Transformer-based model to decode regulatory instructions across cell types and species

Andrea Paz Rodríguez Delherbe

Green Templeton College

University of Oxford



A thesis submitted for the degree of

Doctor of Philosophy

Supervised by Jim Hughes, Christopher Yau, Simone Riva,
Emine Gür, and Tatjana Sauka-Spengler

Trinity 2026

Acknowledgements

I would like to sincerely thank my supervisors, James Hughes, Christopher Yau, Emine Ravza Gur, and Simone Riva, for their guidance, encouragement, and support throughout this project, particularly during uncertain and challenging periods of my doctoral training. Their kindness, understanding, and continued support helped me through some of the most difficult moments of this journey. Their insight and feedback have been invaluable both to this thesis and to my development as a researcher.

I would especially like to thank Edward Sanders and Ron Schwessinger, whose previous work in the lab established a strong foundation for many of the analyses presented in this thesis. I am also deeply grateful for their guidance, technical support, and willingness to answer questions throughout this work.

I am very grateful to Jeanne Inchauspe, Sally SiSi Qu, and Tom Wilson, whose work overlapped with mine and fostered a genuine sense of teamwork throughout our trainings. Sharing ideas and complementing each other's expertise made this experience both productive and enjoyable.

I would also like to thank Martin Sergeant for his support with the MSV/MSD visualisations included in this thesis, and Elisa Barrow for bridging the gap between my in silico analyses and in vitro experimentation. More broadly, I am grateful to everyone in the Hughes Lab for creating such a supportive, collaborative, welcoming, and intellectually stimulating research environment.

Outside of the Hughes Lab, many people also influenced and supported my work throughout my D.Phil. In particular, I would like to thank Alastair Smith, Catherine Chahrour, and Lena Morrill Gavarro for valuable discussions and brainstorming sessions on computational methods and analysis strategies.

I am especially grateful to Christina Reith and Alison Franklin for their advice, guidance, kindness, and support throughout my doctoral training. Their encouragement and willingness to listen during difficult periods provided not only practical guidance, but also reassurance and perspective when it was most needed.

I also acknowledge the Agencia Nacional de Investigación y Desarrollo (ANID). Without Becas Chile, this thesis would not have been possible.

To my friends, and colleagues; Julian Mena Urmeneta, Alejandra Vergara, Elisa Sandoval, David Rodriguez, Ignacia Heredia, Andres Gonzalez, Franco Vizcarra, Fernanda Cid, Mary Durin, Aimee Paterson, Lisa Leinhos, Adam Suttle, Sol Fernandez, Stephanie Winkler, Marlene Galaz, Kristian Gurashi, Ricardo Gonzales, Maira Alvarez, Elisa Armbrecht, Felix Radtke, Andrew McIntyre, Matthew Baxter, Lauren Murphy, Florian Laforets, Lebo Moshanyana, Ana Dopico, Seda Ozturk, Zayra Hernandez, Lior Pitowski, Francisco Calderon, Veronica Puga, Miguel Moctezuma, and everyone I have been fortunate

enough to meet and grow close to during my time as a D.Phil. student at the University of Oxford. Thank you all for the encouragement, conversations, friendship, and memorable moments that made this journey so meaningful and enjoyable.

Finally, I would like to thank my loved ones: Julian Mena, Cecilia Delherbe, Ivan Rodriguez, Paula Rodriguez, Felipe Collante, and everyone in the Rodriguez and Delherbe family, for their unwavering support and belief in me over the years. Their patience, encouragement, and love made this thesis possible.

Abstract

This thesis investigates how deep learning can be used to interpret regulatory information encoded within DNA sequence, reconstruct cell-type-specific regulatory landscapes, and uncover generalisable principles of gene regulation across species. Modern sequence-based models can predict epigenetic signals directly from DNA sequence, enabling the study of regulatory mechanisms at high resolution. However, important challenges remain in systematically benchmarking modern architectures, improving biological interpretability, and understanding how regulatory grammar is conserved and transferable across evolutionary distances.

To address these questions, this work develops and evaluates deep learning frameworks for modelling regulatory activity from DNA sequence, with a particular focus on erythroid biology as an experimentally well-characterised biological system. A benchmarking framework is established to compare convolutional and transformer-based architectures for predicting CTCF binding from high-resolution erythroid datasets. Models are evaluated using signal reconstruction, motif recovery, nucleotide importance, and variant effect prediction to assess both predictive performance and biological interpretability. These analyses show that transformer-based architectures, particularly HYPER, more effectively capture long-range regulatory dependencies and high-resolution epigenetic signals.

Building on this foundation, the selected framework is extended to model more complex erythroid regulatory programs, including transcription factor binding and chromatin accessibility. The results show that sequence-based models can accurately reconstruct both activating and repressive chromatin states while capturing combinatorial transcription factor interactions central to erythropoiesis, particularly GATA1-TAL1 regulatory networks. Integration of attribution methods and motif analyses further demonstrates that the learned representations recover biologically meaningful regulatory syntax, including motif spacing, cooperative interactions, and nucleotide-level determinants of regulatory activity.

Finally, this thesis investigates the conservation and transferability of regulatory grammar across species through comparative analyses of human, mouse, and chicken erythroid regulatory landscapes. Cross-species prediction experiments reveal that regulatory activity is more strongly associated with

conserved transcription factor interactions and regulatory sequence organisation than with exact sequence identity. Although predictive performance decreases with phylogenetic divergence, the models retain the ability to identify conserved regulatory elements and distinguish lineage-specific adaptations, highlighting their potential for predicting regulatory activity in species lacking experimental epigenomic data.

Overall, this work provides a systematically benchmarked and interpretable framework for studying sequence-driven gene regulation across cell types and species. By integrating predictive modelling, mechanistic interpretation, and comparative analysis, this thesis advances the understanding of regulatory grammar, its evolutionary conservation, and the application of deep learning in functional genomics.

Keywords— Deep learning - Epigenomics - Sequence-based modelling - Evolutionary conservation

Table of Contents

1 Introduction **1**

1.1 Overview of gene regulation 1

1.2 Epigenetics driven by cell-type-specific contexts 3

1.3 Chromatin organisation and regulatory elements 4

1.4 Experimental assays for capturing gene regulation 7

1.4.1 Using open chromatin to map the genome-wide regulatory activity . . . 7

1.4.2 Mapping the protein binding and the subclassification of open chromatin sites 9

1.5 Computational modelling of gene regulation 10

1.5.1 Machine learning frameworks for regulatory genomics 10

1.5.2 Deep learning approaches for sequence-based modelling 11

1.5.3 Architectural inductive biases in regulatory genomics 12

1.5.4 Limitations and challenges in current computational approaches 13

1.6 Biological model system to develop and validate regulatory prediction approaches 14

1.6.1 Regulatory mechanisms in erythropoiesis 14

1.6.2 Globin loci as erythroid-specific regulatory models 15

1.7 Gene regulation through evolution 17

1.7.1 Sequence conservation in regulatory elements 18

1.7.2 The potential of cross-species prediction 19

1.8 Globin clusters conservation across vertebrates 20

1.9 Challenges 24

1.10 Overall thesis aims 26

2 Methods **29**

2.1 Introduction 29

2.2 Red blood cells data preprocessing 30

2.3 Generation of training datasets 33

2.3.1 DeepSEA-based model preprocessing 33

2.3.2 BPNet preprocessing 34

2.3.3 Basenji preprocessing 35

2.4 Machine learning-related concepts . . . 36
2.5 Deep learning approaches for epigenetics modelling . . . 45
2.5.1 DeepSEA-based model for binding sites classification . . . 45
2.5.2 BPNet for protein binding regression . . . 47
2.5.3 Enformer for gene expression and epigenetics prediction . . . 49
2.5.4 REnformer for chromatin accessibility and protein binding prediction . 50
2.6 Peak calling . . . 54
2.6.1 LanceOtron . . . 54
2.6.2 MACS2 and MACS3 . . . 56
2.6.3 Hybrid log-normal peak calling . . . 57
2.7 Mappability stress test: Fake Fastqs . . . 60
2.8 Feature attribution methods . . . 62
2.9 Motif identification tools . . . 65
2.9.1 TF-MoDISco . . . 66
2.9.2 TomTom . . . 66
2.9.3 FIMO . . . 67
2.9.4 Motif databases . . . 68
2.9.5 Nucleotide importances to find motifs . . . 68
2.10 Distribution similarity metrics . . . 70

3 Comparative evaluation of deep learning architectures for predicting CTCF binding 76
3.1 Introduction . . . 76
3.2 Description of the DL architecture . . . 78
3.3 Available datasets . . . 80
3.4 Models comparison framework . . . 83
3.4.1 Local evaluation/assessment of CTCF binding predictions . . . 84
3.4.2 Fine-scale modelling of CTCF binding sites . . . 87
3.4.3 Generalisation to an unseen CTCF genetic variation . . . 90
3.4.4 Comparisons summary . . . 93
3.5 Extended genome-wide validation . . . 94
3.5.1 CTCF signal peaks similarity . . . 95
3.5.2 HYPER for genome-wide SNP generalisation . . . 101

3.6 Discussion and conclusion 113

4 Evaluation of HYPER for predicting chromatin accessibility and transcription factor binding signals 117

4.1 Introduction 117
4.2 Available datasets 121
4.3 HYPER model for predicting epigenetic signal 125
4.3.1 Model training 125
4.3.2 Evaluation of HYPER for prediction of TF protein binding 126
4.3.3 Evaluation of HYPER for prediction of chromatin accessibility 138
4.3.4 Model limitations 143
4.3.5 Reconstruction of silenced DNA regions 144
4.3.6 Generalisation to unseen sequences 149
4.3.7 Binding motif enrichment 160
4.4 Identification of mutations modulating enhancer activity 162
4.4.1 Most damaging in situ mutations in the alpha-globin locus 164
4.4.2 Most damaging in situ mutations in the alpha-globin MCS-R elements . 168
4.5 Discussion and conclusions 172

5 Transference of regulatory grammar across species 175

5.1 Introduction 175
5.2 Available datasets 180
5.2.1 Mouse erythroid-lineage 180
5.2.2 Chicken erythroid-lineage 183
5.2.3 Ancient hominin genomes 186
5.2.4 Mammal and avian genomes 187
5.3 HYPER model for predicting epigenetic signal in mouse 188
5.3.1 Model training 189
5.3.2 Reconstruction of the murine alpha globin locus 190
5.4 HYPER model for predicting epigenetic signal in chicken 193
5.4.1 Model training 193
5.4.2 Reconstruction of the chicken alpha globin locus 194
5.5 Generalising epigenetic signals in cross-species contexts 197

5.5.1 Human alpha globin track . . . 198
5.5.2 Cross-species nucleotide importance in R2 . . . 200
5.5.3 Species-specific nucleotide importance in the alpha-globin locus . . . 201
5.5.4 Cross-species motif enrichment in R2 . . . 204
5.5.5 Species-specific motif enrichment in the alpha-globin locus . . . 207
5.5.6 Summary of predictions with cross-species models . . . 211
5.6 Regulatory evolution of the alpha globin locus . . . 212
5.6.1 Recent evolutionary changes in the hominin alpha-globin locus . . . 213
5.6.2 Extended analysis throughout vertebrates . . . 217
5.7 Discussion and conclusions . . . 223

6 Discussion **228**
6.1 Overview and objectives . . . 228
6.2 Approach and analysis . . . 229
6.3 Key findings and implications . . . 230
6.4 Future work . . . 231
6.4.1 Experimental validation . . . 231
6.4.2 Data quality and resolution . . . 232
6.4.3 Extending the biological scope . . . 232
6.4.4 Attribution methods for transformers . . . 234
6.4.5 Integration of genomic languages . . . 234
6.4.6 Deeper cross-species analysis . . . 235
6.4.7 Motif-based representation of DNA sequences . . . 236
6.4.8 Clinical translation . . . 238

A Appendix **239**

List of Figures

1.1 Chromatin accessibility scheme in eukaryotic cells 5
1.2 Overview of ChIP-seq and ATAC-seq experiments 8
1.3 Schematic diagram of the human alpha- and beta-globin loci 16
1.4 Structure of the alpha-globin locus in multiple mammal species 21
1.5 Schematic diagram of the murine alpha- and beta-globin loci 22
1.6 Schematic diagram of the chicken alpha- and beta-globin loci 23

2.1 Overview of DeepHAEM architecture 46
2.2 Overview of BPNet architecture 48
2.3 Overview of the REnformer architecture 51
2.4 Diagram of the HYPER architecture 53
2.5 Predictions generated by Enformer and REnformer for the genomic region chr10:22,241,840–22,356,528 in HUVEC cells 53
2.6 Diagram of the backpropagation-like process of DeepLIFT 65

3.1 Diagram illustrating the inputs and outputs of the three models 80
3.2 Open chromatin and TF binding at the murine β-globin locus 81
3.3 CTCF peaks and signal properties 82
3.4 Validation workflow for comparative evaluation of the three ML models 84
3.5 Predictions of the three CTCF models across the genomic region chr16:40,961–155,649 86
3.6 Predicted nucleotide importances of the motif CTCF_C2 89
3.7 Effect of the SNP at chr16:8,655,299 on CTCF protein binding 91
3.8 Predicted effects of the SNP at chr16:8,655,299 on CTCF binding 92
3.9 Workflow for signal segmentation on predicted tracks 95
3.10 Distributions of predicted and real binned coverage values 97
3.11 Genome browser view of the region chr16:1,300,000–2,000,000 displaying CTCF tracks 99
3.12 Called CTCF peak sets comparison and statistics 100
3.13 Divergence in binding coverage between control and mutant alleles across all SNPs 104
3.14 Heatmaps with the frequency of mutations across the three variant effect classes 105

3.15 Distribution of HYPER-predicted variant damage for CTCF SNPs 106
3.16 Evaluation of model performance in detecting variant effects of SNPs 108
3.17 Evaluation of SNPs' p-value influence in the prediction of variant effects 110
3.18 Predicted versus real skew of variant effects under optimised parameters 111
3.19 Heatmaps with the frequency with which variant effect classes are correctly and incorrectly predicted 112

4.1 Open chromatin and TF binding alpha-globin tracks in erythroid lineage cells . 123
4.2 Epigenetic profile heatmaps centered on peaks 124
4.3 JASPAR's PWMs for GATA1, NF-E2, and TAL1 motifs 124
4.4 Predicted alpha-globin protein binding signals using HYPER 127
4.5 Distribution of reconstructed GATA1 protein binding coverage 129
4.6 UCSC tracks of GATA1 protein binding and signal peaks across the beta-globin locus 131
4.7 Distribution of reconstructed NFE2 protein binding coverage 133
4.8 UCSC tracks of NF-E2 protein binding and signal peaks across the beta-globin locus 134
4.9 Distribution of reconstructed TAL1 protein binding coverage 135
4.10 UCSC tracks of TAL1 protein binding and signal peaks across the beta-globin locus 137
4.11 Predicted alpha-globin open chromatin signal using HYPER 139
4.12 Distribution of the reconstructed open chromatin signal 140
4.13 UCSC session with chromatin accessibility signal peaks across the beta-globin locus 141
4.14 Chromatin accessibility tracks in erythrocytes at the alpha-globin locus 144
4.15 Real and predicted epigenetic profiles across the centromeric region of chromosome 16 146
4.16 Real and predicted epigenetic profiles across the pericentromeric region of chromosome 16 148
4.17 Effect of the Vanuatu SNV on chromatin accessibility 150
4.18 Predicted effect of the Vanuatu variant on erythrocyte epigenetic profiles 151
4.19 Nucleotide importance scores and motif discovery at the Vanuatu SNP functional site 152

4.20 PPP1R3B locus showing combinations of SNP presence in chromatin accessibility tracks . . . 153
4.21 PPP1R3B locus with combinations of SNPs in predicted epigenetic profiles . . 154
4.22 Nucleotide importance in peak 1 (loss-of-function SNP) . . . 155
4.23 Nucleotide importance in peak 2 (gain-of-function SNP) . . . 156
4.24 Heatmap of motif contribution scores for PPP1R3B SNPs . . . 157
4.25 Heatmap of motif contribution scores for alpha-globin MCS-R elements . . . 161
4.26 Workflow illustrating the identification of the most damaging SNPs . . . 163
4.27 UCSC Genome Browser visualisation of alpha-globin variant effects . . . 165
4.28 Heatmaps illustrating the most damaging variant substitutions . . . 166
4.29 Nucleotide importance profiles for the top 11 highest gain-of-function variants . 168
4.30 Heatmaps displaying the most damaging variant substitutions within MCS-R elements . . . 169
4.31 Nucleotide importances of the most damaging loss-of-function SNPs in the MCS-R elements . . . 170

5.1 Open chromatin and TF binding at the murine β-globin locus . . . 182
5.2 Mouse epigenetic profile heatmaps centred on peaks . . . 183
5.3 Open chromatin and CTCF binding at the chicken beta-globin locus . . . 184
5.4 Chicken epigenetic profile heatmaps centred on peaks . . . 185
5.5 Timeline illustrating the estimated ages of the remains of the nine ancient hominins . . . 187
5.6 Evolutionary relationships between the 26 species . . . 188
5.7 Predicted epigenetic profiles across the mouse alpha-globin locus generated using both mouse and human models . . . 192
5.8 Predicted epigenetic profiles across the chicken alpha-globin locus generated using both chicken and human models . . . 196
5.9 Predicted epigenetic profiles across the human alpha-globin locus generated using both cross-species models . . . 200
5.10 Nucleotide importance profiles for the R2 enhancer across species . . . 201
5.11 Nucleotide importance profiles across alpha-globin MCS-R regulatory elements 203
5.12 Motif enrichment of the R2 enhancer across species using cross-species predictions . . . 205

