000 cells (Figure 4.22) [98]. While it lacks many of the advantages of MCC, such as footprinting and allele specific analysis, this method could provide a workable approach to link open chromatin regions with genes in early populations where MCC is currently not possible. I was keen to test this possibility as part of this work and started this experiment and made restriction based 3C libraries from days 0, 5 and 10 CD34 derived samples ready for capturing. Though I ran out of time and so could not complete this experiment, I used the same variant analysis scripts from Chapter 2 on the ATAC data of the early cells to find regions that are variable across individuals and that change during differentiation to design Tiled-C capture designs. While not as elegant or as powerful as MCC and with the caveat of the degree of heterogeneity in differentiated d5 and 10 samples detected in the multiome data, I felt that this approach could be a valuable addition to the tool kit described here to link variable regulatory elements with genes in these challenging rarer cell types. I hope to complete these experiments subsequent to my thesis.

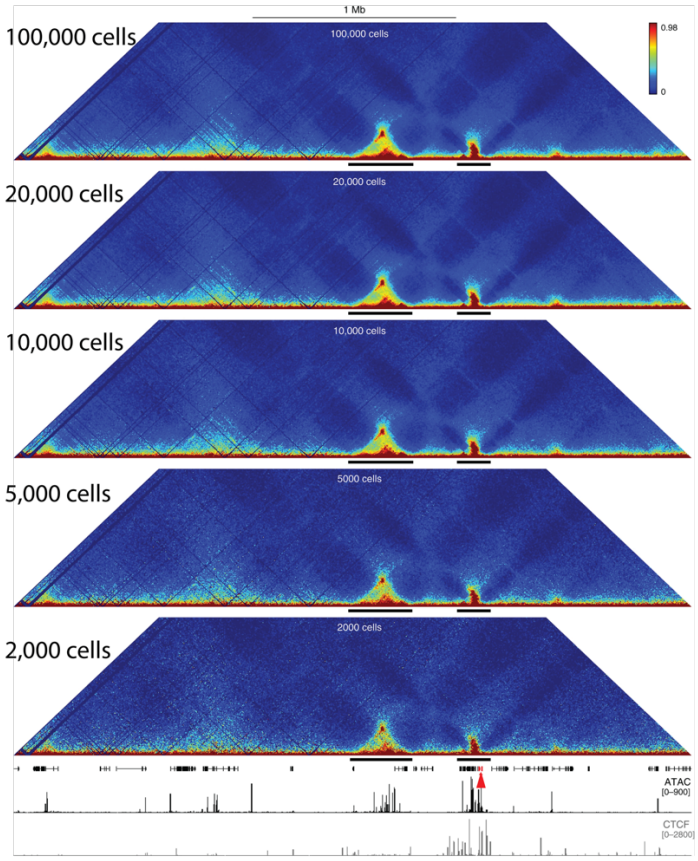


Figure 4.20: Tiled-C can detect chromatin interactions in as few as 2,000 cells.
All-vs-all interaction profiles generated in mouse at the *Hba* locus (red arrow) show strong interactions from 100,000 cells. These profiles can also be seen in maps generated from 2,000 cells. (Adapted from Oudelaar *et al.* [98])

Acknowledgements

In this chapter I optimised cell lysis steps and worked alongside others on library generation. I prepped and pooled final libraries for sequencing. I performed initial analysis and visualisation of data, which was then built upon by Ravza Gur. I worked alongside Simone Riva for further data analysis and visualisation.

Members of the WIMM single-cell facility provided help and support with experiment design, optimisation and library generation. Neil Ashley helped with the earlier experiments, including library generation, and Maria Greco advised on later experiments. Caroline Scott cultured the day 5 and day 10 samples used in this experiment. Christina Simoglou Karali provided technical advice and lead the library generation of the PBMC and day 0 samples. Kevin Clark of the FACS facility ran the PBMC sorting. Tim Rostron of the sequencing facility supported the sequencing runs. Ravza Gur and Jennifer Herrmann provided support with CellRanger analysis. Ravza Gur provided further support with trajectory mapping, cluster generation and data visualisation. Simone Riva provided deepHaem and enrichment profiles.

Chapter 5 – Discussion

Since the completion of the human genome project in 2003 vast advances into our understanding of human genetics and genomics have been made [2]. In particular, the sequences outside of protein coding regions that were thought by some to be 'junk DNA' have in fact been shown to harbour regulatory elements that have been crucial in mammal diversification and human evolution [42, 45, 49, 50, 203]. These regulatory elements are potent evolutionary targets that can have profound phenotypic impacts, such as the loss of limbs central to snake evolution, and have been shown to have driven vital processes that have made us human, such as increased cognition [48, 204]. Consequently, previous research has focussed on finding mutations within these regulatory regions that drive disease phenotypes and thus many mechanisms that break these regulators, disrupt transcriptional output and cause disease have been found [11, 43, 137].

By exploring the genomic and epigenomic profiles of 50 individuals, I have identified further examples of variation within the regulatory genome. In particular, I have described and presented examples of gain of function variants, whose regulatory roles are currently not well recognised and thus underappreciated in the field. To link variable regulatory elements to the genes they control I have performed MCC in a primary human erythroid differentiation system. Through the generation of high-resolution interaction profiles, MCC can unambiguously link regulatory elements. Splitting MCC by genotype has allowed me to generate allele-specific interaction profiles, which refine our understanding of how variants can influence regulatory interactions. Moreover, the footprinting capabilities of MCC combined with allele-specific tracks has enabled me to interrogate TF binding at sites of variation. I have further studied the erythroid differentiation model using the scMultiome system, capable of generating ATAC and gene expression libraries from single cells simultaneously. This has allowed me to track the regulatory and transcriptional trajectory of erythropoiesis to better understand when variants are likely to exert a mechanistic effect and functionally impact the cells. Combined, these approaches have highlighted the extent of variation that can be found in even a small population sample, and have aimed to characterise how these variants may be affecting genome architecture and regulation.

Identifying new variants in the genome, and their evolutionary and disease implications

As previously described, GoF variants have been observed but are currently less defined or well understand than the more extensively studied LoF variants [90, 146, 147, 205] (Figure 5.1). This project not only demonstrates the extent of variation that exists at REs between individuals of a population, but also identified novel cases of these GoF sites. Though the GoF examples presented here have not been fully validated mechanistically, their presence and our ability to locate them is still important.

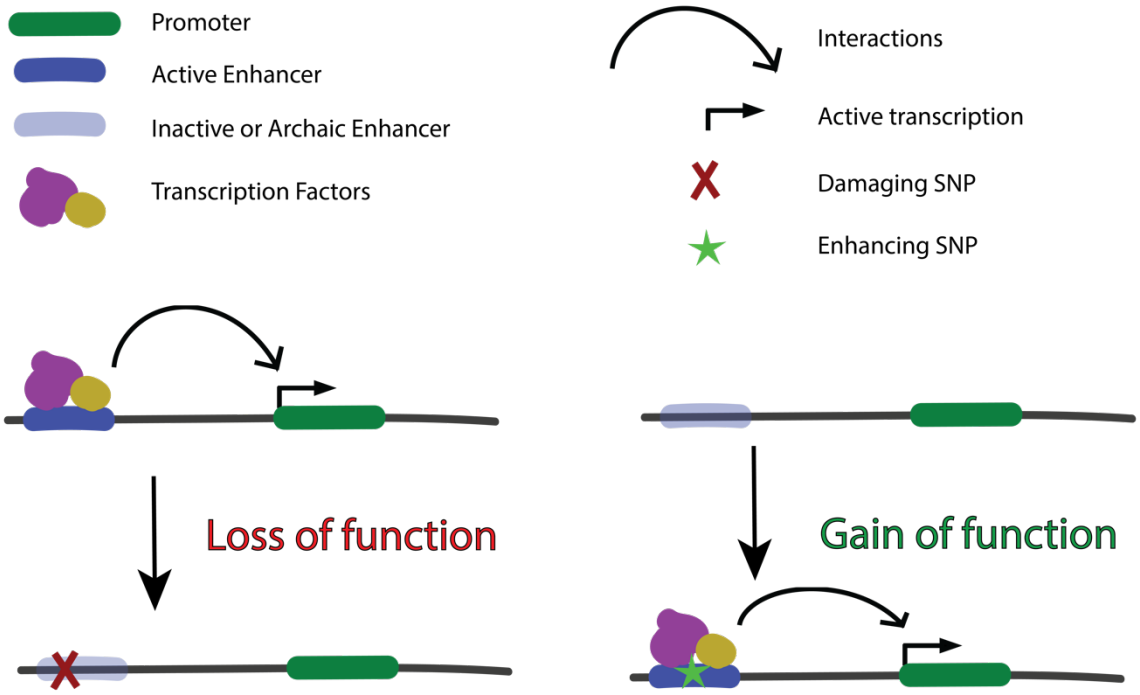


Figure 5.1: Loss and gain of function variants.

The exact process by which new enhancers arise is yet to be fully established. Several mechanisms of element genesis were described in chapter 1, such as the impact of TEs, duplication of existing elements, mutations in archaic elements or the *de novo* formation of elements. Of these mechanisms, GoF variants could either form *de novo* elements or alter archaic elements with pre-existing motifs. A novel GoF in a single donor on a unique haplotype was found in an intron of *PTPRK*. This region often exhibited weak open chromatin in all other donors, despite being functionally inactive in these samples, which suggests priming or the pre-existence of weak, cryptic or archaic motifs. ENCODE DNase hypersensitivity was also assessed across 95 different cell types with none showing open

chromatin at this peak, suggesting that this site does not merely represent an RE with activity in other cell types becoming active in these erythroid cells (Supplementary Figure 13). This does however raise an interesting possibility that cell type specific enhancers could acquire SNPs that alter their activity so that they become active in other cell types. Because of the importance of enhancers in mammalian and human evolution, it is likely that GoF variants have played a key role in this process. In particular, the refined regulation of key neurological genes has been pivotal in human evolution [49, 204].

In addition to their potential role in the progression of human evolution, regulatory elements have also been shown to be involved in disease mechanisms. A notable example is that of a single base change in the human *HBA* cluster that creates a promoter with an associated unidirectional transcript. This GoF variant is able to downregulate *HBA* expression which results in α-thalassaemia [146]. Additionally, the SNP rs17713054 G>A , has been identified as a risk allele in the context of COVID-19, whereby a GoF is associated with a poor prognosis [90]. The methods described in this thesis can be used as tools to help find further examples of disease-associated GoF variants that may have been missed in previous studies.

The importance of discerning the effects of these SNPs is highlighted by the development of novel therapies that target regulatory networks. If a certain gene or pathway can be shown to be altered due to regulatory variation, then drugs or other therapeutic modalities can be developed to restore healthy protein levels with the potential to alleviate symptoms or slow disease progression. Additionally, advances in genome therapy treatments are being developed based on the precise understanding of how genes are regulated [206]. For example, our understanding of the mechanism that underlies the persistent expression of foetal haemoglobin in adults (*HBF*), caused by a naturally occurring SNP within an enhancer of the *BCL11A* gene, has led to a genome editing strategy for treating patients with sickle cell disease and thalassaemia [207].

BCL11A is a transcriptional repressor that acts on *HBG*. In turn, *BCL11A* expression is reduced by a SNP in one of its enhancers, and the downregulation of this repressor leads to increased expression of *HBG* and subsequent production of foetal haemoglobin (Figure 5.2) [208, 209]. In this example, the SNP alters a GATA site, thus preventing the binding of GATA1, a key erythropoietic transcription factor. Utilising genome editing to knock-down BCL11A levels consequently results in an upregulation of *HBF*,

which in turn can fill the gap left by the fault in adult haemoglobin. By using the approaches described in this thesis to decipher further disease SNPs, both LoFs and GoFs, this could therefore open up more potential treatment options based on the understanding of regulatory variation.

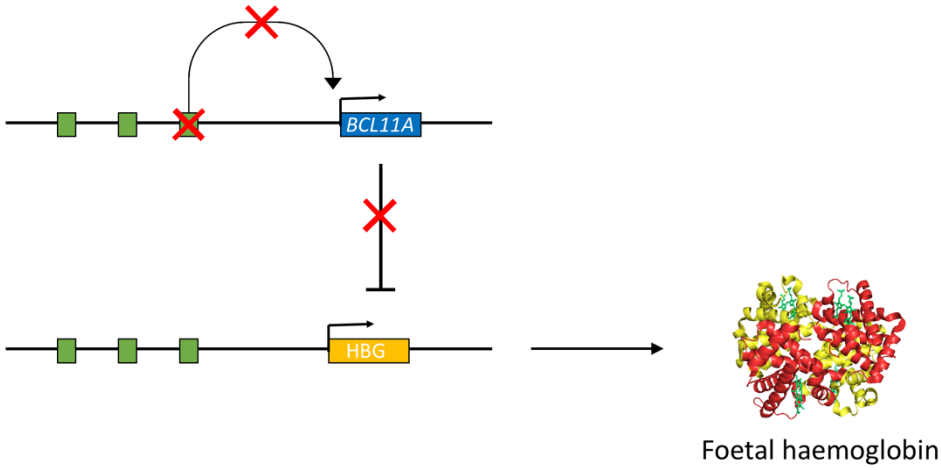


Figure 5.2: A SNP in an enhancer of *BCL11A* causes the derepression of *HBG*, and the subsequent production of foetal haemoglobin.

Challenges in defining variants, interpretation and data visualisation

Many of the reasons why GoF variants are currently under-appreciated in the field have been discussed in previous chapters. Broadly, regulatory SNP prioritisation usually depends on intersecting variant data with pre-existing open chromatin data from a limited number of individuals, and often from only a single individual. GoF SNPs will frequently lie outside of mapped peaks, unless by chance the reference mapped to also has the variant or active peak, and will therefore be excluded from further study. This highlights another issue when defining SNPs and their presence: as the human genome is comprised of data from several individuals, it does not represent an actual individual or the most common genotype of the population. Therefore, when considering a sequence as either 'reference' or 'alternate' this provides no indication as to its actual frequency in the population.

Furthermore, when considering gene expression within a single nucleus, each allele will behave independently, yet this is not reflected in the way we currently present data. The UCSC browser is a

versatile tool for visualising multiple genomic assays simultaneously across many species. However, in the case of diploid organisms, the single tracks displayed do not reflect the true nature of the biological activity taking place within cells. As demonstrated with the allelic ATAC skew data and the allele specific MCC interactions and footprinting data, each allele can and does behave differently. Following on from this observation, further investigations will aim to build on these findings and uncover the extent to which these variants and their changed interaction profiles can impact expression. This will rely on phasing data, whereby SNPs are used to link entire alleles, which in turn can be used in conjunction with allelic MCC and RNA-seq to determine how SNPs are influencing gene expression. By phasing the SNPs in an RE to nascent RNA-seq data where introns also contain SNPs, measurement of the transcriptional output of a gene in relation to the presence of a variant will directly show the variant's effect. This represents ongoing work after this thesis.

When considering these data, it can be assumed that in heterozygous alleles each is subjected to local polygenic influences, and so this differential behaviour will have an effect on phenotype. When amalgamating data from two alleles, the subtle differences between them are diluted. This is unimportant for homozygous regions with mirrored activity, however at heterozygous regions the lack of activity on one allele will be obscured by the active one. Several examples in this project have shown this gradual decrease in open chromatin and activity data from a functional homozygous to a heterozygous to an inactive homozygous. Although techniques such as allele skew analysis can unpick these regions, in terms of visualisation having each allele visualised separately might be more informative. Historically, sequence depth has been a problem for genomic assays, due to low cell numbers producing low yield libraries and/or the cost of sequencing, especially when performed genome-wide. However, as assays become more efficient at working with low cell numbers, and sequencing becomes cheaper and of higher throughput, the depth of data obtained per run is often ample to view each allele separately. Of course, due to alignment biases or the lack of SNPs across the whole genome, certain regions would be harder to do this for.

These issues were also highlighted in the LoF and GoF defining analysis shown in Figure 2.4. With this approach, the true extent of LoFs will have been missed, as a LoF on a single allele would not be considered a loss. Instead, in order for a peak to be considered a loss, it would need to be absent on

both alleles. By running the same analysis again but with each allele mapped separately, the true extent of LoFs in a population would become apparent.

It has been shown that regulatory elements can induce local activity [22]. Gene editing experiments in a mouse model have shown that inserting the strongest enhancer of *HBA* (R2) into a neutral region on the X chromosome (that shows no pre-existing open chromatin or histone modifications) can alter the genomic landscape. Once inserted into mouse embryonic stem cells, cells were differentiated down an erythroid lineage, and not only did the inserted sequence become accessible, but several other active open chromatin regions arose downstream of the ectopic enhancer. Moreover, this change in landscape was only observed in erythroid differentiated cells and not the embryonic stem cells, and so suggests that erythroid specific TFs present during the differentiation were able to open the ectopic enhancer which in turn led to altered downstream activity. In the examples provided in Chapter 2, I have shown that multiple peaks can become accessible simultaneously in the same landscape. Elements are known to interact synergistically, and this can further induce gene expression [102]. However, when considered allelically, it is possible that these elements are not present on the same allele, and instead the open chromatin landscape could be depicted elements on separate alleles. With 3C studies we know that *trans* interactions between different alleles are rare, and so two REs on separate alleles are unlikely to interact with each other at an influential frequency [91]. Conversely, if the elements are on the same allele, this will influence their interactions and ability to express genes. Enhancers in particular have been shown to exhibit distinct roles in a locus; R1 and R2 at the *HBA* locus are the main drivers of expression, whilst R3 and R4 play a facilitator role [103].

Though certain loci of essential genes such as *HBA* show strong conservation, this may not be true for more mutation tolerant regions. Instead, where the alleles at strongly conserved regions are likely to mirror each other, this is less likely at regions that can be more flexible in their output. It is these loci and the SNVs within them that drive the diversity of a species. Through subtle changes in RE activity that can affect multiple components of signalling pathways, this can have clearly detectable effects on phenotype. Additionally, in these affected loci there is a chance that variation is affecting peaks of each allele separately, either by damaging or creating them on a single allele, and in an amalgamated track these subtle differences would be lost. If there were also resultant gene expression changes, the true cause of these would be hard to determine. For example, if a landscape displays two enhancer

peaks, from an amalgamated track it is not clear if these exist on the same allele or if each belongs to a separate one, nor is it clear how this will impact gene expression. This is also in part due to the fact that local enhancers within the same regulatory networks can influence each other and the output of nearby genes, for example either additively, synergistically or with hierarchical logic [12]. The variation such as this is evident at the *PPP1R3B* locus, whereby three prominent haplotype structures with varying degrees of activity were observed.

In addition to SNPs, variation can occur through larger mutations such as deletions, insertions, inversions, duplications and translocations, known collectively as Structural Variants (SVs). As these are more difficult to characterise using typical genome sequencing, their prevalence within the human population is less well characterised. The quantification of SV frequency is challenging due to the shortcomings of short-read based sequencing methods, which struggle to identify SVs. The identification of these variants is also challenging from a computational perspective as the pre-made reference genome is required in the alignment process, and if the reference does not contain the SV, the data is often not aligned. For example: if an individual has an insertion, the reads that come from this region will have nowhere to align to; if an inversion or translocation is present, the aligner will assume the sample matches the reference and align accordingly; if a duplicated region is not reflected in the reference, all reads from the affected regions will align to a single region. This is further complicated by the fact that diploid data is mapped to a single reference, thus if a heterozygous SV is sequenced, an aligner will only report whichever matches the reference. In the case of heterozygous deletions, this means that there will be coverage over the deleted region, originating from the allele without the deletion. These analyses are made even more challenging as large-scale genome rearrangements often have their breakpoints in repetitive regions, which do not align in short read sequencing and so the informative novel junctions caused by SVs are lost during the alignment process. Despite these difficulties, efforts have been made to use current technologies to detect these SVs, such as looking at coverage to find deletions, or break points to find inversions [210, 211]. Additionally, long read sequencing technologies, such a Nanopore [112] and PacBio [113], or optical mapping technologies such as Bionano [212] are being used, often in combination, to overcome such problems. With long sequencing the length of reads is sufficient to span such repetitive regions and allow for the detection and reconstruction of complex SVs and in the Bionano technology, based on fluorescent labelling of the DNA strands, no sequence alignment is required.

With these advances in discoveries, examples of SVs affecting REs and gene regulation have been brought to light. SVs can have a profound effect on gene regulation if they delete, insert, duplicate or shuffle regulatory elements, boundary elements or genes. A well characterised example by Franke *et. al* describes three distinct duplications affecting two adjacent TADs that separate *KCNJ2* and *SOX9*: an intra-TAD duplication causing a sex reversal phenotype; an inter-TAD duplication containing a boundary element that creates a neo TAD with no new phenotype; and an inter-TAD duplication containing a boundary element, enhancers and *KCJN2*, forming a neo-TAD with a duplicated gene, leading to Cooks syndrome [185]. In both of the examples where a phenotype is observed, this is directly linked to the misexpression of the genes present. This example also highlights the importance of boundary elements in regimenting the organisation of the genome to avoid ectopic interactions.

We also believe that SNPs may be marking SVs by linkage disequilibrium. Due to the inherently disruptive nature of SVs, they can inhibit normal recombination during meiosis, a process that shuffles alleles, and relies on the sequences matching. By inhibiting recombination, SNPs that are close to each other will remain linked over generations, forming distinct haplotype blocks and marking the region. If the SV present both enhances the linkage disequilibrium and causes a detectable phenotype by effecting the genome organisation or a regulatory element, any neighbouring SNPs found by GWAS may be marking the SV. This could therefore help explain why certain GWAS hits currently show no obvious mechanism in terms of the LoF and GoF mechanism described in this thesis.

The human genome is highly variable

Despite these challenges, we have been able to demonstrate some of the variation in a population of healthy individuals. Though this thesis has only scratched the surface of the amount of variation present, it provides a valuable starting point to further find and define variants. In this population, the highly conserved *HBA* locus unsurprisingly showed no significant differences across the individuals, which mirrors the lack of an anaemia phenotype in these individuals. However, several loci, many of which are associated with erythroid GWAS SNPs, showed variation in their genomic landscapes.

By searching for differentially active regions in the genome and visualising them across 50 individuals I have demonstrated how much the genome and REs can change within a population. From these regions several patterns emerged, with many affecting enhancer elements. These enhancers were often the easiest to link to the genes they control by using MCC. Future work to fully decipher the impact of these enhancer variants will involve phasing the variants to allelic RNA-seq outputs. By using this technique we hope to further clarify the mechanisms by which these SNVs and GWAS SNPs are functioning.

Additionally, a variant downstream of *FERMT1* was associated with the production of a novel transcript. The true nature of this element is ambiguous, and although it showed promoter-associated histone modifications and produced both nascent and stable transcripts, it is unannotated in the reference genome. It is also well-recognised that enhancers can produce RNA, though these are often much smaller (~3-500bp) than the 80kb transcript seen here. Additionally, studies have shown with genomic editing that ectopic enhancers can act as weak autonomous promoters, further blurring the distinction between REs [213-215]. The exact definition of what constitutes an enhancer or a promoter is ongoing. Regardless of this distinction, the appearance of this novel transcript could shed light on the processes around the genesis and evolution of promoters.

CTCF sites were also shown to be impacted by variation, and their interactions with other sites abolished. Allele-specific MCC at several CTCF sites showed the extent to which variants can disrupt CTCF interactions. As structural proteins forming the base of loop structures, the interactions between these CTCF sites are highly evident in MCC data. As a result, the effect of disrupting these sites is also easily visualised. Though profound in their effect, assessing the downstream effects of these changes is difficult for several reasons. CTCF sites are often not cell type-specific and instead show similar binding patterns across multiple cell types. It is therefore not obvious as to which cell types a CTCF change will be affecting. Importantly, these changes can either disrupt an existing site or create a new one, which can cause ectopic interactions or block interactions, respectively [185, 216]. Since CTCF sites have forward and reverse directionality that dictate loop formation, the orientation of any new sites also will have an impact [217]. As evident with SV examples, the disruption of the landscape and the ectopic expression of genes can lead to disease [185]. Conversely, inserting an ectopic CTCF site between the *HBA* promoter and enhancers caused the downregulation of this gene [218]. As MCC is

able to link REs, if the elements around a variable heterozygous CTCF site are studied using this method, then the impact of CTCF variation on the regulatory interactions between other elements can be determined. This is dependent on working in the affected cell type, and so several cell types might need to be explored to determine the true size of effect of a CTCF deletion. It has also been shown that although disrupting CTCF changes the genomic landscape, it does not always affect gene expression [178, 179]. Combined, these factors make unpinning the effects of CTCF variants difficult.

Mapping data to the right place at the right time

The reference genome is an invaluable resource used to map genomic data. However, with the issues presented around individuals not matching the reference, either with small SNVs or larger SVs, and the effect that these can have on the genomic landscape, a strong case for implementing personalised genomes can be made. Mapping reads to custom genomes is a common practice when studying other species. However, at present it is extremely laborious in humans when considering the issues that arise when determining the following: how much coverage is required for confident base calling; how many individuals across a population are required to fully cover human variation; and the analysis required to generate fully phased alleles. The technology for all this exists in its infancy, but is currently expensive and labour-intensive. However, this would present a gold standard approach to ensure no data is lost during alignment, and that all SNVs and SVs are defined and linked allelically.

As well as recognising the importance of considering the genome allelically, this project has also highlighted the importance of assessing genome variants in the actually affected cell types. Cell type-specific REs have been well established in terminally differentiated cells, and the scMultiome data presented here has shown the dynamic nature of the genome across differentiation. It has shown that elements can be active early in a cell lineage and then inactive later, and that erythroid GWAS interaction is enriched across the whole of erythropoiesis. Combined, this highlights the importance of studying the impact of genomic variation across cell lineages, and the scMultiome system provides the tools to do so. Allelic skew analysis of ATAC from early cell populations can be used to infer causality of SNPs in relation to activity. Moreover, by combining these insights with phasing data and RNA outputs, SNPs can also be linked to gene expression profiles. Though this project has focused on

erythroid differentiation, with ample material other differentiation systems can also be assessed in this way.

Through considered selection of cell stages during erythropoiesis, we aimed to capture all the cell intermediates of this process. However, trajectory analysis suggests that a proportion of this lineage was missing from our analysis and so further experiments will be required to fill in this gap. Additionally, due to the rare nature of some of the variants explored here, assessing the changes to these variants was limited to those that were present in the donors assayed. It would require considerable effort and expense to generate more data from individuals with known rare variants.

One potential issue when deciphering the impact and staging of novel transcripts in early populations is the ability to do nascent RNA-seq. Though powerful, our analysis of the data from the scMultiome assay shows the data to be heavily 3' biased and represent mature transcripts. Without intronic SNPs phasing of RNA data is challenging. However, this can still provide valuable information as to which cell stages should be assessed with more nascent transcriptomic approaches.

Novel techniques have helped define variant mechanisms and their timings

In this thesis, I have demonstrated the power of MCC by implementing the technique for use in primary human cells. This method is able to generate high resolution 3C profiles by utilising MNase, allowing for the design of capture probes to anywhere in the genome and the generation of base pair resolution interaction profiles. More importantly, I have generated allele-specific MCC profiles that are able to show the impact of heterozygous SNPs on genome organisation. Interaction profiles from variants of all classes of elements have shown a range of effects including: complete loss of binding and also interactions; partial loss of activity and interactions; no impact on interaction profiles; and no discernible landscape at all. MCC also generates footprinting data which in turn can be viewed allelically. From these profiles it is clear to see that variants can have an effect on specific motifs, altering TF binding and a region's accessibility.

Using the scMultiome method I have highlighted the importance of looking at earlier cell stages when considering the impact of GWAS SNPs linked to erythropoiesis. Though the MCC protocol currently

requires too many cells to reveal the interactions of earlier cell populations, work is being carried out to lower the required cell number. In the meantime, Tiled-C is an appropriate 3C method to work with low cell numbers and I will continue to explore the use of this method to link regulatory variants with their affected genes.

Conclusion

Our understanding of how variation affects the regulatory genome is ever growing and this project has sought to further explore these variants to better understand the mechanisms through which they have impacted evolution and can cause disease. A population of 50 individuals was shown to have profound amounts of variation, but with insights gained from MCC and scMultiome analysis, we can better understand these variants. This increase in our knowledge of regulatory elements provides a foundation to decipher additional variants and to use this knowledge in the design of treatments and therapies.

Chapter 6 - Methods

Oxford Biobank donors

Donors from the Oxford Biobank were chosen with no selection criteria to represent a random cohort of the Oxford population. Gender split: 38 female, 12 male. No other information (ethnicity, age) recorded.

The ethics approval and consent forms for the 50 donors were approved by the NRES Committee North West, Haydock, HRA NRES Centre, Manchester under the title 'Establishing the molecular basis for atypical forms of thalassaemia' (reference number MREC: 03/8/097).