5.13 Motif enrichment of the R2 enhancer across species using cross-species predictions . . . 208
5.14 Motif contribution similarities for human, mouse, and chicken enhancers predicted with the human model . . . 210
5.15 Predicted alpha-globin chromatin accessibility in ancient hominin sequences . . 215
5.16 Motif contribution similarities for human, mouse, and chicken enhancers predicted with the human model . . . 216
5.17 Predicted TAL1 binding profiles across the alpha-globin locus in multiple species 219
5.18 Predicted open chromatin profiles across the alpha-globin locus in multiple species . . . 221

6.1 UMAP projection of k-mer importance–based representations for regulatory element sequences . . . 237

S1 Distribution of log-transformed CTCF binding signal values in the training and test datasets . . . 239
S2 Histograms of TF binding and chromatin accessibility profiles across HYPER dataset splits . . . 240
S3 Density estimates of signal values for the mouse epigenetic tracks in training and test sets . . . 241
S4 Distribution of signal values for chicken tracks in training and test datasets . . . 242

List of Tables

3.1 Similarity metrics between predicted and experimental whole-tracks 87

3.2 Models' performance metrics for predicting the nucleotide importance of a CTCF_C2 instance. 90

3.3 Summary of the comparison results across the three validation levels. 93

3.4 Performance metrics of HYPER for predicting variant effects. 107

4.1 Similarity metrics between predicted and experimental ChIP-seq tracks for GATA1, NF-E2, and TAL1 128

4.2 Summary of genome-wide peak analysis metrics 138

4.3 Summary of the 11 SNPs with the highest damage scores 166

4.4 Summary of the most damaging loss-of-function SNPs in MCS-R enhancers . . 169

5.1 Functionally and sequence-conserved elements within the alpha-globin locus . 222

S1 Summary of the hominin individuals for the cross-species analysis 243

S2 List of species genomes used in the cross-species analysis 244

S3 UCSC genome browsers with cross-species predictions 245

List of Abbreviations

ATAC-seq assay for transposase-accessible chromatin followed by sequencing. v, 7–9, 19, 30, 31, 46, 49, 56, 79, 81, 118, 121, 124–126, 143, 155, 157, 158, 160, 161, 181–185, 189, 193, 194, 196, 212, 213, 215, 216, 241, 242

AUPRC area under the precision-recall curve. 44

AUROC area under the ROC curve. 44

BCE binary cross entropy. 43, 70

ChIP-seq chromatin immunoprecipitation followed by sequencing. v, ix, 8, 9, 27, 30, 31, 33–35, 46–49, 56, 57, 68, 79, 81, 86, 95, 96, 98, 99, 102, 103, 117, 118, 121–124, 126, 132, 143, 154, 180–185, 219

CNN convolutional neural network. 11, 12, 25, 27, 29, 39, 40, 45–47, 55, 76–79, 82, 84, 87–94, 113–116, 118, 119, 177, 229, 230

CRE cis-regulatory elements. 179

CWM contribution weight matrix. 66

DL deep learning. 6, 10–14, 24–27, 29, 30, 36, 49, 50, 54, 57, 60, 62, 70–74, 76–78, 113, 118, 143, 144, 147, 163, 177, 228–230

DNase-seq DNase I hypersensitive sites following sequencing. 8, 9, 35, 49, 79, 118, 121

EB embryoid bodies. 180

ECDF empirical cumulative distribution function. 73, 74

EPO erythropoietin. 233

FN false negatives. 43, 44

FP false positives. 43, 44

GAN generative adversarial network. 232

GELU Gaussian Linear Unit. 38, 42, 52

GM-CSF granulocyte-macrophage colony-stimulating factor. 15

GPU graphical processing unit. 50

GWAS genome-wide association study. 152, 162, 163, 166, 168, 170, 173, 174

Hi-C High-throughput Chromosome Conformation Capture. 45

HS hypersensitive site. 16, 17, 22, 23, 131, 134, 137, 141, 182, 184

IG integrated gradients. 62–64, 66

IL-3 interleukin-3. 15

JSD Jensen–Shannon divergence. 74, 75, 96, 98, 130, 132, 135, 136, 138, 140, 206

KL Kullback–Leibler. 72, 74, 75, 100, 115, 132, 134, 136, 138, 142

KNN k-neirest neighbours. 10

KS Kolmogorov–Smirnov. 73, 74

LAD lamina-associated domain. 5, 145, 146, 148, 149, 172, 174, 225

LBK Linearbandkeramik. 243

LCR locus control region. 15–17, 22–24

LCR median absolute deviation. 32

LD linkage disequilibrium. 105

LLM large language model. 234, 237

MAE mean absolute error. 42

MCS multi-species conserved sequence. iii, vii, ix, 15, 122, 123, 160, 161, 165, 168–171, 179, 194, 196, 197, 203, 207, 208, 211, 224

mES mouse embryonic stem. 180

ML machine learning. v, 9, 10, 26, 28–30, 36, 45, 50, 54, 60, 62, 63, 76, 78, 94, 175, 176, 228, 232

MLE maximum likelihood estimation. 58

MSE mean squared error. 42, 52, 70, 128

NGS next-generation sequencing. 7–9, 145

NN neural network. 11, 36–39, 63

NRMSE normalised root mean squared error. 86, 87, 93, 114, 128, 138

PBMCs peripheral blood mononuclear cells. 81

PCC Pearson correlation coefficient. 71, 90, 93

PDF probability density function. 58, 59

PFM position frequency matrix. 68

Pol II RNA polymerase II. 16

PWM position weight matrix. vi, 66–70, 87, 90, 114, 123

QTL quantitative trait locus. 102

RBC red blood cell. 14, 25, 27, 120, 121, 183

ReLU Rectified Linear Unit. 37, 38, 42, 47

RNA-seq ribonucleic acid sequencing. 7, 9, 81

RNN recurrent neural network. 11

ROC receiver operating characteristic. 44

scATAC-seq single cell ATAC-seq. 9, 33, 51, 52, 118

SCF stem cell factor. 14

SD standard deviation. 32

SNP single-nucleotide polymorphism. ii, v–vii, ix, 4, 25, 77, 84, 90–94, 101–115, 120, 121, 151–158, 162–166, 168–171, 173, 174, 213

SNV single-nucleotide variation. vi, 101, 112, 149, 150, 216

SVM support vector machine. 10, 176

TAD topologically-associated domain. 4, 5, 20, 49, 117

TF transcription factor. iii, v–viii, 1, 2, 4–12, 14, 15, 17–19, 22, 25–27, 41–43, 45, 48, 50, 68–70, 76, 77, 79, 80, 87, 88, 102, 103, 117–121, 123, 125, 126, 128, 137–139, 141–143, 146, 149, 150, 152–154, 156, 158–162, 168, 171–174, 179–182, 189, 190, 192–194, 196, 198, 203, 204, 206, 207, 209–213, 217, 225, 226, 228, 229, 231, 232, 235, 236

TFBS transcription factor binding site. 6, 20, 39, 65, 159, 173, 175, 182, 225, 228, 231

TN true negatives. 43, 44

TP true positives. 43, 44

1 | Introduction

1.1 Overview of gene regulation

Multicellular organisms contain identical genetic material in nearly every cell of the body [1]. Despite this shared genome, cells adopt highly specialised structures and functions, ranging from neurons and muscle cells to hepatocytes and erythrocytes. These differences arise from selective gene expression, whereby only specific subsets of genes are activated in each cell type [2]. During development and differentiation, multipotent stem cells undergo programmed cellular transitions, progressively acquiring transcriptional programs that define their identity and function in a tightly controlled spatiotemporal manner [3].

A central challenge in molecular biology is therefore to understand how this diversity of cellular phenotypes is encoded within a single, static DNA sequence. When the Human and C. Elegans genomes were sequenced, it was surprising to find that each organism contained roughly the same number of genes (~ 20000) [4], even though C. Elegans contains only $\sim 1,000$ somatic cells, whereas humans contain ~ 3 trillion cells that specify completely new organ systems. Therefore, the complexity of an organism does not scale with the number of genes; rather, it scales with the complexity with which they can be combined in different patterns of expression. Cell-type-specific gene expression patterns are determined by a complex regulatory system driven by the combinatorial expression of transcription factors (TFs) in the context of differential chromatin accessibility, epigenetic modifications and nuclear 3-dimensional interactions [5]. Together, these components form a multilayered regulatory landscape that governs when and where genes are expressed.

At the core of this system are cis-regulatory elements, which are generally classified as promoters, enhancers, and insulators, which integrate inputs from patterns of TFs, both cell-specific and general TFs, to control transcriptional output [6]. At specific gene loci, these elements operate as interconnected regulatory landscapes rather than in isolation [7]. For example, multiple enhancers can cooperate to regulate genes over long genomic distances through chromatin looping, while insulators constrain these interactions within defined domains and prevent misregulation of adjacent genes [8]. As a result, gene regulation depends not only on the presence of regulatory elements, but also on their number and specific function as well as their spatial

organisation and interactions within the genome [9].

An important question is what determines where cell-specific elements are established in the genome, and consequently, which genes are activated. Although TF binding motifs are abundant across the genome, only a small fraction are functionally bound *in vivo* [10], making it difficult to distinguish regulatory sites from non-functional sequence patterns across the genome. This is because TFs often bind cooperatively, stabilising each other at specific combinations of binding sites. DNA sequence grammar, the combinatorial arrangement of TF binding motifs through their composition, spacing, orientation, and contextual interactions, plays a central role in shaping gene regulation [11]. How this sequence grammar shapes the resulting regulatory landscape is referred to as the *regulatory grammar*. This is further influenced by the epigenetic environment, including methylation and chromatin state [12], as well as nuclear context, such as TF concentration and competition for binding sites [13].

Deciphering this system has been extremely challenging because each cell type operates with a distinct regulatory grammar that both shapes and is shaped by the pattern of TF expression [14]. This interplay determines where regulatory elements are established in the genome for that cell type, which in turn controls gene expression patterns and ultimately defines the cell's identity and function.

Each cell type is thus characterised by a unique combination of active TFs and the resulting pattern of regulatory elements, which must be either dynamically established during differentiation or stably maintained through cell divisions [15]. Although epigenetic mechanisms contribute to their formation, it is evident that regulatory information is also encoded directly in the DNA sequence as an evolved, cell–type–specific arrangement of TF binding motifs [16]. There has been a long effort to use computational methods to learn, or at least recognise, these codes and thereby predict functional elements from genome sequence, but the sheer complexity of the problem has limited their success.

A core aim of this thesis is to investigate whether next-generation computational approaches can learn these cell–type–specific codes from experimentally derived, cell–type–resolved epigenetic data. The ultimate objective is to predict which sequences can serve as effective platforms for cell–type–specific TF activity, regardless of their origin or species, and to identify the motifs within these sequences that are essential for that activity.

1.2 Epigenetics driven by cell-type-specific contexts

Epigenetics is the study of how cells regulate gene expression without altering the underlying DNA sequence, primarily through mechanisms that control when and where genes are activated or repressed [2]. These regulatory processes enable cells to respond to environmental and stimulatory factors, leading to changes in gene expression that can influence a wide range of biological functions [17]. At the same time, epigenetic regulation plays a fundamental role in establishing and maintaining cell identity. Different cell types perform specialised functions by activating distinct sets of genes that define their identity as they differentiate [18].

During cell differentiation, an immature, unspecialised stem cell develops into a mature and functionally specialised cell. Tightly controlled changes in gene expression govern this process [19], which drive the synthesis of specific proteins required for each cellular function. Once established, these expression patterns must be stably maintained across cell divisions, allowing daughter cells to retain the identity and function of their progenitors. The coordination of gene expression across differentiation, proliferation, and other biological processes relies on precise temporal and spatial regulation of transcriptional activity [20].

At a general level, epigenetic regulation operates through molecular mechanisms that influence the accessibility of different regions of the genome to the cellular machinery responsible for gene expression [21]. Chemical modifications to DNA and associated proteins can alter the local structural environment of the genome, thereby facilitating or restricting transcriptional activity [22].

Furthermore, epigenetic regulation provides a mechanism for cellular memory [23]. Once a gene expression program is established, epigenetic instructions that determine regulatory signals can be propagated through cell division, ensuring the stability of cell identity over time [1]. At the same time, these marks remain reversible, allowing cells to adapt to changing conditions or, in some cases, undergo reprogramming [24]. This balance between stability and plasticity is a defining feature of epigenetic control and is essential for both normal development and disease processes.

Taken together, epigenetic mechanisms provide a dynamic framework for establishing, maintaining, and modulating cell-type-specific gene expression. These processes enable cells to

respond to developmental and environmental cues while preserving stable transcriptional programs [25]. Understanding how epigenetic regulation is determined by and interacts with genome organisation and regulatory elements is therefore essential for deciphering gene regulatory mechanisms. Importantly, this framework also determines how genetic variation influences gene expression, as even small sequence changes, such as single-nucleotide polymorphisms (SNPs), can alter TF binding or chromatin accessibility in a cell-type-specific manner [26]. Consequently, a key challenge is to understand how regulatory systems respond to sequence variation and to what extent these effects are predictable across different genomic contexts.

1.3 Chromatin organisation and regulatory elements

In eukaryotic cells, genetic material is organised as DNA sequences wrapped around histone proteins. Together, DNA and histones form nucleosomes, which, along with structural RNAs and DNA-binding proteins, make up chromatin. Within chromosomes, chromatin exists in two structural states depending on its level of compaction [27]: heterochromatin and euchromatin. Heterochromatin is tightly condensed and is typically inactive, silenced, and closed. Conversely, euchromatin is more loosely packed and often transcriptionally active, accessible, and open. DNA contains coding and non-coding regions. Coding regions are segments that get translated into proteins, whereas non-coding regions do not encode proteins. Figure 1.1 illustrates the different structural states of chromatin, ranging from highly condensed forms to more open and accessible DNA, where proteins can bind to promote or repress transcription and other molecular processes.

While genes occupy less than 2% of the genome, the remaining non-coding DNA contains the majority of regulatory elements [28]. These elements encode instructions that initiate, enhance, or repress the expression of their target genes. Some of these instructions involve the recruitment of TFs that bind to specific DNA sequences to form enhancer-like elements [29], while others, such as CTCF, bind to chromatin and contribute to the structural organisation of the genome [30] or influence post-transcriptional processes [31].

The genome is organised into higher-order structures that help regulate these interactions. Topologically-associated domains (TADs) are regions of the genome within which DNA elements are more likely to interact with each other, thereby facilitating communication between

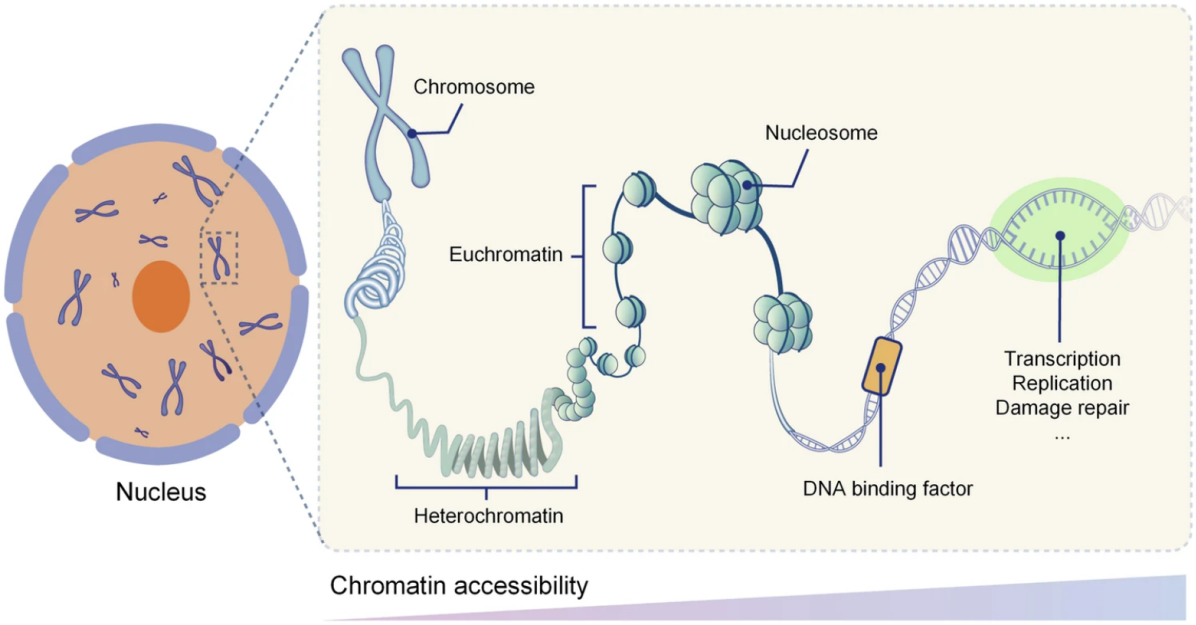


Figure 1.1: Chromatin accessibility scheme in eukaryotic cells. Most of the chromatin is inaccessible, whereas the accessible part is dynamically bound by histones, TFs, and other molecules, which are essential for maintaining genetic information and expressing it as proteins. *Source:* Reproduced from Figure 1 of Chen, 2024 [27] under the Creative Commons Attribution 4.0 License.

regulatory elements and their target genes [30]. In contrast, lamina-associated domains (LADs) are genomic regions that interact with the nuclear lamina and are generally associated with transcriptionally inactive or repressed chromatin. Together, these structural features contribute to the general spatial organisation of the genome and influence gene regulatory activity [32].

Within TADs in active regions, the combined action of multiple classes of cis-regulatory elements determines the genome's transcriptional output. These elements are currently thought to include promoters, that initiate transcription; enhancers, that increase transcriptional activity; silencers, that repress gene expression; insulators, that restrict regulatory interactions; and facilitators, that support stronger enhancers [33]. Although these elements are traditionally defined by their individual roles, they operate together, often in a position-and sometimes orientation-, dependent manner, to determine expression levels [34].

Predicting cell-type–specific gene expression from sequence has long been a central goal in science, but it remains highly challenging due to the system's complexity. One potential approach is to decompose this overarching problem into more manageable subproblems. In particular, TF binding, combinatorial motif interactions, and long-range regulatory dependence represent key features that fundamental computational models must capture together.

These properties suggest that computational architectures designed for regulatory genomics must be capable of modelling interactions across large genomic distances while preserving local sequence resolution. This requirement is particularly relevant for enhancer-promoter communication, cooperative TF binding, and chromatin domain organisation, where regulatory activity often depends on relationships between distal sequence elements rather than purely local motif content. In this context, transformer-based architectures are especially attractive because self-attention mechanisms allow the model to dynamically weight interactions between distant genomic regions, potentially capturing aspects of regulatory organisation analogous to higher-order chromatin interactions observed *in vivo*.

This raises the possibility that architectural properties of DL models may influence not only predictive performance, but also the types of regulatory relationships that can be represented and inferred from genomic sequence.

A further complexity is that the elements, such as enhancers and boundary elements, increasingly appear to come in different flavours, which standard genome annotation cannot account for [35]. Recent studies suggest that enhancers can be further subdivided into enhancers and facilitators and can also exist as single elements or in clusters known as locus control regions or super-enhancers [36, 37]. Likewise, these elements can form transcriptional hubs [38], compartments that contain a dense amount of TFs and cofactors, and binding sites. It is thought that transcription is activated by the recruitment of the pre-initiation complex to the promoter, triggered by transcriptional hubs and promoters coming into proximity [39].

These properties are directly relevant to this thesis, which aims to benchmark modelling approaches for TF binding, characterise the regulatory grammar underlying these binding patterns –including the identities and combinations of TFBSs, their organisation into regulatory elements, and the higher-order arrangement of these elements– and evaluate how these learned regulatory rules generalise across contexts and species. This suggests that the most suitable computational architectures for predicting binding and epigenetic states should be capable of learning from sufficiently long stretches of genomic sequence, capturing both interactions between elements and the sequences within the elements themselves.

Beyond predicting TF binding to DNA, it would also be highly desirable for computational models to capture the emergent outcomes of this binding, such as chromatin accessibility and histone modifications, which provide a broader view of genome activity than any single factor's

binding alone.

These regulatory signals cannot be directly observed within the nucleus and must instead be inferred through experimental assays that capture different facets of gene regulation [40]. As a result, these data types typically serve as the foundation for training computational models. Advances in high-throughput sequencing technologies have made it possible to profile TF binding, chromatin accessibility, and gene expression across the genome in diverse cell types. Together, these approaches generate large-scale, information-rich datasets that provide indirect readouts of regulatory activity, underpinning both biological interpretation and computational modelling.

1.4 Experimental assays for capturing gene regulation

Next-generation sequencing (NGS) technologies enable the parallel processing of millions of DNA fragments [41] and are typically applied to large, homogeneous cell populations, commonly referred to as bulk samples. Among these methods, ribonucleic acid sequencing (RNA-seq) is widely used to measure RNA molecules within a biological sample, allowing the identification of expressed genes and the quantification of their expression levels [42].

1.4.1 Using open chromatin to map the genome-wide regulatory activity

Open chromatin can be used to map general regulatory activity in a cell–type–specific manner at the genome scale [43]. Because chromatin accessibility is a biophysical property closely associated with protein binding to DNA in its chromatin context, measuring the degree of openness provides a genome-wide, cell–type–specific map of chromatin binding. Transposases are enzymes that cut and insert DNA sequences, and they do so in a manner that depends on chromatin accessibility. The assay for transposase-accessible chromatin followed by sequencing (ATAC-seq) [44] is a NGS technique that uses a hyperactive mutant of the Tn5 transposase with reduced sequence bias to probe genome-wide chromatin accessibility. In addition to fragmenting accessible DNA, Tn5 inserts sequencing adapters into these regions in a process known as tagmentation, enabling their subsequent amplification and sequencing. ATAC-seq thus provides a global profile of chromatin accessibility for a population of cells, independent of the

specific proteins bound to the DNA. A brief overview of the inputs and outputs of ATAC-seq experiments is shown in Figure 1.2B.

However, genetic mapping studies have shown that, unlike previous assumptions, Tn5 does not integrate adapters completely randomly into open DNA regions [45]. Instead, this enzyme exhibits some preference for certain DNA sequences, chromatin structures, and positioning [46]. For instance, Tn5 has been reported to favour GC- or AT-rich regions [47] with flexible structures [46]. Such intrinsic bias is thought to distort the representation of chromatin accessibility, posing challenges for advanced analyses, such as footprinting, that aim to detect protein-binding footprints within hypersensitive regions [48]

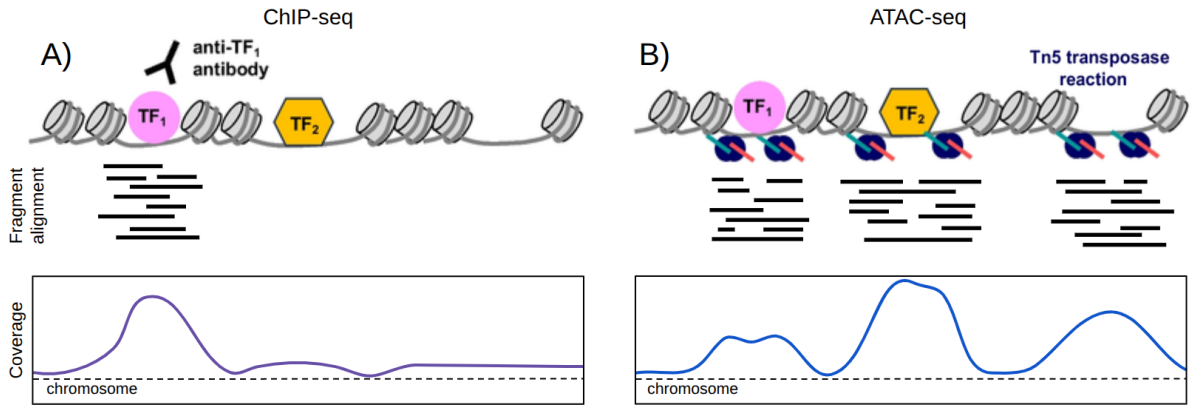


Figure 1.2: Overview of ChIP-seq and ATAC-seq experiments. A) ChIP-seq experiments employ an antibody targeted to a specific TF. After the protocol, the output data can be visualised as a snapshot in blue, where the peaks represent a high enrichment of binding for the targeted protein. B) ATAC-seq experiments employ Tn5 transposase to identify all locations in the sample genome where the chromatin is open. After the protocol, the output data can be visualised as a snapshot in orange, where the peaks represent a high enrichment of general chromatin accessibility. *Source:* Adapted from Figure 1 of Hojo, 2023 [49] under the Creative Commons Attribution 4.0 License.

DNase I, an endonuclease, is another enzyme that cuts DNA in a chromatin-accessibility-dependent manner [50]. This is employed in another NGS technique to detect open chromatin called DNase I hypersensitive sites following sequencing (DNase-seq). Unlike Tn5, DNase I does not incorporate sequencing adapters; therefore, DNA ends need to undergo end repair and adaptor ligation before amplification and sequencing, which increases the number of cells required [51]. Although DNase I exhibits some sequence preferences for short sequences near its cleavage sites, its bias has been reported to have a lesser impact on footprinting performance than that of Tn5, and so is advantageous for footprinting-based analyses [52]. Nevertheless, ATAC-seq offers a substantially simpler workflow and requires fewer input cells

(around 5000) [53], whereas DNase-seq typically requires millions of cells [54]. This technical simplicity has also facilitated the rapid and widespread development of ATAC-seq single-cell approaches [55].

Single-cell ATAC-seq (scATAC-seq) is a high-throughput sequencing approach that characterises genome-wide chromatin accessibility in thousands of individual cells [56], and it can be combined with RNA-seq in the same cell (multiome). Its key advantage is that it can be applied to heterogeneous cell populations, which can then be computationally separated into distinct cell clusters; integrating RNA-seq data further enhances this resolution and interpretability [57]. This approach removes the need for physical purification of cell types from tissues or complex mixtures, often difficult or impossible, and enables the identification and characterisation of previously unknown cell types within such populations.

1.4.2 Mapping the protein binding and the subclassification of open chromatin sites

While chromatin accessibility assays reveal which genomic regions are accessible, they do not directly identify the specific proteins bound to those regions or even their likely function. They therefore don't retrieve information on individual TF bindings or histone modifications that would indicate the class of cis-element detected. However, other NGS techniques allow scientists to map the binding sites of specific target proteins or chromatin modifications. This method is known as chromatin immunoprecipitation followed by sequencing (ChIP-seq) [58], which involves using an antibody to enrich for chemically crosslinked and fragmented chromatin bound by a target protein or nucleosomes modified in a specific manner. A brief overview of the inputs and outputs of ChIP-seq experiments is depicted in Figure 1.2A.

These technologies continue to advance at a fast pace, now accounting for the sequencing of 100,000 terabases per year [59]. In addition, the creation of new experimental protocols and the inherent complexity of these high-dimensional, sparse datasets pose a significant challenge for researchers. To address these issues, high-performance computing approaches, including machine learning (ML), have become essential tools for the efficient and scalable analysis of such data. These techniques facilitate the identification of complex patterns and statistical relationships, ultimately assisting scientists in translating their experimental results into meaningful

biological insights.

However, modelling gene regulation depends not only on the accuracy, resolution, and consistency of high-throughput datasets, but also on cellular context. Because regulatory grammar varies between cell types, TF expression patterns differ, leading to distinct regulatory element landscapes and motif usage. As a result, the interpretation of phenomena such as non-coding genetic variation must be carried out in a cell-type–specific manner: a sequence variant may influence TF binding or chromatin accessibility in one cell type while having little or no effect in another. This context dependence complicates the inference of general regulatory principles and underscores the importance of studying gene regulation within well-defined, biologically coherent systems. Two key requirements, therefore, guide the framework of this thesis: first, the use of a well-characterised biological system with thoroughly studied gene loci for testing and validation; and second, the selection of an appropriate computational architecture as a foundation for methodological development.

1.5 Computational modelling of gene regulation

1.5.1 Machine learning frameworks for regulatory genomics

Computational approaches play a central role in modern genomics by enabling the integration and interpretation of large-scale molecular data. In particular, ML and more recently deep learning (DL), provide powerful frameworks for modelling gene regulation directly from DNA sequence and epigenomic data. By capturing nonlinear relationships and higher-order interactions, these methods offer the potential to bridge the gap between sequence-level information and functional regulatory outcomes.

The development of ML approaches in genomics has progressed alongside advances in sequencing technologies and computational infrastructure. Early applications relied on classical algorithms such as SVMs, random forests, KNN, and Bayesian methods. These models were successfully applied to tasks including gene function prediction, sequence annotation, and microarray analysis [60–63]. However, they depended heavily on handcrafted features and prior biological knowledge, limiting their ability to capture complex, unsuspected hierarchical relationships directly from raw genomic data. As sequencing technologies rapidly expanded the

scale and dimensionality of available datasets, these limitations became increasingly apparent, motivating the development of more flexible and expressive modelling frameworks.

The introduction of neural networks (NNs) marked a transition toward representation learning, enabling models to automatically learn and extract informative features from data. Early NN applications in genomics, typically based on shallow architectures such as multilayer perceptrons [64, 65], demonstrated improved predictive capabilities but remained constrained by limited depth and continued reliance on highly engineered and annotated inputs. These constraints restricted their ability to model the combinatorial and context-dependent nature of gene regulation.

1.5.2 Deep learning approaches for sequence-based modelling

A major advance came with the adoption of deep NNs, particularly convolutional neural networks (CNNs), which are well suited to learning local sequence features such as TF binding motifs. Models such as DeepBind [66] and DeepSEA [67] demonstrated that regulatory features could be predicted directly from raw DNA sequence, while later approaches such as BPNet extended this framework to base-resolution prediction of TF binding profiles [68]. These developments established DL as a powerful paradigm for modelling regulatory genomics in a data-driven manner.

Despite their success, CNN-based models have inherent limitations. Their reliance on local receptive fields (amount of sequence the model can see) restricts their ability to capture longer-range regulatory interactions, which are fundamental to gene regulation through mechanisms such as super-enhancer clusters, enhancer–promoter looping, and CTCF interactions. Sequential models such as recurrent neural networks (RNNs) were introduced to address this limitation by modelling dependencies across longer genomic regions [69,70]. However, these approaches treat the data as purely linear, whereas DNA operates in three dimensions, where important regulatory interactions can occur between distant regions that are brought into proximity. This limits their ability to capture key long-range relationships, and their high computational cost has further restricted their widespread use.

More recently, attention-based architectures have emerged as a solution to these challenges [71]. By allowing models to dynamically weight interactions between distant regions of the

genome, attention mechanisms enable the capture of long-range and context-dependent regulatory relationships [72]. These approaches can be broadly divided into hybrid CNN–attention models and transformer-based architectures that rely primarily on self-attention mechanisms. Transformer models such as Enformer [73] have demonstrated the ability to predict gene expression and epigenetic features across large genomic contexts, while genomic foundation models such as DNABERT [74] aim to learn general representations of regulatory sequence through large-scale self-supervised training.

1.5.3 Architectural inductive biases in regulatory genomics

Different DL architectures make distinct assumptions about how regulatory information is organised within genomic sequence. These assumptions, commonly referred to as inductive biases, influence both predictive performance and the types of biological relationships that models are able to capture.

CNNs are particularly effective at identifying local sequence features such as TF binding motifs because their receptive fields operate over relatively short contiguous regions of DNA. As a result, CNN-based architectures implicitly assume that regulatory activity can largely be inferred from local motif composition and nearby sequence context.

Architectures such as BPNet extend this framework through the use of stacked dilated convolutions, enabling the hierarchical integration of motif interactions across broader genomic regions while preserving base-pair resolution. This introduces an inductive bias favouring progressively more complex combinations of local regulatory features and cooperative motif interactions.

In contrast, transformer-based architectures rely on self-attention mechanisms that allow interactions to be modelled between distant sequence elements across large genomic contexts. These models therefore assume that regulatory activity depends not only on local motif syntax, but also on distributed and long-range contextual relationships analogous to enhancer–promoter communication, chromatin domain organisation, and higher-order regulatory interactions observed *in vivo*.

Understanding these differing architectural assumptions is important for interpreting model behaviour and benchmarking predictive performance in regulatory genomics tasks.

Consequently, differences in model performance may reflect not only optimisation capacity,

but also the extent to which a model's inductive biases align with the biological organisation of regulatory systems.

1.5.4 Limitations and challenges in current computational approaches

Each modelling strategy varies from other approaches in its ability to capture specific features of the data. Architectures focused on local sequence features are highly effective at identifying motif-level signals [75], whereas models with larger receptive fields are better suited to capturing distal regulatory interactions [73]. However, increasing model complexity and context size often comes at the cost of interpretability and computational efficiency [76]. As a result, there is currently no consensus on a single optimal modelling approach for gene regulation, and performance comparisons across studies are complicated by differences in datasets, evaluation strategies, and biological systems.

A key challenge lies in the generalisability and interpretability of these models. Many approaches perform well on held-out test data (same data which is not included in the original training data) but remain biased toward sequence patterns similar to those encountered during training, limiting their ability to predict regulatory activity in very novel genomic contexts or across species. Furthermore, while DL models can achieve high predictive accuracy, extracting biologically meaningful insights from their internal representations remains an ongoing challenge.

A further challenge concerns the interpretation of attribution and feature importance analyses derived from DL models. Although methods such as saliency mapping, integrated gradients [77], and DeepLIFT [78] can highlight sequence features associated with model predictions, these representations do not necessarily correspond to causal biological mechanisms. Attribution scores may be unstable across model initialisations, sensitive to small perturbations in the input sequence, and influenced by correlations between regulatory features present in the training data. Furthermore, multiple distinct sequence configurations can produce similar predictive outputs, making it difficult to determine whether a model has learned biologically meaningful regulatory syntax or statistical patterns that are only indirectly associated with regulatory activity. Consequently, interpretability analyses should be viewed as hypothesis-generating tools rather than direct evidence of mechanistic understanding. This distinction between predictive utility and mechanistic interpretability represents a central challenge in modern regulatory

genomics.