Donor	Number of SNPs	ATAC peaks
Don001	5,639,466	30,961
Don002	5,688,016	41,428
Don003	5,712,371	43,981
Don004	5,928,242	21,725
Don005	6,163,623	22,199
Don006	6,095,412	23,336
Don007	6,012,038	22,717
Don008	6,052,386	34,408
Don009	6,090,551	19,605
Don010	6,177,971	20,656
Don011	6,208,354	18,780
Don012	5,720,239	24,655
Don013	5,636,609	21,891
Don014	6,143,840	21,226
Don015	6,099,136	20,262
Don016	6,206,342	26,633
Don017	6,225,298	30,810
Don018	6,116,820	29,418
Don019	-	21,315
Don020	-	25,085
Don021	5,752,297	23,856
Don022	5,868,331	24,126
Don023	5,825,935	24,527
Don024	6,120,029	29,875

Don025	6,584,461	26,166
Don026	6,365,762	31,526
Don027	6,198,394	28,629
Don028	5,714,531	25,889
Don029	-	21,952
Don030	5,813,948	26,112
Don031	5,759,145	26,371
Don032	5,791,212	26,266
Don033	5,825,601	21,590
Don034	5,730,375	33,365
Don035	5,677,408	34,998
Don036	5,537,771	35,432
Don037	5,933,618	31,587
Don038	5,805,316	32,472
Don039	6,015,669	29,025
Don040	5,781,315	21,382
Don041	5,703,720	27,768
Don042	5,984,975	24,993
Don043	6,146,075	27,001
Don044	6,122,402	32,094
Don045	5,532,098	20,204
Don046	5,607,329	30,284
Don047	5,909,041	30,154
Don048	5,699,288	22,750
Don049	5,676,493	27,747
Don050	5,951,834	29,907

Table 6.1 Summary statistics of donor SNPs and called ATAC peaks.

CD34 isolation

Primary $CD34^+$ cells were isolated from healthy adult donors. 100ml of blood was collected from human donors into anti-coagulation EDTA coated tubes. All following steps conducted in tissue culture hood. Blood was mixed at a 1:1 ratio with PBS and PBMCs were isolated by histopaque layering and centrifugation (630xg, 30 minutes, RT, deceleration at 0). Interphase layer of PBMCs was collected and washed once with PBS and twice with MACS buffer (PBS supplemented with final 0.5% BSA and 2mM EDTA). Cells were counted and $CD34^+$ Haematopoietic Stem Cells (HSCs) isolated with the Human CD34 Microbead Kit (Miltenyi Biotec 130-046-702). Cells were resuspended in 300ul MACS buffer, 100ul FcR Blocking reagent and 100ul CD34 Microbeads per 10^8 cells, and incubated on a roller at 4^oC for 30

minutes. Cells were washed with 10ml MACS buffer and centrifuged at 300xg, 15 minutes, and resuspended in 500ul MACS buffer per 10^8 cells. MACS columns for separation were primed with 3ml MACS buffer before cells were added. Columns were washed thrice with 3ml MACS buffer, $CD34^-$ PBMC flow through was collected. $CD34^+$ cells were eluted from the column with 5ml MACS buffer into a second column, washes and elution repeated. Cells were counted with Luna cell counter (LUNA-FL™ Dual Fluorescence Cell Counter, Logos Biosystems) and put into culture or frozen in 1ml freezing solution (10% DMSO, 90% FCS)

CD34 differentiation

CD34 base media was made by supplementing Iscove's Modified Dulbeccos Media (IMDM) with Glutamine with 3% Human AB Serum (heat inactivated at 56°C for 30 mins, spun at 3500rpm for 5 mins to separate), 2% FBS, 10ug/ml Recombinant Human Insulin (Sigma Aldrich I9278), 3 U/ml Heparin (H3149-10KU, syringe filtered), 200ug/ml Holo-transferrin (Sigma: T0665) and 1X Pen/Strep. Base media was further supplemented for culture stages. Phase I supplemented with 3U/ml Erythropoietin (Peprotech: 100-64), 1ng/ml Interleukin-3 (Peprotech: 200-03-10) and 10ng/ml SCF (Peprotech: 300-07-10). Phase II supplemented with 3U/ml Erythropoietin and 10ng/ml SCF. Phase III supplemented with 3U/ml Erythropoietin and 500ug/ml Holo-transferrin (200ug/ml from base media, 300ug/ml added).

Frozen cells thawed by gradual dilution in 9ml RT PBS, spun at 400xg, RT for 5 minutes.

During the first stage of cell expansion, cells were plated at 1-3x10^5 cells/ml in Phase I media and incubated at 37 °C with 5% CO_2 for 7 days, with fresh media added daily to maintain density. At day 7, cells were spun at 400xg, 5 minutes, RT and resuspended in Phase II media at a density of 3 x 10^5 cells/ml. Cells were then harvested at day 10, or transferred to Phase III media at day 11. Cells in Phase III plated at 1-2 x 10^6 cells/ml, until day 17.

ATAC-seq

ATAC-seq was performed on CD34+ cells as per the published protocol [80]. Cells were lysed in cold lysis buffer (10mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630) and Tn5 transposase (Illumina Tagment DNA Enzyme and Buffer 20034197) was added for 30 minutes at 37°C. DNA was purified using a Qiagen MinElute kit (Qiagen 28004) and libraries were amplified and indexed using NEBNext 2X Mastermix (NEB M0541S) and custom indexing primers. Library quality was assessed by D1000 Tapestation and libraries were quantified using the universal library quantification kit (KAPA 07960140001). ATAC-seq libraries (4nM) were sequenced on the Illumina NextSeq platform using a High-Output v2 75 cycle kit.

ATAC Custom Primers:

	Name	Primer Sequence
	Ad1_noMX	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG
N701	Ad2.1_TAAGGCGA	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGT
N702	Ad2.2_CGTACTAG	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGT
N703	Ad2.3_AGGCAGAA	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT
N704	Ad2.4_TCCTGAGC	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT
N705	Ad2.5_GGACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGT
N706	Ad2.6_TAGGCATG	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT
N707	Ad2.7_CTCTCTAC	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGT
N708	Ad2.8_CAGAGAGG	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGT
N709	Ad2.9_GCTACGCT	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT
N710	Ad2.10_CGAGGCTG	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGT
N711	Ad2.11_AAGAGGCA	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGT
N712	Ad2.12_GTAGAGGA	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGT
N713	Ad2.13_GTCGTGAT	CAAGCAGAAGACGGCATACGAGATATCACGACGTCTCGTGGGCTCGGAGATGT
N714	Ad2.14_ACCACTGT	CAAGCAGAAGACGGCATACGAGATACAGTGGTGTCTCGTGGGCTCGGAGATGT
N715	Ad2.15_TGGATCTG	CAAGCAGAAGACGGCATACGAGATCAGATCCAGTCTCGTGGGCTCGGAGATGT
N716	Ad2.16_CCGTTTGT	CAAGCAGAAGACGGCATACGAGATACAAACGGGTCTCGTGGGCTCGGAGATGT
N717	Ad2.17_TGCTGGGT	CAAGCAGAAGACGGCATACGAGATACCCAGCAGTCTCGTGGGCTCGGAGATGT
N718	Ad2.18_GAGGGGTT	CAAGCAGAAGACGGCATACGAGATAACCCCTCGTCTCGTGGGCTCGGAGATGT
N719	Ad2.19_AGGTTGGG	CAAGCAGAAGACGGCATACGAGATCCCAACCTGTCTCGTGGGCTCGGAGATGT
N720	Ad2.20_GTGTGGTG	CAAGCAGAAGACGGCATACGAGATCACCACACGTCTCGTGGGCTCGGAGATGT
N721	Ad2.21_TGGGTTTC	CAAGCAGAAGACGGCATACGAGATGAAACCCAGTCTCGTGGGCTCGGAGATGT
N722	Ad2.22_TGGTCACA	CAAGCAGAAGACGGCATACGAGATTGTGACCAGTCTCGTGGGCTCGGAGATGT
N723	Ad2.23_TTGACCCT	CAAGCAGAAGACGGCATACGAGATAGGGTCAAGTCTCGTGGGCTCGGAGATGT

N724	Ad2.24_CCACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTGGGTCTCGTGGGCTCGGAGATGT

Table 6.1: ATAC-seq custom primers.

ChIP-seq

Chromatin Immunoprecipitation protocol adapted from Millipore protocol (Millipore: 17-10085) and optimised by Emily Georgiades. Cultured cells were washed in PBS and aliquoted in $1x10^7$ replicates. Cells were fixed in 1ml PBS with 2mM Di(N-succinimidyl) glutarate (DSG) (Sigma 80424-50MG-F) whilst rotating at room temperature for 30 minutes (DSG was prior resuspended in DMSO and aliquoted). Cells were spun down and resuspended in 1ml PBS with 1% final formaldehyde (Sigma: 252549) whilst rotating at room temperature for 30 minutes. Fixed cells were washed in PBS and frozen at -80°C.

Fixed cells were thawed and resuspended in 120ul SDS lysis buffer (1% SDS (Severn-Biotech 20-4002-02), 10mM EDTA (ThermoFisher AM9260G), 50mM Tris-HCl (pH 8) (Life Tech 15568025), in water) with cOmplete Protease Inhibitor Cocktail (PIC) (Sigma P8340) and syringe passaged through a 27G needle into a covaris tube. Samples left on ice for 10 minutes before being sonicated with Covaris Sonicator ME220 for 10 minutes (75 power, 25% Duty Factor, 1000 cycles per burst, 4°C) to yield average fragment size between 200 and 400 bp. Sonicated samples were spun and the supernatant diluted to 1.3ml with ChIP dilution buffer (0.01% SDS (Severn-Biotech 20-4002-02), 1.2mM EDTA (ThermoFisher AM9260G), 16.7mM Tris-HCl (pH 8) (Life Tech 15568025), 1.1%Triton-X (Sigma T8787), 16.7mM NaCl (ThermoFisher 207790010) in water, filter sterilised) with PIC. 200ul removed to act as input, stored at 4°C. Protein A/G Dynabeads (InVitrogen:100-01D/100-03D) were prepped by mixing at 1:1 and with a magnetic rack rinsed and resuspended in ChIP dilution buffer. 10ul of Dynabeads per 1ml of sample used to pre-clear samples, rotating at 4°C for 30 minutes. Supernatant saved as cleared chromatin.

The following antibodies were added:

Histone Modification	Supplier	Code	lot	Concentration	Volume used
H3K27ac	abcam	Ab4729	G83205523-I	0.3mg/ml	15ul
H3K4me1	abcam	Ab8895	GR3206285-I	0.3mg/ml	1ul
H3K4me3	Merck	07-473	2664283		1ul

Table 6.2: ChIP antibodies.

Antibodies added to cleared chromatin samples and incubated overnight rotating at 4°C. 20µl of the 1:1 rinsed Protein A/G Dynabeads added and incubated for a further 5 hours. Samples washed with 1x RIPA buffer (50mM Hepes-KOH (pH 7.6) (Sigma HO527), 500mM LiCl (Sigma 310468), 1mM EDTA (ThermoFisher AM9260G), 1% NP-40 (IGEPAL) (Sigma I8896), 0.7% Na-Deoxycholate (Sigma D6750), in water) for four washes using a magnetic stand. Then washed with TE (Sigma 93283) with 50mM NaCl. Bound chromatin was eluted from the beads by the addition of 105ul of elution buffer (1% SDS (Severn-Biotech 20-4002-02), 10mM EDTA (ThermoFisher AM9260G), 50mM Tris-HCl (pH 8) (Life Tech 15568025), in water) and 30 minutes in a thermomixer at 65°C, shaking at 1000rpm. Input sample diluted with elution buffer 1:1 and incubated at 65°C overnight. Samples were RNase (Roche 12849100) treated (5 ug/ml) at 37°C for 60 minutes, then proteinase K (BioLabs P81075) treated (200 ug/ml) at 45°C for 60 minutes. DNA fragments were purified using a Zymo ChIP Clean and Concentrator kit (Zymo Research D5205) with final elution in water.

ChIP libraries were generated from immunoprecipitated DNA fragments using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs: E7645) as per the manufacturer's instructions. Before library preparation, library concentration was quantified with Qubit dsDNA High Sensitivity Assay (Invitrogen, ThermoFisher: Q32854) to determine adaptor concentration needed and number of PCR amplification cycles required. If DNA input is < 100 ng, NEBNext Adaptor for Illumina (15 µM) diluted 10 fold in 10 mM Tris-HCl with 10 mM NaCl to a final concentration of 1.5 µM. PCR cycles based on initial input: 1 µg input, 4 cycles; 50 ng input, 7-8 cycles; 5 ng, 12 cycles. Library profiles were visually assessed on a D1000 tape on a Tapestation (Agilent Technologies) and quantified using a universal library quantification kit (KAPA Biosystems, Roche 07960140001). ChIP-seq libraries were sequenced on the Illumina Nextseq platform using a 75-cycle paired-end kit (NextSeq 500/550 High Output Kit v2.5 20024906).

ATAC and ChIP analysis

The alignment of ATAC and ChIP data from new donors introduced in this project was done manually with the following tools and flags:

bowtie2/2.4.1
samtools/1.10
trim_galore/0.6.5
deeptools

trim_galore --paired -o trim_galore --stringency 2
bowtie2 –threads 2 -x Bowtie2Index/genome
Converted to bams with samtools, sorted, indexed, rmdup
bamCoverage -b $1_mapped_filtered.bam -o $1_RPKM_normalised.bw -bs 20 -p 3 --normalizeUsing RPKM
bamCoverage -b $1_mapped_filtered.bam -o $1_RPKM.bg -of bedgraph -p 3 --normalizeUsing RPKM
bowtie2 alignment, using -F4 with samtools to only keep aligned reads
bowtie2 --threads 2 -x /t1-data/databank/igenomes/Homo_sapiens/UCSC/hg38/Sequence/Bowtie2Index/genome -1 trim_galore/$1_R1_val_1.fq.gz -2 trim_galore/$1_R2_val_2.fq.gz | samtools view -b -F4 -@ 2 | samtools sort -@ 2 -o $1_mapped_sorted.bam

Bigwigs visualised and presented with UCSC browser outputs.

Peakcalling done with LanceOtron [149] at https://lanceotron.molbiol.ox.ac.uk/

MCC

MCC libraries were generated from day 10 samples of primary CD34 differentiation cultures of seven human donors.

$1x10^7$ cells were fixed in 2% Formaldehyde (Sigma: 252549) for 10 mins, quenched with 1M glycine (Sigma: G7126) and washed in PBS. Cells were pelleted, permeabilised in 0.005% digitonin and snap frozen on dry ice. Mnase (NEB M0247) concentration was titrated individually for each sample, and then digested in a volume of 800ul at 37°C for 1 hour. MNase digestion was stopped with 5mM EGTA (Merck 324626) and washed in PBS with 5mM EGTA. A control was taken and extracted with Qiagen DNeasy Blood & Tissue Kit (Qiagen 69504) column to check digestion profile. Sample washed with PBS and ligated with ligation master mix (DNA ligase buffer 1X (Life Tech: EL0013), dNTPs 10mM each (NEB N0447), EGTA 500mM, T4 PNK 10U/ul (NEB M0201), Klenow DNA Pol 1 Large Fragment (NEB M0210L) 5U/ul, T4 DNA ligase 30U/ul (Life Tech: EL0013)) overnight at 37°C for 1 hour, 22°C for 8 hours and then kept at 5°C. Cells were washed and DNA extracted with Qiagen DNAeasy columns. Digestions and library profiles were visually assessed on a D1000 tape on a Tapestation (Agilent Technologies)

DNA was sonicated using a Covaris Sonicator ME220 for 140 seconds (70 power, 20% Duty Factor, 1000 cycles per burst, 4°C) to acquire final peak size of 180-200bp and >80% of DNA <300bp, timings could be changed to achieve correct level of sonication. Samples were cleaned with Ampure XP SPRI beads (Beckan Coulter A63881) and quantified with Libraries were generated with Qubit dsDNA Broad Range Assay (Invitrogen, ThermoFisher: Q33230). 2ug of sonicated DNA was indexed using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs: E7645) as per the manufacturer's instructions. Amplification PCR done with Agilent Herculase II Fusion Polymerase kit (Agilent: 600677) and NEB indexing reagents (NEB E7500, E7335) (NEB Universal primer 5ul, NEB unique index primer 5ul, Herculase II 5× buffer 10ul, dNTPs 0.5 µL, Herculase II polymerase 1ul) for a total of 12 PCR cycles. For each MCC sample, multiple indexing reactions were performed with different indices to increase quantity of input DNA and complexity in final pools. Libraries quantified by Qubit and visualised by tapestation.

Libraries to be captured together were pooled with 1-2ug per individual. Pooled libraries were prepped for hybridisation with KAPA HyperCapture Reagent Kit (Roche, KAPA 09075828001) reagents. As per manufacturers protocol, libraries were mixed with 5ug COT DNA per 2ug of input library and

speed dried in Savant SpeedVac SPD1030 until dry. Library resuspended in hyb mixture (13.4 µL Universal Enhancing Oligonucleotides, 28 µL hybridization buffer and 12 µL of Hybridization Component H per library), split into the appropriate amount of tubes, and denatured in a thermocycler at 95 °C for 10 min. Capture oligos (xGenTM Custom Hyb Panel-Accel 16rxn) were diluted 1:2 and 4.5 µL of diluted oligonucleotides per library heated in a thermocycler at 47°C. Denatured libraries were added to the oligos and kept at 47°C (with the lid set to 57°C) for 72 hours. Bead washes were performed with KAPA HyperCapture Reagent Kit as per protocol. 100 µl of M-270 Streptavidin Dynabeads (Invitrogen: 65305) per library were prepped with wash buffer. Hybridised library was added to the prepped Dynabeads and wash buffer I and incubated at 47°C for 45 minutes. At a consistent 47°C libraries were washed twice with stringent wash buffer. Wash buffers I, II and III washes followed at room temperature. Dynabeads resuspended in 20ul water per library and amplified with 25 ul KAPA HiFi HotStart ReadyMix and 5 ul POST-LM_PCR Oligo 1&2 for a total of 11 cycles. Captured library was hybridised and captured again for a double capture, as previously but treating the new captured library as a single library for volumes.

Libraries quantified with KAPA Library Quantification Kit (KAPA 07960140001). Sequencing was performed on the Illumina Nextseq platform using a 300-cycle paired-end kit (NextSeq® 500/550 High Output Kit v2.5 20024908).

MCC Probe designing

First probe set

Probes were designed to a range of regions, including previously known variant regions, differential ATAC regions described in Chapter 2, and loci with active erythroid genes with known enhancer interactions as controls.

Probe set 1 design was done manually, by selecting 120bp regions of interest within regulatory peaks and using the UCSC BLAT tool to test for uniqueness. Combined ATAC-seq from the 50 donors was aligned to the hg38 human genome build and visualised for peak and probe selection. Designs that passed were then order through the IDT xGenTM Custom Hyb Panel-Accel process (Table 6.3).

When designing probes over a variant, when a SNV was present a single probe matching the reference was designed. Prior tests have shown probes abilities to enrich for regions despite having variants, and sequence aligners also tolerate a degree of variation during analysis [175].

Probe name	Source	Coordinates	Sequence
PPP1R3B1	Known GWAS	chr8:9,118,075-9,218,074	TGAGTGGCTGAGCCCTATCCTTATCTGCAGCCTCTCTTCATGGCCCCCAGCCAGGACTGGCACTTGTCACATGGCATGTGAAGCAAGTGACTCTTGCAAGGAGAGAGTGTGATCAGGACG
PPP1R3B2		chr8:9,118,075-9,218,074	TGGTTAACTTGAAAACAGTTATGACTGTCCCTAAATGATAAACCAGAGATCAGTGCTGGGTCATACATTTCTTCAAATGTTGTGCCTTAAATAGCTGCAGAAGGACACAGTAAGGCAAGG
BZW2		chr7:16,585,352-16,735,351	ATAAGTGAATAGGTTGCTGAGGAAGGTGGAGGTTCATATGAAGGGAACATATTTCTTGTCCATGAGGCTGTGGGTTGTCTATAGATTAGCTCTATCTTGGCTAAGCCTGGCTTAGTTCCA
JAK2		chr9:4,704,151-5,132,309	CTCTACCACGCGCTACCAGGAGACCGTGGAAATGGTATTTAGCATCTTGGAGCCTTGGTGTCATTTGTAGATAAGGTGCGTGGTCCTGAAGTCCTGGACTTGTGGAGATGGAACGAAGCT
TFRC		chr3:196,022,148-196,448,893	TCTTTGCAGGCCCGAGACACCCACAAGAGTAACAGCAGCGCCTCATTCAGGAGGCTGAGGACTAAGTGAGCAAACGCAGGCCGAGTGCTTCCCCAGGCCTGGACCTCCATTAGCATTCAA
CCND3		chr6:41,899,786-42,093,408	CAGCCCTAACCAATTTTATCTCAGCCCTAGCACTCCAGCCCACCTCTGGGCCTTATCACCCACCGCTGGCATCCTGCTTCTTTCTCCAAATCCAGCCTCTTCCCTTAAGGCCTAGCTCCT
FAM123B		chr2:237,780,868-238,540,242	GGGAGATCAGCCAGCAGGTGGTGCAGTTGACCCTCTAAGGCACAAGGAAGCACTGAGCAGCGCTGAGGGCCCTGGGCCCCTGAACCCTTTCTCCGCCCAACCACAGGCTCCTCGGCACCG
HBA1R3		chr16:75,967-184,240	CATCTGTGCCCTTCTGTCTCCTTACTGGCAATGCTGATGCCTCCTGCCTGGGGAGATCTGGATAAGGGAGTGTCACCTGCAGGGGTGCTGTGGGTGGTGCTGACTCTTATCTCAGAGTCC
AP2	Differential ATAC peaks	chr6:79614959-79615078	AGTTTACTCCCTTAGAGGGGGTTCTGCTGTAGATAAGCACTTCTGAGTTCCCTCTCAGGGCTCAGCTTATAGGTTGCCTGCATAAGCGACTGATAAGGCACTAAGCAGAGGAAGCCAGAG
AP3		chr3:108406739-108406858	TGGAGAAAATGCGAATAGAATGTGTTGCCATCTAGTGGTACTAAGTGGTTTGAGCATTATGGCCTTCCTGTCATTCTCACCCTTGACCCACGTGATGATCGACTGAATACACAAAGATGC

AP5		chr19:19417966-19418085	GGGAAACAGGGCAGGGTGTCATGATGCAGACATCTGGGGCACTCAGCTAAAGGGCGGGCAGGAGCGGTGTCGTAGCAGTGGCAGCTGAATGACGACTGGAGGGTGACATAGGCCCTCAGA
AP6		chr4:47463645-47463764	GCTCCATCCCTGCGCGAAGCTGGGGCTTGGGTCACGTGTCAAGGCTGGCCGCTTGTCTAAAGCTACGACTCGCCCACGTGACTCTCGCCGGCGGGGAGGGGCGTGGACCCAAGAAGTCCC
AP11		chr7:157418909-157419028	GGAGCTCTGTGCTGCAGATAGAGCCCGCGGCTGCTTATGCCAGCGTCTGCTGAGGGGGTGGATGCTGCAGATAGAGCCCGCGGCTGCTTATGCCAGCGTCTGCTGAGGGGGTGGAGAGGG
AP12		chr4:76233690-76233809	GGGCTAGGAAAACATCCAGATACCAACCACAGCAGGCTCATCCCCTAAGGCAGGATTGCTGAACAAAGCCCTATCAAGCAGTTACAGGCAGTGCCCCCAATATGTAAAACCAGATGAAGG
AP14		chr20:24507634-24507753	CCTTTGGCTGCTGAAGGTCTGTGGGTGTGGGGGAGCCTTCCCTGCTAGAGATAAGAAACTGAGCAGTATCGGGTGTGAGCAAGGCATCTGAGCTAGGCCGAGGCTCCTCCTCCATTTAAC
AP15		chr22:44055956-44056075	CTTCACAGGCACGCCCAGAATAATGGTTAGCCAAGCACCAGGCCGCTCCCCACCCAGTCCAGGCAGCTAGTCAGCAGGCCCTGCCCCACTTGCTTGTCTTTCCTTGCGTTTTGCACACGC
ERFE	Regulatory elements associated with erythroid genes	chr2:238158868-238158987	GCAGCGCGCACGGGAGCCGGGAGGGGGCGCGGGCGGGCGCGCGGTCTCGGGCGAGGCGGGGGATTCAAGCGCGTTATAAGGCGAGGAGCCCGGGGCGTCCGAGACGCCGCGCTCGGAGCC
EnhDel		chr5:178730696-178730815	GCGCCGGGGACCAGGAGGGGCGGGGCAGCCGCTGAGGCCGCTGGGAAATGGAGTCCACGAGAGGACACCTGCATTCGGGGAAGTGTAGTTCCATGCGTTTGAGGCGGGGTCCCGGCTGAG
8p23L		chr8:8386444-8386563	GGGCCCGGCGGCCGGGTGGCAGCTGTGAGCGCGGCTCCGCCCCCTCCGCCTCCGGTGACGCCGCCCGCCCCAGCTCACGCCGCACCTCCGCAGTCCCTGCCGCCGCCCGGCCCCGCCTGG
8p23R		chr8:12755528-12755647	CGGCGGGGCTGCGCGCGCCCGGGAAGCCCCGAGGGCGGGGCCTGGCGAGGGCGGGGCCGGGCTCGGCGACGGCGCCCGGGACTCCCCCACGCCGCCCGCGGCCGCGCTCCCGAGGAGCCA
17q21L		chr17:46194018-46194137	GCCCCAGCACCGCGCTCCGCAGCCACCAGAGTTGGGCGGGGTGGGGAGGGTGCACGTAGTCACTGCGCAGGCCCCAGCCAGGGCCCGCGCCGGGCCAGCCCGGGAAGGGACAGCCGGGAG
17q21R		chr17:46923051-46923170	GCTAGTAAGGCGCGCGTGCGGGCAGCCCTGGCCGGCAGGGGAGGGAGGGTCCTGCGACCGGAAGCCGGAAACCGGAAGGGGGGCTGTGAGGACGTGTTCCGAGGAAGCCAGAGCCGGAGC

HBA		chr16:176611-176730	AATGAGCGCCGCCCGGCCGGGCGTGCCCCCGCGCCCCAAGCATAAACCCTGGCGCGCTCGCGGCCCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCC
HBB		chr11:5227114-5227233	GCTCCTGCCCTCCCTGCTCCTGGGAGTAGATTGGCCAACCCTAGGGTGTGGCTCCACAGGGTGAGGTCTAAGTGATGACAGCCGTACCTGTCCTTGGCTCTTCTGGCACTGGCTTAGGAG
Slc25a37		chr8:23528736-23528855	CGTCCGGTTGCTTGGAAATGTGCACGGGACGGTGACGTCACACAGGGAGGCGGGTCCCCACTGTCCCGCACACCAATGGCGCGGCGGCCGCCGGCCTCTTGGTTCCCGGCGATTGGTGTC
CD47		chr3:108091102-108091221	GCGCACGCGCACTCGGCCCCGCCCCGCAACCGCTGCTGTTGACGCGGCGCCCTCCCCAGTCGCAGGCTCCAGACCCGGGGCCCGCCTGCCCGCTCCACGGAGGCCGGATGCCTGCTCCAG
PNPO		chr17:47941422-47941541	GCCGCTGGGTTCCTCTGTTCTCGCTCTAATTGGCCTCGTTACCGAAAAGCTCCAACCAATGACCTCTCCGTGTCTGTCTCCCTCCTCCGATCCTGGAAACTCATTGGCTCCGAGGACTTA
KLF1		chr19:12887173-12887292	TCCTCTTCCTCGGCTGCCTCGTGAACTCTGAGGCTGTGATAGCCCCTTCGAGGGCTCCTCTCTGTCCTTAGCTGATTGGCTGCAGCCTCTGATAAGGCAAAGCAAGGCAAGGCGGCGGGG
ALASctcf		chr3:52220543-52220662	ACAGGCCTTCCCCTCCCGACCCTGAACCAGCCTTCTCCGCTTGGCACAGGCATGGCCACCAGAGGGCGCAATCACCCCTCCACCTTTGCTGCAATTCATTCTACCGTTTTGGAATGAACA
BIRC3		chr11:102317361-102317480	AGTCCTAAAAGGAAAGCACCAGTGCACATGCAAACCACTGGGAGGAGTGCGGAACGCCTGGTACAGATAGGGGTGGGGATTTGGGTGACGCATTTAAAAGACAGCGTGAGACTCGCGCCC
HDAC6		chrX:48801900-48802019	CAAGGGCGGAGTTTGAGAAAGGGGCTGCGTCCAATGAGTGGAGCGGTGAGTACAGCGCGTCGACGGCGGAGGCGGGAAGGGGGCGGAGGCCGACCGAAGGCGGCTAGGGACTGGCTGAAG
JUND		chr19:18281585-18281704	CGGCCGCAGCGCCCGGCCCTCCACTGGCGGCGGCTCCTGCGCCGCGCCACTTGTGCGCCCTCTTATAGCCTCGGCTCGGCGCGATGAGCTCACCGCCCCCGTTCGCCATTGGGTGCGGGT
ACTB		chr7:5530637-5530756	CGCCGCGGCCGCTCGAGCCATAAAAGGCAACTTTCGGAACGGCGCACGCTGATTGGCCCCGCGCCGCTCACTCACCGGCTTCGCCGCACAGTGCAGCATTTTTTTACCCCCTCTCCCCTC
PGKctcf		chrX:77899207-77899326	TTGAAGCTAGTTCCTTAGACGATTCGGTCGAAGATCTGTCTCTCGGCTGCCCTCTATAGAGCTCTAAACCTGGCTTCAAGGAGTTCTTTCATTCTTCATTAGATACCACCCACCATACGT

AHSPenh		chr16:31522238-31522357	CTGATAGTCGCAGAATGGCATCTCTGAGGCCAAACCATGTCTGAATTGACACTCCATGTTATCTAACTTTTTCACTCTCCTCTGCTCTGCAGTGTGGGGTGGAGTTCCCCAAAGCAGAAG
B2M		chr15:44711387-44711506	AAACGGGAAAGTCCCTCTCTCTAACCTGGCACTGCGTCGCTGGCTTGGAGACAGGTGACGGTCCCTGCGGGCCTTGTCCTGATTGGCTGGGCACGCGTTTAATATAAGTGGAGGCGTCGC
HBS1Lctcf		chr6:135075104-135075223	TCTTACTTTGTAAATGGATAATTAATTGCTCTTCACCTGACTTTTCTCACTTGTGGTTGACCTTGGCGTGACTTTTTCATGACACTAGGTGGCAGGAATAGAAGTTTTCCTTCCTTCCCT
PHF6		chrX:134373196-134373315	CATGCGCTGTGGCTGTGCGGCTTCAGCTGCCCCCTCTCTGCGGCCAATCAGGGCCCGCGGCGCGCTCGGGACGTATCAACGCTCTGTGGGTCGTGTGCGTGCGAGGGGGGCGACGTAAGG
GIN1		chr5:103120117-103120236	TCGCGCTGCCGGAAGCGGAACTACCCGGGTCCTTTAGAGGGCAGTCCTTTGGGGAACCTTGGAACCACGGTCAGAAAGGTTAGGACTGATCATTTGGCTCTTGCCGGAAATCGTTCCGCC
SLC25a5enh		chr19:58509858-58509977	CAGCCGAGGATCCCAACGACAAGTCCCATCACGTGGTACAGAGGCAGGACCGTGTAAACCACATCATCAGCTGTGGCCCCAGATAAGGACAGCATCTTGCTCATCTGCAGTACCCGCTCA
AVENenh		chr15:33956530-33956649	GTTAAAACACATCATATATATTGGAAACTAGTTCTCTGTAATGAGTCACCATCCTATCTCCCCTGCCCACACACACATAATCATTTATCTTTTGGCCTTGGTTATGGTGGGTTTTGCCAT
GPIenh		chr19:34367009-34367128	CTTTGGGGTGTGCCTTCCACAAGATTGGTTTGGTGAGCTCCTTGGAATGGTTCCCAGTGTCTGCCTGGGAGGAGCATATCTCATACCCACAGGGAGAGTTTCTCAAGGCTGACATTTTCT
PSMB2enh		chr1:35619419-35619538	TTACAGTTGTGTTACTTGAGATAGGCCACTTGGAGTCTTGAGTGCCCCATCTGTAATGTGGAGGACAGTACTGCTTGCTTGACAGAGCTGTGAGGAATTCAAGTGCTCACTCGTGTTAAA
POGZctcf		chr1:151444007-151444126	ACAGTAGCAGTTTGCAAGAAGCCCAACAGATGGCATTGTTTCGCTTGAAAATTGAAAACAATCTTGCAGTGCCCCAGTGAGTCAGCTATAGCCAGGGGCACTCTGGGTTTTAATGGGTAA
RAB7Actcf		chr3:128784368-128784487	CCATCCACCTTTACTTTCTCCCCATTCTCTTCTGCTTTCACACTGGTCTAGTGCAGTCCTCTCTTGCCTGAAGGTGGCGCAGCATCCTCACTGTCCACCGCAGGGTCTCCTACCCTCCCC
APC		chr5:112737750-112737869	GAGAGAGAAGCAGCTGTGTAATCCGCTGGATGCGGACCAGGGCGCTCCCCATTCCCGTCGGGAGCCCGCCGATTGGCTGGGTGTGGGCGCACGTGACCGACATGTGGCTGTATTGGTGCA

SNRPD3		chr22:24555902-24556021	CGGAGGCGAGACCGCGCTCAACACTGGGGAAAGGAAGGAGGCGGGCCCTGCGCGCAGCATTTTGGGAAAGAGTGGAATTCTGGGTGTTAGGCCCGCCATTCGCTTGACTCACGCCTTCGC
FAM214B		chr9:35111510-35111629	CCCTCCCCGGGGCCGCCCCGGCTCCGCGCCGCCGACTCCACTGCCCCTCTGACTCACAGGCCATTTCCTACCAGCTGATGGGGCCGCCCGCCCCGCCCAGTGCCCCACCACCCGCGCCGC
Nowhere1	Controls	chr2:23247633-23247752	GGAACTCATTGCTCTCTTAGGGCTAGTTGGAGCTCATGGATCTGAAGGTCAAGGTATCTTTTGTTCATTATGATAAAAACGTATTTATGATGTCAGGGAAAAATTATGACACCCCAAATC
Nowhere2		chr2:221887173-221887292	GGCTGAACACAGTAAAGAGTTTATTTCTTAATCCTGTTACAATCCAATTGGACATGTGGTGAGTGGTCTTCCACGTAGTGGTTCTGATATCTAACTCCACTATTTTCTAGGTACTCACAG

Table 6.3: MCC set 1 probes

Second probe set

Second probe set was comprised of: regions highlighted by the differential ATAC and H3K27ac peaks from the analysis in Chapter 2; extra variants of interest; promoters of skewed genes; and controls. Differential regions of activity were manually assessed and a list of regions of interest was curated. Additional regions of interests were provided by Ravza Gur and represent previously known variants. Skewed genes were determined using the allelic RNA data from the 50 individuals.

Second probe set design was done as previously with manual selection of regions of interest. Regions were then BLATed with the following script written by Jon Kerry and based on CapSequm [6]: design_interator_slum.sh to output the density score; a probe with a density score >30 was excluded from the final design. Designs that passed were then order through the IDT xGenTM Custom Hyb Panel-Accel process (Table 6.4).

Probe name	Source	Coordinates	Sequence
rs4332427	Ravza analysis	chr10:24569401-24569521	TGATAGGGTTTCCAGTAAGTACCCCGCCCCACCCAAATTGCTCCAGCCGGTGGCCAGAGATAAGAACTTGGAAGTGCCTCTCCCGCCTAGCAGGCGGGACACCCTGCTTGTGAGGTGGCA

rs3752531		chr12:120719988-120720108	TAATGCACAACAAGGCTGATTATGCCTATAATGGAGGCCAGTAAGTGCTGCAAGAGTAGATAGGGGAGGTGCTTTGCCGCGGCCACAAGATCCTGGCACACGGAGATGATCGAAGCTGCA
rs554122		chr11:65266343-65266463	TTTTCCATGATTCTTTAATTGGTCCCCCTGATTCTGCCCTTATCCTCTTTTGTCTCCCCTCAGTGGTGAGCTTTGTAAGGACAAGGATGCTTCTGTTTTCTCATTGCCCAAAACAGTGCC
rs10758656		chr9:4852539-4852659	TACCAGGAGACCGTGGAAATGGTATTTAGCATCTTGGAGCCTTGGTGTCATTTGTAGATAAGGTGCGTGGTCCTGAAGTCCTGGACTTGTGGAGATGGAACGAAGCTTAGATTAAGTCAG
rs330946		chr8:9164607-9164727	GCCCTATCCTTATCTGCAGCCTCTCTTCATGGCCCCCAGCCAGGACTGGCACTTGTCACATGGCATGTGAAGCAAGTGACTCTTGCAAGGAGAGAGTGTGATCAGGACGGGGGGACAGTT
rs3748136		chr8:9172590-9172710	ATGGTTAACTTGAAAACAGTTATGACTGTCCCTAAATGATAAACCAGAGATCAGTGCTGGGTCATACATTTCTTCAAATGTTGTGCCTTAAATAGCTGCAGAAGGACACAGTAAGGCAAG
rs696281		chr7:16691357-16691477	TTGCTGAGGAAGGTGGAGGTTCATATGAAGGGAACATATTTCTTGTCCATGAGGCTGTGGGTTGTCTATAGATTAGCTCTATCTTGGCTAAGCCTGGCTTAGTTCCACATGCCTTTTCAT
rs76374459		chr3:45859112-45859232	ATGATACACAACATGTAAGGCAGAAATGGGAACTAATCTCTTCCTGTTTTTAAAAATTTGCTATGCAGAGATAATTTTCATGTAATAATGAACTTAAAAGTCTGGTCTAAATATTCTACT
rs11248850		chr16:113539-113659	CATCTGCCCAAGCCAAGGGTGGAGGCATGCAGCTGTGGGGGTCTGTGAAAACACTTGAGGGAGCAGATAACTGGGCCAACCATGACTCAGTGCTTCTGGAGGCCAACAGGACTGCTGAGT
rs826415		chr1:117611295-117611415	CACAGCTGCTCAGCTAGAGCTAGCACCGCTGCTATCTCCTGTGCCCAGGCTGTTTGTCATCTTTTTACCTCCTCTGCTTCTGAGGCTGACTTTAGCTGTGGGGAGACCTGGGCTCTGGTC
rs4889599		chr16:30957208-30957328	CCCCTAGTCCCAGGCTCTGCACGCCCTGGCCCCGCCCCTTGACTCGGCCCCGCCCACAGCGGAATCCGCAGATTCGCCAGGTCGGATCCTCAGAATTCCTCGGGTCCCTCGATACTCGGC
VHL		chr3:10141648-10141768	CAACGGCCTACGGTGCTGGAGGATCCTTCTGCGCACGCGCACAGCCTCCGGCCGGCTATTTCCGCGAGCGCGTTCCATCCTCTACCGAGCGCGCGCGAAGACTACGGAGGTCGACTCGGG
JAGN1		chr3:9890497-9890617	CGCTCCACAAGAGGCGGGGCTGGGCGGGGCGGGGGCGAGAGGGAACGAGGGCGCCTGCGCGGAGAGGAGTGACCATAGAGTCGGTCCTCCAGAGACAGAGGGGCCGGAAGTTCTCTTCAC

IRAK2		chr3:10164796-10164916	TGGAGCCGGCGGGGCCGTCGTGCCGCCCGTGGCCGCTGGGGACGCTGGTGCCCTCCCGCCCCGCCCACTTCCCCGCCGGTCCGCCCCTGGCTAGAGAAGCCGCAGCCCGCAGTGTCCGAC
HBB		chr11:5227068-5227188	TGTAAGCAATAGATGGCTCTGCCCTGACTTTTATGCCCAGCCCTGGCTCCTGCCCTCCCTGCTCCTGGGAGTAGATTGGCCAACCCTAGGGTGTGGCTCCACAGGGTGAGGTCTAAGTGA
SLC25A37		chr8:23528788-23528908	GTCCCCACTGTCCCGCACACCAATGGCGCGGCGGCCGCCGGCCTCTTGGTTCCCGGCGATTGGTGTCTCTGGCCGCTGCGCTCCACCCCAGCCCGCCATGGCGTCAGGCGCCGCGGCCCC
CD47		chr3:108091079-108091199	CTGCTCCCCGCCCGCGAGAGCCGCGCACGCGCACTCGGCCCCGCCCCGCAACCGCTGCTGTTGACGCGGCGCCCTCCCCAGTCGCAGGCTCCAGACCCGGGGCCCGCCTGCCCGCTCCAC
KLF1		chr19:12887190-12887310	TCGTGAACTCTGAGGCTGTGATAGCCCCTTCGAGGGCTCCTCTCTGTCCTTAGCTGATTGGCTGCAGCCTCTGATAAGGCAAAGCAAGGCAAGGCGGCGGGGGGGCACTGTTTCTGGGGC
BCL11A		chr2:60553620-60553740	AACTTGAACGTCAGGAGTCTGGATGGACAGAGACACACAAAACATGGGCAGGGCGAGCAGGAGAGAAGGGGAGGAGGGAAGGGGAAGCTCACACCAATGGACACACATCAGGGGCTGGAC
MYB		chr6:135181182-135181302	GGCGGCAGCCAGGGCGACCGCGGAGGCGGCGGGCAGGGCGCGTGCGCACTGCAGGGGCGCCAGATTTGGCGGGAGGGGGAGTGTCCAAAGCTCTTTGTTTGATGGCATCTCTGTTTACAG
ZNF410		chr14:73886538-73886658	AACTACTCCCTCCGCTACTGCGGGGTGGCAGTGACGTCAACGAATTAGCTGACCTGTTTTTTCGGGCTTCGCGGACATAGGAGTCCCGAAGGCCGAGAGTGAGTGAGTGAGTGGGAGCGG
LRF1		chr19:4066836-4066956	CAGTAACTGTACAGTACAGCCAAAGCCCCGTATGATGCTCGCGACTCCCTTCCTGGTTTCTCCGAGGGCAAGAGGGGTCCTCCCTTGTCTTAAAGGTGCCGCAGCCGCCGGCCCCCCCAC
JI1	Jeanne analysis - differential ATAC peaks with differential H3K27Ac marks	chr5:159102472-159102592	TGCCTGAACCTCCATTTTACTCTCCAGTAAACAGTCCTCAACTGCAGGACTCCCAGCTGAGTCTGGTGGGGTGTCTCCAGTGGTGAGGGACACAAAATAAGGGAGCAGATCACTGCCAAA
JI2		chr1:75991318-75991438	CTCCCATGTTTGCTCGGCTCAGATACCTACTAAAGTGCCTGCTGCATAACAGACAATTTGCAGATTAATTGGAAAACCAGAAAAATACCCAGGATGCCTGACTTCTATTCTGGTACCCCT
JI4		chr5:31707541-31707661	CCCAGTCTGCAGTATTTCATTATGGCTAATATGAGTAATACAGTCTGTATAGAAAGATAGAGCCACACCCACGTGGAGCTACTGCAAAAGTGAACTTGCTTTTTTTATCGGGGCCTCACT

JI5		chr14:33951080-33951200	GGAGACGCTGAGGTCCGGCCCTTAAGCCGGCCGACTCCGCGCGTCACGGGCTCCGCCCCGGGACCGCCTCCGGGACCGCCCTTTCCCCCGCCCCGCTGCCAGTCACCCCCGCCTGCTCTG
JI6		chr6:29927154-29927274	GCTGACCGCGGGGGCGGGGCCAGGGTCTAACACCCTCCAGATAATGTATGGCTGCGACTTGGGGCTGGAAGGGCGCCTCCTCCGCGGGTATGAACAGCACGCCAACGATGGCAAGGATTA
JI7		chr9:129178315-129178435	TGAGAAGAGCCAATAGGAGCTCTGAGAGTCTCCTGAGTGACAGGCGCAGCCCGCCCCGGGGCTACTCGGCTGGCCAATCAGACCAAGGCGGTTCCTTTGTGCTATTCAAATACCGAACGG
JI8		chr6:168302764-168302884	ACTGAGTGAATGAAAGAACACACCCTGTGTATGGCTTTCACCAGACAAGCCCCCTTTGAAAAGGGCTGTGTCTGCTGGATAGTGGCGGGGCCCATGTGACGTGATTGAGGGCCGCGGTGT
JI9		chr7:81813169-81813289	AAGCTTATCCATTGTGCCAACATCACTCTAGGACAATCAAATATGCCATCTGCACTTGAGAGATATGCCAGCCCTCTGACTAAACAGTTCATATTCTTATCTCCTGCCGCCTCATGCCTC
JI10		chr6:79764812-79764932	AGCATTTCACATTGTTAACAGCAAATGACGCAGGGGAATATTCGGCATGAGTTTACCAAACAGTGAACATCAGAGTTTTGGATCAGGCCCAGGGGTGTGCTGCTAATGATAGTGGCAGGA
JI11		chr11:57237559-57237679	TCAGTGCTCTCCGCCCTCCTGTTCTCACCCCCTTCCATCCAATCTAAATGGGACACATTATGCAGATTGCAATCCAGCTTGAGCTGGAGCCAGGGCTGGTAGGGAGCAAGGAGGGTGTAA
JI12		chr3:38717778-38717898	CATGCCTCACCCTGCTTGCAGAAACCCTCAGCTATGCGGCAGTGAACAGATAACTCAGTTAACCATGCTGCAAAGCAAAGAAGGTAGTGGCCTAGACAGTGATAGTGCAATGTGGTGGTT
JI13		chr12:94699961-94700081	CTGCAAAGGGCTCATAAGAAATGGGAGCTATAACCCTGTGACTTCCCTTTGCAGTAAACACATGACCCTGGTAATCAGTCACTCTCCAACTTGGCTGGTCCGGTGTGGTAGAACACACCC
JI14		chr2:231604798-231604918	GGAGAGGCTGAGCCGGTTGCTGCGCGACGGCCCAGTGACGTCAGACAATTGAGTTTGGTTCCCAAGGCAACTGGCCGGGCAGCTGGGCCGGGAACTATTTGTGTCTGTCCCCAGAGGGGC
JI15		chr10:6920433-6920553	AAAACAAGTTGTACTCTTTGTCTGCGATTGTTGCTTCTAACATATGTTGGGCTCTTATCGTTTCCCAGCAGATGCAGGCACAAGGACGAAACTTGCTCACTCTCCCTATAATCAGTTAAG
JI16		chr12:63130068-63130188	AGATGGCTCACTCAGAGTTGTGTTGATGCTGGCTATTATCTGGGAACCCAGCTGGGGTTGTGGGCCAGAGTTTCTATACAGTCTCTCCGTGGTATCTTGGGTTTCTATCTGCATGGCAGC

KCND3		chr1:111920353-111920473	CACTCCCCACACTGAAGAAACACAGTCAACAGATTCTGGTTTCCTAGATACGCCAGTGGCTTCTAAAAATAGCCAGGCAATGACAAAGCAAATTGGCATATGGACTGCAAATAATTCCAG
PTPRK		chr6:128256061-128256181	AATATTTGCTGGACTAATATCTGACAGAAAGAAATACCTAATTTTTGGTCCCTGTCCTCAAGAGAGTCACAATGTGGCAGAAGAGATAAGCATGGTAGGAAATAACTATAGCACAGTTAT
XIRP2		chr2:167135768-167135888	TTTTACCTGAAGCATATTGAAAACATCCCTCATGACAGAAATCCCTCCCTCTTTCATTTCTGCCTTCTAACCCCCCTAAAAAACACATCTGTTTCCTACTGATAGCTTTTAACATACAAC
LIPI1		chr21:14063789-14063909	TCTGGTGGCCGCGAAGTGCCCTTGCCAGCACCCGATCTCCCTTCCGAGGAGAAGCAGGGCGGGCCGCAAGGCCAGACAGGCCATCCTCCTCAGGCCGGGCTGGCTGAGCGCGTGCGATTC
ZMYND6		chr20:47315832-47315952	CAACCCTCTGCCTCTCACCCATCCCCCAAAAATGAGTGGTTCTGAATGAGTTAACGTGGGTGAGGGAGGAATGCAGCCTTATGGAGAGGACAGAGCAGTGACTCAGGAAAACAGCTTGGC
GLDC		chr9:6672155-6672275	AGCAGTGAAACAAAAAGTGGCTTTAATCAAAACGGTCTTGAAGAAAACCAAGCATACAGGAAAAGACTTAAGAGAAAAATAACAAATAATCAAGCAGACTAAGCAGGGCCAGACTTAGGT
OXTR		chr3:8798536-8798656	TCCCACCCACTAACTGTGCAGATAAGAATACTACAACCTGAAAGGCCAAGATTAGGCTCATATTCTCCAGCTCTCAACGTGCATTTCATGATACCAGCACAAACTCCTCTCAGATCATTT
IL6R		chr1:154465414-154465534	GATGATAAAACACAAACGGGCTCAGCAAAAGATGCTTCTCACTGCCATGCCAGCTTATCTCAGGGGTGTGCGGCCTTTGGCTTCACGGAAGAGCCTTGCGGAAGGTTCTACGCCAGGGGA
HSPA9		chr5:138601584-138601704	AGAGAAGCAGGAAGAACCAAGAACTGAAGGCAGACTCTTGTGCTAAGTCAGAGTCCGACTCTTGTGCTGAGTCAGAGTCCCACTCTTGCATCAAGCCATGTTCCTGCAATTTGCCAGATT
FERMT1		chr20:6,261,729-6,261,849	GTGATATAGAACTGGGTTTATTTAATCTGGAGAACGTAGTAGGGAAGGATTTTATTATGGTATTTGTATTAGTCAGCTATGCCATAGCAAAGTTGGGTGACAAACCATCCCCAACTGTGA
ASIC4		chr2:219526004-219526124	TCAGCATCTCCTGGCTGCTTCTTATCTCCTGGAGGCCCAGGGCCCTCCCCCTGCCACAGCTCTGAGCCCTGGCCCCGCTAATTGCCTGTTAACCCTGTGGGCTCTGCGAGGGAGGGAGGA
CLNK		chr4:10584133-10584253	TCTCCCTGTGTGAACTCGTATTTCCCACTTTCACTTTTGCTTTTGCCCAGTTCTCCCAGCTGCTGGGTGGGACATGCTGAGTGCTGGTTTTGTACTCCCACATGCTCTATAGTGAAACTT

SPON2		chr4:1177756-1177876	CTGGGGCGCGGCACAGGCACCCAGGCCTGGCCCAGCAGCATGGAGCTCCTGGCCACATCTGCTGACACCACACCAGCTGACTCAGACGCGGAGTACCCACCCACGGGAGCGGCCGTCTCC
GATAD2A		chr19:19417998-19418118	CTGGGGCACTCAGCTAAAGGGCGGGCAGGAGCGGTGTCGTAGCAGTGGCAGCTGAATGACGACTGGAGGGTGACATAGGCCCTCAGATTGTGGATCTTTGGTTTTAGGGGTAGTGAGCTA
SPIN2B		chrX:57117658-57117778	AAAAAGCAAATGGGATCAAAGCAAGTCAAAGGGGCAGTTTTTCCAGGGTGTGGCCCCAGTTACAAGTGCAAACGCTTCCAGGGAACGACTGTGCATTTTATTAGAGTAAAGCAGGCCAGC
SLC25A48		chr5:135730098-135730218	AAAATCACCATGATTTTGGGGGAGATTACTAGAGTCCAAGTGCTATTCTTTTTCTTTATGACTATTCTCTAAATTAATTCCTATAACAAACCTCTGCTGTTACTGTTCCAATTTGATATG
UGT2B		chr4:68864563-68864683	AGTCCAAAGCCCTTGTCTTCTGTGCAACCAGTGAGCTCCGCCTACACACGGTGGCCTTTGGTGAGCACATTCCTTTTCTTTCTGGACACACTCAGGTAAAGGAACTTGCACAGGGTGCTT
CDC7		chr1:91578807-91578927	CAAAACAGCCAAGGCCGTTCCCAGGCATCACCACTTTCCTTTGCATCTTAATCTCCTGCCCTGTAAACATCTTGGCTGGGTGTGGCGGGGAGTGTTAGCATGAAGGGAAAGGTGGTAAAT
SCPEP		chr17:56982452-56982572	AGAGCTGAGGCTCAGCCCCAGGTGCGGTTATGATAATGCAGAAAAAGAACCCTTGTATGTCAGCCTGGGGAGCCCTGTTCACCTCCGGTGGCTGGAGACAAAGGTGCGGCACAATTGGCA
SHB		chr9:38089886-38090006	CTCTTCTGTAGGTGAAGGTTTAGCAGCTGCTTCTCTCACTGCTGATAACCTCTGGAAGGCCCAGCCTGGGTGGAAGCTAGACACGCCCCAAGAGTTGGCACAAAATGCCACCAGCCTGTA
RHD		chr1:25272367-25272487	ATCATAGTCCCTCTGCTTCCGTGTTAACTCCATAGAGAGGCCAGCACAACCAGCCTTGCAGCCTGAGATAAGGCCTTTGGCGGGTGTCTCCCCTATCGCTCCCTCAAGCCCTCAAGTAGG
SH3BGRL2		chr6:79614953-79615073	CTAGAAGTTTACTCCCTTAGAGGGGGTTCTGCTGTAGATAAGCACTTCTGAGTTCCCTCTCAGGGCTCAGCTTATAGGTTGCCTGCATAAGCGACTGATAAGGCACTAAGCAGAGGAAGC
SH3BGRL2_ins		chr6:79614953-79615073	CTAGAAGTTTACTCCCTTAGAGGGGGTTCTGCTGTAGATAAGCACTTCTGAGTTCCCTCTCAGGGCTCAGCTTATAGGTTGCCTGCATAAGCGACTGATACAGGCACTAAGCAGAGGAAG
SCARB2		chr4:76233695-76233815	GGAAAACATCCAGATACCAACCACAGCAGGCTCATCCCCTAAGGCAGGATTGCTGAACAAAGCCCTATCAAGCAGTTACAGGCAGTGCCCCCAATATGTAAAACCAGATGAAGGCTTCAT

KCNK1		chr1:233665682-233665802	GGGGGCAGAGCCTTCCCGGCAAATGCCAAGTGCTAATCACAGTTGACAGCAAGGCAGTGTCAGTTGAGATGGCCGTTTTGAGCTACCAATGTCCTGCTCTTTTTATGGCCACAGAGTCCT
CHFR		chr12:132835527-132835647	GGTGTCCTCGCAGGAAGAGATTAGGGCACAAACACGCATGAAGAGAAGATGACGTGAGAACTCAAGGAGATGCTGCCCACGAGCCAAGGAGACAGACCAAAGAGGAACCGGCCCCGCTGA
RMI2		chr16:11297609-11297729	CCACCACCTCCTTCTTCGCCTTGCATCGGTACGATAAGGCACTTGCTTGACGGGAAAGAGAAACTCAGCTGCCAGCTGGGGTTCATTTGCACTTTCCCCCGCCTGGTCTGCGGTCTGGCT
RMI2_ins		chr16:11297609-11297729	CCACCACCTCCCTTCTTCGCCTTGCATCGGTACGATAAGGCACTTGCTTGACGGGAAAGAGAAACTCAGCTGCCAGCTGGGGTTCATTTGCACTTTCCCCCGCCTGGTCTGCGGTCTGGC
RNVU1		chr1:143720589-143720709	AAACTTAGGACGAATCACTTTGGTAACCATAGGATTTCCCCTGCCAGATAAGTGGCAGTCCACATACCTCCGCCCTGCCACAGCTCCATATGCCTCACCCTTTACACCCACGGTTACTTG
NUP58		chr13:25421755-25421875	TTGATCTATATAGCCTGCAGTGGCCCTACGTGGTGCCAAGCAAATTGCTGCAGATGAAGCAACCGTGTGTATGTATTTGATGCCAAATTCCACGGCTTGGCAGGACTGATAAAGGCAACA
HPS3		chr3:149151836-149151956	ATATCTGTCTGGAAACTGGAATACCTTTTTCCAGAAGGATAAGCCTTCAGTAATGGTGGTTATCTTAACAAGCCAGCCCCGACTTAGTGCCTAATGTGTCTGAGATACTCTTTCACGGTC
SYNDIG1		chr20:24507643-24507763	CTGAAGGTCTGTGGGTGTGGGGGAGCCTTCCCTGCTAGAGATAAGAAACTGAGCAGTATCGGGTGTGAGCAAGGCATCTGAGCTAGGCCGAGGCTCCTCCTCCATTTAACCAGAAGAGAA
HTR4		chr5:148574252-148574372	CAGCTACAGACAGGACAGCAAATGGTTGGGTGAGAAAATGCACTGCTATTCACCCAAGACTGTTAGGCTAATCAGCAATGGAAAAAACATCTGAAAACACCTGATAAAAGGGCAAATATG
MAP2K3	Haplo Skewed genes	chr17:21284481-21284601	GGGCGGAGCCTGACCGGACCTGGGTTTTTCGGGCGGCTGGCGGACTTGACAGAGAGCGGGAGGGGCCCGGATCCTGGGCCTAAGGAGGGTGGGGCGTGTGCTTGACAGGCAGGGACAGGC
CBWD5		chr9:65675886-65676006	GCCGGGAGTGGCGTGCTGGGCGTGCGCGGCTGCGGTACGGCGTGTTGGTCCCAGCGGTTCAGCTGAGGTAGGGACGTGCCGTAGGCCGGAATGTTACCGGCTGTTGGATCTGTGGATGAG
ANKRD36		chr2:97113548-97113668	GCCGGCTAAGGTGCGCGTGCTCGCTGGTTCTAACCCTTCTGTTGGGCGTTTCTGCTGAGAGGCGGGAGGCGCTGAGAGTCTGTGCGGAGGTCCGTGGACAGACTGCTTTGCTCGTTGTTG

KCNJ12		chr17:21376320-21376440	GCGCGGGGCGGGCGTGCTCTCCGCGGAGAGTTAGAGGAGTTGCCGAGCTGAGCTCCGGTGGAGCGAGGCGCGGAGCTGGGTGCGCGCCGAGCCTCTGCCTCCCGCCGCCTCCTCGCCCTC
DUSP22		chr6:292105-292225	CCTAGTGCCTCCCACGCCCGGCGGCCGCGAGCCGGGGTCCGCGAGGGCGGAGTGGGGCGCGGCAGCCAGGAACCCGACTACGAATCCCAGGGTGCGGGCGGGCGGAGCGAGGAGGGACGC
MTCH2		chr11:47642650-47642770	GCTCGCGAGGGGGCGCTGAGGTCGGGGCGGGGCTTAAGGGACAGAGGCGCATGCGTTCTGCCGGGGGCTCGCCCTTCACCAGGAATGACGGTTTCTCTTGCACATGCGCTTCCAGGCCGC
DAPK1		chr9:87497584-87497704	GTGTCACAGGTGGGGCGCCCGCGTTCCGGGCGGACGCACTGGCTCCCCGGCCGGCGTGGGTGTGGGGCGAGTGGGTGTGTGCGGGGTGTGCGCGGTAGAGCGCGCCAGCGAGCCCGGAGC
EIF4EBP1		chr8:38030438-38030558	ACTACAAATCCCAGGGGCGTGGGGCGGGAGAGGCGGAAGGGGCGTCCTGGGGCGGGGCGGCACGGGGCGAGGCGGAGCGAGGCTGGAGGCGCGGGAGGGCAGCGAGAGGTTCGCGGGTGC
RSU1		chr10:16817602-16817722	GAGGATAGTTTATCCCGGTGGCGTCCACCCCACGGTTTGGGACGCTGATTGGGCCGAAGTTGGCGCGTGCGCAGTAGACTTGGCGCAAGCTCAGTTGATGGGTCGAAGCCCAGAGGGCTG
46_1	Extras	chr9:5040712-5040832	GCCCAACAGGTGGGCCCCTGAGCGGGCGTCCGGGACCGTGGTGTGCCAGGCGCCTTCCGCCCTTAACATGCGGTGGCTGGCCCAGGCGGTGCAGGAGCTGGAGCGGGGCCCGAGTGGGGC
RUNX1		chr21:35049281-35049401	GTTTATGAGGCCCAAAGAAGTTTTCACACAACCCAAATTACAAATTTAACTGTTCCCCTTTCCACAGCCCATCTCAATTGGTTCTTGCCAATCATGTGACTTAAGTGATGTCAATTTTTT
AP6		chr4:47463645-47463765	CTCCATCCCTGCGCGAAGCTGGGGCTTGGGTCACGTGTCAAGGCTGGCCGCTTGTCTAAAGCTACGACTCGCCCACGTGACTCTCGCCGGCGGGGAGGGGCGTGGACCCAAGAAGTCCCT
AP11		chr7:157418909-157419029	GAGCTCTGTGCTGCAGATAGAGCCCGCGGCTGCTTATGCCAGCGTCTGCTGAGGGGGTGGATGCTGCAGATAGAGCCCGCGGCTGCTTATGCCAGCGTCTGCTGAGGGGGTGGAGAGGGA
H3C1		chr6:26020352-26020472	ATAGACCAATCAGAGTGTGTAACGTCATATTTAGCGTCTTCTATCATCCAATCACTGCACTTTACACACTATAAATAGAGCAGCTCATGGGCGTATTTGCGCTAGTGTTGGGTGTTCCGC

Table 6.4: MCC set 2 probes

Third probe set

Third set of probes were designed to ATAC peaks that exhibit allelic skew in our erythroid samples in donors 1, 2, 3 and 16 (donors with MCC libraries). Probes were also selected for when containing a known GWAS SNP (associated with erythroid development and other phenotypes).