1.6 Biological model system to develop and validate regulatory prediction approaches

For validating predictions from DL approaches, it is highly beneficial to train models on a well-characterised system with extensive functional annotation and orthogonal datasets for cross-validation of both predictions and the features underlying them. We therefore selected the erythroid cell system due to its comprehensive epigenomic characterisation, including chromatin modifications and TF binding, as well as the availability of a large, unique in-house dataset of validated regulatory variants that alter epigenetic activity. This enables a more thorough assessment of the TF motifs identified in these analyses, and allows the models to predict the effects of sequence variation, helping to evaluate their ability to generalise to unseen data. Importantly, this cell type includes two key gene loci model systems widely used in fundamental gene regulation research, providing a rich foundation of detailed studies for comparison.

1.6.1 Regulatory mechanisms in erythropoiesis

Red blood cells (RBCs), or erythrocytes, contain haemoglobin, the protein responsible for transporting oxygen to tissues and removing carbon dioxide, thereby supporting aerobic cellular respiration. RBCs are produced through erythropoiesis, a developmental process that proceeds through three stages: primitive, transient, and definitive. Primitive erythropoiesis occurs in the embryonic yolk sac and generates large, nucleated erythroblasts. This is followed by transient erythropoiesis in the fetal liver, which produces enucleated RBCs. Finally, definitive erythropoiesis takes place primarily in the bone marrow, where haematopoietic stem cells differentiate through a series of progenitor and erythroblast stages before forming reticulocytes and ultimately mature erythrocytes [79].

Although all cells share the same genome, distinct gene expression programs drive their differentiation into specialised cell types, making the erythroid lineage a well-established model for studying gene regulation. This process is tightly controlled by a combination of cytokines, growth factors, and TFs. Early erythroid commitment is supported by factors such as stem

cell factor (SCF), interleukin-3 (IL-3), and granulocyte-macrophage colony-stimulating factor (GM-CSF), while erythropoietin acts as a key regulator of later differentiation stages [80]. TFs, including GATA-1 [81], TAL-1 [82], NFE-2 [83], FOG-1 [84], or KLF1 [85] coordinate stage-specific gene expression, and essential micronutrients support haemoglobin synthesis.

Haemoglobin A, the predominant adult form, is composed of subunits encoded by the alpha- and beta-globin gene clusters, which are tightly co-regulated during erythropoiesis. These loci provide contrasting but complementary model systems for studying gene regulation. The alpha-globin locus resides in a constitutively transcriptionally active region and is regulated by long-range interactions with a cluster of enhancers located within the body of a neighbouring housekeeping gene, and is therefore activated within a complex but generally permissive genomic environment [86]. Whereas the beta-globin locus lies in a more repressed and gene-poor region that undergoes highly stage-specific activation mediated by a locus control region and long-range interactions [87]. Together, their distinct regulatory architectures and dynamics highlight both the complexity of erythroid gene regulation and build on their unique characterisation as model systems for understanding how precise, context-dependent transcriptional programs are established.

1.6.2 Globin loci as erythroid-specific regulatory models

The human alpha-globin locus is located near the telomere of chromosome 16 and spans approximately 26 kb. It contains three functional genes, HBZ, HBA1, and HBA2, alongside additional genes of unclear function and several pseudogenes [88]. During development, HBZ is expressed in the embryonic stage with HBA1 and HBA2, while HBA1 and HBA2 are solely expressed thereafter [89], with HBA2 contributing more strongly despite encoding an identical protein [90]. Despite being embedded within a transcriptionally active genomic region populated by housekeeping genes, expression of the alpha-globin genes is tightly restricted to erythroid cells [91], highlighting the need for precise regulatory control mechanisms.

Regulation of the alpha-globin locus is mediated by an upstream locus control region (LCR) composed of four enhancer-like elements (R1–R4), also referred to as multi-species conserved sequence (MCS) elements. Among these, R2 is the dominant contributor to gene expression [92, 93], with R1 playing a secondary role [94] and R3/R4 acting as facilitators [34]. These elements recruit key erythroid TFs, including GATA1, TAL1, NF-E2, KLF1, and associated

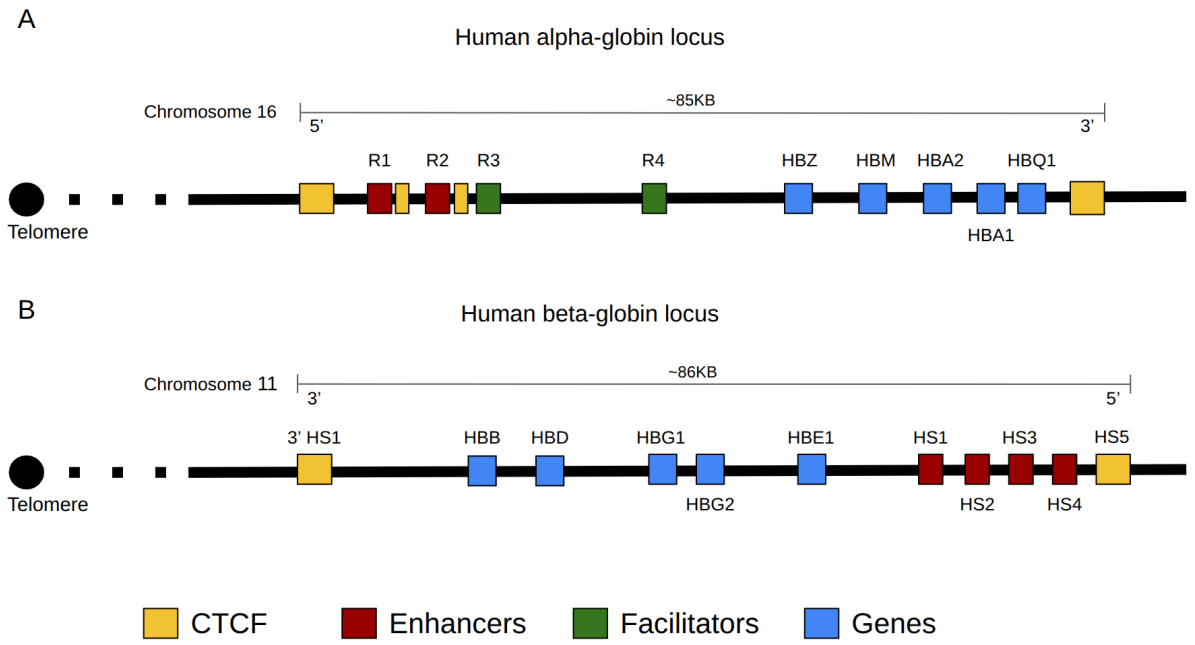


Figure 1.3: Schematic diagram of the human alpha- and beta-globin loci. A) The human alpha-globin locus is located in the short arm of chromosome 16. It contains five genes: HBZ, HBM, HBA2, HBA1, and HBQ1. The four distant enhancer-like elements R1 to R4 are located upstream of the HBZ gene. B) The human beta-globin locus is located in the short arm of chromosome 11. It contains five genes: HBB, HBD, HBG1, HBG2, and HBE1. The four distant enhancer-like elements, 5' hypersensitive sites HS1-HS5, are located upstream of the HBE1 gene. The entire locus is flanked by two CTCF sites, 3' HS1 and 5' HS5.

cofactors such as LMO2 and LDB1 [95]. Together, these factors coordinate transcriptional activation, with NF-E2 implicated in recruiting Pol II [96] and TAL1 contributing to erythroid lineage commitment [82, 97]. Chromatin architecture further refines this regulation: CTCF and cohesin define domain boundaries and mediate looping interactions that constrain enhancer activity within the locus, ensuring erythroid-specific expression [91]. A scheme of the alpha-globin organisation between its regulatory elements and genes is depicted in Figure 1.3A.

The beta-globin locus, located on chromosome 11, spans approximately 70 kb and comprises five genes, HBE1, HBG2, HBG1, HBD, and HBB, arranged according to their temporal activation during development. This organisation underpins globin switching, whereby embryonic, fetal, and adult globin genes are expressed sequentially during erythropoiesis [98]. In contrast to the alpha-globin locus, the beta-globin cluster is largely transcriptionally silent in non-erythroid cells and becomes progressively activated during differentiation, making it a well-defined model of stage-specific gene regulation.

Control of the beta-globin locus is mediated by a LCR located $\sim$ 60 kb upstream, comprising

five DNase hypersensitive sites (HS1–HS5), whose organisation is depicted in Figure 1.3B. These elements are essential for proper gene expression [99], with HS2 and HS3 playing particularly critical roles [100]. Key TFs such as GATA-1 and NF-E2 bind both the LCR and gene promoters, recruiting the transcriptional machinery [101] and facilitating long-range chromatin interactions [102]. GATA-1 also assembles multiprotein complexes including TAL1, FOG-1, and LDB1 to drive looping between the LCR and target genes, while NF-E2 contributes to chromatin accessibility and histone modification [83]. CTCF binding sites further define the chromatin domain, enabling coordinated regulation of the locus [103].

Mutations affecting the alpha- and beta-globin loci give rise to haemoglobinopathies, including alpha- and beta-thalassemia and sickle cell disease. These disorders result from imbalances in globin chain production, leading to defective haemoglobin assembly, ineffective erythropoiesis, and anaemia [104]. Alpha-thalassemia is frequently associated with mutations that impair gene expression or enhancer accessibility, particularly within the alpha-globin LCR, whereas beta-thalassemia typically arises from mutations in coding regions [105], although LCR deletions can also occur [106]. Mouse models have been instrumental in studying these conditions, demonstrating conserved regulatory mechanisms and enabling functional testing of mutations. Notably, while deletion of key enhancers severely disrupts alpha-globin expression in both species, the beta-globin locus shows species-specific differences, with the murine locus retaining partial activity even in the absence of its LCR [107].

1.7 Gene regulation through evolution

From an evolutionary perspective, differences in gene regulation are a key driver of phenotypic divergence between species. Analysing the regulation of the same gene in the same cell type across different species can therefore provide insight into the mechanisms underlying these differences. It also enables the exploration of multiple functional configurations that can produce similar regulatory outcomes, helping to reveal the fundamental principles of gene regulation and its degrees of inherent flexibility.

Regulatory changes tend to occur in non-coding elements such as enhancers and promoters rather than in TF proteins themselves [108], likely because they enable more precise, context-specific effects with fewer widespread consequences. As a result, evolutionary modifications at

the level of cis-regulatory elements allow advantageous changes in gene expression without altering the broader TF expression programs that define cell identity [109]. This enables specific aspects of cellular function to evolve without disrupting global gene regulatory networks.

1.7.1 Sequence conservation in regulatory elements

Genomic sequences are subject to continual drift due to replication errors, but changes that disrupt important functional elements are typically selected against. As a result, these regions are under stronger evolutionary constraint and tend to be more highly conserved across species [110].

However, different forms of conservation should be conceptually distinguished. Sequence conservation refers to the preservation of nucleotide identity across species, whereas motif conservation describes the maintenance of specific TF binding sites or sequence patterns. Regulatory conservation instead refers to the preservation of regulatory activity, such as chromatin accessibility or TF recruitment, while functional conservation reflects the maintenance of a broader biological outcome, for example, similar gene expression patterns or cellular phenotypes. Importantly, these forms of conservation do not necessarily coincide. Regulatory elements may retain similar functions despite substantial sequence divergence through mechanisms such as motif turnover, redundancy between enhancers, or compensatory changes in regulatory architecture. Conversely, highly conserved sequences may exhibit different regulatory behaviours depending on cellular or evolutionary context. Understanding how these distinct forms of conservation relate to one another represents a central challenge in comparative regulatory genomics.

This principle has previously been widely used to identify potential regulatory elements based on sequence conservation [111]. However, the approach has important limitations: conserved sequences may have a regulatory function, and regulatory elements can retain their function despite substantial sequence divergence, making them difficult to detect through conservation alone [112]. Furthermore, equivalent gene regulation can be achieved through different sets of enhancers due to the evolutionary turnover of regulatory elements [113], meaning functionally similar elements may show little or no detectable sequence conservation. However, the major limitation of this approach is that it provides no information about the cell type in which a regulatory element is active [114].

Determining cell–type–specific activity requires epigenetic assays such as ATAC-seq, but such data is generally only available for a limited number of species and cell types. This constrains cross-species comparisons and limits our understanding of how regulatory mechanisms are conserved or diverge over evolutionary time. Developing computational models capable of learning conserved, cell–type–specific regulatory principles and transferring them across species would therefore be highly valuable, enabling the *in silico* prediction of regulatory activity in multiple genomes using data available only in a single species.

1.7.2 The potential of cross-species prediction

Once validated for accuracy, such an approach would enable comparison of gene regulation at specific loci across equivalent cell types in any species with a sequenced genome. This would make it possible to investigate how regulatory programs adapt to environmental pressures and, over longer evolutionary timescales, how they diverge while preserving similar functional outputs. Beyond its value for studying the evolution of the non-coding genome, this approach could also reveal multiple regulatory configurations that achieve the same outcome, offering important insights into which aspects of gene regulation are constrained and which are inherently flexible.

However, current cross-species models tend to focus on well-conserved sequences, making it harder to detect species-specific regulatory elements and limiting their ability to generalise across larger evolutionary distances [115]. This leads to a key question: if the same TFs are active in equivalent cell types, to what extent can regulatory principles be transferred between species? Studies comparing human and mouse, where such direct comparisons are possible, have provided some important preliminary insights [116].

First, although many regulatory processes are conserved, for example, genes are often expressed with similar cell-type-specificity across species, the underlying DNA sequences can differ substantially [117]. In some cases, regulatory elements retain similar functions despite sequence divergence, indicating that multiple sequence configurations can produce equivalent regulatory outcomes. Conversely, conserved sequences may drive different regulatory behaviours depending on cellular or species-specific context [118]. This balance between conservation and divergence complicates the identification of general principles of gene regulation and limits the direct extrapolation of findings from model organisms to humans.

Second, the degree of sequence conservation varies across regulatory elements, particularly enhancers, and depends on their biological role [119]. Enhancers involved in developmental processes tend to be more highly conserved, whereas those associated with more dynamic or context-dependent functions, such as immune responses, often show much greater sequence variability [120].

Taken together, these observations point to a central biological challenge: determining how many different genomic configurations can produce the same functional output when experimental data mapping active elements are available for only a small number of species. Validated and accurate computational models therefore offer a unique opportunity to use evolution as a form of large-scale mutagenesis, enabling this question to be addressed without the practical constraints of extensive genome engineering or generating data across a wide range of evolutionarily distant species.

1.8 Globin clusters conservation across vertebrates

Another key advantage of the erythroid system as a model is that its evolution and sequence conservation, down to the level of TFBSs, has been extensively characterised. The structural organisation of the alpha globin cluster has been conserved amongst vertebrates for over 500 million years [121]. Figure 1.4 highlights the evolutionary conservation of the alpha-globin locus, with a similar structural organisation observed across multiple mammalian species. In contrast, while the overall organisation of the beta-globin cluster is also conserved, it exhibits greater evolutionary divergence, particularly in gene content and regulatory elements [122], reflecting differences in developmental regulation across species.

At the mouse alpha-globin locus on chromosome 11, four CTCF-binding sites define its chromatin organisation, including an erythroid-specific subTAD of ~ 54 kb nested within a larger ~ 165 kb domain. This subTAD contains the embryonic globin gene (HBAX) and the adult alpha-globin genes (HBA1 and HBA2), along with two theta-globin genes of unclear function. Gene expression is regulated by five upstream enhancer-like elements [123], with R1 and R2 acting as the main enhancers and R3, Rm, and R4 serving as facilitators [34]. Most of these regulatory elements are conserved in humans, although Rm, HBQ2, and some CTCF binding sites are mouse-specific. Figure 1.5A illustrates the structure of this locus in mouse.