Third probe set was designed with a custom script by Nicholas Denny that further built on the design_interator_slum.sh script. (Table 6.5)

Probe name	Source	Coordinates	Sequence
rs2505887	Allelic skew	chr10:42645210-42645330	GAGCCTTATCTGTTCTCACCAGTTTCACATTCCTTGAGGCTCAGTGAGTTCCTGCGCACCTCCCTAGCACAGCTGCAAAGTTACAAGGTTGATATGCATATGTTACAGAAACATAGTTTC
rs10141865		chr14:106633843-106633963	CAGCCGTTGCAAAAGTAAATGTCAAATGTAGCTCTGAAGAGATGCACAGACATCTCCACAATCAAGGCCCAGCACAGATAGGGTGTGGTCAAATACAGGCAGAAATGCACAAAATAAAAT
rs484228		chr18:63025852-63025972	CCACAAGCACCCAGCAAGCAACTTCCCTTATCTCTCTGCAGCTTGCTGGTGGGGCATTTATCTTGGCCTATGGGTGGGGAGTGGTTATCCAAACCTCAGACTGAGTCTTCCAGTTGGGAG
rs200366671		chr2:113499918-113500038	TGCTGCACACGCCGCCCTGCACAACCCCCGCCCAGGCCCTCTGCTTGGGGCCCCTGCCCTGCCGCAGCCAGTCCCGGGGGCCTACCCCAACACCGCCCCCGGGAGACGCCCTTACGCTCT
rs11891196		chr2:227865414-227865534	TGCCCAAGATCGCATGTCTAGAAAACGGTAGGCTTCAAGTCAGGCATTCTGCTCTAATGACTCAGTGGCAAAGCTGTTTATCAGCTGTGCCAGCTGTTTATCAGCCTGCCATGTACTGGC
rs78069315		chr4:68864599-68864719	TCCGCCTACACACGGTGGCCTTTGGTGAGCACATTCCTTTTCTTTCTGGACACACTCAGGTAAAGGAACTTGCACAGGGTGCTTGCCTAAGACAGACCTGTAGCTGTATAGATAAGGAAA
rs7448500		chr5:159102424-159102544	GGGTCCCATATAGCCCTGGCAAAAGGCCTAGGCCTCCTTCTCTTTATCTGCCTGAACCTCCATTTTACTCTCCAGTAAACAGTCCTCAACTGCAGGACTCCCAGCTGAGTCTGGTGGGGT
rs330946		chr8:9164607-9164727	GCCCTATCCTTATCTGCAGCCTCTCTTCATGGCCCCCAGCCAGGACTGGCACTTGTCACATGGCATGTGAAGCAAGTGACTCTTGCAAGGAGAGAGTGTGATCAGGACGGGGGGACAGTT
rs3122724		chr9:66153188-66153308	ACCGCGGTTGCTGCCGGGCTCCCGGAAGACGCTCCCTAGAGGCGCGCGCCAGCAGGTGGGGCTGCAGCTCCGGGCGTGCGCCGATGGGCTCGCCGGTTCTCCTGGGATCGCCCGGGCGGC
rs1309687540		chrX:2630106-2630226	ATTGTGACACTTTCTGGTTCATAGTGACATTGCCACACTTCCTGGTCCATGTGGATAATAGTAGCACTTCCTGGTTCGTAGGGACCTTACCACACTTCCTGGTTCATCCAGATAATTGTA

rs2236055		chr1:11982144-11982264	GGATTTGAGCTGAGTGGACAGGGGTGTGCTCTCCTGGACCTCTGGTGGCAAGTCTCTTGAATGGCATCCTCTTTGTATAGAGCCACAGCTTTCTAACTTGTGTTGAAAAACTACCGTATC
rs4912314		chr1:56502745-56502865	TTTAAATGCTGTGTCATATTTCCCCTTCTAGACTGAGCTAGCTACTCAAGGGCATGGATCGTGTTGAGTCATCCCTGTTTTCAGAGTTCAAAAACTGTAGCTCTTAATAGAGCTCCACCA
rs666779		chr1:84944801-84944921	GTGTGGGTGTGAATGAGCCTCTGAACTGTTAGAAATAGTGTAAGTGTCCCACCAGATGTCGCCCTTGTCCTTGCCCACTTCCCCCAAAGCCAAGGCAATCTCAATCCACCAGACAGCTTC
rs1002437		chr1:88681016-88681136	ATTTAAGATTCTCTAAACACATTTGTAAATAGGGGCCCAATACAACCCATCTGAATCAACCTGATATGTTTCCATTAAAACTTTTTCAAGTGTGCTTCAGTCTGATAAGTATGATCCATG
rs72694951		chr1:150146927-150147047	TAGCAGTCGCGTCTTATCCTCAGATGTTGTTAGAGAGGTACCCCCATCATGAGGTCCTCCCAGTGTCTTATCTTCACCACAGACAGAAACTGGCTGAGAGTGATTTCTGTGGTTTATGTG
rs2229238		chr1:154465360-154465480	ACTACTTCTTCCCCAGATAGCTGGCTGGGTGGCACCAGCAGCCTGGACCCTGTGGATGATAAAACACAAACGGGCTCAGCAAAAGATGCTTCTCACTGCCATGCCAGCTTATCTCAGGGG
rs3820176		chr1:175173635-175173755	TGGGGTGTTCATCCAACTATTCATCTATCAACATGTAAAAGCCTGCCACGTGCGTCACACCCTGTGGGTGCACTCTTGATTACGATGCCTTCTGTGACCTCAAGGAGCTCATGTTTTAAT
rs7553035		chr1:211518304-211518424	TCAGCCTCTGGGCATACCATGTTTGGAGCCCAAGGGCAGTTTTCCATGCAACGAGGCACGTATCACTCTCTACCACCTTGTCACCTTCTGGGACCTGCGTGGACTCCAAATCCCTCTCTC
rs12219804		chr10:3668374-3668494	AGTTTCAACTAAATGACTAAGCAAATTAGATCTTGAGTTGCTTTTCTGGAATGCGGAGAGTTGGAGAATCAGATAAAAGTCGAGGTATCTCTCTCTTCAGCAGCTTCCAGATCGGGGGCT
rs4332427		chr10:24569401-24569521	TGATAGGGTTTCCAGTAAGTACCCCGCCCCACCCAAATTGCTCCAGCCGGTGGCCAGAGATAAGAACTTGGAAGTGCCTCTCCCGCCTAGCAGGCGGGACACCCTGCTTGTGAGGTGGCA
rs174573		chr11:61832795-61832915	CCATAAAGTTCTGGTCCGCAGGGCTGCCACCTACTGCACTGTGTGCCCTGTGCTTGCCCGTGCTGCACTTCCCAAATCCTAGGCAAAGTGCCCAGTTGTGCAGCATTCGATCAGAGATTA
rs285575		chr12:41527803-41527923	GCACAGCTGCCAGCATTTGCAGCCTGTTTACCTCTAGACTGATTCTGTCTCACTCTGAGGTTTTTCTATTTCCACCCAAATAGCCCCCAAAACACATTGAGGCTATGCCTACCTATGTGA

rs4760682		chr12:48118442-48118562	TTATCTCCTAGTCACTTGCATTTGTATTGTTTGTGCCAGAATGTGCAGAGGTAGAGATACAAGTAAGCAAGAGGAAGAGGGCAGTGGTAAAAGCAAGAGGAGACCCGACTGTGGAGACAA
ASr892		chr12:53047378-53047498	TGGCCGCGCCGCGCTGGGCGGGCGGCGGTCTCTGCGTCACGGCCGCGCGGGCAGTGCCAGGGGCGGGCGCCAGGGGGCGCGGGGCCGCCGGAGGTCGAGAGGGCGCGGGGCGCGGGGCGC
rs148015593		chr12:122085702-122085822	AACACAACCTTTGTGGAATGGTGCTTTAGGGGACATTGGCGTGGAGTGGGACAACAAGATAAGTGCTTTTGCTGGATGGTGTTCGTATCGTGGCTTGCTTGGACAGGTGTGGTGGGCAGA
rs73179655		chr13:44917943-44918063	GAACCCGGGCCGCCTGGGTGAAAACCAGGAATCCTAACCGCTAGACCATGTGGGAGACGGCAATAGCGACTCCAAGCCTAGACAAATTGAGTCTTCTCGGTCGGCTTCCGCCCACTCCAT
rs3752105		chr13:114281770-114281890	GCCGCCAACTTCGTCCTCCGGTTGCGGGGGCGGTGCGGGCAAACCTCGCGAGGAGAAGAGGACGGCCCTGAGCAAGGTGGGGACGGGGGCGGGGGCGCGCAAACCTCGCGAGGAGAGGACG
rs1760930		chr14:20486791-20486911	TGGCATGGTAATCAAAAGTTTCTTTTAAGGCTGTCACGTTTGAAAAGCTAGACCTCCAGATGGACATGTGGATTACACAGTTGGGTATATGAGTTTGGAGTCCAAGGGAAGAGGTTTAAG
rs4369588		chr14:103137099-103137219	AGGCTCTGCTCCTCCCTGGGCCTGAAAGTTCCTTCTCAGCCTGAGAGGGGGCCCTTCGGACTCAGGCATGACTCAGCCCGGCTGATGCCTCTGCAGTGCTGAGTCAGGATTTGGGGCCGG
rs66480687		chr16:87852879-87852999	TCCACGACACCCCTCACTGCCTGCTGCACTACCCTTCACTCCTCTGCTCACGAGGTTATCTGTCGCTGTCAGCAGCTTGAAATCTGCCAGTGGGAACGTCTACACCCCAGAAGTGGGCAA
rs16947455		chr18:618141-618261	TGCATGTTGGTTGTCCTGCTGGCGTTTATAGTGAGTCGCAGTTGAGAGATAACCATATTCGCTGTTTTCACGGTGAAACGTTCTCAAGGCGCTTAAACCAGGTCATCCTGACGCCAAACA
rs8092508		chr18:31681317-31681437	CACAAAGGTAGACACAGAATAAGAAATAATTTATGGCTCTGGCCTCTAGGGGTGGGCTACAGCAATCTGTCCATGGACTTCCTATGGTGTGGTGATGGGGCAGGTGTCTGCCATCCAGTT
rs897804		chr19:12766090-12766210	CCAGGCGCTCACCGGTGCTGCGTCTCCAACCCGTGGCGCGCGCGGTTGGCATCCTCCAGGTGGCGCTGCAGCTCCTCCTGCCGCTCCCGGTCGGCCGCCTCCTGCCTGCACAGCCGCTTG
rs1469712		chr19:19417952-19418072	TAGAGCCTCGTGGGGGAAACAGGGCAGGGTGTCATGATGCAGACATCTGGGGCACTCAGCTAAAGGGCGGGCAGGAGCGGTGTCGTAGCAGTGGCAGCTGAATGACGACTGGAGGGTGAC

rs10401347		chr19:48589254-48589374	AAGGGTGACTCCAAGGTTTCGGCGGAAGCAACTGTCAGGGTGGGGC AACTGGGAAGGGGAGGAGGGAGGGGAGCAGGCGGGGGAGGAAGA CACGGGCTCTGCGGCGGCCCATCCCACGTC
rs8109429		chr19:53333549-53333669	GCCTCCAAAGCTCGAAAACAAGAGGCGCTGGGAGCGAGGTTGAGAG ACGCACGGGTCCCACCCCGGCCCCGCCTTCTGCCAATTCTAAAGGGG AAGGTCTCTCCGCTTCGCGCCCCGCCC
rs2630744		chr2:8324637-8324757	CCAGCAGTCAGTTAAGGGCCCCTGCAAGGCGCTTATCTTGTTTCTGGC TGTTGGGGTAATGATGTGTCACTGACAGCTGTTGCCCACATTTGCTCC TTCACTCTCCCTTCTATGTAAACA
rs12692386		chr2:9555717-9555837	CCTCTTTTCCCTCCCGCGCCGCCTACTGGGAAGATTCTACCGCCAGGCT CGACGCCCCCAGAAGTGCAGGTGGCGTTACCAAAGGCCGGCTCAGCC AGCTTCCGGCCGCCGCTGTCCCCC
rs4953353		chr2:46340077-46340197	TCCTGTGTACACACAGCTTTTCCCAGCTGGGCTGGGGAGTCCACGTGT CCTCCGGTTACGTAAGGAGCTGGGTTATCTGGAGCCCCAGATGGGCT GCTTAGTAATCGAATCTGCTGCTCT
rs7561278		chr2:48727706-48727826	ACTCATTATTTTGGATTTCTCTGTAATTGGTGAGTAGTCTCACTTCCCT GGCTACACTGTCATCTCATGGTGGACATTTATATTTTATCTCTTCTGTA TTCCAACAGCATTTGGCCCTGT
rs55930424		chr2:54081800-54081920	CCCATGTGTGGCCACCTGGAGGGGTGACTGGCCACAGGATGATCTGG GATGGCCTCACCGTCATGACTGGCGGTTGGCTGGCTTATCAGCTGAG CTGAGGTAACAGGGATGATTAGGCCA
rs7589901		chr2:113235707-113235827	AGTCTGGGCTGGGGAGAGCCGGGGCCCACGAGGCGTGGCAAGGCG GGGAGAGAGAAGCTAGACCTCCCTGCTGCGCTTCTGCAGCGAAAATG GAGAGACCCGGAAGGCGTGTGTGCGCCC
rs6720763		chr2:167135716-167135836	GTAAACTAAGCGTTCAGATATATGTTTGCAGTTCTTGGCAAAGTTATC AACATTTTACCTGAAGCATATTGAAAACATCCCTCATGACAGAAATCC CTCCCTCTTTCATTTCTGCCTTCT
rs3749034		chr2:170816905-170817025	GTGCGAGAGGGAACTAGCGAGAACGAGGAAGCAGCTGGAGGTGAC GCCGGGCAGATTACGCCTGTCAGGGCCGAGCCGAGCGGATCGCTGG GCGCTGTGCAGAGGAAAGGCGGGAGTGCC
rs13397673		chr2:218182999-218183119	TCTTCCAATTGCCAAAACTCATGGACTCATCGCTGATGCATCTTTATCT CTTACACTCAGCAGTCAATCCACCAGAAACTGCTGCTGACTCTGCCAA AGTAAATTGCAGCTCCTGTTTTC
rs113329814		chr2:218406420-218406540	GGGCTCCGGCGCCTGGCTCTGCCGGGCGGCCTGGGTGGGGCCGGCG CCGGGGCTCGGCCCCCCCCGCCCCTCTGCGGCCTCTGAGCAGCCATTG GCCGCGCCCCCGCCCCACTTCCCGCC

rs1356334		chr2:235460774-235460894	TGAACATGATGAGTCAGCGGAGGAGATGCGCTAGATGGAATATTCCATAAACACTGGTGGGAATTACATGCTATCTAGGGACACAGCTCAGAGTCATGCTGCCAGGGGGTGAAGATAAGA
rs17726337		chr20:47315855-47315975	CCCCAAAAATGAGTGGTTCTGAATGAGTTAACGTGGGTGAGGGAGGAATGCAGCCTTATGGAGAGGACAGAGCAGTGACTCAGGAAAACAGCTTGGCATGGAGCGGAAAGCAGCAGAATT
rs2839127		chr21:46153576-46153696	CTCCCCCGCCTTTGCCGCCACCGCCCAGGGCTGGCCACTTAGAGATAAGCTTCTCATCCATCCACCCCTCAGGATCACTGTGGATCCCTGGCCTCTCATGAGCGGCCCACAGCCCAGCCC
rs2605393		chr3:36368049-36368169	TAAAATGTTGGCTCCACTGCTGATATGAACTAATTGCCTGATGCTGCCCCTCCCTTTTTTTGGTTTCAGTGCAACAGATGACCAGTATTCCTTCCCGATAAGAGACCAAGGACCACCAAG
rs4677147		chr3:72337029-72337149	TCCTGCAGAAAGAGTCCACACCCAGGAGGACACTGTGCTCGGGCTTATGCCAGGAGGCCGCCCCCTACCCCAGCAAGGGCGGCGACAGGAGGCGGAAACCCTGGGAGCAGAGGGGGAGGG
rs3804749		chr3:123114096-123114216	GGACAACTGGCCGACTTATCCTCTGTTTCAGGGTGTTATTGCTGCAGGCTGCAAACAGGCGAAACAGGAAGAGAGAGGCCACACTGTGGAGACAGATAGTAAATAAAGTTCTTGATAGAT
rs11928162		chr3:157109378-157109498	CAGGGAGTAAATAAAACCCATGGCAATGAACAAGTCACCAACAACTCAAGCATCTGAGACACAAGAGGCGAGAGATAAAATTTTCCTAGGTGTCTGCAACTTTGGAGATAAGGGCAGTGA
rs4894797		chr3:172065696-172065816	AAGTTTTGGCCTCATGCCTGGAGTAGCTCCAGATGGTCAGAAAAGATTTTTGTACTGAAGTCAGAAAGTATATTGCAGTGATAGAATTGGGCTTGTCATGTTTATATTAGAACACTGGCT
rs10559726		chr4:144119858-144119978	CTCTCCCACATCCCTCCCAGCAGCGGCTCCAGAATTTCTTCTTCTCCCCTGAAAACTAACAAGAGAGACTGCCACCAGATTACTCTTATCTCAGCACCCAGCTTTGCTTATTGTTTATCA
rs56235845		chr5:177370939-177371059	TCTCAGCCAAGCCCACCAAGCGGCTGCAGGAGGCGCTGCAGCCCATTCTGGAGAAGCACGGCTTGAGCCCGCTAGAGGTGGTGCTGCACCGGGTGAGCTTCCGGGCCGCGGGGCGGGGCG
rs390764		chr6:27814697-27814817	CCTTGGCGCGAGCTTTGCCTCCCTGCTTACCACGCCCAGACATGGCAAAAGGTCTATTACCTTTACGGTCAAGAAAGACTGAAATGAAATTGGAAAAACGTACATTTTATAGCCTCTACT
rs9366927		chr6:37059396-37059516	TGACCGGCAGGGGGAACCAGGGCATGTCCCTGGATGACTTTGGGGGTGTTCCCTGTGTCCTTGTGAATTTCCCTGAGTCACCATTACTCACTGCCCACAATCAGGGCTCTGCTACCTTCC

rs7764016		chr6:125749543-125749663	GCGCGGGGGCGGGTGAGGTGGCCCCACCCGCTCGAGGGGGGGGCGTGGCCGGCGCCGGCTCTTGCGGCCGAGCAGAGTTGCGGCGTGGGAAAGAGCCGCTAGGAGCAGACCGCGCCGCCG
rs13197384		chr6:135497699-135497819	CCGCCAGCGACCCGTGTCCTGAAAGCAAGCCGCGGCTCTGTGCCGCAGTGCGGACAGCGCGCCGCCTGCTGACGTCAGCGCCTGCCTACGCCGCCGTTCCCATGGCGACCGATGCGTCCC
rs2817441		chr6:156617505-156617625	GCCAGGGCCCCAGACAGCGCCCGCAGACCAAAGTCTACATGGCCTGCATTCCGGCAGCACTCTCTCTCATTTCCTAATTCCAAATGACTGGCAAAGAGCCAAGCCATGCAAGACTTAAGT
rs17132594		chr7:2735408-2735528	GATGCCCCAGCCAGTGCTCTTATCACCACAGGACACAAAGCCCTGGACACACGTGCTCCGCCCAGATAGTACGGACGCAAGGGGAAAAGCACCACAGAGGGCACAATCAGGACAAGCTCG
rs696281		chr7:16691357-16691477	TTGCTGAGGAAGGTGGAGGTTCATATGAAGGGAACATATTTCTTGTCCATGAGGCTGTGGGTTGTCTATAGATTAGCTCTATCTTGGCTAAGCCTGGCTTAGTTCCACATGCCTTTTCAT
rs2723514		chr7:17774864-17774984	AAGCTGACTCGTGGCAAGCTTATCTCCATGACTGCAGCCCAAAAGAAGTAAGTAACTCCATAAATGGAAGGCTATATCATTGCACTTCACAGCACCAAGACCAACGGATGGGCTCTAAAT
rs6974533		chr7:21357770-21357890	TTAGCTGTTTAAGTTGTCTTGACTTAGACGTGACAAAAATTTCTGCTCACATTCCATTGTCAAGAACTGGTTAAATGGCCCCACTAGATAAAAGAGTAAATACAGGTGCTTTTGAGTGAT
rs17323934		chr7:26864651-26864771	GGGGAGATGGGTGGGAAGGGACACGAAGGGCCTGAGGGGTCCAACTGCGCATGTGTATTCCTCGGTTTTCCCGGCCCCAGAAGAAAGAGCCTGTGGGCAAGCCACGCCCACCGCTACGCC
rs2071646		chr7:87475783-87475903	GCGAGGGCCGCGGCCCGCGAGAGCAGCTCCCGCCGCCCGCCCAGCTCTCTCCCCAGGCTAGGCGCGGGCCAGGGGCGCAGCCTGCCGCTCTTGCCGCCGGGGCGCCCGGCTGGGGCCGCC
rs35680304		chr7:131288676-131288796	GGGCCCTTCAGTAGTCAGTATCTGTTGCTTTGCACTCTAATTATCTGCTAGTATTGTGACTCTCCACTGCCCTTGCTGTTCAGTACTTTCTATTCTTTTCTCATCTCACAGCTTCTATTG
rs1618110		chr8:11809047-11809167	GCCCTTCCAGGCTCTTTGCCATCTAGTAGAGTCCCTGCGGGCCCAGCCTTTCAGAGAAGAGGGGGGAGGGGGTGATGTTTATTAACTTTTTTTAGTCTTGGCAGCTGAACCTGCCTGTGA
rs541977430		chr8:12224239-12224359	TTGAGATTGTGACATCACAGGCATCTGAACCGCGGGAGAAGGAGTTGGTGGCCATCGTGGTCTAAACTCTGATGGCCCAAATAATTCTGGACCTCCTTCCCCCAATTTTGATGATTCAGG

rs13272013		chr8:41799261-41799381	CCCCGAAGGACAGATGATCTCACTTGGCAACACATACTGGGCCAGGCTCACTCATGAGGGGGCAGCGTGGCATGCACCCCACTCAGGGGCAAGAATGAAGCTTAAATATTAGCTGACTCA
rs6996898		chr8:51870037-51870157	CTGGAGGATTAAGCACTGACAAGTTAAAGGAGGTAGGAAGCAGACCAATGGCAGACAGATGTTGCAAGGTGGGAGGAGAACAAAATGAATGCTGACTCACTTTCCTTTCCTCCTAGCACC
rs881492		chr8:100472350-100472470	CCTGAGCAGTTCCTTCTTTAAAGGCTCTGACCTGATAAAGTAGCTTTTGGCCCCATCAGCCTGCTGGCTAGGTGAGATAAAGCGATCCCTGAACCATTTGCTCCTTTTTCTCAGACCTAC
rs67811951		chr8:141242599-141242719	GCAGCTCCCGTGGTACATCGGTGAGTGCCCCAGATAACAAGCCCAACACTGACCCTGGACAAGTCGCTGAGTCAAACTCCTCCATTAGGCAGCTTGCCACTCCCTGCTCTCACTCACAGG
rs11783772		chr8:143934215-143934335	GCTTGGGTTCCCCCGCCGGGGGCGGGGCGGGGAGGGGGGTTTCCTTCCCAGGCAGTGACACACCCTCGGGTGGTTCCCCTGCCACCACAGGGCCCCACCCACCTCTCAAACTCGCTGCGG
rs10758656		chr9:4852539-4852659	TACCAGGAGACCGTGGAAATGGTATTTAGCATCTTGGAGCCTTGGTGTCATTTGTAGATAAGGTGCGTGGTCCTGAAGTCCTGGACTTGTGGAGATGGAACGAAGCTTAGATTAAGTCAG
rs1570247		chr9:35748957-35749077	TTGAGCCGACGATTAGCTCGCCCGGCCAAAGGGCGACTGGGCCCGCAGAGAGGGGAGGAGCCGCGGGGCCAGCCCACCAGGCACCGCCCCGGGACGGGCCCGCAAGGCACCGCCCCCGGG
rs3758354		chr9:73149619-73149739	TCTGTTTGTAAAGTAAGTGGAGCCGATGAATCATCATCTTCTAGAAACAGCAAGAGTGGTGAGTCATTTCTCTGGAGTATGTGGCATTTCACCCACAGAAAGCATTTAACCCATGATTAT
rs67545317		chrX:72181317-72181437	TTGTCTGGCGGCCCCGCGGGATTGGCCAGAGCTGGCGCCGAGGGACTGCAGAGCCACCCACCCCCGTGGCTCCTACTGCGCACGCGTGAAAAAAGCCAGGTTGAGGGGAACGGGGGAGCT
rs2236054	Neutrals	chr1:11980190-11980309	GGGAGTCCGAGCCTCTGCGTCGTCCGCTTGGGACGCGCCGGCGGAGGAGTGGCGCGCGGAGGAGTGGCGCGCTGAGACGCCGCTCGAAGCGCCGAGTCGCGGGGCAGCAGAGGCGTAAGG
rs111798358		chr1:54764620-54764739	GCCCCACGCTAGGCGCGCGTGAGAGGTTTTCCGGAAGTGTAGCCAAAGAGCGAGTACCCGGCTTTCCCCAAGATCGTTGATTGGCCAGGGCAGCCTTGAGTAAATCACCGTTCCGGAAGT
rs41279886		chr10:25016470-25016589	CGGCTGAGACCACGAGGACGCAAGGCCAGCGCCGCCAGGGAGACGAGGGAGGGGAGGAGCCTGAGGAGATCGGAGTGCTCGGATAGGCTACAGTGAGAAAGGAGAGTCACGTGCCCTGCG

rs9325444		chr10:89701435-89701554	GCGGCCCACGCCTGCTTCCGCCCATTGGCTGCCCGCGTTTCAAGACCCGCCCCCATATTGGAGAGCTTCTTCCGGTAGCCACTGGAAAGCCCGCAGCGCGAGCCGGCTACGCTCCCTGGC
rs12291986		chr11:95789848-95789967	TCCTCCCAACTCCCAACTTCCGCGACGCAGGAAGCGCCCACAAAGGGAGGATGTCGCCGGCCAATGACATCCCGGAGCTCTCAGCGCGCGAAATCCGCTTTGAACCTGCCCCGCCCAAAG
rs11248850	Extras	chr16:113599-113659	CATCTGCCCAAGCCAAGGGTGGAGGCATGCAGCTGTGGGGGTCTGTGAAAACACTTGAGGGAGCAGATAACTGGGCCAACCATGACTCAGTGCTTCTGGAGGCCAACAGGACTGCTGAGT
FAM123B		chr2:238160486-238160605	GGGAGATCAGCCAGCAGGTGGTGCAGTTGACCCTCTAAGGCACAAGGAAGCACTGAGCAGCGCTGAGGGCCCTGGGCCCCTGAACCCTTTCTCCGCCCAACCACAGGCTCCTCGGCACCG
TFRC		chr3:196192041-196192160	TCTTTGCAGGCCCGAGACACCCACAAGAGTAACAGCAGCGCCTCATTCAGGAGGCTGAGGACTAAGTGAGCAAACGCAGGCCGAGTGCTTCCCCAGGCCTGGACCTCCATTAGCATTCAA
SCARB2		chr4:76215136-76215255	AAATATCAGGGTATGGATTTCAGCAGAGCTCAGCAGTTGAGGCGGAGCCATTGGTCAGTTATGCTATCTACACTGCAAGATCCAAGAAGTACATTTTCTTGGCAACTATAGTGTATTATC
CCND3		chr6:41957357-41957476	CAGCCCTAACCAATTTTATCTCAGCCCTAGCACTCCAGCCCACCTCTGGGCCTTATCACCCACCGCTGGCATCCTGCTTCTTTCTCCAAATCCAGCCTCTTCCCTTAAGGCCTAGCTCCT
BZW2		chr7:16691345-16691464	ATAAGTGAATAGGTTGCTGAGGAAGGTGGAGGTTCATATGAAGGGAACATATTTCTTGTCCATGAGGCTGTGGGTTGTCTATAGATTAGCTCTATCTTGGCTAAGCCTGGCTTAGTTCCA
PPP1R3B		chr8:9150736-9150855	TCGCACCCGGGCGGGGCCGGCAAAGGGGCAACCCCGGCCGGAGCAAGCAGCGCGACACCAGCACCCCCGCGCGGCCGCCTGGCCTCCGCCAGCCCGGGACGGGGATCGTGCTAGGTCCCG

Table 6.5: MCC set 3 probes.