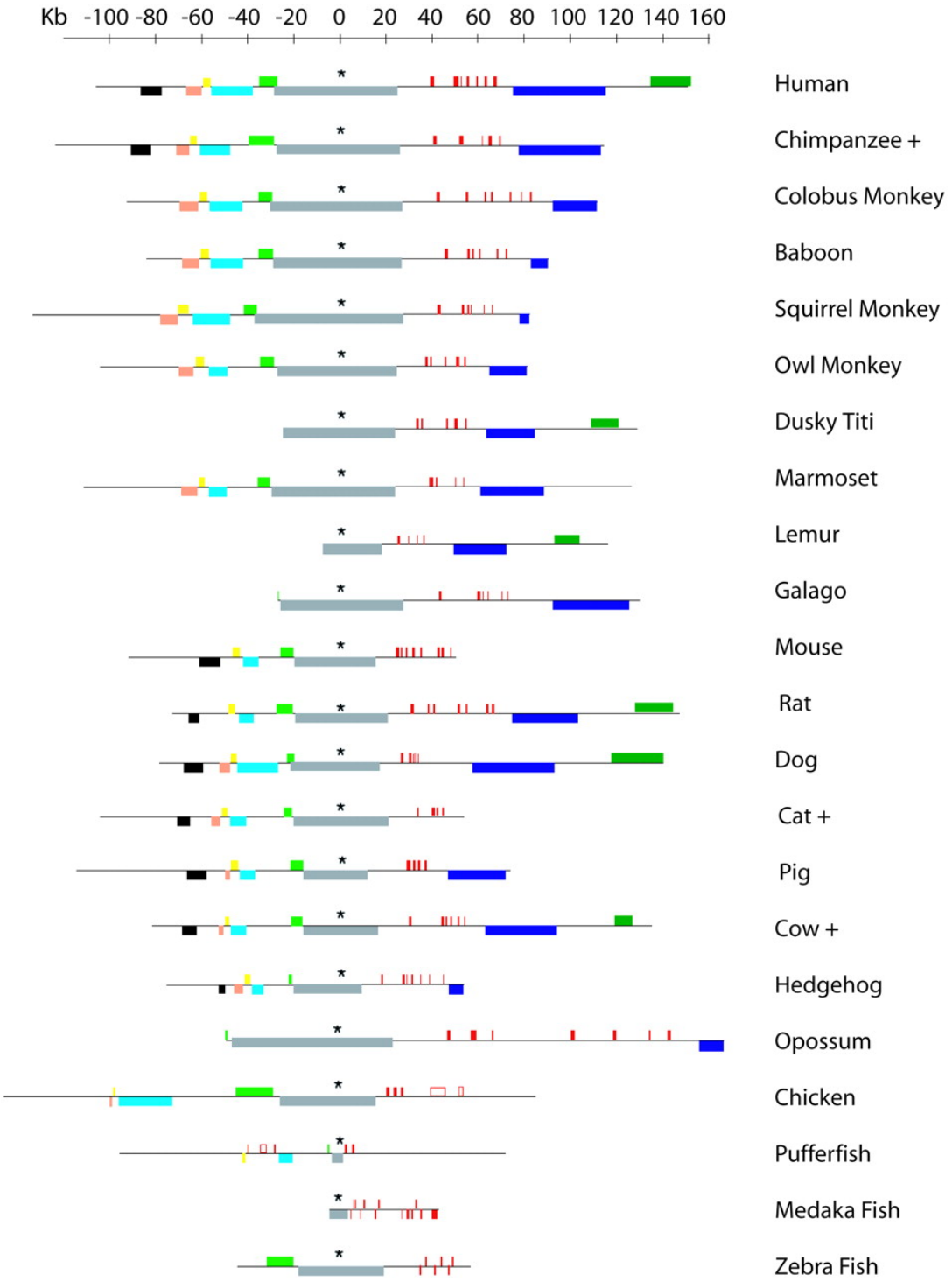


Figure 1.4: Structure of the alpha-globin locus in multiple mammal species. The alpha-globin cluster is aligned to the sixth exon of the C16ORF35 gene (indicated by an asterisk) and the flanking globin genes (colour red). The amount of DNA sequenced is denoted by the black box, while all the other coloured boxes indicate non-globin genes. *Source:* Reproduced from Figure 1 in Hughes, 2005 with permission of the authors [111].

The murine beta-globin locus contains four functional genes that are expressed in a developmentally regulated manner: HBB-Y and HBB-BH1 are active during embryonic stages, while HBB-BT and HBB-BS become the predominant adult genes [124, 125]. Regulation of this locus is controlled by a LCR located $\sim$ 60 kb upstream, comprising six DNase hypersensitive sites (HS1–HS6) [126]. Of these, HS1–HS4 are well-conserved between mouse and human, with HS2 acting as the strongest enhancer, followed by HS3.

These regulatory elements are enriched for conserved TF binding sites, including GATA-1, NF-E2, AP-1, and TAL1, although HS5 shows lower conservation. Functional studies indicate that HS5 and HS6 are largely dispensable for transcription and chromatin organisation, with the remaining elements maintaining locus activity [127]. The locus is flanked by CTCF-binding sites that define its chromatin domain [128]. Notably, unlike in humans, regulation of the murine beta-globin locus appears less dependent on strict developmental staging, with enhancer elements contributing in a more additive manner to gene expression [129]. Figure 1.5B illustrates the organisation of the murine beta-globin locus.

Another important model organism in haematological research is the chicken, whose gene

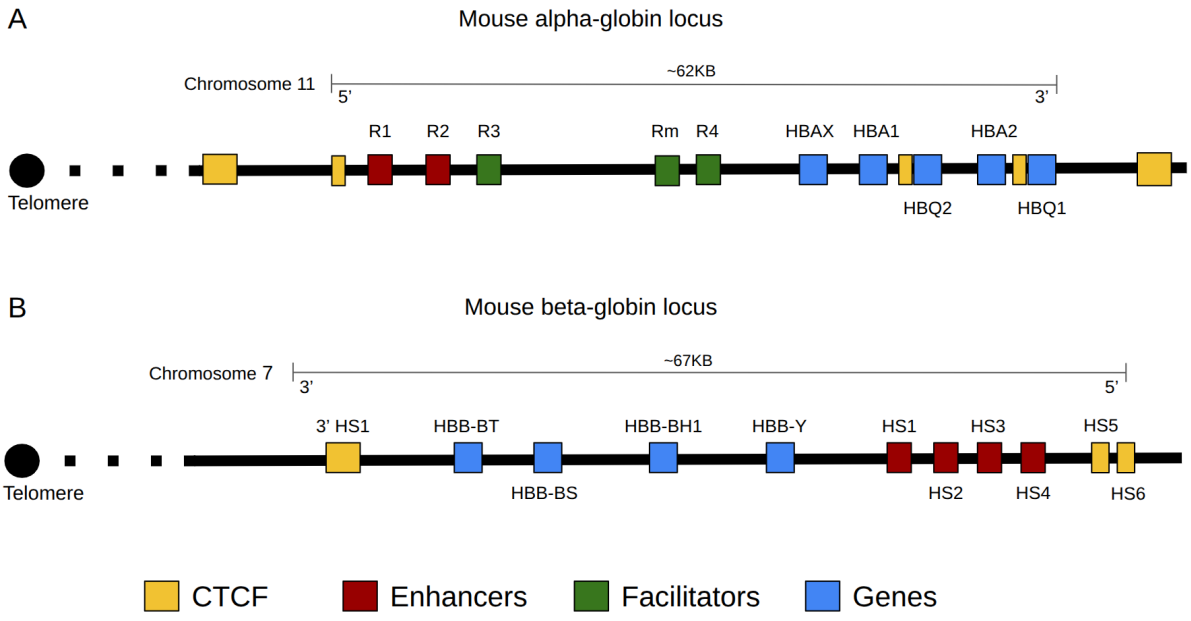


Figure 1.5: Schematic diagram of the murine alpha- and beta-globin loci. A) The murine alpha-globin locus is located on chromosome 11. It contains five genes, HBAX, HBA1, HBQ2, HBA2, and HBQ1. The five distant enhancer-like elements, R1, R2, R3, Rm, and R4, are located upstream of the HBAX gene. B) The murine beta-globin locus is located on chromosome 7 and contains four genes: HBB-BT, HBB-BS, HBB-BH1, and HBB-Y. The four distant enhancer-like elements, HS1-HS4, are located upstream of the HBB-Y gene.

content and genomic organisation resemble those of humans more than rodent models do [130]. Although avian erythrocytes remain nucleated, the fundamental mechanisms underlying haematopoiesis, including stem cell migration and organ transitions, are broadly conserved between chickens and humans [131]. In the chicken, the alpha-globin locus comprises three genes: one embryonically expressed (HBZ) and two adult-expressed genes (HBAD and HBA1) [132].

The chicken beta-globin locus consists of four genes, including two embryonic genes (HBE and HBBR), one gene expressed until hatching (HBE1), and one definitive adult gene (HBBA) [133]. Chicken globin LCRs are composed of development-specific regulatory elements. The alpha-globin LCR contains one DNase I HS orthologous to the enhancer R2 [111, 134, 135]. This region is flanked by two CTCF-bound HSs. Similarly, the beta-globin LCR comprises three regulatory HSs flanked by two CTCF-associated HSs [134, 136]. The chicken globin organisations are depicted in Figure 1.6.

In most mammals and avian species, the alpha-globin gene cluster is located within a GC-rich genomic region and is transcriptionally regulated by an upstream super enhancer. In contrast,

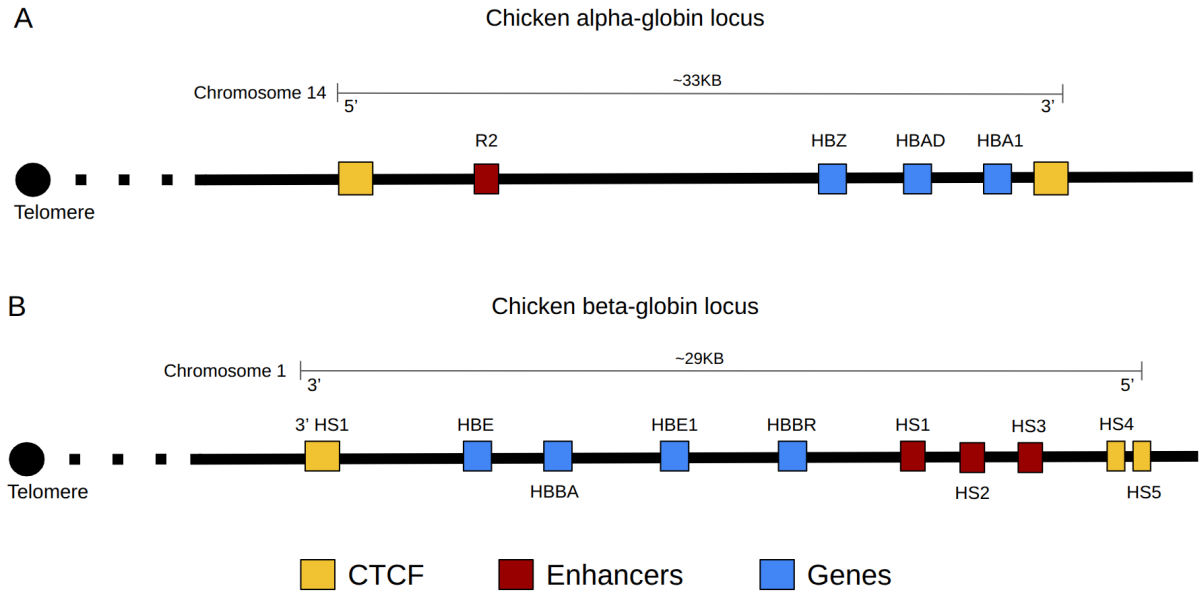


Figure 1.6: Schematic diagram of the chicken alpha- and beta-globin loci. A) The chicken alpha-globin locus is located on chromosome 14. It contains three genes, HBZ, HBAD, and HBAa. The distant enhancer element R2 is located upstream of HBZ. B) The chicken beta-globin locus is located on chromosome 1 and contains four genes: HBE, HBBA, HBE1, and HBBR. The three distant enhancer-like elements, HS1-HS3, are located upstream of the HBBR gene.

the beta-globin gene cluster is generally found in a gene-sparsed region with a lower GC content and is often relatively independent of its upstream LCR [137]. Although alpha- and beta-globin genes exhibit a similar genomic distribution across these species, the beta-globin locus has undergone greater evolutionary divergence and presents lower levels of conservation.

In summary, comparative genomics across species has revealed both conservation and divergence of regulatory patterns within globin loci throughout evolution. Importantly, this cross-species conservation of regulatory mechanisms suggests that aspects of the underlying regulatory logic may be shared, raising the possibility that principles learned in one organism could inform our understanding of gene regulation in another. However, the extent to which such regulatory rules are transferable remains unclear, particularly given species-specific differences in sequence and regulatory context, highlighting a key challenge in comparative and computational genomics.

1.9 Challenges

Advances in high-throughput sequencing technologies have enabled comprehensive, genome-wide profiling of regulatory landscapes across diverse cell types and organisms. Large international consortia such as ENCODE have generated extensive epigenomic maps, providing unprecedented resources for computational modelling of gene regulation. In parallel, DL has profoundly transformed computational biology. Contemporary sequence-based models have expanded receptive fields, incorporated long-range genomic interactions, and achieved predictive resolution at the nucleotide level. These approaches demonstrate impressive performance in predicting epigenomic features directly from DNA sequence, offering a powerful framework to investigate how regulatory information is encoded in sequence and how functional elements can be distinguished from non-functional patterns.

Despite this progress, important conceptual and methodological challenges remain.

First, the majority of existing models are trained and benchmarked on aggregated, heterogeneous datasets that are often biased toward a limited set of well-characterised cell lines. Although reported performance metrics are often strong, there is a lack of systematic evaluation on rigorously curated, cell-type-specific gold-standard datasets representing experimentally validated regulatory elements. In particular, relatively few studies assess model perfor-

mance using high-resolution binding data or extensively replicated primary-cell datasets. This limitation constrains our ability to reliably distinguish functional regulatory elements from non-functional sequence patterns and to assess whether models truly capture biologically meaningful regulatory signals.

Second, there is currently no publicly available last-generation DL framework specifically designed and comprehensively validated for RBC regulatory programs. Erythropoiesis represents a uniquely tractable and biologically well-characterised system for computational investigation. It is governed by a relatively well-defined set of master TFs whose cooperative interactions, motif syntax, and regulatory hierarchies have been extensively studied experimentally. This makes it an ideal system for dissecting the combinatorial rules of regulatory grammar that underlie cell-type-specific gene expression, and for developing interpretable sequence-to-epigenome models grounded in strong biological priors.

Third, several DL methods, particularly CNN-based architectures, have sought to extend predictions beyond the sequences encountered during training. A common way to assess generalisation is to test performance on unseen variants, such as SNPs. However, datasets of truly validated regulatory variants are scarce and dispersed across different cell types. As a result, even recent models such as Enformer and AlphaGenome remain only weakly validated and poorly calibrated for accurately predicting the effects of unseen variants. Although models typically do not exceed their original performance on held-out test data, they tend to generalise best to conserved regions or sequences closely resembling those seen during training. To date, model performance appears to be biased toward familiar sequence patterns, while comparatively little progress has been made in capturing regulatory function independent of sequence similarity. This limitation directly impacts our ability to use sequence conservation to infer regulatory function and to identify conserved regulatory mechanisms across divergent genomic contexts.

Fourth, although recent studies have begun to explore cross-species transfer learning and phylogenetically informed modelling strategies, our understanding of how predictive regulatory grammars generalise across evolutionary distances remains limited. It is unclear which components of regulatory logic are conserved and which are species-specific, and how predictive performance decays as phylogenetic divergence increases. These challenges are compounded by the limited availability of high-quality epigenetic data across species, which restricts sys-

tematic comparison of gene regulatory mechanisms and robust validation of cross-species predictions. Consequently, it remains an open question whether ML models can learn regulatory patterns that generalise across species and accurately predict regulatory activity in the absence of experimental data.

Addressing these gaps is essential not only for improving predictive accuracy but also for advancing mechanistic understanding of gene regulation. Moving beyond performance metrics toward biologically interpretable, generalizable, and rigorously benchmarked models is critical for establishing computational genomics as a predictive and explanatory discipline.

The broader implications of such a framework extend beyond basic research. A robust and transferable model trained on a well-characterised organism, such as human or mouse, could enable informed computational extrapolation to related organisms, cell types, or tissues, even in the absence of experimental data. This has potential ethical and practical benefits, including minimising the use of animal models and accelerating hypothesis generation. At the clinical level, accurate prediction of cell-type-specific regulatory activity could support the interpretation of noncoding variants, improve molecular diagnostics, and guide precision medicine strategies by identifying regulatory disruptions associated with disease. In this way, advances in interpretable and generalizable epigenomic modelling may ultimately bridge fundamental genome biology and translational healthcare applications.

1.10 Overall thesis aims

The central hypothesis of this thesis is that a systematically benchmarked, interpretable DL model trained on high-confidence erythroid regulatory datasets can accurately reconstruct TF binding and chromatin accessibility profiles, reveal cell-type-specific regulatory grammar, and uncover generalisable regulatory principles that enable the prediction of regulatory activity across species, even in the absence of experimental data.

Aim 1: Decoding regulatory information in DNA sequence and identifying functional elements

This aim addresses how regulatory information is encoded in DNA sequence and how functional regulatory elements can be distinguished from non-functional sequence patterns. It seeks

to systematically evaluate diverse DL architectures for predicting CTCF binding from DNA sequence. The benchmarking will include CNN-based classification and regression models, as well as transformer-based architectures capable of modelling longer genomic contexts.

All models will be trained on a well-characterised in-house CTCF ChIP-seq dataset to ensure high data quality and experimental consistency. Evaluation will be performed at multiple levels, including peak classification accuracy, signal track reconstruction, quantitative signal similarity, nucleotide-level importance scoring, and generalisation to held-out genomic regions.

Performance will be assessed using both global metrics and fine-grained profile-level comparisons. The best-performing architecture, based on robustness and interpretability, will be selected for subsequent downstream analyses.

Aim 2: Defining combinatorial regulatory grammar underlying cell-type-specific gene regulation

This aim addresses the combinatorial rules that define regulatory grammar and enable cell-type-specific gene regulation. While CTCF binding provides a relatively well-defined and structurally constrained binding paradigm, it represents a comparatively simple regulatory grammar. This aim will extend the selected modelling framework to more complex epigenetic profiles, including cell-type-specific TF binding and chromatin accessibility in RBCs.

Validation will employ similar quantitative measures as in Aim 1, including signal reconstruction, motif recovery, and nucleotide attribution consistency. Additionally, models trained on equivalent epigenetic datasets from different species will be systematically compared to begin disentangling conserved versus context-specific regulatory features.

By increasing regulatory complexity, this aim will test the scalability and robustness of the modelling approach beyond single-factor binding events.

Aim 3: Interpreting model representations to derive mechanistic regulatory principles

This aim further addresses the nature of regulatory grammar by moving beyond predictive performance toward a mechanistic understanding. It will apply interpretability techniques such as DeepLIFTShap, *in silico* mutagenesis, and motif discovery from learned convolutional filters or attention maps. These analyses will aim to: (i) identify sequence motifs captured by the model; (ii) characterise motif spacing, orientation, and cooperative interactions; (iii) infer

higher-order regulatory syntax specific to erythrocyte regulatory elements.

Model-derived insights will be validated against experimentally well-characterised regulatory regions, including the alpha-globin locus. Concordance between model-inferred regulatory logic and established biological knowledge will provide evidence for mechanistic validity.

Aim 4: Inferring regulatory conservation and enabling cross-species prediction

This aim addresses the extent to which sequence conservation can be used to infer regulatory function, how gene regulatory mechanisms can be compared across species given limited epigenetic data, and whether ML models can generalise regulatory patterns across species. It will systematically investigate model generalisation across species and cell types.

Predictions will be generated for conserved genomic regions across multiple mammalian species, enabling quantitative assessment of predicted regulatory activity as a function of phylogenetic distance. By analysing performance decay and conservation patterns, this aim seeks to: (i) identify regulatory elements whose activity is evolutionarily conserved; (ii) distinguish species-specific adaptations in erythroid regulatory grammar; (iii) characterise the extent to which sequence-based regulatory rules are transferable across evolutionary scales; and (iv) evaluate whether models can accurately predict regulatory activity in species lacking experimental data.

This comparative analysis will establish a principled framework for cross-species inference and provide insight into the conservation and divergence of regulatory mechanisms.

2 | Methods

2.1 Introduction

ML constitutes a major branch of artificial intelligence concerned with the development of computational models capable of learning patterns from data. These models can be understood as high-dimensional functions that are iteratively optimised to approximate the relationship between inputs and outputs through a process of training. During this process, model parameters are progressively adjusted to minimise prediction error, thereby improving performance over successive iterations. Once trained, such models can be used to generate predictions or infer patterns relevant to a specific task [138]. In the context of genomics, these approaches provide a powerful framework for capturing complex and often non-linear interactions between epigenetic features across the genome.

Early DL models, such as DeepSEA, established a foundation for predicting chromatin features directly from sequence data, demonstrating the potential of CNNs in regulatory genomics. Since then, rapid advances in ML have led to the development of a wide range of models tailored to epigenetic analysis tasks. These models differ substantially in their underlying architectures, input representations, and optimisation strategies, reflecting the diversity and complexity of biological data. Many of these approaches have been pre-trained on large-scale datasets, requiring extensive computational resources and time. Leveraging such pre-trained models enables the reuse of learned representations and facilitates the application of advanced methodologies without the need for prohibitively expensive training procedures [139].

The methodological framework adopted in this work builds upon these developments by integrating both established and contemporary computational approaches for the analysis of epigenomic data. Particular emphasis is placed on evaluating their applicability to the identification and characterisation of regulatory regions, as well as their capacity to generalise across different data types and experimental conditions.

This chapter provides a comprehensive description of the methodology underpinning the *in silico* experiments presented in subsequent chapters. It begins by outlining the motivations and objectives guiding the selection of methods. This is followed by a detailed account of the data preprocessing pipeline, including data acquisition, cleaning, and transformation steps. The

chapter then describes the ML models and algorithms employed, along with their configuration and implementation. Finally, the post-processing procedures and evaluation strategies used to analyse and interpret the results are presented.

Chapter aims

- To describe the tools behind robust and scalable bioinformatics pipelines for the preprocessing, prediction and analysis of epigenetic regulatory programmes.
- To ensure that the proposed methodology is reproducible, transparent, and amenable to systematic evaluation across different datasets and experimental settings.
- To assess the performance and limitations of both classical and DL–based approaches in the previous studies in the context of epigenomic signal analysis.

2.2 Red blood cells data preprocessing

CATCHUP is a bioinformatics pipeline designed to preprocess and clean bulk ATAC-seq and ChIP-seq data in preparation for downstream analysis and visualisation [140]. My preprocessing workflow was based on the functionalities of this pipeline, with some adaptations. From a biological perspective, these preprocessing steps are essential to ensure that the resulting signals accurately reflect true chromatin accessibility and protein binding events, rather than technical noise or experimental artefacts. This enables more reliable identification of regulatory elements and supports meaningful interpretation of gene regulatory activity.

Read quality check

FastQC [141] is employed to assess the quality of both bulk and single-cell ATAC-seq and ChIP-seq datasets. Adapter trimming was omitted in cases where overall data quality was high and/or when read lengths were short (≤ 50 base pairs), as the impact of residual adapter sequences under these conditions was considered negligible.

Adapter sequences trimming

During library preparation of ATAC-seq and ChIP-seq experiments, short adapter sequences are inserted into both ends of DNA fragments. If not removed, these adapters can lead to false alignments. Trimmomatic [142] is therefore used to trim adapter sequences, reducing the false-positive rate during genome alignment.

Read alignment

Raw sequencing reads are subsequently aligned to a reference genome to link sequencing signals back to specific regions of the genome, allowing them to be interpreted in a functional context. This alignment step is performed using Bowtie2 [143], which constructs an index of the reference genome and efficiently maps reads by identifying the highest-scoring alignments based on sequence similarity.

Read filtering

To improve the accuracy and reliability of the inferred regulatory signals, SAMtools [144] was used to filter out unmapped reads and reads with a mapping quality score below 30. Duplicate reads were identified and marked, and the resulting BAM files were subsequently sorted. BAM files corresponding to the same sample or sequencing run were then merged and indexed for downstream analysis.

Signal coverage

Signal quantification in sequencing data is achieved by computing genome-wide coverage, defined as the number of reads overlapping each genomic position. This is performed using the `bamCoverage` function from deepTools [145], generating raw (unnormalised) bigWig files at base-pair resolution.

Normalisation

Normalisation standardises epigenetic signals relative to the genome-wide background, enabling the identification of regions with significant regulatory enrichment while reducing the influence of technical variability across samples.

Developed by my research collaborator, Mr Tom Wilson, ZEN is a Python package designed for this purpose, enabling the evaluation of pairwise differential accessibility or binding based on zone Z-scores, which quantify how much the coverage of a genomic region deviates from the mean. This approach allows the signal to be analysed directly from the coverage without relying on peak calling.

ZEN was run using the ENCODE blacklist regions [146]. This adapted version of the Z-score is better at handling skewed datasets and outliers in comparison to the traditional Z-score.

The normalisation workflow is performed separately for each chromosome and consists of six steps. First, blacklisted regions and long stretches of consecutive zeros are masked, and the remaining signal is smoothed using a triangular kernel to reduce noise while preserving real peaks or enriched regulatory regions.

Second, a robust Z-score is computed to model chromosome-specific background noise. Instead of using the mean and standard deviation (SD), the signal is centred at the median and scaled using a robust estimate of spread. If the median absolute deviation (LCR) is non-zero, the scale is $1.4826\times$ LCR; otherwise, the mean absolute deviation scaled by 1.2533 is used. This does not make the data normal but provides a stable way to measure how extreme each value is relative to the background.

Third, a strict threshold corresponding to three deviations from the median is applied to balance detection of true regulatory regions while limiting false positives, where background noise is extremely unlikely to lie. Regions exceeding this threshold for at least 35 consecutive base pairs are retained, rounded to $1,000$ bp bins, and merged into candidate signal zones.

Fourth, low-coverage zones are removed using a quality threshold defined as the maximum of 1.5 times the chromosome mean, 1.5 times the global mean, and ten times the minimal non-zero fragment signal.

Fifth, for each sample, signal values within the retained zones are extracted. Zeros and the top

and bottom 10% of values are removed to reduce technical artefacts, and the remaining signal is standardised using the traditional Z-score.

Finally, since the scATAC-seq signal is non-negative, negative Z-score values are shifted upward by subtracting the minimum value, producing the final normalised signal.

2.3 Generation of training datasets

After data preprocessing and before training a model, the data must be transformed into a format that reflects a true biological signal and that the model can interpret correctly. This preparation stage varies depending on the type of model being used and is crucial for ensuring that the model can learn effectively from the data, achieve accurate results, and generalise well to new, unseen input DNA sequences.

2.3.1 DeepSEA-based model preprocessing

After compiling genome-wide CTCF binding profiles for 33 distinct cell types, the raw ChIP-seq FASTQ files were processed using the CATCH-UP pipeline. Sequencing reads were aligned to the human reference genome (GRCh38/hg38), and strand-specific coverage tracks were generated and exported in BigWig format.

Peak calling was subsequently performed independently for each cell type using LanceOtron [147]. The resulting peak coordinates were then merged across all cell types using `bedtools merge`, yielding a unified set of candidate CTCF binding regions representing the union of peaks observed across the 33 cell types.

To refine this master peak set, a quality-based filtering procedure was conducted using LanceOtron peak scores, which range from 0 (lower confidence) to 1 (higher confidence). Multiple score thresholds were evaluated to distinguish high-confidence binding events from ambiguous regions. Based on this assessment, regions with LanceOtron scores < 0.2 and > 0.8 were excluded, retaining peaks with scores in the interval $[0.2, 0.8]$. This filtering step reduced the total number of candidate regulatory elements (peaks) from $58,064$ to $22,695$.

In addition, regions exhibiting ambiguous LanceOtron scores in any cell type were removed

from the peak sets of all cell types, ensuring consistency across the merged dataset. The resulting set of $22,695$ high-confidence CTCF regions was subsequently used for model training.

2.3.2 BPNet preprocessing

To generate the training dataset for BPNet, a structured preprocessing pipeline developed by BPNet's authors was applied [148], starting from the aligned ChIP-seq reads in BAM format. Strand-specific 5' coverage tracks were computed using `samtools` and `bedtools genomecov` with the -5 option, ensuring that only the 5' end of each aligned read was counted. This approach captures precise cut-site information. Separate coverage tracks were generated for the positive and negative strands and exported as bedGraph files.

These bedGraph files were sorted and subsequently converted into BigWig format for efficient downstream processing. Peak calling was performed using MACS2 with a p-value threshold of 0.01, a default value to identify significantly enriched genomic regions. Reads were extended to a fixed fragment length of 147 bp, without inferring fragment size from the tag distribution. This generates a BigWig file for the positive strand and another for the negative strand.

Subsequent preprocessing steps were carried out using BPNet-specific utilities. First, the called peaks were filtered for outliers using `bpnet-outliers`. Peaks overlapping genomic regions listed in the blacklist file were excluded, as were peaks with read counts exceeding the 99th percentile of the signal distribution. For signal evaluation, $1,000$ bp windows centred on peak summits were extracted (`-sequence-len 1000`). The parameter `-global-sample-weight 1.0` was applied to avoid reweighting during filtering. This step produced a BED file containing high-confidence inlier peaks.

Next, the `bpnet-gc-reference` script was used to generate a genome-wide GC content reference track for bias modelling and GC-matched background sampling. In this procedure, non-overlapping $1,000$ bp windows (`-stride 1000`) were tiled across the reference genome. For each window, the central $2,114$ bp sequence, corresponding to BPNet's required input length, was extracted, and the GC percentage was calculated and stored.

GC-matched negative regions were then generated using `bpnet-gc-background`. This step samples enriched epigenetic signal sites that do not overlap, called peaks, but exhibit a GC content distribution similar to that of the positive (inlier) peaks, thereby minimising GC bias

during training. For GC matching, windows of $2,114$ bp were used (`-flank_size 1057`). A fourfold excess of negative regions relative to positive regions was generated for the training set (`-neg_to_pos_ratio_train 4`). The resulting GC-matched background regions were saved in BED format.

Model training requires the strand-specific BigWig coverage files (positive and negative strands), the filtered inlier peak regions, and the GC-matched negative regions.

2.3.3 Basenji preprocessing

Both the Enformer and HYPER training datasets were generated using the preprocessing framework provided by Basenji [149]. This pipeline converts raw functional genomics signal tracks (e.g., ChIP-seq, DNase-seq, CAGE) stored in BigWig format into HDF5 training files containing fixed-length DNA sequence inputs paired with dense target signal profiles. The preprocessing workflow comprises four principal steps.

First, a metadata file (targets.txt) is constructed, in which each row specifies an individual signal track. Unlike BPNet, which focuses on peak-centred regions, Basenji reads the corresponding BigWig files, extracts numerical signal values, and sanitises the data by removing undefined bases (Ns) and non-numerical entries. Then, Basenji samples approximately $200,000$ non-overlapping 1 Mb genomic intervals distributed evenly across the genome. Rather than excluding blacklisted regions, Basenji assigns all bases within such regions a fixed baseline coverage value, typically close to background signal levels (e.g., 0.01–0.05), thereby preventing spurious signal inflation while maintaining consistent genomic coverage.

To exclude validation loci from training, such as the alpha-globin locus, I included an additional parameter (`-exclude`) in the Basenji preprocessing script. This parameter accepts a BED file of genomic coordinates, and any sampled interval overlapping these coordinates is strictly removed from the dataset. The resulting $\sim 200,000$ genomic intervals are subsequently partitioned into training, validation, and test sets, and their coordinates are written to BED files.

In the second step, each genomic interval is used to extract a $131,072$ bp DNA sequence from the reference genome. These sequences constitute the model inputs.

Third, for each corresponding genomic interval, Basenji retrieves signal values from all specified BigWig tracks within the same 131 kb window. The continuous signal is then aggregated

into $1,024$ bins of 128 bp each. For broader histone modification signals (e.g., H3K27ac), signal values are summed within each bin, whereas for sharper signals (e.g., CAGE), the maximum value per bin is retained. A `log1p` transformation is subsequently applied to stabilise the dynamic range of the signal values. This procedure yields aligned pairs of DNA sequences and multi-track signal targets for each genomic region.

In the fourth and final step, data augmentation strategies are applied. First, translational augmentation is performed by generating eight shifted versions of each sequence–signal pair (± 1, ± 2, and ± 4 bp), promoting positional robustness in model learning. Second, reverse-complement augmentation is implemented by reversing and complementing half of the dataset, ensuring strand invariance. The processed data are then stored in large HDF5 files (approximately 100 GB each), corresponding to the training, validation, and test partitions. These files serve as the input for subsequent model training.

2.4 Machine learning-related concepts

Neural networks and learning methods

NNs are composed of multiple interconnected **layers** of **artificial neurons**. Each neuron contains two values: a weight and a bias. **Weights** measure the importance of the input data feature, while **bias** represents the flexibility of the neuron. Neurons receive an input, apply a weighted transformation followed by an activation function, and produce an output.

A typical NN begins with an input layer, which receives the raw data. This information is then propagated through one or more hidden layers, where neurons process the inputs in parallel and progressively extract higher-level features. As the number of hidden layers increases, the architecture becomes deeper, and such models are referred to as deep neural networks, forming the basis of DL. At each layer, the network adjusts its weights and biases during training to minimise an error or loss. Finally, the processed information reaches the output layer, which produces the model's prediction. This entire process is governed by forward propagation, while learning occurs through backpropagation and optimisation algorithms that iteratively update the model parameters.

In ML, two primary learning paradigms are commonly used: supervised and unsupervised

learning. **Supervised learning** relies on labelled data, where each input is associated with a known output. The model learns a mapping function between inputs and labels during training, enabling it to make predictions on previously unseen data. This approach is widely used for tasks such as classification and regression.

In contrast, **unsupervised learning** operates on unlabelled data, where no explicit target is provided. The model instead seeks to uncover underlying patterns, structures, or relationships within the data. This paradigm is typically applied to tasks such as clustering, density estimation, and dimensionality reduction, where the goal is to gain insight into the intrinsic organisation of the data.

Neuron activation functions

Activation functions act as non-linear decision mechanisms within a NN. They transform the input received by a neuron and determine the signal passed to the next layer, enabling the network to model complex patterns beyond simple linear relationships. In this sense, they play a role analogous to logic gates in electronic circuits, controlling how information flows through the system.

One of the most widely used activation functions is the **Rectified Linear Unit (ReLU)** [150], primarily due to its simplicity and computational efficiency. ReLU outputs zero for any negative input and returns the input unchanged if it is positive, effectively implementing the function `max(0, x)`, where `x` is the input feature. Sparse activations produced by functions such as ReLU are particularly useful in regulatory genomics because biologically informative sequence features are typically distributed sparsely across large genomic regions. This property helps introduce non-linearity while maintaining stable gradients during training.