Low cell MCC

For low cell MNase titration, digestion was done with 10% volumes and all material was extracted for digestion visualisation to determine optimum digestion concentration. An aliquot of $2x10^6$ cells was fixed as per standard MCC protocol in a smaller volume. Six digestions each of $0.3x10^6$ were then

digested in a volume of 80ul with a range of MNase concentrations. Concentrations reflected a 10x dilution.

Libraries generated

Bio reps	Tech fix reps		Digestion rep	Library ident			Indexed Round 1				Indexed Round 2		
Donor	Cells date	Library batch	Mnase U		index	name	ug/ml	ul	ug		ug/ml	ul	ug
Don2	Dec-21	7	5	A	1	Don27A1	116	30	3.5		174	25	4.4
		7	10	B	2	Don27B2	108	30	3.2		138	25	3.5
	Dec-21	8	5	C	3	Don28C3	103	30	3.1		133	25	3.3
		8	5	D	4	Don28D4	111	30	3.3		135	25	3.4
	Dec-21	9	5	E	5	Don29E5	111	30	3.3		79.3	25	2.0
		9	5	F	6	Don29F6	108	30	3.2		151	25	3.8
Don3	Dec-21	7	5	A	7	Don37A7	123	30	3.7		133	25	3.3
		7	10	B	8	Don37B8	116	30	3.5		144	25	3.6
	Dec-21	8	5	C	9	Don38C9	119	30	3.6		130	25	3.3
		8	5	D	10	Don38D10	98.3	30	2.9		144	25	3.6
	Dec-21	9	5	E	11	Don39E11	114	30	3.4		141	25	3.5
		9	5	F	12	Don39F12	127	30	3.8		156	25	3.9
Don16	Nov-22	9	15	A	13	Don169A13	122	30	3.7		171	25	4.3
		9	17.5	B	14	Don169B14	137	30	4.1		157	25	3.9
		9	20	C	15	Don169C15	101	30	3.0		120	25	3.0
Don54	Aug-21	5	5	A	16	Don545A16	121	30	3.6		43.7	25	1.1
		5	5	B	27	Don545B27	164	30	4.9		145	25	3.6
		5	10	C	18	Don545C18	85.1	30	2.6		146	25	3.7
Don55	Jul-21	5	5	A	19	Don555A19	118	30	3.5		110	25	2.8
		5	10	B	20	Don555B20	148	30	4.4		128	25	3.2
	Jul-21	6	8 col	C	21	Don556C21	141	30	4.2		127	25	3.2
		6	8 phe	D	22	Don556D22	46.2	30	1.4		50.5	25	1.3
	Jul-21	9	5	E	23	Don559E23	115	30	3.5		127	25	3.2
		9	5	F	25	Don559F25	159	30	4.8		113	25	2.8
Don1	Feb-23	10	5	A	13	Don110A13	158	30	4.7		148	30	4.4
	Feb-23	11	5	B	14	Don111B14	182	30	5.5		137	30	4.1
		11	5	C	15	Don111C15	114	30	3.4		114	30	3.4
	Feb-23	12	5	D	16	Don112D16	151	30	4.5		140	30	4.2
		12	5	E	27	Don112E27	155	30	4.7		138	30	4.1
		12	5	F	18	Don112F18	81.7	30	2.5		75.3	30	2.3

Table 6.6: MCC libraries generated

	MCC Probe Set		
Donor	1	2	3
Cone	X	X	
Don1		X	X
Don2		X	X
Don3		X	X
Don16		X	
Don54		X	
Don55		X	

Co-Captured

Table 6.7: MCC capture experiments and pools

MCC optimisations

The main issue I was facing was a low yield from the digests, and the ligation stage (figure). I first checked the permeabilisation by checking viability on a cell counter and by assessing the cells visually. To try and remedy this problem I tried: increasing the spin speeds to prevent pellet loss, comparing the outputs from column and phenol based extractions and finally reducing the shaking. Altering the spin speeds had no effect. Phenol and column extractions were comparable in DNA yield and fragment distribution. Not shaking the digestions or ligations improved the yields and was used for subsequent library preparations.

MCC analysis

MCC data were analysed using a published pipeline[100] aligned to hg38 and visualised using the UCSC Genome Browser.

MCC allelic scripts

MCC allelic scripts provided by Joseph Hamley.

MCC_SNP_splitter_runner_2.py
MCC_Wig_combiner_runner.pl
MCC_Wig_combiner_track_hub_generator.pl

All MCC data visualised and presented with UCSC.

scMultiome

Single Cell Multiome preparations followed 10X resources Chromium Next GEM Single Cell Multiome ATAC + Gene Expression User Guide (CG000338 Rev B) and Nuclei Isolation for Single Cell Multiome ATAC + Gene Expression Sequencing demonstrated protocol (CG000365 Rev B and Rev C).

scMultiome libraries were generated from primary CD34 differentiation cultures of two human donors at days 0, 5 and 10. Data from PBMCs of the same donors were also generated (optimisation for this project, data used for other projects).

Nuclei Isolation

Prior to washes, PBMCs were FACS sorted on forward and side scatter to remove granulocytes and NETs as per FACS plot (Figure 6.1).

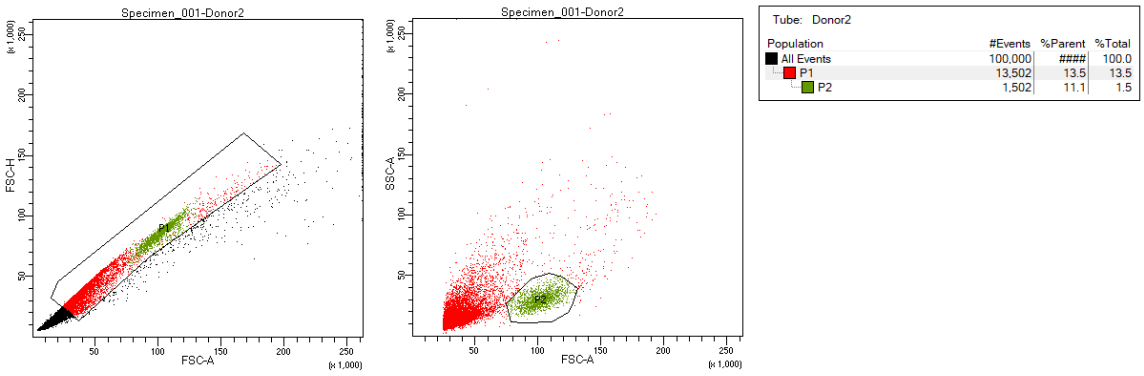


Figure 6.1: FACS gating for removal of granulocytes and NETs prior to PBMC lysis.

Data generated from days 5 and 10 used the standard cell number protocol. CD34 cells were isolated and washed in 1ml PBS with 0.04% Bovine Serum Albumin (MACS BSA Stock Solution) (Miltenyi Biotec 130-091-376) for a total of 2 washes. Cell were then counted on the Luna counter (Luna FL Automated Fluorescence Cell Counter). For nuclei isolation cells were washed twice in PBS with 0.04% BSA and then resuspended in 100ul of chilled Lysis buffer (Tris-HCl (pH 7.4) 10mM (Sigma T2194), NaCl 10mM (Sigma 59222C), $MgCl_2$ 3mM (Sigma M1028), Tween-20 0.1% (Bio-Rad 1662404), IGEPAL CA-630 0.1% (Sigma 74385), Digitonin 1% (Thermo Fisher BN2006), BSA 1% (Miltenyi Biotec 130-091-376), DTT 1mM (Sigma 646563), Rnase inhibitor 1U/ul (Sigma 3335402001), in water at 4°C) for 5 minutes. Lysis concentration was used at 1x, with 0.1x and 0.01x also being used for optimisations. Lysis times were tested from 2 to 6 minutes with 5 minutes being optimal. Lysis efficiency was determined using AO/PI stain on the Luna counter, with 'viability' being used a proxy for lysis. Nuclei quality was assessed by eye under the microscope (Figure 6.2). Nuclei were then washed three times with cold wash buffer

(Tris-HCl (pH 7.4) 10mM (Sigma T2194), NaCl 10mM (Sigma 59222C), $MgCl_2$ 3mM (Sigma M1028), Tween-20 0.1% (Bio-Rad 1662404), BSA 1% (Miltenyi Biotec 130-091-376), DTT 1mM (Sigma 646563), Rnase inhibitor 1U/ul (Sigma 3335402001), in water at 4°C). Based on prior cell concentrations and assuming ~50% nuclei loss during cell lysis, nuclei were diluted in diluted nuclei buffer (Nuclei Buffer 1X (10x Genomics 2000153/2000207), DTT 1mM (Sigma 646563), Rnase inhibitor 1U/ul (Sigma 3335402001), in water at 4°C) for a final concentration that matched cell requirements.

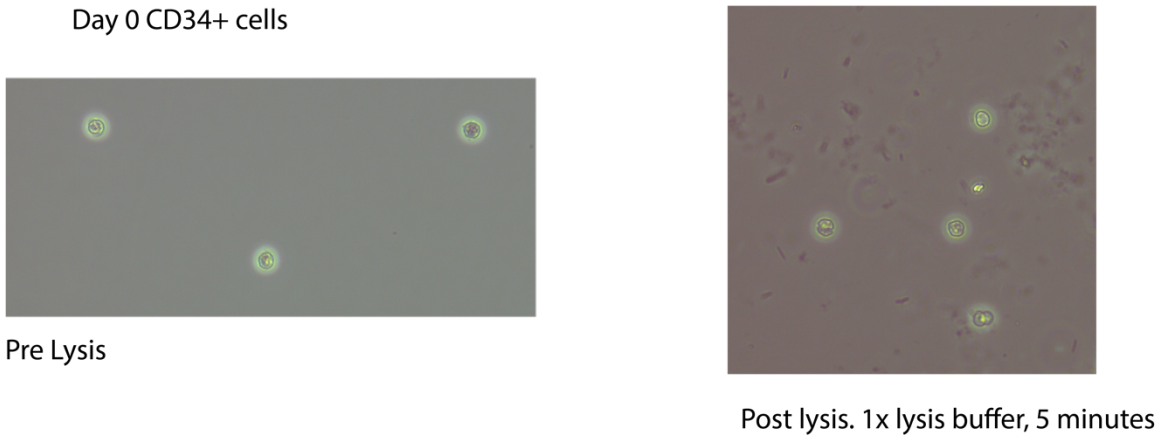


Figure 6.2: Cell and nuclei integrity during scMultiome lysis optimisations.

An atypical treatment with ACD-A Anticoagulant Citrate Dextrose of the PBMCs prior to receiving the cells for the second round of lysis optimisations may have affected their integrity and therefore the lysis. It was therefore concluded that during optimisations it was important to make sure the cells were treated in the exact same way they would be treated for the real experiment, to ensure parity.

Data generated from PBMCs used the low cell input protocol. Initial counts and washes as described for standard protocol. Post counting, cells were washed in PBS with 0.04% BSA and resuspended to a volume of 50ul in PBS with 0.04% BSA. 50ul of wash buffer (Tris-HCl (pH 7.4) 10mM (Sigma T2194), NaCl 10mM (Sigma 59222C), $MgCl_2$ 3mM (Sigma M1028), Tween-20 0.1% (Bio-Rad 1662404), BSA 1% (Miltenyi Biotec 130-091-376), DTT 1mM (Sigma 646563), Rnase inhibitor 1U/ul (Sigma 3335402001), in water at 4°C) was added and the cells spun. Supernatant was removed, leaving a small amount behind as not to disrupt the cell pellet. Cells lysed in 45ul of cold lysis buffer buffer (Tris-HCl (pH 7.4) 10mM (Sigma T2194), NaCl 10mM (Sigma 59222C), $MgCl_2$ 3mM (Sigma M1028), Tween-20 0.1% (Bio-

Rad 1662404), IGEPAL CA-630 0.1% (Sigma 74385), Digitonin 1% (Thermo Fisher BN2006), BSA 1% (Miltenyi Biotec 130-091-376), DTT 1mM (Sigma 646563), Rnase inhibitor 1U/ul (Sigma 3335402001), in water at 4°C) for 5 minutes. 50ul of wash buffer was added and nuclei spun for a total of one wash. Supernatant was carefully removed and pellet was resuspended in diluted nuclei buffer (Nuclei Buffer 1X (10x Genomics 2000153/2000207), DTT 1mM (Sigma 646563), Rnase inhibitor 1U/ul (Sigma 3335402001), in water at 4°C) for a final concentration that matched cell requirements.

Indexing

Indexing of all samples was done according to the 10X resources Chromium Next GEM Single Cell Multiome ATAC + Gene Expression User Guide (CG000338 Rev B). Reagents were all sourced from Chromium Next GEM Single Cell Multiome ATAC + Gene Expression Reagent Bundle (10x 1000285) and GEM formation was done using Chromium Next GEM Chip J Single Cell Kit (10x 1000230). Day 5 and day 10 samples were processed by Neil Ashley of the WIMM single cell facility. PBMCs and day 0 samples were processed by Christina Simoglou Karali (Adam Mead lab).

Nuclei were mixed with 10ul transposition mix (ATAC Buffer B 2000193 and ATAC Enzyme B 2000265/2000272) and incubated for 1 hour at 37°C. GEMs were then generated using a chromium Next GEM Chip J in a Chromium controller as per the protocol. GEM incubation was cleaned and DNA eluted. Pre-amplification PCR was done as per the protocol and product was split for use in ATAC indexing and cDNA generation.

ATAC indexing as follows:

	ATAC	final indexed library mass(ng/ul)	PCR cycles used	Plate N Set A indices used				
Don002 Day 5	ATAC_1	47	7	SI-NA-H1	AAACTCAT	CGGGAGTA	GTCACAGG	TCTTGTCC
Don003 Day 5	ATAC_2	9.1	9	SI-NA-H2	AACGGTCA	CCGAACTC	GGTCCAAG	TTATTGGT
Don002 Day 10	ATAC_7	2.68	7	SI-NA-H7	AACGGGTG	CTAATTCT	GCTTCAAC	TGGCACGA
Don003 Day 10	ATAC_8	3.23	7	SI-NA-H8	AAGAGCGG	CTTGTTAT	GGCCCATC	TCATAGCA

Table 6.8: scMultiome ATAC indexing for donors 2 and 3, days 5 and 10.

Final libraries ready for sequencing. Libraries eluted in Qiagen Buffer EB (Cat.No.19086)											
Sample ID		**Library type**	**Average Library size (bp)** *as determined using the Agilent Bioanalyzer 2100 High Sensitivity DNA Assay*	**Concentration (ng)**	**i7 Index (Orientation: P5-P7', 5'-3')** *as stated on the 10x Chromium Shared Sample Indexes Plate file*		**index2_workflow_a(i5)**	**index2_workflow_b(i5)**	**Indexing kit & index used**	**Number of cycles used for indexing PCR** (see pg.49 for ATAC libraries and pg.64 for GEX libraries on CG000338	**Volume left** *(1µl of original product used for QC)*
Sample 1(a,c)	**D2 PBMC**	Chromium Single Cell Multiome ATAC Library	389	0.62	SI-NA-A1a	AAACGGCG	P5 Unknown (Single Index)	P5 Unknown (Single Index)	1000084: Chromium i7 Multiplex Kit N, SI-NA-A1	7 for 1a, 9 for 1c	19µl
					SI-NA-A1b	CCTACCAT					
					SI-NA-A1c	GGCGTTTC					
					SI-NA-A1d	TTGTAAGA					
Sample 2	**D3 PBMC**	Chromium Single Cell Multiome ATAC Library	460	121.08	SI-NA-A2a	AGCCCTTT	P5 Unknown (Single Index)	P5 Unknown (Single Index)	1000084: Chromium i7 Multiplex Kit N, SI-NA-A2	7	19µl
					SI-NA-A2b	CAAGTCCA					
					SI-NA-A2c	GTGAGAAG					
					SI-NA-A2d	TCTTAGGC					
Sample 3	**D2 CD34+**	Chromium Single Cell Multiome ATAC Library	498	1,884.85	SI-NA-A3a	AAAGCATA	P5 Unknown (Single Index)	P5 Unknown (Single Index)	1000084: Chromium i7 Multiplex Kit N, SI-NA-A3	7	19µl
					SI-NA-A3b	CTGCAGCC					
					SI-NA-A3c	GCCTTTAT					
					SI-NA-A3d	TGTAGCGG					
Sample 4(a,c)	**D3 CD34+**	Chromium Single Cell Multiome ATAC Library	440	52.59	SI-NA-A4a	AGAACGCC	P5 Unknown (Single Index)	P5 Unknown (Single Index)	1000084: Chromium i7 Multiplex Kit N, SI-NA-A4	7 for 4a, 9 for 4c	19µl
					SI-NA-A4b	CATGGCAG					
					SI-NA-A4c	GTCTTTGA					
					SI-NA-A4d	TCGCAATT					

Table 6.9: scMultiome ATAC indexing for donors 2 and 3, CD34 day 0 and PBMCs.

cDNA was amplified as per the protocol and library construction and indexing was performed as follows:

	cDNA	cDNA mass (ng/ul)	PCR cycles	Final indexed GEX cDNA mass (ng/ul)	Final indexed size (bp)	DUAL index plate TT set A indices used			
Don002 Day 5	cDNA_1	88	7	35	398	SI-TT-H1	ACAATGTGAA	CGTACCGTTA	TAACGGTACG
Don003 Day 5	cDNA_2	19	9	7.5	450	SI-TT-H2	TAGCATAGTG	CGGCTCTGTC	GACAGAGCCG
Don002 Day 10	cDNA_7	5.2	7	14	497	SI-TT-H7	ACCTCGAGCT	TGTGTTCGAT	ATCGAACACA
Don003 Day 10	cDNA_8	6.7	7	45	483	SI-TT-H8	ATAAGGATAC	ATAGATAGGG	CCCTATCTAT

Table 6.10: scMultiome GEX indexing for donors 2 and 3, days 5 and 10.

Final libraries ready for sequencing. Libraries eluted in Qiagen Buffer EB (Cat.No.19086)													cDNA libraries			
Sample ID		**Library type**	**Average Library size (bp)**	**Conc (ng)**	**i7 Index (Orientation: P5-P7', 5'-3')** *as stated on the 10x Chromium Shared Sample Indexes Plate file*		**index2_workflow_a(i5)**	**index2_workflow_b(i5)**	**Indexing kit & index used**	**Number of cycles used for indexing PCR** (see pg.49 for ATAC libraries and pg.64 for GEX libraries on CG000338 protocol)	**Volume left** *(1µl of original product used for QC)*	**Number of cycles used for cDNA amplification** *(see pg.49 on CG000338 protocol)*	**Average Library size (bp)**	**Conc (ng)**	**Volume left** *(25% of original product used for indexing plus 1µl for QC)* - to be used within 3months
			as determined using the Agilent Bioanalyzer 2100 High Sensitivity DNA Assay										*as determined using the Agilent Bioanalyzer 2100 High Sensitivity DNA Assay*			
Sample 1a	**D2 PBMC**	Chromium Single Cell Multiome Gene Expression Library 1a	461	11.91	SI-GA-C10a	GTAACATGCG	AGTGTTACCT	AGGTAACACT	1000215: Chromium Dual Index Kit TT, SI-TT-A1	14	34µl	7	766	21.92	None
Sample 2	**D3 PBMC**	Chromium Single Cell Multiome ATAC Library	497	12.03	SI-GA-C11a	GTGGATCAAA	GCCAACCCTG	CAGGGTTGGC	1000215: Chromium Dual Index Kit TT, SI-TT-A2	14	34µl	7	738	44.32	80µl
Sample 3	**D2 CD34+**	Chromium Single Cell Multiome Gene Expression Library	511	10.81	SI-GA-C12a	CACTACGAAA	TTAGACTGAT	ATCAGTCTAA	1000215: Chromium Dual Index Kit TT, SI-TT-A3	14	34µl	7	694	33.13	80µl
Sample 4a	**D3 CD34+**	Chromium Single Cell Multiome Gene Expression Library 4a	363	10.81	SI-GA-C12a	CTCTAGCGAG	TATCTTCATC	GATGAAGATA	1000215: Chromium Dual Index Kit TT, SI-TT-A4	14	34µl	7	712	7.77	None
Sample 1b	**D2 PBMC**	Chromium Single Cell Multiome Gene Expression Library 1b	365	11.91	SI-GA-C10a	GTAACATGCG	AGTGTTACCT	AGGTAACACT	1000215: Chromium Dual Index Kit TT, SI-TT-A5	16	34µl				
Sample 4b	**D3 CD34+**	Chromium Single Cell Multiome Gene Expression Library 4b	404	10.81	SI-GA-C12a	CTCTAGCGAG	TATCTTCATC	GATGAAGATA	1000215: Chromium Dual Index Kit TT, SI-TT-A6	16	34µl				

Table 6.11: scMultiome GEX indexing for donors 2 and 3, CD34 day 0 and PBMCs.

Shallow sequencing was performed in house to determine library quality and complexity. Sequencing requirements fulfilled as per protocol on with NextSeq 500/550 Mid Output Kit v2.5 (150 cycles) (Illumina 20024904)

ATAC libraries were sequenced as follows: Read 1N, 50 cycles; i7 index, 8 cycles; i5 index, 24 cycles; Read 2, 49 cycles. Custom run was required for i5 read and was ran as 8 dark cycles and 16 nt cycles.

GEX libraries were sequenced as follows: Read 1, 28 cycles; i7 index, 10 cycles; i5 index, 10 cycles; Read 2, 90 cycles.

Deeper sequencing was performed in a NovaSeq instrument (PE-50 and PE-150) by Novogene (Novogene (UK) Company Limited, 25 Cambridge Science Park, Milton Road, Cambridge, CB4 0FW, United Kingdom).

scMultiome analysis

Raw reads were processed with 10x Genomics Cell Ranger 7.1.0 [195, 219] ARC using default settings and hg38 as the reference genome. 10x Genomics Loupe Browser 6.5.0 software was used for initial clustering and gene expression patterning.

Further analysis including clustering and trajectory generation was performed with ArchR [197] by Dr Ravza Gur (Hughes Lab).

The cellranger-arc count pipeline (10X Genomics, v2.0.0) with default parameters was used to analyse scATAC-seq data, and the human genome hg38 was used as the reference genome. 10X pipeline generates a sorted and indexed BAM file. Peak calling was then performed via LanceOtron [149], called peaks were filtered by peak score (> 0.5), and ENCODE blacklisted regions [220] were excluded from the peak file. Finally, the coverage track was generated using deepTools BamCoverage. Data were visualised on the UCSC genome browser.

After preprocessing the data using cellranger-arc count pipeline, the ArchR tool was used to analyse the data more comprehensively [197]. ArchR analysis used a fragment file resulting from the 10X count pipeline as an input to create an arrow file with the following parameters:

- filterTSS = 4;
- filterFrags = 1000;
- addTileMat = TRUE;
- addGeneScoreMat = TRUE.

The ArchR project was created after calculating and removing doublets by filtering generated doublet scores (using the addDoubletScores function with default parameters). Workflow was as follows.

1. The iterative LSI dimensionality reduction was implemented using the addIterativeLSI function on the tile matrix.
2. UMAP embedding was added via the addUMAP function.

3. Seurat's graph clustering with default parameters were used to call clusters via addClusters() function.
4. The plotEmbedding function was then used to visualise data in 2D space and
5. Cells were coloured by cluster information.
6. The addClusters function was used to form clusters.
7. Gene Scores matrix was imputed using MAGIC in ArchR.
8. The plotEmbedding function was then used to visualise data in 2D space and cells were coloured by gene activity scores of genes of interest: *CD34, PTPRC, NKX2-5, GATA1, KLF1, NFE2, TAL1, HBA1, HBA2, HBB, SLC25A37, GYPA, EPB42.*

For a full explanation of how the ArchR tool analyses scATAC-seq data, see their website, https://www.archrproject.com/bookdown/index.html.

There is no need for data integration since gene expression (nuclear) and chromatin information are from the same cells.

DeepHaem analysis ran by Simone Riva (https://github.com/rschwess/deepHaem)

Enrichment analysis "Cell Type Prioritisation Platform" in Figure 4.20 used the following traits from the associated GWAS, ran by Simone Riva. GWAS catalogues from van der Harst *et al*. [71], Astle *et al*. [70] and Vuckovic *et al*. [72]

van de Harst 2012	Astle 2016	Vuckovic 2020
Red blood cell traits	Mean Corpuscular volume	Mean corpuscular volume
	Mean corpuscular haemoglobin	Mean corpuscular haemoglobin
	Reticulocyte fraction of red cells	Reticulocyte fraction of red cells
	Red cell distribution width	Red cell distribution width
	Reticulocyte count	Reticulocyte count
	Red blood cell count	Red blood cell count
	High light scatter reticulocyte count	High light scatter reticulocyte count
	High light scatter reticulocyte percentage of red cells	High light scatter reticulocyte percentage of red cells
	Immature fraction of reticulocyte	Immature fraction of reticulocytes
	Heamoglobin concentration	Haemoglobin
	Hematocrit	Hematocrit
	Mean corpuscular haemoglobin concentration	Mean corpuscular haemoglobin concentration
		Mean reticulocyte volume
		Mean spherical Corpuscular volume

Table 6.12: GWAS traits used in enrichment analysis.

Data analysis and bioinformatics

The following tools and software packages were used throughout data analyses and visualisation: Python, Jupyter, ucsctools [13], bedtools [150], samtools [174], Bowtie2, deeptools [221], trim galore and the UCSC genome browser [13].

Differential peak analysis

Initial peak comparisons done with following python commands in Jupyter

```
import numpy as np
import pandas as pd
Multicov = pd.read_csv('atac_coverage.bed', sep='\t')
Peaks=Multicov[['Don001','Don002','Don003']]minPeak = Peaks.min(axis=1)
maxPeak = Peaks.max(axis=1)
Multicov['minPeak']=minPeak
Multicov['maxPeak']=maxPeak
peakDiff=Multicov['maxPeak'] - Multicov['minPeak']
Multicov['peakDiff']=peakDiff
percentDecrease=(100 - ((Multicov['minPeak'] / Multicov['maxPeak']) * 100))
Multicov['%decrease']=percentDecrease
Multicov.to_csv('/Users/mgosden/Documents/Bioinformatics/MACS/Multicov/Multicov_comparison.bed',
sep='\t', index=False)
```

Dendrograms

Dendrograms were generated by Edward Sanders with Python

- Biopython package modules, TreeConstruction models

Distance matrix computed in a simple way based on number of different variants between haplotype pairs. Trees made with UPGMA method to construct rooted dendrograms. Tree branches coloured based on consistent ATAC skew for each haplotype using common ancestors. No inference based on branch length, events not mapped based on evolutionary context. Branches instead show how a single haplotype can be present in multiple donors, and how many confident haplotypes there are. Branch distance depicts the acquisition of SNPs.

Motif analysis

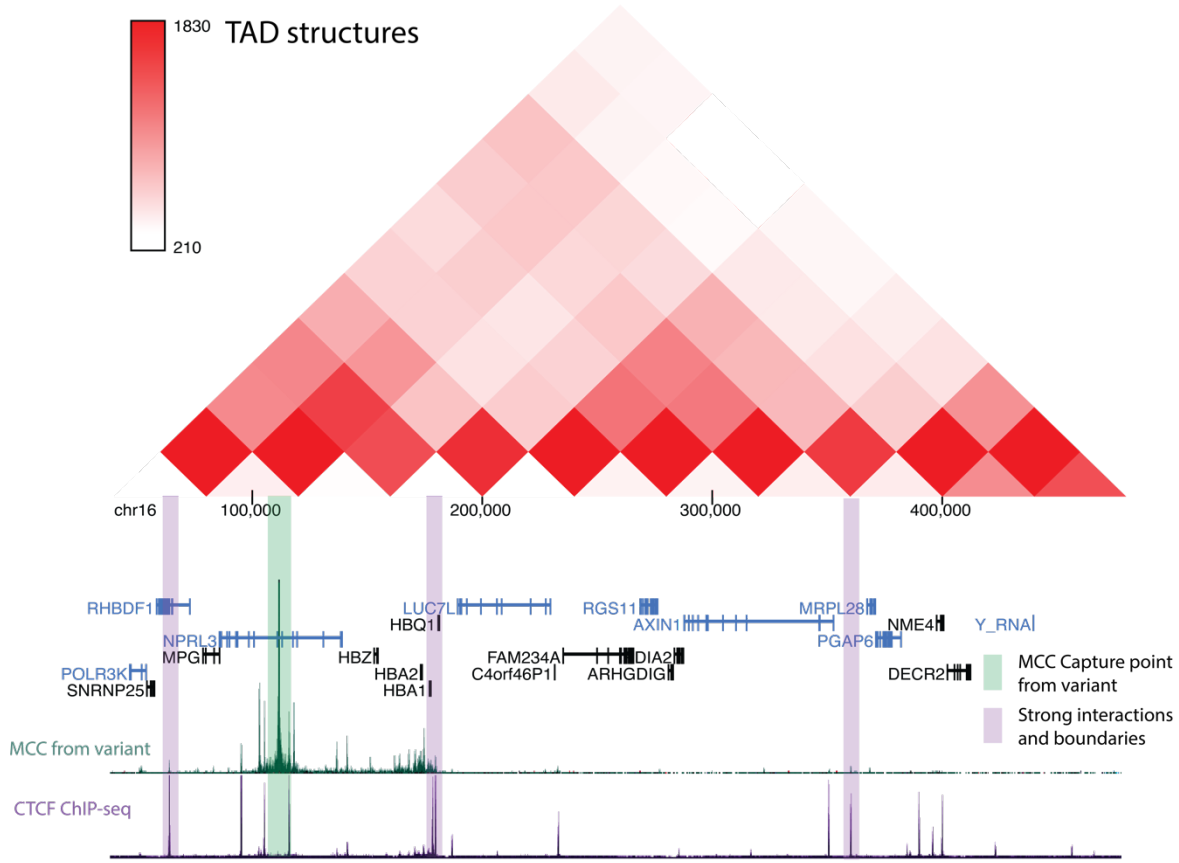
Motif analysis was done with homer v3.0 and ran with default parameters [158].

Sasquatch

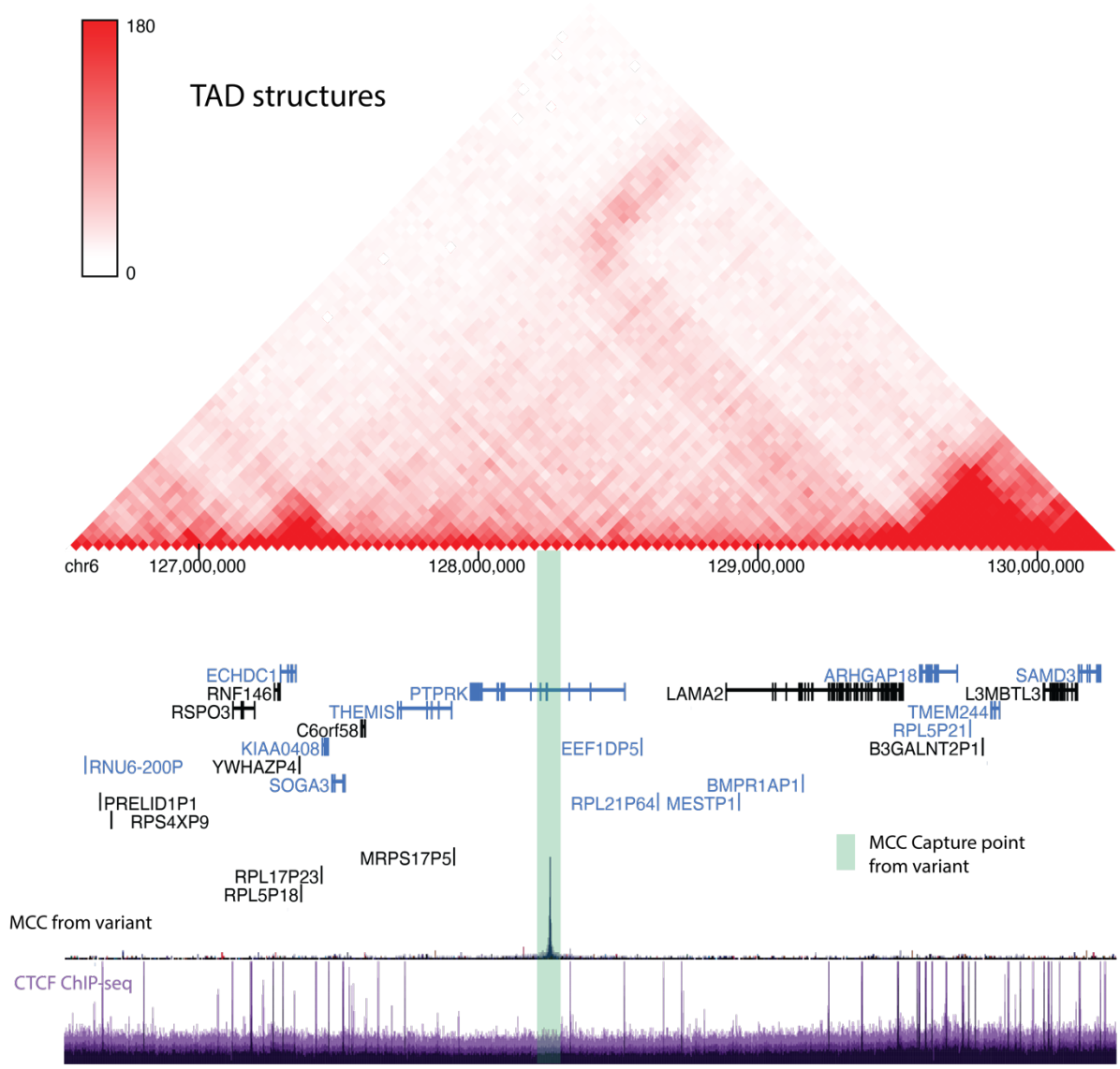
Initial Sasquatch ran with hexamer sequences at https://sasquatch.molbiol.ox.ac.uk/cgi-bin/foot.cgi

Later analysis ran by Martin Sargeant

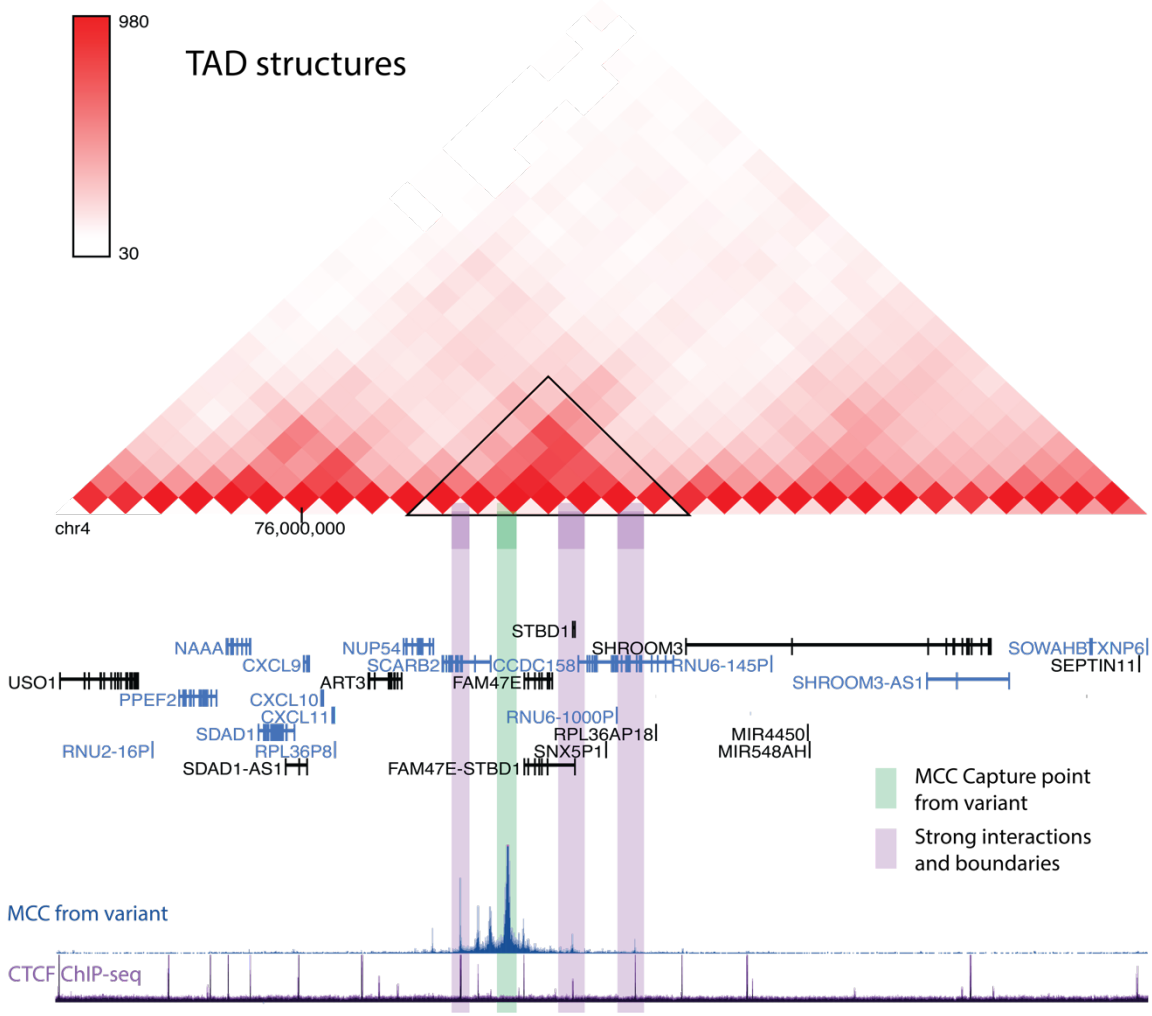
Supplementary material



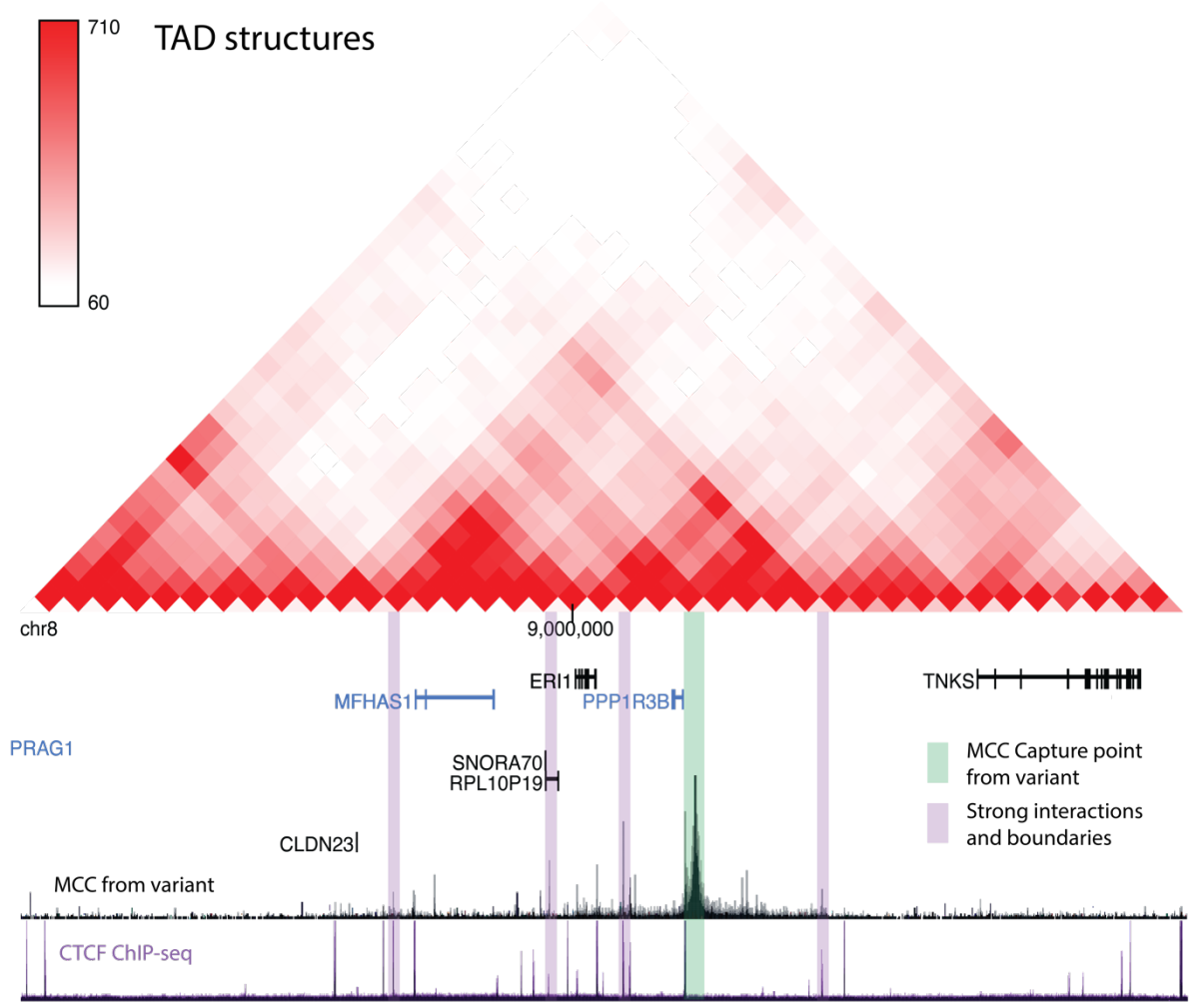
Supplementary Figure 1: TAD structure at *HBA*.
Hi-C data from K562 cells (Rao *et.al* [5], visualised with the 3D genome browser [222]) shows TAD structures at the *HBA* locus. MCC track (described and presented in Chapter 3) shows local interactions from the R2 enhancer. CTCF ChIP-seq from late erythroblast cells. Green highlight shows site of variant described in Chapter 3, and site of MCC capture. Purple highlights show strong local interactions, that can be seen to align with the overall TAD structures in the Hi-C data.



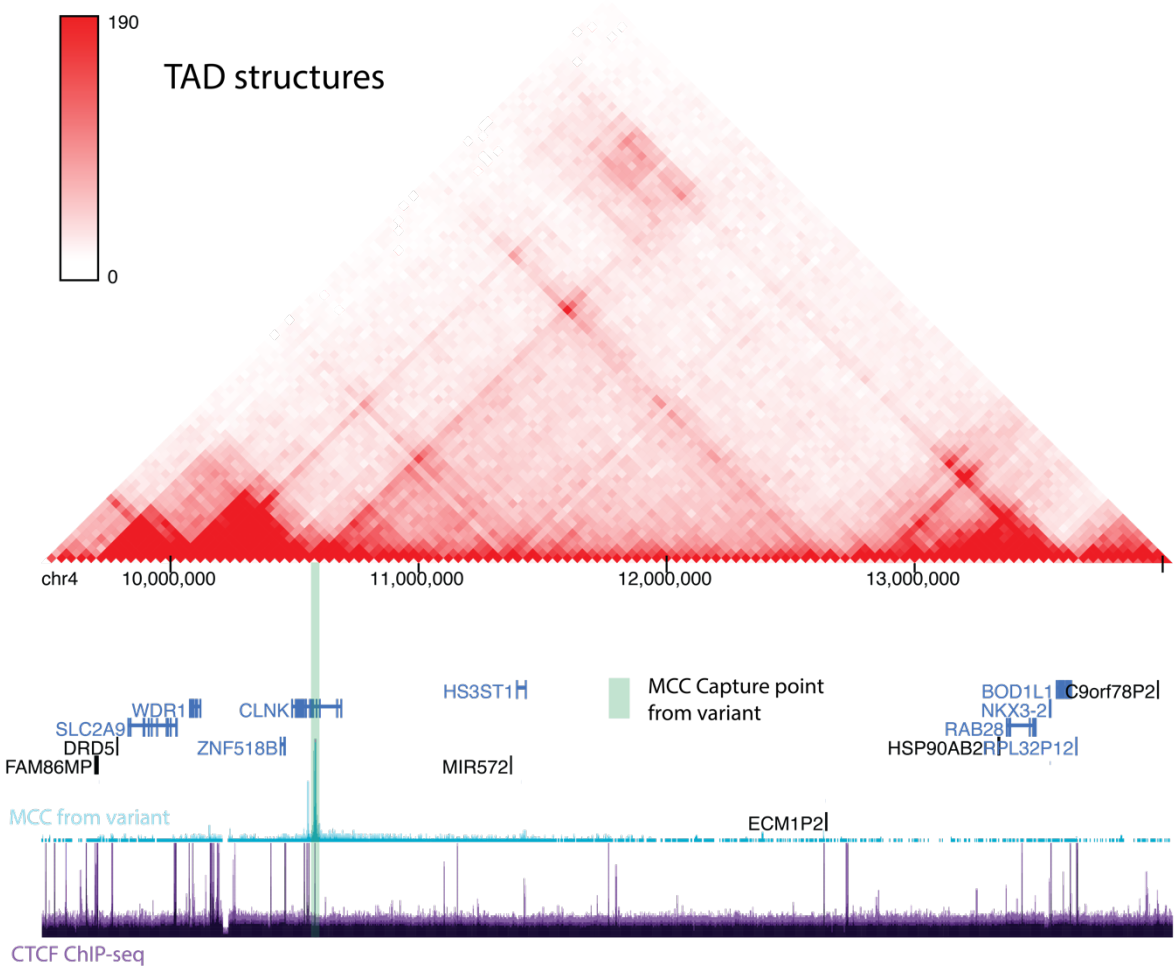
Supplementary Figure 2: TAD structure at *PTPRK.*
Hi-C data from K562 cells (Rao *et.al* [5], visualised with the 3D genome browser [222]) shows TAD structures at the *PTPRK* locus. MCC track (described and presented in Chapter 3) shows local interactions from the variant region. CTCF ChIP-seq from late erythroblast cells. Green highlight shows site of variant described in Chapter 2, and site of MCC capture.



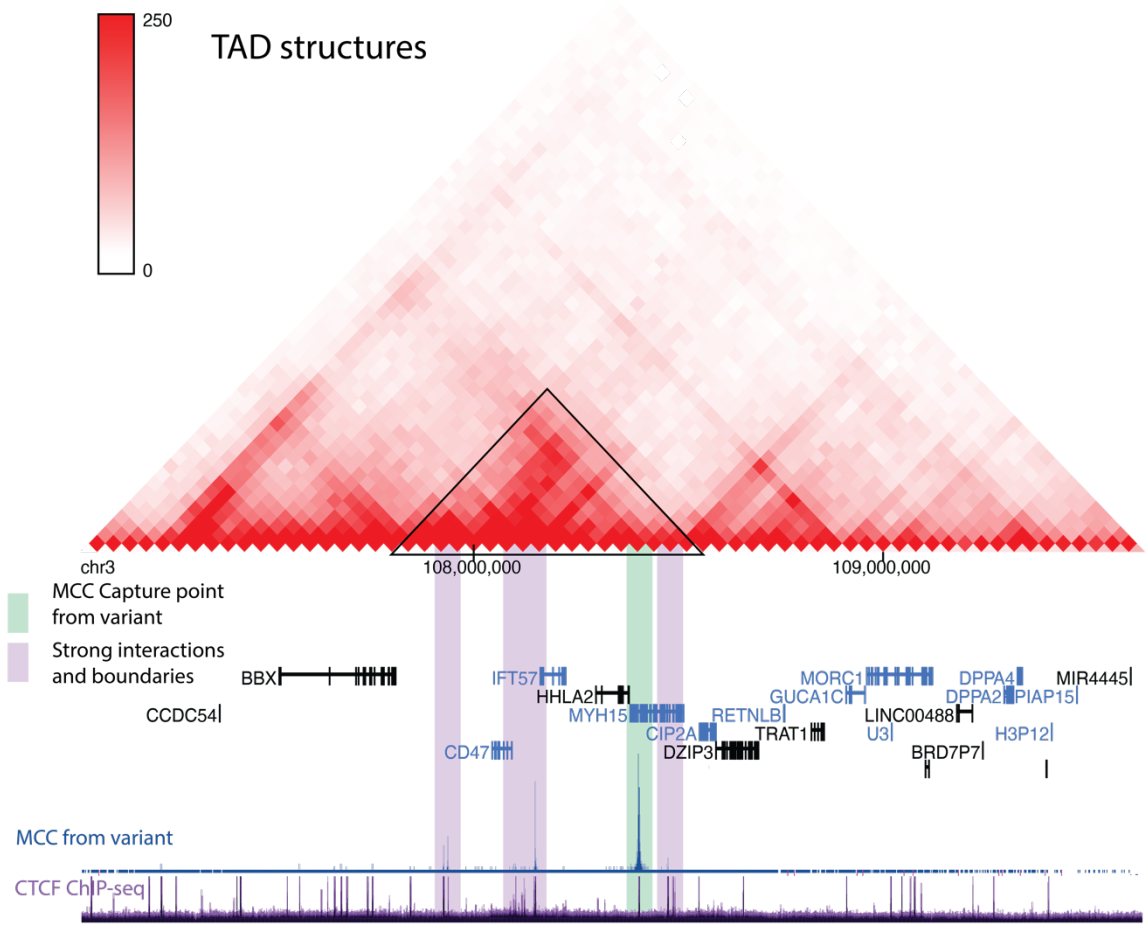
Supplementary Figure 3: TAD structure at *SCARB2*.
Hi-C data from K562 cells (Rao *et.al* [5], visualised with the 3D genome browser [222]) shows TAD structures at the *SCARB2* locus. MCC track (described and presented in Chapter 3) shows local interactions from the variant enhancer containing rs11097297. CTCF ChIP-seq from late erythroblast cells. Green highlight shows site of variant described in Chapter 2, and site of MCC capture from Chapter 3. Purple highlights show strong local interactions, that can be seen to align with the overall TAD structures in the Hi-C data.



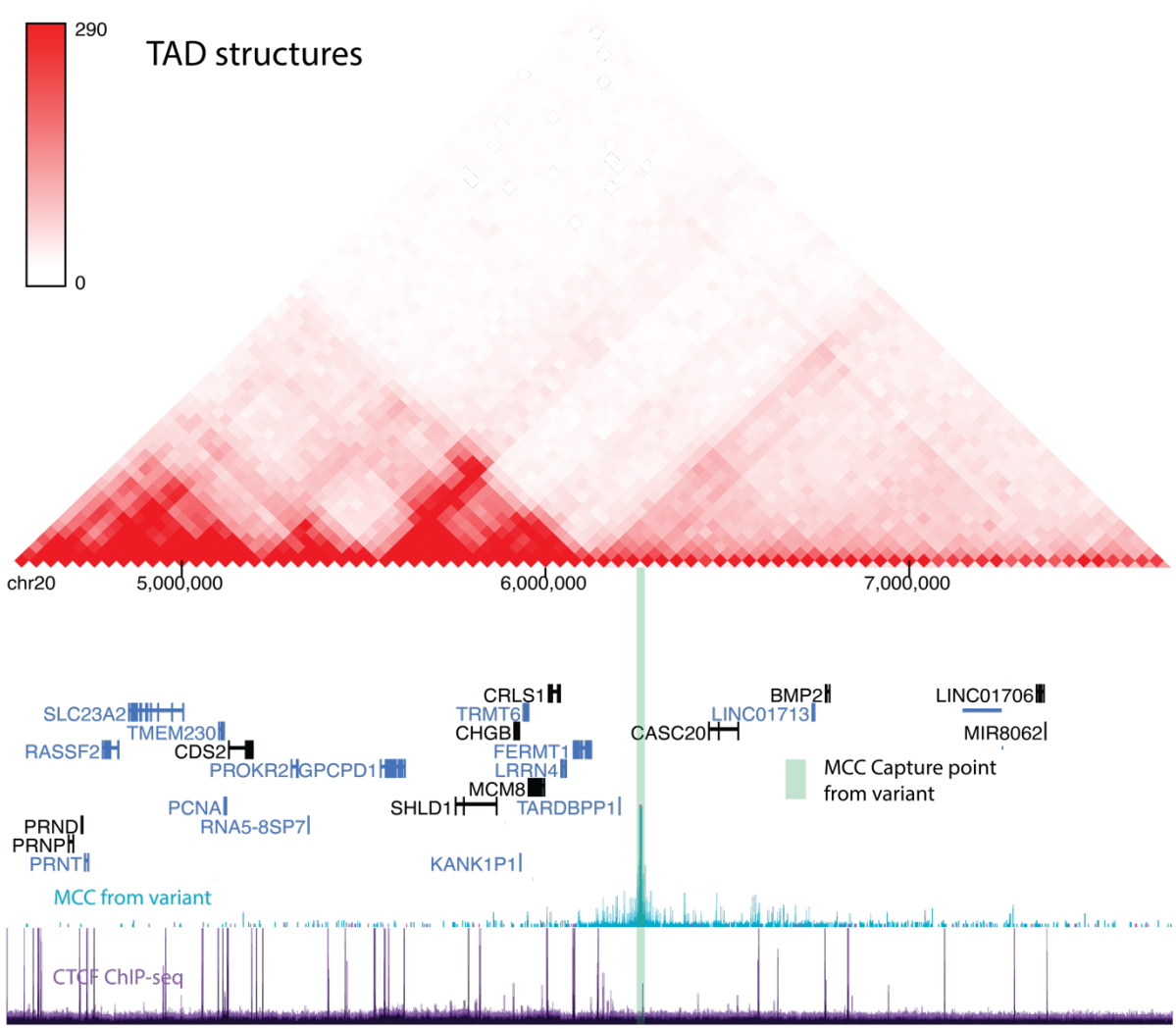
Supplementary Figure 4: TAD structures at *PPP1R3B.*
Hi-C data from K562 cells (Rao *et.al* [5], visualised with the 3D genome browser [222]) shows TAD structures at the *PPP1R3B* locus. MCC track (described and presented in Chapter 3) shows local interactions from the variant enhancer 'peak 1' containing rs330946. CTCF ChIP-seq from late erythroblast cells. Green highlight shows site of variant described in Chapter 2, and site of MCC capture from Chapter 3. Purple highlights show strong local interactions, that can be seen to align with the overall TAD structures in the Hi-C data.



Supplementary Figure 5: TAD structures at *CLNK*.
Hi-C data from K562 cells (Rao *et.al* [5], visualised with the 3D genome browser [222]) shows TAD structures at the *CLNK* locus. MCC track (described and presented in Chapter 3) shows local interactions from the variant enhancer region. CTCF ChIP-seq from late erythroblast cells. Green highlight shows site of variant described in Chapter 2, and site of MCC capture from Chapter 3.



Supplementary Figure 6: TAD structures at *MYH15.*
Hi-C data from K562 cells (Rao *et.al* [5], visualised with the 3D genome browser [222]) shows TAD structures at the *MYH15* locus. MCC track (described and presented in Chapter 3) shows local interactions from the variant CTCF peak containing rs936266. CTCF ChIP-seq from late erythroblast cells. Green highlight shows site of variant described in Chapter 2, and site of MCC capture from Chapter 3. Purple highlights show strong local interactions, that can be seen to align with the overall TAD structures in the Hi-C data.



Supplementary Figure 7: TAD structures at *FERMT1*.
Hi-C data from K562 cells (Rao *et.al* [5], visualised with the 3D genome browser [222]) shows TAD structures at the *FERMT1* locus. MCC track (described and presented in Chapter 3) shows local interactions from the variant novel promoter containing rs71182124. CTCF ChIP-seq from late erythroblast cells. Green highlight shows site of variant described in Chapter 2, and site of MCC capture from Chapter 3.