Another commonly used activation function is the **sigmoid** function, which maps input values to a continuous range between 0 and 1 (excluding the exact endpoints). Large positive inputs are compressed to values close to 1, while large negative inputs approach 0. Because of this probabilistic interpretation, sigmoid is often used in output layers for binary classification tasks.

The **softmax** function is commonly employed as an activation function in the output layer of NNs designed for classification tasks. It maps a vector of real-valued inputs to a probability distribution over a predefined set of classes, such that the resulting values are non-negative and

sum to one. This is achieved by exponentiating each input and normalising it by the sum of all exponentiated values within the output layer. Softmax is particularly well-suited to mutually exclusive multi-class classification problems, where each input is assigned to a single class, and a probabilistic interpretation of the model's predictions is required.

More recent architectures often employ variants of ReLU to address the "dead neuron" problem, where neurons receiving consistently negative inputs stop updating and become inactive. One ReLU variant is the **Gaussian Linear Unit (GELU)** [151], which introduces a smooth, probabilistic gating mechanism. Instead of abruptly zeroing out negative values, GELU weights inputs based on their magnitude: strongly positive values pass through largely unchanged, strongly negative values are suppressed, and values near zero are scaled smoothly. This allows a small amount of negative information to propagate, improving gradient flow and overall model performance.

Error linear unit functions provide strong stability and learning capacity for capturing nonlinear relationships during training; therefore, at least one of them, such as GELU or ReLU, is implemented in the models used.

Forward- and Backpropagation

Forward propagation refers to the process by which input data are successively transformed as they pass through the layers of a NN, from the input layer to the output layer. At each layer, neurons perform an affine transformation, combining inputs through weighted sums and bias terms, followed by the application of a nonlinear activation function. This sequential transformation enables the network to map input features to a final predictive output.

Backpropagation provides the mechanism through which the network's parameters are iteratively refined in response to prediction error. Specifically, it evaluates the contribution (C) of each weight (w) to the overall loss by computing gradients of the cost function with respect to the network parameters. These gradients are obtained via the chain rule, reflecting the layered dependency (L) of each parameter on intermediate activations (a) and input features (z).

$$\frac{\partial C}{\partial w_L} = \frac{\partial C}{\partial a_L} \cdot \frac{\partial a_L}{\partial z_L} \cdot \frac{\partial z_L}{\partial w_L}$$

The resulting gradients are then used to update the model parameters, typically through a gradient-based optimisation scheme in which each weight value (w_{old}) is updated to a new value (w_{new}), the adjustment of which is proportional to its contribution to the error and moderated by a learning rate (η).

$$w_{new} = w_{old} - \eta \frac{\partial C}{\partial w}$$

The **learning rate** serves as a scaling factor that controls the magnitude of parameter updates, thereby ensuring stable convergence.

During each backpropagation step, the parameters of the network are updated simultaneously. This procedure is iteratively repeated by performing forward propagation on different, typically randomly selected, subsets of the training data (**mini-batches**), enabling the model to encounter diverse input patterns and improve generalisation. A full iteration over the entire dataset is referred to as one **epoch**.

Convolutional neural networks

CNNs were initially developed for image analysis tasks; however, their application has since extended to a wide variety of data modalities. Central to CNNs is the use of **filters** (or **kernels**, which are small, learnable matrices that are convolved across the input data. As these filters traverse the input, they detect local patterns by computing responses to specific numerical configurations, thereby enabling the extraction of meaningful features. In regulatory genomics, convolutional filters frequently learn sequence motifs corresponding to TFBSs or higher-order combinations of regulatory features.

The parameters of these filters are not fixed but are optimised during training. Typically, filters are initialised with random values and are progressively refined through iterative learning. During forward propagation, each filter produces feature maps that emphasise regions of the input relevant to the task. Backpropagation subsequently adjusts the filter parameters, improving their ability to capture salient patterns. Over successive training epochs, individual filters become specialised in detecting distinct features or structures within the data.

Unlike traditional NNs, CNNs restrict connections to local regions of the input space instead of connecting every neuron to all inputs. This reflects the assumption that meaningful patterns are

often spatially local. Furthermore, the same set of filter parameters is applied across different regions of the input, a property known as parameter sharing. This significantly reduces the number of trainable parameters and improves generalisation by enforcing consistency in feature detection across spatial locations.

CNNs also exhibit a hierarchical structure of feature learning. Lower layers typically learn to detect simple and generic patterns, such as edges or basic textures. As information propagates through the network, deeper layers combine these low-level features into increasingly complex and abstract representations.

Pooling layers are commonly incorporated to reduce the spatial dimensionality of intermediate representations. By summarising local regions, often through operations such as maximum or average selection, pooling reduces the resolution of feature maps while retaining their most salient information.

In standard convolution, filters are applied as contiguous, or "solid", kernels over the input. In contrast, **dilated convolution** introduces gaps between the elements of the filter, effectively expanding its coverage [152]. This modification enables an increase in the receptive field without a corresponding increase in computational cost, as the number of parameters and operations remains the same. Dilated convolutions are particularly useful in mitigating the loss of fine-grained information that can occur due to repeated pooling or downsampling between convolutional layers.

Transformers

Transformers represent the current state-of-the-art architecture in many areas of artificial intelligence, particularly for tasks involving sequential data such as natural language, genomic sequences, or financial time series. In contrast to convolutional architectures, which process inputs through local receptive fields, transformers operate on entire sequences simultaneously. This global processing is enabled by the self-attention mechanism [71], which allows the model to determine the relative importance of different elements within the input sequence.

To process discrete sequential data, inputs must first be transformed into a suitable numerical representation. This is achieved through embeddings, where tokens are mapped into a high-dimensional continuous vector space. To preserve information about the order of the sequence,

positional encodings, either fixed or learned, are incorporated into these embeddings, typically using sinusoidal functions or trainable vectors.

The core operation of the transformer is the **self-attention mechanism**, which constructs context-aware representations of the input. In regulatory genomics, this property is particularly important because regulatory activity frequently depends on interactions between distal sequence elements rather than purely local motif composition. For example, enhancer–promoter communication, cooperative TF binding, and chromatin domain organisation may involve regulatory elements separated by thousands of base pairs in the linear genome but brought into functional proximity through higher-order chromatin structure. Self-attention mechanisms are therefore conceptually well suited to modelling these long-range contextual dependencies, as they enable the contribution of distant sequence regions to be dynamically weighted during prediction.

Each token embedding is linearly projected into three distinct matrix representations: queries, keys, and values. Queries capture the syntactic role of a token in relation to others, keys represent the contextual identity of tokens, and values encode the associated information content. These representations are learned during training.

Attention scores are computed by taking the dot product between queries and keys, reflecting the degree of relevance between tokens. To ensure numerical stability, these scores are scaled by the dimensionality of the key vectors. A softmax function is then applied to obtain a probability distribution over the sequence, which is used to weight the value vectors. This process yields a refined representation of each token, commonly referred to as **scaled dot-product attention**.

Unlike convolutional architectures with fixed receptive fields, self-attention allows potentially relevant interactions to emerge adaptively from the data, which may be advantageous for modelling heterogeneous regulatory architectures across the genome.

Transformers further extend this mechanism through **multi-head attention** [153], where multiple attention operations are performed in parallel. Each head learns different relationships within the data, enabling the model to capture both local dependencies and broader contextual patterns.

Following the attention mechanism, each token representation is passed through a position-wise feed-forward network. This component consists of fully connected layers with nonlinear

activation functions, such as ReLU or GELU, applied independently at each sequence position. Residual connections are incorporated around both the attention and feed-forward sublayers, facilitating gradient flow and enabling the effective training of deep architectures.

The self-attention property is particularly relevant to regulatory genomics, where enhancer–promoter interactions and combinatorial TFs binding may depend on relationships between distant genomic regions.

Loss functions

The loss function quantifies the discrepancy between a model's predictions and the true target values, thereby providing a measure of how well the model is performing. It takes as input the predicted output ($\hat{y}$) and the corresponding ground truth (y), and produces a scalar value representing the prediction error. A higher loss indicates poorer model performance and a greater need for adjustment, whereas a lower loss suggests that the model is successfully approximating the target.

For regression tasks, a widely used loss function is the **mean squared error (MSE)**, which is particularly suited to predicting continuous values. It measures the average of the squared differences between predicted values and the true targets, thereby penalising larger deviations more strongly.

$$MSE = \frac{1}{n}\sum(y - \hat{y})^2$$

Mean absolute error (MAE) is another commonly used loss function closely related to MSE. It quantifies prediction error by computing the average of the absolute differences between predicted and true values. Compared to MSE, MAE is generally more robust to outliers, as it does not disproportionately penalise large deviations.

$$MAE = \frac{1}{n}\sum|y - \hat{y}|$$

The **Huber loss** function combines the characteristics of both MSE and MAE, achieving a balance between sensitivity to small errors and robustness to outliers. It is defined as a piecewise function that behaves quadratically for small error values and transitions to a linear regime for larger errors. Given a threshold parameter δ, which determines the threshold point (commonly

set to 1), the loss function is defined as follows:

$$L_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{for } |y - \hat{y}| \leq \delta \\ \delta\left(|y - \hat{y}| - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases}$$

Binary cross entropy (BCE) is a widely used loss function for binary classification tasks, where the objective is to distinguish between two discrete classes. It evaluates the discrepancy between the predicted probability and the true label by comparing their respective probability distributions. Let $\hat{y}$ denote the predicted probability and y the ground truth label. For a single data point, the loss function is defined as follows.

$$L = -[\hat{y} \log y + (1 - \hat{y}) \log 1 - y]$$

Regression losses are especially relevant for epigenomic signal reconstruction tasks, where models predict continuous chromatin accessibility or TF binding profiles.

Performance metrics

Model performance is assessed using evaluation metrics, which quantify how effectively a model accomplishes its intended task. While loss functions guide the learning process during training, some of them can also serve as performance measures for model evaluation.

In classification tasks, several standard metrics are widely employed. These metrics are defined based on the following outcomes: **true positives (TP)**, where positive instances are correctly identified; **true negatives (TN)**, where negative instances are correctly identified; **false positives (FP)**, where negative instances are incorrectly classified as positive; and **false negatives (FN)**, where positive instances are incorrectly classified as negative.

Accuracy measures the proportion of correctly classified instances among all predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision quantifies the proportion of correctly predicted positive instances out of all instances

predicted as positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall, also referred to as sensitivity, measures the proportion of actual positive instances that are correctly identified by the model.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Specificity, in contrast, evaluates the proportion of actual negative instances that are correctly classified.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The **F1-score** is defined as the harmonic mean of precision and recall, providing a balanced measure that is particularly informative in the presence of class imbalance.

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In addition to these threshold-dependent metrics, curve-based evaluation methods assess model performance across all possible classification thresholds. The **receiver operating characteristic (ROC)** curve plots the true positive rate (sensitivity/recall) against the false positive rate (FPR, $1 -$ specificity), and the area under this curve (AUROC) provides a threshold-independent, scalar summary of performance.

$$\text{AUROC} = \int_0^1 \text{Sensitivity}(FPR)\ \partial\text{FPR}$$

Similarly, the **precision–recall curve** illustrates the trade-off between precision and recall across varying thresholds, with the area under this curve (**AUPRC**) offering a complementary evaluation, particularly in imbalanced settings.

$$\text{AUPRC} = \int_0^1 \text{Precision}(\text{Recall})\ \partial\text{Recall}$$

2.5 Deep learning approaches for epigenetics modelling

The present study investigates three distinct model architectures selected to represent different modelling strategies for regulatory genomics, each characterised by distinct assumptions regarding the organisation of regulatory information within genomic sequence. These include a DeepSEA-based CNN for classification (referred to as classification CNN or class. CNN for short), a DeepSEA-based CNN adapted for regression (BPNet), and HYPER, an enhanced version of REnformer that builds on a pretrained Enformer model. CNN–based classification models, such as DeepSEA, are primarily designed to detect local sequence patterns and motif combinations associated with regulatory activity. In contrast, regression-based architectures such as BPNet extend this framework by reconstructing quantitative binding profiles at base-pair resolution, enabling more detailed modelling of local motif syntax and TF occupancy. Transformer-based architectures, including Enformer-derived models such as REnformer and HYPER, further expand the receptive field through self-attention mechanisms capable of modelling long-range contextual dependencies across large genomic regions. These architectures therefore differ not only in predictive objectives but also in their capacity to capture local versus distal regulatory interactions, reconstruct continuous epigenetic profiles, support biological interpretability, and balance predictive performance against computational complexity. Comparing these approaches provides a framework for evaluating how architectural assumptions influence both predictive performance and the biological relationships that can be represented from genomic sequence. The architectural details of these models are described below.

2.5.1 DeepSEA-based model for binding sites classification

DeepSEA is a ML architecture designed to classify the functional effects of noncoding DNA variants directly from sequence [67], addressing the question of whether DNA sequence alone encodes regulatory activity. It has previously been trained on TF and histone modifications, protein binding and DNase I hypersensitivity. A former collaborator of the Hughes lab, Dr Ron Schwessinger, adapted the DeepSEA architecture to implement DeepHAEM, which extracts chromatin features from chromatin accessibility profiles. Part of his work involved transferring these features into a second DeepSEA-based network trained on High-throughput Chromosome Conformation Capture (Hi-C) data to predict chromatin architecture from DNA se-

quences [154]. The work of Dr Schwessinger laid the foundation for many of the computational pipelines developed within the Wellcome Group, and has since been further extended by our collaborators Dr Edward Sanders and Mr Simone Riva. Their contributions form the basis of the work presented in the following chapter.

The training dataset consists of DNA sequences of $1,000$ bp long, encoded in matrices of shape $1,000 \times 4$, where the columns represent the nucleotides A, C, G, T. Each cell in the matrix represents a basepair in the sequence and corresponds to an A, a C, a G, or a T, encoded in a 1 as a True statement of 0 as False. This way, the bp A would be encoded as $[1, 0, 0, 0]$, C as $[0, 1, 0, 0]$, G as $[0, 0, 1, 0]$, and T as $[0, 0, 0, 1]$.

The model adopts a CNN architecture composed of five convolutional layers with kernel sizes of 20, 10, 8, 4, and 8, respectively. Each convolutional layer is followed by a pooling layer, and a fully connected layer is placed before the final output layer. The model is trained using a learning rate of 0.0001 and a weight decay of 0.00001. The latter penalises large weight values to reduce overfitting, thereby preventing the model from learning unrealistic motifs and bio-

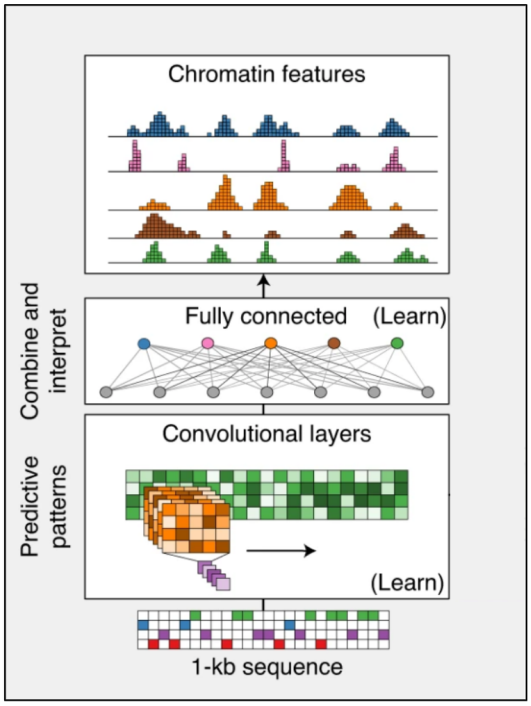


Figure 2.1: Overview of DeepHAEM architecture. Just like our classification model, DeepHAEM consists of multiple convolutional modules followed by a fully connected layer. It is designed to extract regulatory patterns from ATAC-seq and ChIP-seq profiles across 1 kb DNA sequences. *Source:* Reproduced from Figure 1A in Schwessinger, 2020 with permission of the authors [154].

logical artefacts. Training is optimised using the binary cross-entropy loss function. Figure 2.1 illustrates the DeepHAEM architecture, on which our model is based, including both training and prediction workflows.

The input data consists of cell-type-specific chromatin accessibility or protein binding signals, which represent the training classes. These signals are usually extracted using a peak-calling tool such as LanceOtron [147]. During training, the model learns to classify whether or not a given 1 kb DNA sequence contains cell-type-specific signal peaks corresponding to each class.

Since this model is classification-based, its output consists of binary or probabilistic predictions (values between 0 and 1), indicating whether a given 1 kb input sequence exhibits epigenetic activity. In this way, it addresses the question of whether a genomic region is likely to contain functional regulatory elements.

2.5.2 BPNet for protein binding regression

Similar to DeepHAEM, BPNet [68] is also based on CNNs but is designed to predict protein binding profiles at base-pair resolution from 1 kb DNA sequence inputs. However, BPNet is a regression model, meaning that its output is the predicted read count at each nucleotide position, instead of probabilities.

BPNet consists of a stack of dilated convolutional layers, which enable the model to capture dependencies across larger genomic regions. By introducing gaps between kernel elements, like a "colander", dilated convolutions effectively sample the input over an expanded receptive field while preserving base-pair resolution. Unlike many conventional architectures, BPNet does not use pooling layers.