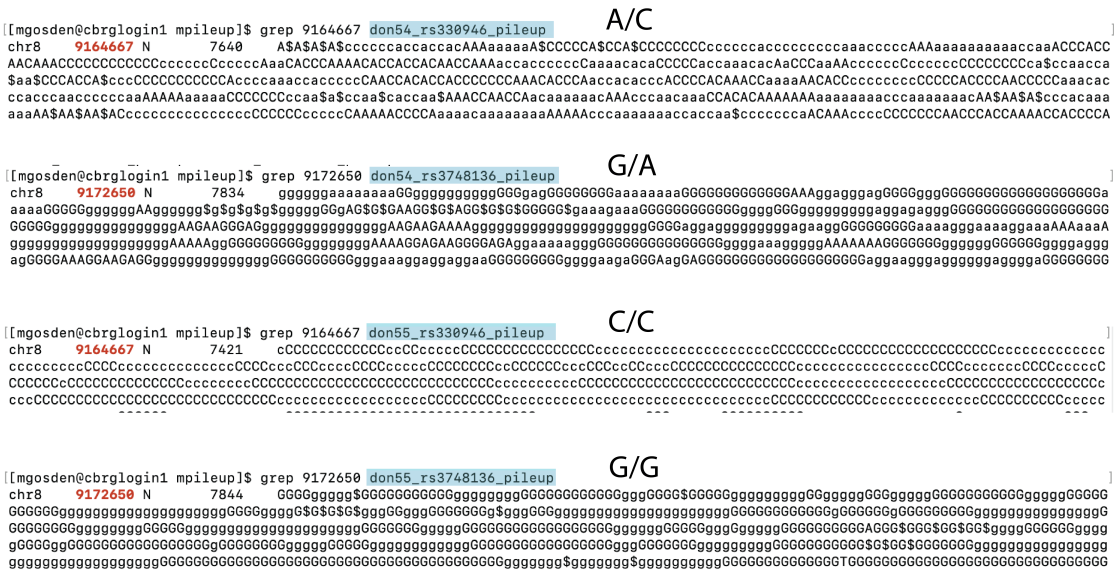
Locus	Element information	SNP position	Base change	SNP rs	MAF
HBA	R2 enhancer	chr16:113599	G>A	rs11248850	0.295327
PTPRK	Novel peak	chr6:128256094	A>G	-	-
SCARB2	Enhancer	chr4:76233712	C>T	rs11097297	0.48143
SCARB2	Enhancer	chr4:76233804	T>G	rs76504716	0.041334
PPP1R3B	Enhancer 'peak 1'	chr8:9164667	A>C	rs330946	0.438099
PPP1R3B	Enhancer 'peak 2'	chr8:9172650	G>A	rs3748136	0.158746
PPP1R3B	Enhancer 'peak 2'	chr8:9172593	G>C	rs3748137	0.019369
MYH15	CTCF	chr3:108406767	C>T	rs936266	0.207268
FERMT1	Novel promoter	chr20:6261964-6261971	CAGACCCA>CA	rs71182124	0.14397

Supplementary Table 1: Summary of SNPs introduced in Chapter 2.

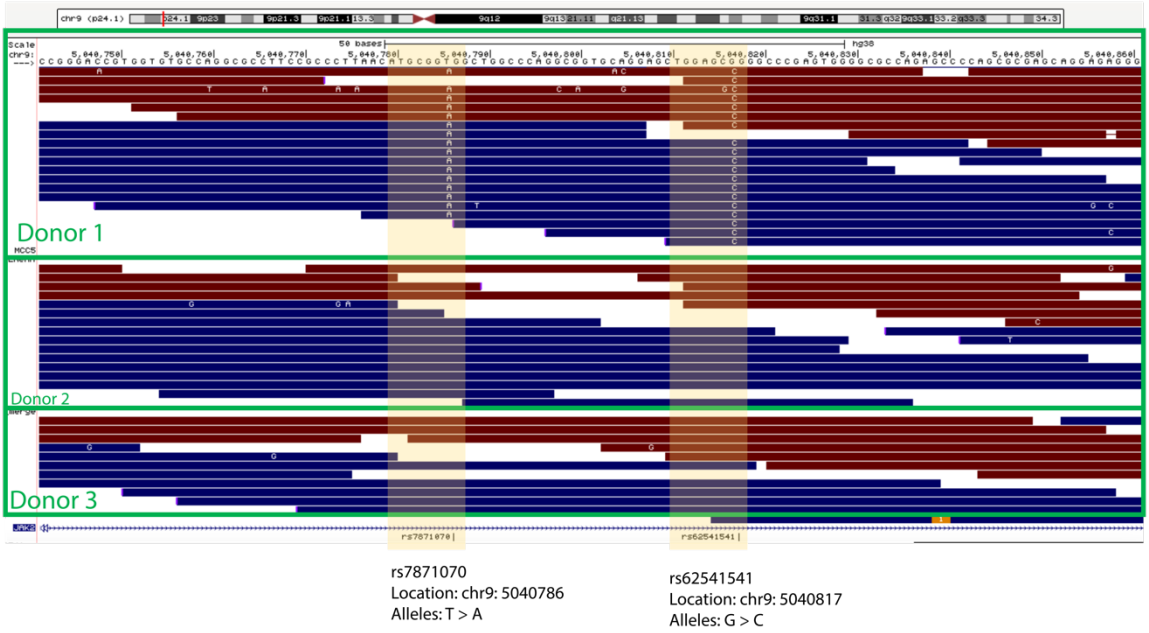
a)



b)



Supplementary Figure 8: Exemplar Mpileups.
Output of Mpileup can be used to determine genetics of an individual at a selected location a) Mpileup of an enhancer in *RCL1* in an uncharacterised donor. b) using Mpileup to determine the genetics of donors 54 and 55 at the *PPP1R3B* enhancers.



Supplementary Figure 9: Exemplar of using BAMs to determine genotype.
Three donors show BAMs at the site of two SNPs. Donor 1 shows different bases to the reference and these are visible on the BAMs.

rs11097297		**nascent RNA**			
		Haplotype observations		Paired Samples Wilcoxon Test	
Individual	**Genotype**	Hap1 count	Hap2 count	p	N
Don043	het	106	22	2.12E-08	60
Don013	het	18	104	4.72E-08	53
Don018	het	13	96	9.41E-08	53
Don028	het	2	35	1.80E-05	27
Don020	het	69	16	0.0001159	45
Don026	het	25	0	0.0001431	19
Don017	het	46	3	0.0002707	19
Don035	het	56	9	0.0003637	25
Don030	het	16	42	0.01443	40
Don045	het	26	10	0.09344	16
Don025	het	3	4	0.8125	7
Don011	het	-	-	-	-
Don046	het				
Don048	het				

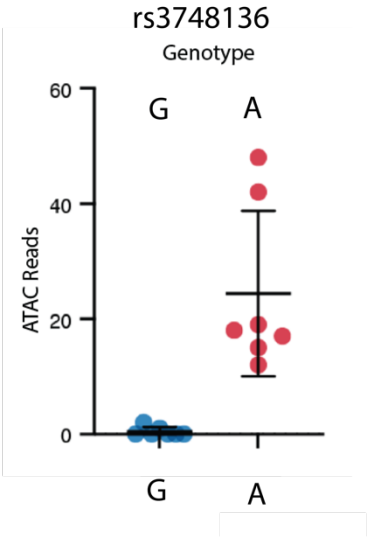
Don033	hom ref	140	61	2.51E-05	81
Don022	hom ref	38	97	0.0001516	64
Don031	hom ref	1	24	0.0002294	21
Don034	hom ref	11	79	0.001824	23
Don041	hom ref	0	13	0.001953	10
Don029	hom ref	10	0	0.03125	6
Don008	hom ref	2	9	0.1094	7
Don027	hom ref	12	4	0.1094	8
Don019	hom ref	16	34	0.1138	21
Don016	hom ref	21	15	0.2699	25
Don001	hom ref	20	29	0.2748	26
Don049	hom ref	11	17	0.3507	20
Don037	hom ref	6	9	0.4609	8
Don003	hom ref	5	7	0.625	10
Don009	hom ref	5	7	0.625	10
Don042	hom ref	7	5	0.6523	9
Don036	hom ref	8	10	0.6951	18
Don039	hom ref	4	3	0.8125	7
Don004	hom ref	4	1	-	5
Don005	hom ref	1	4	-	5
Don012	hom ref	-	-	-	-
Don015	hom ref	1	4	-	5
Don021	hom ref	-	-	-	-
Don024	hom ref	3	2	-	4
Don032	hom ref	0	2	-	2
Don038	hom ref	3	0	-	3
Don040	hom ref	0	0	-	-
Don044	hom ref	4	5	-	5
Don047	hom ref	0	0	-	-
Don050	hom ref	0	0	-	-
Don006	hom ref				
Don007	hom ref				
Don014	hom ref				
Don023	hom ref				
Don002	hom var	76	84	0.4538	72
Don010	hom var	127	120	0.6884	71

Supplementary Table 2: Nascent allelic RNA transcription of *SCARB2* across the 50 donor.

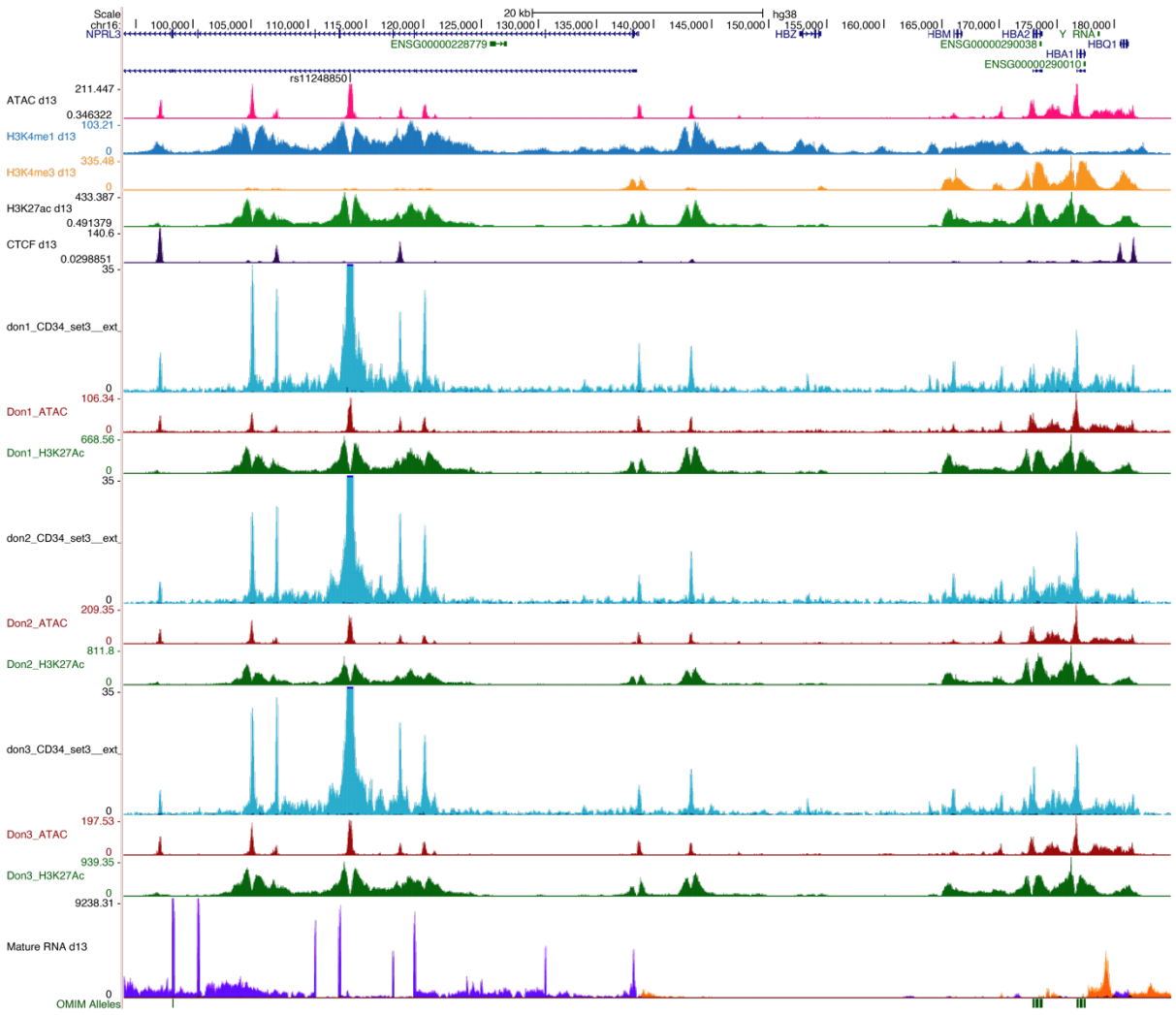
rs11097297		mature RNA			
		Haplotype observations		Paired Samples Wilcoxon Test	
Individual	**Genotype**	Hap1 count	Hap2 count	p	N
Don043	het	32	9	0.002235	24
Don013	het	6	40	0.0002808	25
Don018	het	9	34	0.005507	22
Don028	het	1	14	0.003418	13
Don020	het	30	10	0.03703	21
Don026	het	11	2	0.07422	9
Don017	het	16	2	-	5
Don035	het	8	3	-	5
Don030	het	2	3	-	5
Don045	het	36	5	0.06372	15
Don025	het	1	0	-	1
Don011	het	6	37	0.000102	26
Don046	het				
Don048	het				
Don033	hom ref	68	13	0.0006189	43
Don022	hom ref	15	30	0.2668	18
Don031	hom ref	0	13	0.001953	10
Don034	hom ref	3	23	0.01074	14
Don041	hom ref	2	3	-	5
Don029	hom ref	1	0	-	1
Don008	hom ref	0	6	0.03125	6
Don027	hom ref	1	1	-	2
Don019	hom ref	8	18	0.1514	12
Don016	hom ref	18	7	0.07204	17
Don001	hom ref	24	14	0.2066	18
Don049	hom ref	10	7	0.5186	12
Don037	hom ref	3	5	0.5469	8
Don003	hom ref	1	2	-	2
Don009	hom ref	5	11	0.1055	10
Don042	hom ref	3	7	0.1562	6
Don036	hom ref	5	5	-	5
Don039	hom ref	2	3	-	2
Don004	hom ref	0	0	-	-
Don005	hom ref	0	1	-	1
Don012	hom ref	8	11	0.4922	10

Don015	hom ref	2	4	-	5
Don021	hom ref	0	1	-	1
Don024	hom ref	0	0	-	-
Don032	hom ref	0	0	-	-
Don038	hom ref	3	0	-	3
Don040	hom ref	1	0	-	1
Don044	hom ref	5	1	0.1562	6
Don047	hom ref	-	-	-	-
Don050	hom ref	1	0	-	1
Don006	hom ref				
Don007	hom ref				
Don014	hom ref				
Don023	hom ref				
Don002	hom var	55	67	0.3212	44
Don010	hom var	40	39	0.7464	31

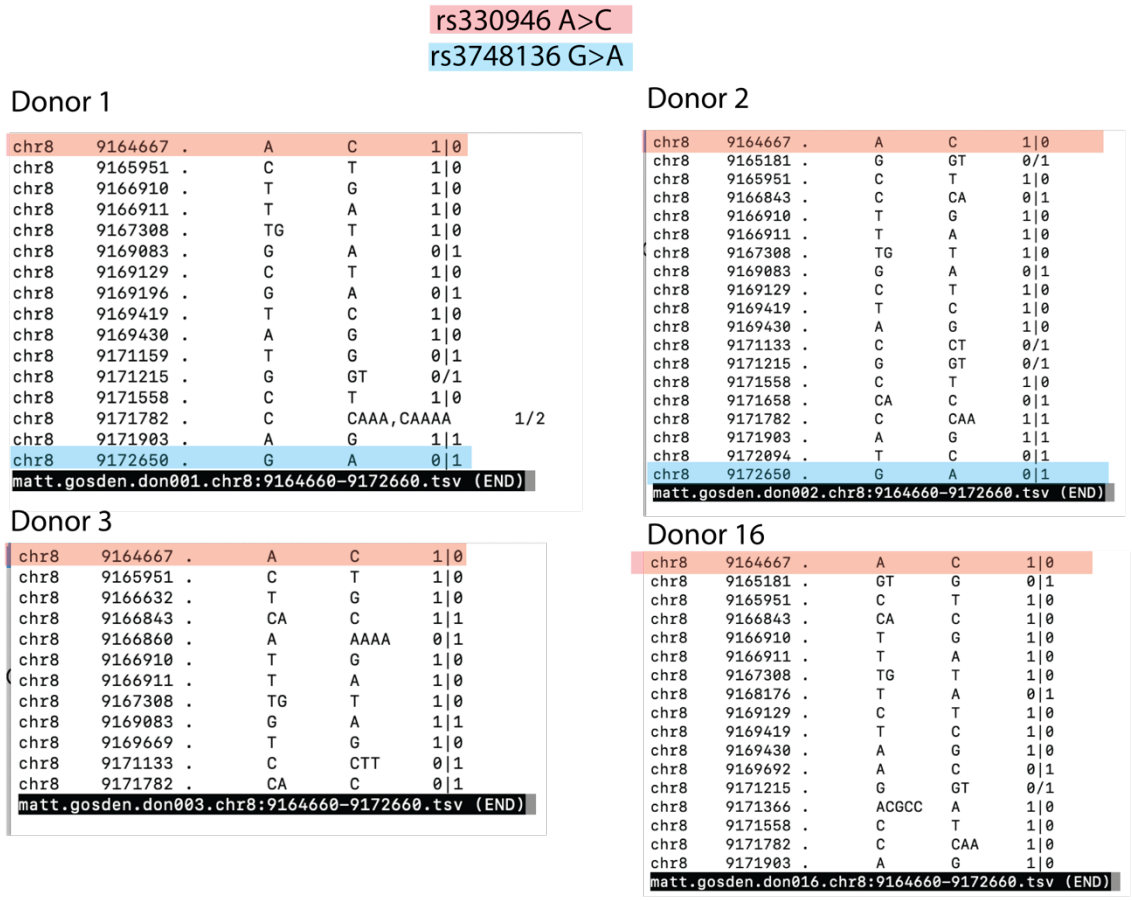
Supplementary Table 3: Mature allelic RNA transcription of *SCARB2* across the 50 donors.



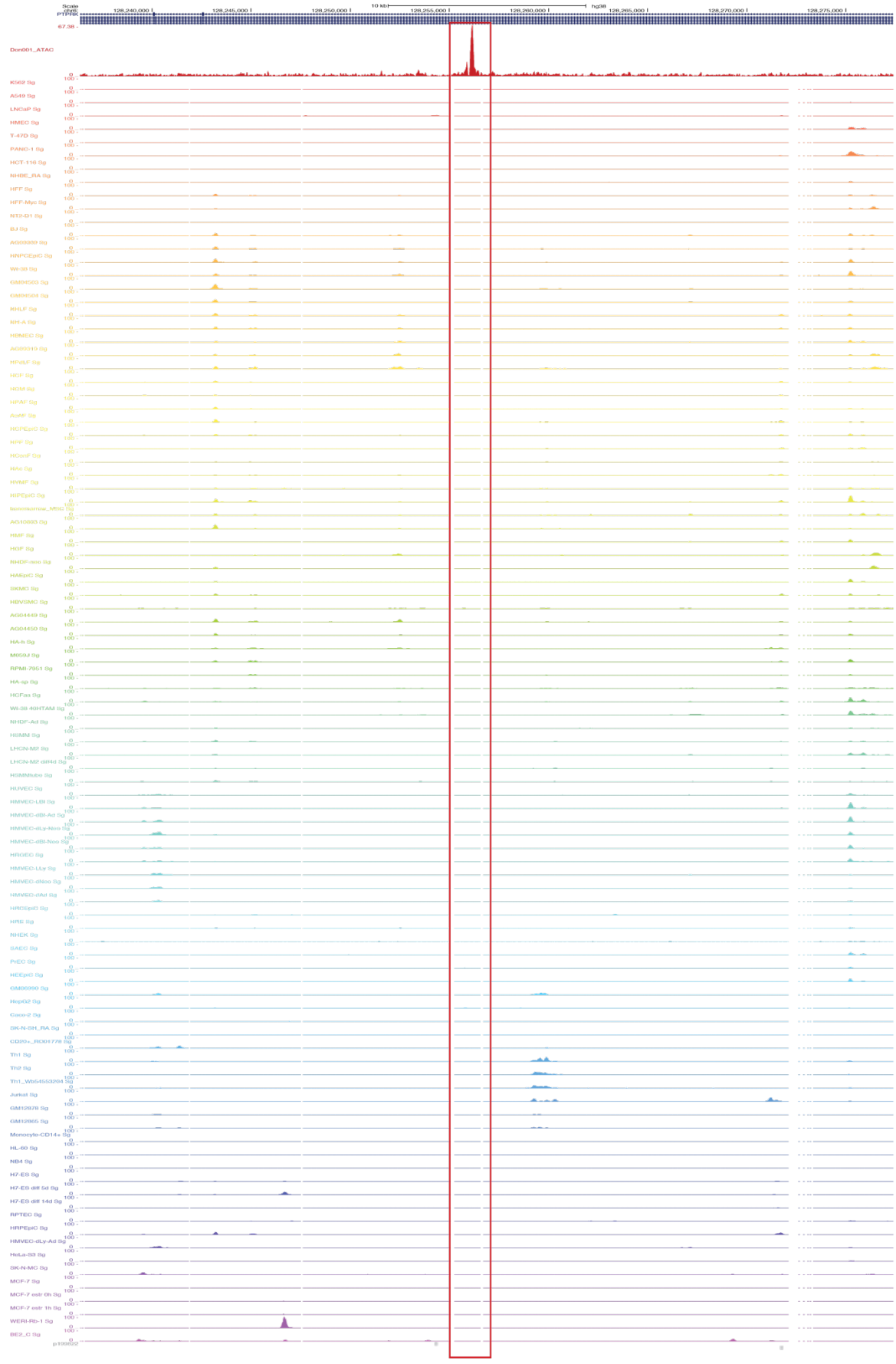
Supplementary Figure 10: ATAC Skew at rs3748136



Supplementary Figure 11: MCC validation of probe set 3 at *HBA*.



Supplementary Figure 12: Phasing links SNPs on alleles at *PPP1R3B* locus.


Scale
chr6
128,240,000
128,245,000
128,250,000
10 kb
128,255,000
128,260,000
hg38
128,265,000
128,270,000
128,275,000
PTPN6
67.38
Don001_ATAC
K562 Sg
A549 Sg
LNCaP Sg
HMEC Sg
T-47D Sg
PANC-1 Sg
HCT-116 Sg
NHBE_RA Sg
HFF Sg
HFF-Myc Sg
NT2-D1 Sg
BJ Sg
AG09309 Sg
HNPCEpiC Sg
WI-38 Sg
GM04503 Sg
GM04504 Sg
MHLF Sg
NH-A Sg
HBMEC Sg
AG09319 Sg
HPdLF Sg
HGF Sg
HDM Sg
HPAF Sg
AoSF Sg
HCPEpiC Sg
NPF Sg
HGonF Sg
HAc Sg
HVMF Sg
HIPEpiC Sg
AG10803 Sg
HMF Sg
HGF Sg
NHDF-neo Sg
HAEpiC Sg
SKMC Sg
HBVSMC Sg
AG04449 Sg
AG04450 Sg
HA-h Sg
M059J Sg
RPMI-7951 Sg
HA-sp Sg
HCFaa Sg
WI-38 40HTAM Sg
NHDF-Ad Sg
HSMM Sg
LHCN-M2 Sg
LHCN-M2 diff4d Sg
HSMMtube Sg
HUVEC Sg
HMVEC-LBl Sg
HMVEC-dBl-Ad Sg
HMVEC-dLy-Neo Sg
HMVEC-dBl-Neo Sg
HRGEC Sg
HMVEC-LLy Sg
HMVEC-dNeo Sg
HMVEC-dAd Sg
HRCEpiC Sg
HRE Sg
NHEK Sg
SAEC Sg
PrEC Sg
HEEpiC Sg
GM06990 Sg
HepG2 Sg
Caco-2 Sg
SK-N-SH_RA Sg
CD20+_RO01778 Sg
Th1 Sg
Th2 Sg
Th1_Wb545532R4 Sg
Jurkat Sg
GM12878 Sg
GM12865 Sg
Monocyte-CD14+ Sg
HL-60 Sg
NB4 Sg
H7-ES Sg
H7-ES diff 5d Sg
H7-ES diff 14d Sg
RPTEC Sg
HRPEpiC Sg
HMVEC-dLy-Ad Sg
HeLa-S3 Sg
SK-N-MC Sg
MCF-7 Sg
MCF-7 estr 0h Sg
MCF-7 estr 1h Sg
WERI-Rb-1 Sg
BE2_C Sg
p1998932

Supplementary Figure 13: Open chromatin profiles from 96 cell types at the gain of function peak in PTPRK
ATAC data from erythroid showing the gain of function peak present in a single donor in erythroid cells, red box. ENCODE DNase data from 95 additional cell types shows that there is no activity in any other cell types [203].

References

1. Bianconi, E., et al., *An estimation of the number of cells in the human body.* Ann Hum Biol, 2013. **40**(6): p. 463-71.
2. Consortium, I.H.G.S., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **21;431(7011):931-45**.
3. Chen, N., et al., *WormBase: a comprehensive data resource for Caenorhabditis biology and genomics.* Nucleic Acids Res, 2005. **33**(Database issue): p. D383-9.
4. Symmons, O., et al., *Functional and topological characteristics of mammalian regulatory domains.* Genome Research, 2014. **24**(3): p. 390-400.
5. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.
6. Davies, J.O., et al., *Multiplexed analysis of chromosome conformation at vastly improved sensitivity.* Nat Methods, 2016. **13**(1): p. 74-80.
7. Deng, W., et al., *Reactivation of developmentally silenced globin genes by forced chromatin looping.* Cell, 2014. **158**(4): p. 849-860.
8. Benabdallah, N.S., et al., *Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation.* Molecular Cell, 2019. **76**(3): p. 473-484.e7.
9. Zhang, T., S. Cooper, and N. Brockdorff, *The interplay of histone modifications - writers that read.* EMBO Rep, 2015. **16**(11): p. 1467-81.
10. Lee, J.E., et al., *H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation.* Elife, 2013. **2**: p. e01503.
11. Long, H.K., et al., *Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder.* Cell Stem Cell, 2020. **27**(5): p. 765-783.e14.
12. Long, H.K., S.L. Prescott, and J. Wysocka, *Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution.* Cell, 2016. **167**(5): p. 1170-1187.
13. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
14. Holwerda, S.J. and W. de Laat, *CTCF: the protein, the binding partners, the binding sites and their chromatin loops.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1620): p. 20120369.
15. Yang, J., et al., *Structures of CTCF-DNA complexes including all 11 zinc fingers.* Nucleic Acids Res, 2023. **51**(16): p. 8447-8462.
16. Clarkson, C.T., et al., *CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length.* Nucleic Acids Res, 2019. **47**(21): p. 11181-11196.
17. Bell, A., A. West, and G. Felsenfeld, *The protein CTCF is required for the enhancer blocking activity of vertebrate insulators.* Cell, 1999. **6;98(3):387-96**.
18. Davidson, I.F., et al., *CTCF is a DNA-tension-dependent barrier to cohesin-mediated loop extrusion.* Nature, 2023. **616**(7958): p. 822-827.
19. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-380.
20. Denker, A. and W. de Laat, *The second decade of 3C technologies: detailed insights into nuclear organization.* Genes Dev, 2016. **30**(12): p. 1357-82.
21. Vos, E.S.M., et al., *Interplay between CTCF boundaries and a super enhancer controls cohesin extrusion trajectories and gene expression.* Molecular Cell, 2021. **81**(0): p. 1-14.
22. Georgiades, E., et al., *Active regulatory elements recruit cohesin to establish cell-specific chromatin domains.* bioRxiv, 2023. **10.13.562171**.
23. Ren, G., et al., *CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression.* Mol Cell, 2017. **67**(6): p. 1049-1058 e6.

24. Kubo, N., et al., *Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation.* Nat Struct Mol Biol, 2021. **28**(2): p. 152-161.
25. Chen, L.F. and H.K. Long, *Topology regulatory elements: From shaping genome architecture to gene regulation.* Curr Opin Struct Biol, 2023. **83**: p. 102723.
26. Hafner, A., et al., *Loop stacking organizes genome folding from TADs to chromosomes.* Molecular Cell, 2023. **83**(9): p. 1377-1392.e6.
27. Lupiáñez, Darío G., et al., *Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions.* Cell, 2015. **161**(5): p. 1012-1025.
28. Ushiki, A., et al., *Deletion of CTCF sites in the SHH locus alters enhancer-promoter interactions and leads to acheiropodia.* Nat Commun, 2021. **12**(1): p. 2282.
29. Laugsch, M., et al., *Modeling the Pathological Long-Range Regulatory Effects of Human Structural Variation with Patient-Specific hiPSCs.* Cell Stem Cell, 2019. **24**(5): p. 736-752 e12.
30. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.
31. Schwessinger, R., et al., *DeepC: predicting 3D genome folding using megabase-scale transfer learning.* Nature Methods, 2020. **17**(11): p. 1118-1124.
32. Schwessinger, R., et al., *Sasquatch: predicting the impact of regulatory SNPs on transcription factor binding from cell- and tissue-specific DNase footprints.* Genome Res, 2017. **27**(10): p. 1730-1742.
33. Villar, D., et al., *Enhancer evolution across 20 mammalian species.* Cell, 2015. **160**(3): p. 554-66.
34. Reitzer LJ and M. B., *Transcription of glnA in E. coli is stimulated by activator bound to sites far from the promoter.* Cell, 1986 **20;45(6):785-92**.
35. N, W., et al., *The yeast UASG is a transcriptional enhancer in human HeLa cells in the presence of the GAL4 trans-activator.* Cell, 1988. **29;52(2):169-78**.
36. Panigrahi, A. and B.W. O'Malley, *Mechanisms of enhancer action: the known and the unknown.* Genome Biol, 2021. **22**(1): p. 108.
37. Whalen, S. and K.S. Pollard, *Enhancer Function and Evolutionary Roles of Human Accelerated Regions.* Annu Rev Genet, 2022. **56**: p. 423-439.
38. Osterwalder, M., et al., *Enhancer redundancy provides phenotypic robustness in mammalian development.* Nature, 2018. **554**(7691): p. 239-243.
39. Kvon, E.Z., et al., *Enhancer redundancy in development and disease.* Nat Rev Genet, 2021. **22**(5): p. 324-336.
40. Visel, A., E.M. Rubin, and L.A. Pennacchio, *Genomic views of distant-acting enhancers.* Nature, 2009. **461**(7261): p. 199-205.
41. Bahr, C., et al., *A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies.* Nature, 2018. **553**(7689): p. 515-520.
42. Zemke, N.R., et al., *Conserved and divergent gene regulatory programs of the mammalian neocortex.* Nature, 2023. **624**(7991): p. 390-402.
43. Lettice, L.A., *A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.* Human Molecular Genetics, 2003. **12**(14): p. 1725-1735.
44. Lettice, L.A., et al., *The Conserved Sonic Hedgehog Limb Enhancer Consists of Discrete Functional Elements that Regulate Precise Spatial Expression.* Cell Rep, 2017. **20**(6): p. 1396-1408.
45. Chan, E.T., et al., *Conservation of core gene expression in vertebrate tissues.* J Biol, 2009. **8**(3): p. 33.
46. Afzal, Z. and R. Krumlauf, *Transcriptional Regulation and Implications for Controlling Hox Gene Expression.* J Dev Biol, 2022. **10**(1).
47. Santini, S., J.L. Boore, and A. Meyer, *Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters.* Genome Res, 2003. **13**(6A): p. 1111-22.
48. Kvon, E.Z., et al., *Progressive Loss of Function in a Limb Enhancer during Snake Evolution.* Cell, 2016. **167**(3): p. 633-642 e11.
49. Chen, H., et al., *Fast-Evolving Human-Specific Neural Enhancers Are Associated with Aging-Related Diseases.* Cell Syst, 2018. **6**(5): p. 604-611 e4.
50. Li S, H.S., Ovcharenko I., *De novo human brain enhancers created by single-nucleotide mutations.* Sci Adv., 2023 Feb.

51. Won, H., et al., *Human evolved regulatory elements modulate genes involved in cortical expansion and neurodevelopmental disease susceptibility.* Nat Commun, 2019. **10**(1): p. 2396.
52. Snetkova, V., et al., *Ultraconserved enhancer function does not require perfect sequence conservation.* Nat Genet, 2021. **53**(4): p. 521-528.
53. Bourque, G., et al., *Ten things you should know about transposable elements.* Genome Biol, 2018. **19**(1): p. 199.
54. Trizzino, M., et al., *Transposable elements are the primary source of novelty in primate gene regulation.* Genome Res, 2017. **27**(10): p. 1623-1633.
55. McEwen, G.K., et al., *Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis.* Genome Res, 2006. **16**(4): p. 451-65.
56. Eichenlaub, M.P. and L. Ettwiller, *De novo genesis of enhancers in vertebrates.* PLoS Biol, 2011. **9**(11): p. e1001188.
57. Mahmudi O, S.B., Arvestad L, Nowick K, Lagergren J., *Gene-pseudogene evolution: a probabilistic approach.* BMC Genomics., 2015. **2015;16 Suppl 10(Suppl 10):S12.**
58. Jonathon R. Stone, G.A.W., *Rapid Evolution of cis-Regulatory Sequences via Local Point Mutations,.* Molecular Biology and Evolution,, 2001. **Volume 18, Issue 9,**.
59. Duque, T. and S. Sinha, *What Does It Take to Evolve an Enhancer? A Simulation-Based Study of Factors Influencing the Emergence of Combinatorial Regulation.* Genome Biology and Evolution, 2015. **7**(6): p. 1415-1431.
60. Nesta, A.V., D. Tafur, and C.R. Beck, *Hotspots of Human Mutation.* Trends Genet, 2021. **37**(8): p. 717-729.
61. Carlson, C.S., et al., *Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans.* Nat Genet, 2003. **33**(4): p. 518-21.
62. Shaikho, E.M., et al., *A phased SNP-based classification of sickle cell anemia HBB haplotypes.* BMC Genomics, 2017. **18**(1): p. 608.
63. Gisler, F.M., et al., *Identification of SNPs in the cystic fibrosis interactome influencing pulmonary progression in cystic fibrosis.* Eur J Hum Genet, 2013. **21**(4): p. 397-403.
64. Steri, M., et al., *Genetic variants in mRNA untranslated regions.* Wiley Interdiscip Rev RNA, 2018. **9**(4): p. e1474.
65. D. P, B.-J., *Short-Range DNA Looping by the Xenopus HMG-Box Transcription Factor, xUBF.* Science, 1994. **264**: p. 1134-1137.
66. Kulkarni, M.M. and D.N. Arnosti, *Information display by transcriptional enhancers.* Development, 2003. **130**(26): p. 6569-75.
67. Jindal, G.A. and E.K. Farley, *Enhancer grammar in development, evolution, and disease: dependencies and interplay.* Dev Cell, 2021. **56**(5): p. 575-587.
68. Bhatia, S., et al., *Disruption of autoregulatory feedback by a mutation in a remote, ultraconserved PAX6 enhancer causes aniridia.* Am J Hum Genet, 2013. **93**(6): p. 1126-34.
69. Miguel-Escalada, I., et al., *Pancreas agenesis mutations disrupt a lead enhancer controlling a developmental enhancer cluster.* Dev Cell, 2022. **57**(16): p. 1922-1936 e9.
70. Astle, W.J., et al., *The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease.* Cell, 2016. **167**(5): p. 1415-1429.e19.
71. Van Der Harst, P., et al., *Seventy-five genetic loci influencing the human red blood cell.* Nature, 2012. **492**(7429): p. 369-375.
72. Vuckovic, D., et al., *The Polygenic and Monogenic Basis of Blood Traits and Diseases.* Cell, 2020. **182**(5): p. 1214-1231.e11.
73. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(23): p. 9362-9367.
74. Slatkin, M., *Linkage disequilibrium--understanding the evolutionary past and mapping the medical future.* Nat Rev Genet, 2008. **9**(6): p. 477-85.
75. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 499-511.

76. Lichou, F. and G. Trynka, *Functional studies of GWAS variants are gaining momentum.* Nat Commun, 2020. **11**(1): p. 6283.
77. Boyle, A.P., et al., *High-resolution mapping and characterization of open chromatin across the genome.* Cell, 2008. **132**(2): p. 311-22.
78. Ingram, R., et al., *PAP-LMPCR for improved, allele-specific footprinting and automated chromatin fine structure analysis.* Nucleic Acids Res, 2008. **36**(3): p. e19.
79. Vierstra, J., et al., *Global reference mapping of human transcription factor footprints.* Nature, 2020. **583**(7818): p. 729-736.
80. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nat Methods, 2013. **10**(12): p. 1213-8.
81. Buenrostro, J.D., et al., *ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.* Curr Protoc Mol Biol, 2015. **109**: p. 21 29 1-21 29 9.
82. Grandi, F.C., et al., *Chromatin accessibility profiling by ATAC-seq.* Nat Protoc, 2022. **17**(6): p. 1518-1552.
83. Maurano, M.T., et al., *Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo.* Nat Genet, 2015. **47**(12): p. 1393-401.
84. Tehranchi, A.K., et al., *Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk.* Cell, 2016. **165**(3): p. 730-41.
85. Lambert, S.A., et al., *The Human Transcription Factors.* Cell, 2018. **172**(4): p. 650-665.
86. Soskic, B., et al., *Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases.* Nat Genet, 2019. **51**(10): p. 1486-1493.
87. Wander, B., K. Broderick, and Z.W. Ulissi, *Catlas: an automated framework for catalyst discovery demonstrated for direct syngas conversion.* Catalysis Science & Technology, 2022. **12**(20): p. 6256-6267.
88. Domcke, S., et al., *A human cell atlas of fetal chromatin accessibility.* Science, 2020. **370**(6518).
89. Liu, Q., et al., *Chromatin accessibility prediction via a hybrid deep convolutional neural network.* Bioinformatics, 2018. **34**(5): p. 732-738.
90. Downes, D.J., et al., *Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus.* Nature Genetics, 2021. **53**(11): p. 1606-1615.
91. Davies, J.O., et al., *How best to identify chromosomal interactions: a comparison of approaches.* Nat Methods, 2017. **14**(2): p. 125-134.
92. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-1311.
93. van de Werken, H.J., et al., *Robust 4C-seq data analysis to screen for regulatory DNA interactions.* Nat Methods, 2012. **9**(10): p. 969-72.
94. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-293.
95. Mifsud, B., et al., *Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C.* Nat Genet, 2015. **47**(6): p. 598-606.
96. Oudelaar, A.M., et al., *Low-input Capture-C: A Chromosome Conformation Capture Assay to Analyze Chromatin Architecture in Small Numbers of Cells.* Bio-protocol, 2017. **7**(23).
97. Oudelaar, A.M., et al., *Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains.* Nat Genet, 2018. **50**(12): p. 1744-1751.
98. Oudelaar, A.M., et al., *Dynamics of the 4D genome during in vivo lineage specification and differentiation.* Nature Communications, 2020. **11**(1): p. 1-12.
99. Hua, P., et al., *Defining genome architecture at base-pair resolution.* Nature, 2021. **595**(7865): p. 125-129.
100. Hamley, J.C., et al., *Determining chromatin architecture with Micro Capture-C.* Nat Protoc, 2023. **18**(6): p. 1687-1711.
101. Su, C., et al., *Restriction enzyme selection dictates detection range sensitivity in chromatin conformation capture-based variant-to-gene mapping approaches.* Hum Genet, 2021. **140**(10): p. 1441-1448.
102. Hay, D., et al., *Genetic dissection of the alpha-globin super-enhancer in vivo.* Nat Genet, 2016. **48**(8): p. 895-903.

103. Blayney, J.W., et al., *Super-enhancers include classical enhancers and facilitators to fully activate gene expression.* Cell, 2023. **186**(26): p. 5826-5839 e18.
104. Li, X., et al., *A high-efficiency and versatile CRISPR/Cas9-mediated HDR-based biallelic editing system.* J Zhejiang Univ Sci B, 2022. **23**(2): p. 141-152.
105. Nambiar, T.S., et al., *Stimulation of CRISPR-mediated homology-directed repair by an engineered RAD18 variant.* Nat Commun, 2019. **10**(1): p. 3395.
106. Kurita, R., et al., *Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells.* PLoS One, 2013. **8**(3): p. e59890.
107. He, Z., et al., *Chromosomal instability and its effect on cell lines.* Cancer Rep (Hoboken), 2023. **6**(6): p. e1822.
108. Porto, E.M., et al., *Base editing: advances and therapeutic opportunities.* Nat Rev Drug Discov, 2020. **19**(12): p. 839-859.
109. Schwalb B, M.M., Zacher B, Frühauf K, Demel C, Tresch A, Gagneur J, Cramer P., *TT-seq maps the human transient transcriptome.* Science., 2016. **3;352(6290):1225-8**.
110. Gardini, A., *Global Run-On Sequencing (GRO-Seq).* Methods Mol Biol, 2017. **1468**: p. 111-20.
111. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
112. Wang, Y., et al., *Nanopore sequencing technology, bioinformatics and applications.* Nat Biotechnol, 2021. **39**(11): p. 1348-1365.
113. Rhoads, A. and K.F. Au, *PacBio Sequencing and Its Applications.* Genomics Proteomics Bioinformatics, 2015. **13**(5): p. 278-89.
114. Edge, P., V. Bafna, and V. Bansal, *HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies.* Genome Res, 2017. **27**(5): p. 801-812.
115. Kravitz, S.N., et al., *Random allelic expression in the adult human body.* Cell Reports, 2023. **42**(1): p. 111945-111945.
116. Bajrami, E. and M. Spiroski, *Genomic Imprinting.* Open Access Maced J Med Sci, 2016. **4**(1): p. 181-4.
117. Dzierzak, E. and S. Philipsen, *Erythropoiesis: development and differentiation.* Cold Spring Harb Perspect Med, 2013. **3**(4): p. a011601.
118. Basak, A. and V.G. Sankaran, *Regulation of the fetal hemoglobin silencing factor BCL11A.* Ann N Y Acad Sci, 2016. **1368**(1): p. 25-30.
119. Sankaran, V.G. and S.H. Orkin, *The switch from fetal to adult hemoglobin.* Cold Spring Harb Perspect Med, 2013. **3**(1): p. a011643.
120. Chen, M.J., et al., *Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter.* Nature, 2009. **457**(7231): p. 887-91.
121. Boisset, J.C., et al., *In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium.* Nature, 2010. **464**(7285): p. 116-20.
122. Manwani, D. and J.J. Bieker, *The erythroblastic island.* Curr Top Dev Biol, 2008. **82**: p. 23-53.
123. Zivot, A., et al., *Erythropoiesis: insights into pathophysiology and treatments in 2017.* Mol Med, 2018. **24**(1): p. 11.
124. T, A., et al., *Isolation of putative progenitor endothelial cells for angiogenesis.* Science, 1997. **14;275(5302):964-7**.
125. CI, C., et al., *Antigenic analysis of hematopoiesis. III. A hematopoietic progenitor cell surface antigen defined by a monoclonal antibody raised against KG-1a cells.* . J Immunol, 1984. **133(1):157-65**.
126. Abrahamsen, J.F., et al., *Variation in cell yield and proliferative activity of positive selected human CD34+ bone marrow cells along the circadian time scale.* Eur J Haematol, 1998. **60**(1): p. 7-15.
127. Scott, C., et al., *Recapitulation of erythropoiesis in congenital dyserythropoietic anemia type I (CDA-I) identifies defects in differentiation and nucleolar abnormalities.* Haematologica, 2021. **106**(11): p. 2960-2970.
128. Dolznig, H., et al., *Establishment of normal, terminally differentiating mouse erythroid progenitors: molecular characterization by cDNA arrays.* FASEB J, 2001. **15**(8): p. 1442-4.
129. Fibach, E., et al., *Proliferation and maturation of human erythroid progenitors in liquid culture.* Blood, 1989. **73**(1): p. 100-103.

130. Migliaccio, G., et al., *In vitro mass production of human erythroid cells from the blood of normal donors and of thalassemic patients.* Blood Cells Mol Dis, 2002. **28**(2): p. 169-80.
131. Miharada, K., et al., *Efficient enucleation of erythroblasts differentiated in vitro from hematopoietic stem and progenitor cells.* Nat Biotechnol, 2006. **24**(10): p. 1255-6.
132. Giarratana, M.C., et al., *Ex vivo generation of fully mature human red blood cells from hematopoietic stem cells.* Nat Biotechnol, 2005. **23**(1): p. 69-74.
133. Ronzoni, L., et al., *Erythroid differentiation and maturation from peripheral CD34+ cells in liquid culture: cellular and molecular characterization.* Blood Cells Mol Dis, 2008. **40**(2): p. 148-55.
134. Bunn, H.F., *Erythropoietin.* Cold Spring Harb Perspect Med, 2013. **3**(3): p. a011619.
135. **M, C.** and **O. SH**, *Regulation of the beta-globin locus.* Curr Opin Genet Dev, 1993. **3(2):232-7**.
136. A, C. and M. P, *Regulation of the globin genes. .* Pediatr Res, 2022. **51(4):415-21**.
137. Brown, K.K., et al., *Deletion of an enhancer near DLX5 and DLX6 in a family with hearing loss, craniofacial defects, and an inv(7)(q21.3q35).* Human genetics, 2010. **127**(1): p. 19-31.
138. McCarroll, S.A., et al., *Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease.* Nature Genetics, 2008. **40**(9): p. 1107-1112.
139. Arbogast, T., et al., *Mouse models of 17q21.31 microdeletion and microduplication syndromes highlight the importance of Kansl1 for cognition.* PLOS Genetics, 2017. **13**(7): p. e1006886-e1006886.
140. Spineli-Silva, S., et al., *Distal deletion at 22q11.2 as differential diagnosis in Craniofacial Microsomia: Case report and literature review.* European Journal of Medical Genetics, 2017(August).
141. Koolen, D.A., et al., *Clinical and molecular delineation of the 17q21.31 microdeletion syndrome.* Journal of medical genetics, 2008. **45**(11): p. 710-20.
142. Koolen, D.A., et al., *Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome.* Nature Genetics, 2012. **44**(6): p. 639-641.
143. Lynch, M., et al., *Genetic drift, selection and the evolution of the mutation rate.* Nat Rev Genet, 2016. **17**(11): p. 704-714.
144. Marstrand, T.T. and J.D. Storey, *Identifying and mapping cell-type-specific chromatin programming of gene expression.* Proc Natl Acad Sci U S A, 2014. **111**(6): p. E645-54.
145. Corradin, O. and P. Scacheri, *Enhancer variants: evaluating functions in common disease.* Genome Med, 2014. **6(10):85**.
146. Bozhilov, Y.K., et al., *A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker.* Nat Commun, 2021. **12**(1): p. 3806.
147. M, D.G., et al., *A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter Enhanced Reader.* Science, 2006. **312(5777)**: p. 1215-7.
148. Wang, X., M.J. Cairns, and J. Yan, *Super-enhancers in transcriptional regulation and genome organization.* Nucleic Acids Res, 2019. **47**(22): p. 11481-11496.
149. Hentges, L.D., et al., *LanceOtron: a deep learning peak caller for genome sequencing experiments.* Bioinformatics, 2022. **38**(18): p. 4255-4263.
150. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
151. Sherry, S., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res., 2001. **29(1):308-11**.
152. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies.* Genome Res, 2010. **20**(1): p. 110-21.
153. Sebastian, A. and B. Contreras-Moreira, *footprintDB: a database of transcription factors with annotated cis elements and binding interfaces.* Bioinformatics, 2014. **30**(2): p. 258-65.
154. Pietzner, M., et al., *Mapping the proteo-genomic convergence of human diseases.* Science, 2021. **374**(6569): p. eabj1541.
155. van Arensbergen, J., et al., *High-throughput identification of human SNPs affecting regulatory element activity.* Nat Genet, 2019. **51**(7): p. 1160-1169.
156. Yang, Z., et al., *A semi-supervised model to predict regulatory effects of genetic variants at single nucleotide resolution using massively parallel reporter assays.* Bioinformatics, 2021. **37**(14): p. 1953-62.
157. Ding, Z., et al., *Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association.* PLoS Genet, 2014. **10**(11): p. e1004798.

158. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Mol Cell, 2010. **38**(4): p. 576-89.
159. Han, Z. and W. Li, *Enhancer RNA: What we know and what we can achieve.* Cell Prolif, 2022. **55**(4): p. e13202.
160. Kowalczyk, M.S., et al., *Intragenic enhancers act as alternative promoters.* Mol Cell, 2012. **45**(4): p. 447-58.
161. A.C, M., et al., *Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs.* Genome Biol, 2013. **14, R131**.
162. Hyle, J., et al., *Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping.* Nucleic Acids Res, 2019. **47**(13): p. 6699-6713.
163. Kadauke, S. and G.A. Blobel, *Chromatin loops in gene regulation.* Biochim Biophys Acta, 2009. **1789**(1): p. 17-25.
164. Vernimmen, D., et al., *Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression.* EMBO J, 2007. **26**(8): p. 2041-51.
165. Julian Banerji, Sandro Rusconi, and Walter Schaffner, *Expression of a β-globin gene is enhanced by remote SV40 DNA sequences,.* Cell,, 1981. **Volume 27, Issue 2, Part 1,**: p. Pages 299-308,.
166. Jubb, A.W., et al., *Glucocorticoid Receptor Binding Induces Rapid and Prolonged Large-Scale Chromatin Decompaction at Multiple Target Loci.* Cell Rep, 2017. **21**(11): p. 3022-3031.
167. Belaghzal, H., et al., *Liquid chromatin Hi-C characterizes compartment-dependent chromatin interaction dynamics.* Nature Genetics, 2021. **53**(3): p. 367-378.
168. Mukherjee S, Erickson H, and B. D, *Detection of DNA looping due to simultaneous interaction of a DNA-binding protein with two spatially separated binding sites on DNA.* Proc Natl Acad Sci U S A, 1988 **85(17):6287-91**.
169. Cullen KE, Kladde MP, and S. MA., *Interaction between transcription regulatory regions of prolactin chromatin.* Science, 1993. **Jul 9;261(5118):203-6**.
170. Acemel, R.D., et al., *Straightforward protocol for allele-specific chromatin conformation capture.* Gene, 2021. **767**: p. 145185.
171. Magoc, T. and S.L. Salzberg, *FLASH: fast length adjustment of short reads to improve genome assemblies.* Bioinformatics, 2011. **27**(21): p. 2957-63.
172. Chereji, R.V., T.D. Bryson, and S. Henikoff, *Quantitative MNase-seq accurately maps nucleosome occupancy levels.* Genome Biol, 2019. **20**(1): p. 198.
173. Barozzi, I., P. Bora, and M.J. Morelli, *Comparative evaluation of DNase-seq footprint identification strategies.* Front Genet, 2014. **5**: p. 278.
174. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
175. Downes, D.J., et al., *High-resolution targeted 3C interrogation of cis-regulatory element organization at genome-wide scale.* Nature Communications, 2021. **12**(1): p. 1-15.
176. Pantier, R., et al., *High-throughput sequencing SELEX for the determination of DNA-binding protein specificities in vitro.* STAR Protoc, 2022. **3**(3): p. 101490.
177. Zhu, F., et al., *The interaction landscape between transcription factors and the nucleosome.* Nature, 2018. **562**(7725): p. 76-81.
178. Xu, B., et al., *Acute depletion of CTCF rewires genome-wide chromatin accessibility.* Genome Biol, 2021. **22**(1): p. 244.
179. Rao, S.S.P., et al., *Cohesin Loss Eliminates All Loop Domains.* Cell, 2017. **171**(2): p. 305-320 e24.
180. Kamitaki, N., C.L. Usher, and S.A. McCarroll, *Using droplet digital PCR to analyze allele-specific RNA expression*, in *Methods in Molecular Biology*. 2018, Humana Press Inc. p. 401-422.
181. Ernst, J. and M. Kellis, *Chromatin-state discovery and genome annotation with ChromHMM.* Nat Protoc, 2017. **12**(12): p. 2478-2492.
182. Einarsson, H., et al., *Promoter sequence and architecture determine expression variability and confer robustness to genetic variants.* Elife, 2022. **11**.
183. Perner, F., et al., *Roles of JAK2 in Aging, Inflammation, Hematopoiesis and Malignant Transformation.* Cells, 2019. **8**(8).

184. Downes, D.J., et al., *An integrated platform to systematically identify causal variants and genes for polygenic human traits.* bioRxiv, 2020.
185. Franke, M., et al., *Formation of new chromatin domains determines pathogenicity of genomic duplications.* Nature, 2016. **538**(7624): p. 265-269.
186. Heinz, S., et al., *The selection and function of cell type-specific enhancers.* Nat Rev Mol Cell Biol, 2015. **16**(3): p. 144-54.
187. Cheng, H., Z. Zheng, and T. Cheng, *New paradigms on hematopoietic stem cell differentiation.* Protein Cell, 2020. **11**(1): p. 34-44.
188. Corces, M.R., et al., *Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution.* Nat Genet, 2016. **48**(10): p. 1193-203.
189. Ludwig, L.S., et al., *Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis.* Cell Rep, 2019. **27**(11): p. 3228-3240 e7.
190. Chapman, O.S., et al., *Circular extrachromosomal DNA promotes tumor heterogeneity in high-risk medulloblastoma.* Nat Genet, 2023. **55**(12): p. 2189-2199.
191. Hao, Y., et al., *Integrated analysis of multimodal single-cell data.* Cell, 2021. **184**(13): p. 3573-3587 e29.
192. Ulirsch, J.C., et al., *Interrogation of human hematopoiesis at single-cell and single-variant resolution.* Nat Genet, 2019. **51**(4): p. 683-693.
193. Koczera, P., et al., *The Ribonuclease A Superfamily in Humans: Canonical RNases as the Buttress of Innate Immunity.* Int J Mol Sci, 2016. **17**(8).
194. Papayannopoulos, V., *Neutrophil extracellular traps in immunity and disease.* Nat Rev Immunol, 2018. **18**(2): p. 134-147.
195. Zheng, G.X., et al., *Massively parallel digital transcriptional profiling of single cells.* Nat Commun, 2017. **8**: p. 14049.
196. Dore, L.C. and J.D. Crispino, *Transcription factor networks in erythroid cell and megakaryocyte development.* Blood, 2011. **118**(2): p. 231-9.
197. Granja, J.M., et al., *ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis.* Nat Genet, 2021. **53**(3): p. 403-411.
198. Caprioli, A., et al., *Nkx2-5 represses Gata1 gene expression and modulates the cellular fate of cardiac progenitors during embryogenesis.* Circulation, 2011. **123**(15): p. 1633-41.
199. Caiulo, *<Caiulo-1991-Mapping-the-gene-encoding-the-human.pdf>.*
200. Lausen, J., et al., *Targets of the Tal1 transcription factor in erythrocytes: E2 ubiquitin conjugase regulation by Tal1.* J Biol Chem, 2010. **285**(8): p. 5338-46.
201. Korsgren, *<korsgren-cohen-1991-organization-of-the-gene-for-human-erythrocyte-membrane-protein-4-2-structural-similarities-with.pdf>.*
202. Heybrock, S., et al., *Lysosomal integral membrane protein-2 (LIMP-2/SCARB2) is involved in lysosomal cholesterol export.* Nat Commun, 2019. **10**(1): p. 3521.
203. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
204. Nord, A.S. and A.E. West, *Neurobiological functions of transcriptional enhancers.* Nat Neurosci, 2020. **23**(1): p. 5-14.
205. Gerasimavicius, L., B.J. Livesey, and J.A. Marsh, *Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure.* Nat Commun, 2022. **13**(1): p. 3895.
206. Mettananda, S., et al., *Editing an α-globin enhancer in primary human hematopoietic stem cells as a treatment for β-Thalassemia.* Nature Communications, 2017. **8**(1).
207. Esrick, E.B., et al., *Post-Transcriptional Genetic Silencing of BCL11A to Treat Sickle Cell Disease.* N Engl J Med, 2021. **384**(3): p. 205-215.
208. Bauer, D.E., et al., *An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level.* Science, 2013. **342**(6155): p. 253-257.
209. Bauer, D.E. and S.H. Orkin, *Update on fetal hemoglobin gene regulation in hemoglobinopathies.* Curr Opin Pediatr, 2011. **23**(1): p. 1-8.

210. Chaisson, M.J.P., et al., *Multi-platform discovery of haplotype-resolved structural variation in human genomes.* bioRxiv, 2018: p. 193144-193144.
211. Sanders, A.D., et al., *Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs.* Nature Protocols, 2017. **12**(6): p. 1151-1176.
212. Shao, H., et al., *Ongoing human chromosome end extension revealed by analysis of BioNano and nanopore data.* Sci Rep, 2018. **8**(1): p. 16616.
213. van Arensbergen, J., et al., *Genome-wide mapping of autonomous promoter activity in human cells.* Nat Biotechnol, 2017. **35**(2): p. 145-153.
214. Nguyen, T.A., et al., *High-throughput functional comparison of promoter and enhancer activities.* Genome Res, 2016. **26**(8): p. 1023-33.
215. Andersson, R. and A. Sandelin, *Determinants of enhancer and promoter activities of regulatory elements.* Nat Rev Genet, 2020. **21**(2): p. 71-87.
216. Hanssen, L.L.P., et al., *Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo.* Nature Cell Biology, 2017. **19**(8): p. 952-961.
217. de Wit, E., et al., *CTCF Binding Polarity Determines Chromatin Looping.* Mol Cell, 2015. **60**(4): p. 676-84.
218. Stolper, R.J., et al., *Loop extrusion by cohesin plays a key role in enhancer-activated gene expression during differentiation.* bioRxiv, 2023.
219. Satpathy, A.T., et al., *Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion.* Nat Biotechnol, 2019. **37**(8): p. 925-936.
220. Amemiya, H.M., A. Kundaje, and A.P. Boyle, *The ENCODE Blacklist: Identification of Problematic Regions of the Genome.* Sci Rep, 2019. **9**(1): p. 9354.
221. Ramirez, F., et al., *deepTools2: a next generation web server for deep-sequencing data analysis.* Nucleic Acids Res, 2016. **44**(W1): p. W160-5.
222. Wang, Y., et al., *The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions.* Genome Biol, 2018. **19**(1): p. 151.