In summary, the architecture consists of a standard one-dimensional convolutional layer with a kernel size of 15, which operates on the one-hot encoded DNA sequence. This initial layer is followed by a stack of dilated residual convolutional blocks comprising 11 layers with exponentially increasing dilations ($1, 2, 4, \ldots$, up to $1,024$). Each layer includes batch normalisation and a ReLU activation. Figure 2.2 illustrates the workflow used for BPNet training and prediction.

BPNet requires matched control tracks for training, such as ChIP input for ChIP-seq or immunoglobulin G antibody control (IgG) for ChIP-nexus. It is designed to model both back-

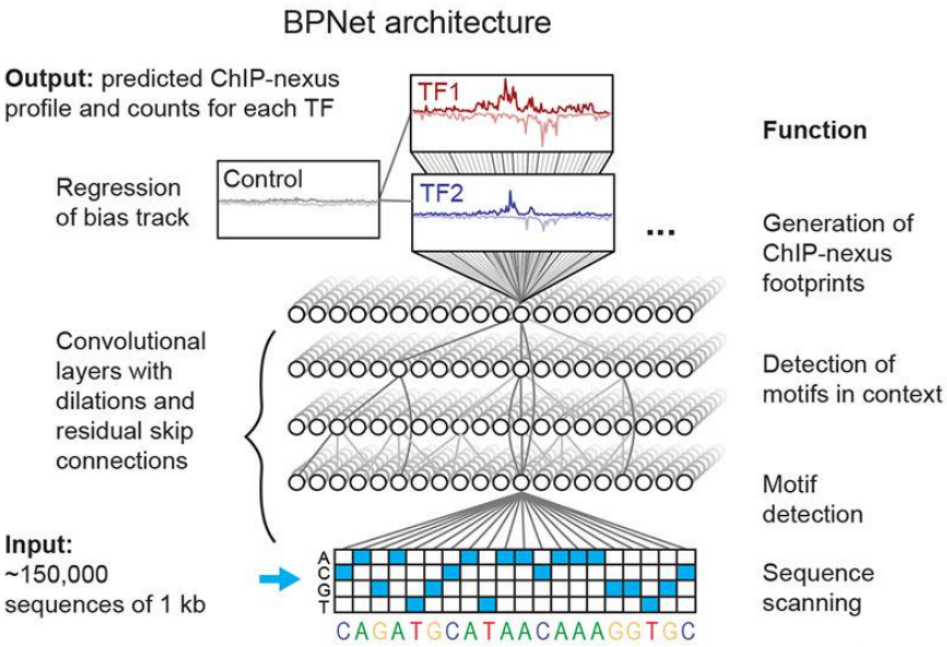


Figure 2.2: Overview of BPNet architecture. The model is trained to simultaneously predict protein-binding read counts from 1 kb DNA sequences, while minimising bias-related signal in its predictions. *Source:* Reproduced from Figure 1D of Avsec, Nature, 2022 [68] under the Creative Commons Attribution 4.0 License.

ground and protein-binding signal through a two-step training. First, a primary model is trained on the control data to capture experimental and assay-specific biases. Second, the predictions from the primary model are incorporated as an additional input track during training on the ChIP-seq, and their contribution to the ChIP profile is regressed out in the loss function. This strategy discourages the model from relearning experimental biases.

Overall, BPNet achieved a Pearson correlation of approximately 0.72 for its ChIP-seq model and up to 0.93 for the ChIP-nexus model when predicting Sox2 binding profiles on the test set. In contrast, performance for total count prediction was lower, with a reported Pearson correlation of 0.62. Based on these results, the authors suggested that incorporating longer input sequences or additional epigenetic information, such as chromatin accessibility data, may be necessary to further improve the accuracy of TF binding predictions.

As a regression-based model, BPNet outputs continuous signal profiles that capture both the precise location and intensity of protein–DNA interactions, closely matching the structure of the input epigenetic data.

2.5.3 Enformer for gene expression and epigenetics prediction

Enformer [73] is a transformer-based DL model developed by DeepMind/Calico. Unlike earlier approaches, it predicts gene expression and multiple epigenetic signals, including CAGE, ChIP-seq, DNase-seq, and ATAC-seq, by modelling long-range regulatory interactions from DNA sequences spanning $196,608$ base pairs at a 128 bp resolution.

Enformer features a dual-tower architecture optimised for genomic sequence modelling, consisting of an initial convolutional stem followed by a deep transformer backbone. The convolutional stem receives an input DNA sequence spanning 196,608 bp, one-hot encoded with 4 channels. It first passes through six residual convolutional blocks. Each block applies dilated convolutions (rates 1, 2, 4, 8, 16, 32) with kernel sizes increasing from 9 to 39 bp, progressively expanding the receptive field while maintaining spatial resolution via residual skip connections. Then, this stem performs downsampling by a factor of 128, converting the raw sequence into a compressed representation of $1,536$ bins $\times$ $1,536$ channels (each bin = 128 bp). The architectural shape becomes $(-1,\ 1,536,\ 1,536)$, where the first dimension represents batch size.

The core processing occurs in 12 transformer layers of the transformer backbone, each layer containing: 8 heads of multi-head self-attention with relative positional encodings, a feed-forward network with an expansion factor of four, a normalisation layer, and residual connections. This block captures interactions across the full genomic context, with attention mechanisms spanning the entire 196 kb input window. Post-transformer, the output tensor maintains dimensions $(-1,\ 1,536,\ 1,536)$.

The output layer performs a central crop of the input sequence to the inner $114,688$ bp region (896×128 bp bins). This ensures all output positions have complete contextual information from both sequence flanks. A final linear layer compresses the spatial dimensionality down to the number of matching target tracks. For multi-task regression (CAGE, DNase-seq, ChIP-nexus), Enformer produces 896 bins per track, where each bin predicts signal intensity across the 114 kb output window. Track-specific output heads enable simultaneous prediction across hundreds of genomic assays from a single 196 kb input sequence.

As a result, the model can learn long-range dependencies across the sequence, achieving an effective receptive field of approximately 114 kb, which is comparable to the scale of TADs

and the regulatory interactions that occur within them.

Enformer was developed as an extension of the Basenji2 project [155], building directly on its training data, prediction targets, loss function, and preprocessing pipelines. The training dataset is composed of $5,313$ human epigenetic tracks and $1,643$ mouse tracks, which include bulk profiles of chromatin accessibility, protein binding (TFs and histone modifications), and transcription start site activity. The datasets were compiled from publicly available resources such as ENCODE [156], the Roadmap Epigenomics projects [157], GTEx (for CAGE/RNA-seq) [158], and other large consortia. The majority of these profiles were generated from immortalised lines (e.g., HEK293, K562, etc.) as well as primary cell types from healthy tissues.

Enformer is also a regression-based model, producing continuous signal values that resemble the input epigenetic profiles. However, unlike BPNet, its outputs are aggregated into binned tracks, where each value represents a genomic region of 128 base pairs.

2.5.4 REnformer for chromatin accessibility and protein binding prediction

Building upon existing DL models has become increasingly advantageous, as modern architectures continue to grow in complexity and training datasets expand in size. Developing and training such models from scratch demands substantial computational resources, particularly high-performing graphical processing units (GPUs). Leveraging pre-trained architectures, therefore, enables more efficient model development, reduces computational costs, and facilitates adaptation to new tasks or data types through fine-tuning rather than full retraining.

REnformer [159] was developed by two research collaborators, Mr Simone Riva and Mr Nicolo Stranieri. It is a ML architecture derived from the pretrained Enformer model and adapted for use with scATAC-seq data. In brief, REnformer retains the core Enformer architecture but replaces the final layer with an extra pointwise convolutional module, which is then fine-tuned on custom single-cell datasets. The model achieves strong performance in predicting regulatory elements, including promoters and enhancers. Figure 2.3 illustrates a summary of the training workflow performed by REnformer.

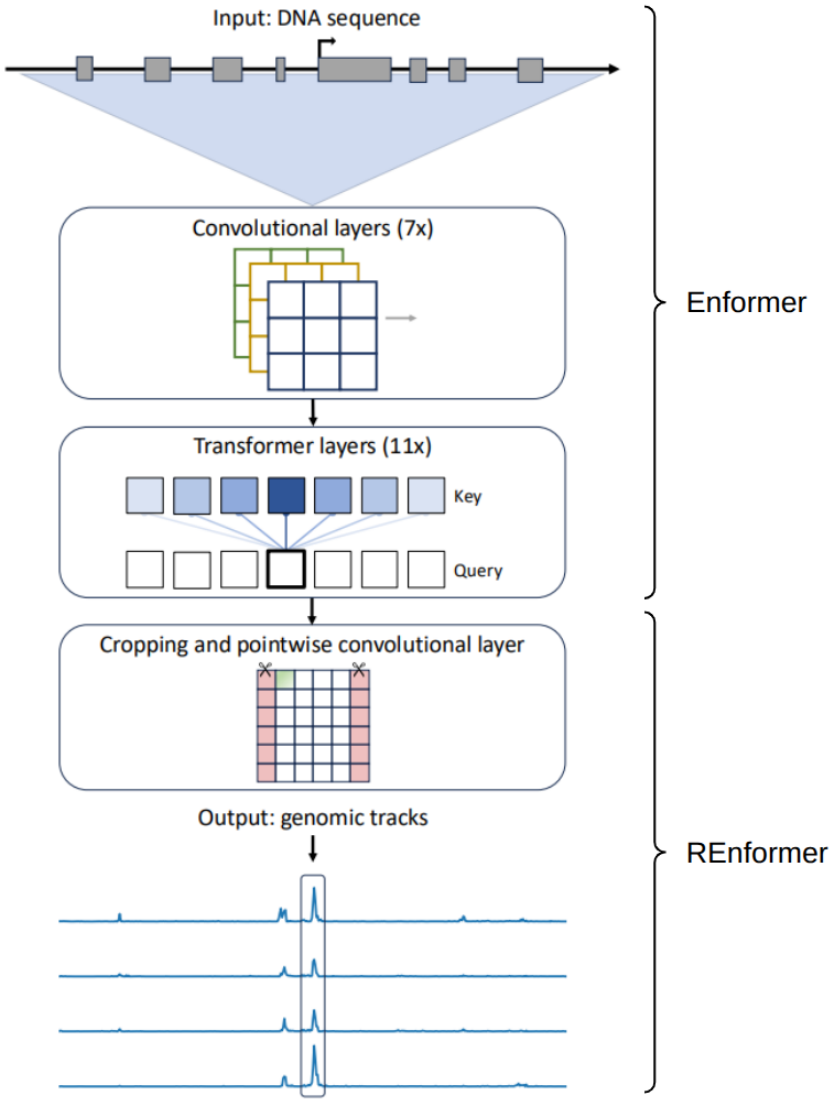


Figure 2.3: Overview of the REnformer architecture. REnformer takes DNA sequences of $196,608$ bp as input, which are progressively compressed and processed through seven convolutional layers. These are followed by 11 transformer layers incorporating attention mechanisms, adapted from the original Enformer architecture. The final layer performs cropping and applies a pointwise operation to generate the reconstructed epigenetic signal. *Source:* Reproduced from Figure 1 of Riva, openRxiv, 2025 [159] under the Creative Commons Attribution 4.0 License.

Since REnformer shares the same network backbone as Enformer, it takes the same ~ 200 kb input sequences and produces outputs in the same format, namely coverage predictions at 128 bp resolution. REnformer was originally trained on 151 scATAC-seq tracks covering almost all tissues in the human body and normalised with the ZEN tool described previously.

Further work on Enformer resulted in a new transformer-based architecture, HYPER, implemented by Mr Simone Riva and Mr Nicolo Stranieri (unpublished). HYPER builds upon the

Enformer framework and similarly consists of two main components: a convolutional stem and a transformer backbone.

In contrast to Enformer, whose convolutional stem includes six residual convolutional blocks, HYPER employs seven convolutional blocks with pooling layers. The transformer backbone comprises 11 multi-head self-attention layers with relative positional encoding, compared to the 12 layers used in Enformer. Despite these architectural differences, the resulting output representation keeps the same shape $(-1, 3,072, 896)$.

In the final prediction layer, the activation function was changed to GELU. In addition, an extra output head was introduced, consisting of a linear layer that expands the feature dimensionality to $3,072$ channels. After the centre cropping, these $3,072$ channels are reduced to the number of matching classes or tracks. An overview of HYPER's architecture is illustrated in Figure 2.4.

Riva *et al.* showed that REnformer achieves improved performance relative to Enformer for the prediction of scATAC-seq profiles, likely reflecting its adaptation to high-resolution chromatin accessibility data and cell-type-specific epigenomic signals.

Figure 2.5 presents a comparison between the predicted and experimentally observed open chromatin signals for both models at the same genomic locus. Visually, REnformer appears to reconstruct the overall structure of the chromatin accessibility signal more faithfully, particularly for narrower peaks and fine-scale accessibility patterns. This qualitative observation is further supported by a substantially lower normalised MSE for REnformer, whose value is around 1, while Enformer's is averaged to 10.

Given the strong architectural similarities between the two models, the observed performance differences are unlikely to arise primarily from differences in model design. Rather, they are more plausibly attributable to differences in the training data. REnformer was trained on high-resolution scATAC-seq data, enabling it to learn cell-type-specific chromatin accessibility patterns with greater precision. In contrast, Enformer was trained predominantly on bulk gene expression and other related genomic signals, which may limit its ability to resolve fine-grained, cell-type-specific open chromatin and protein binding features.

Consequently, HYPER and related Enformer-derived architectures should not necessarily be interpreted as universally superior models, but rather as modelling frameworks specifically

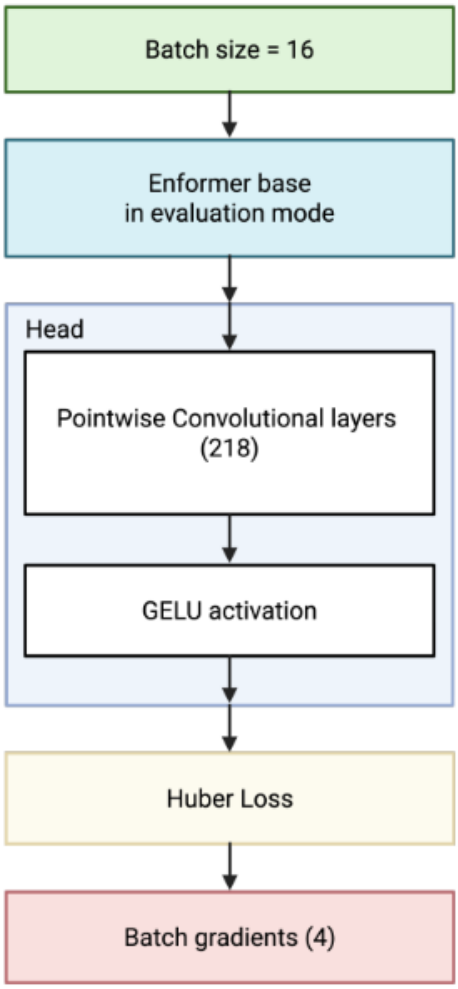


Figure 2.4: Diagram of the HYPER architecture. HYPER consists of the Enformer backbone architecture followed by two additional output layers. During training, the Huber loss is minimised.

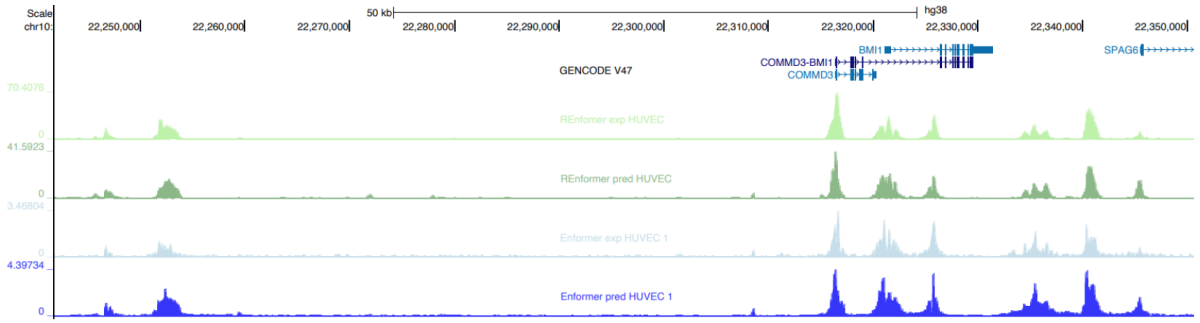


Figure 2.5: Predictions generated by Enformer and REnformer for the genomic region chr10:22,241,840–22,356,528 in HUVEC cells. The light green track represents the experimentally measured HUVEC scATAC-seq signal, while the dark green track shows the corresponding REnformer prediction. The light blue track displays the observed DNase-seq signal, and the dark blue track illustrates the Enformer-predicted open chromatin profile for the same region. *Source:* Reproduced from Figure 2 of Riva, openRxiv, 2025 [159] under the Creative Commons Attribution 4.0 License.

adapted for the reconstruction of cell-type-specific epigenetic profiles from high-resolution chromatin accessibility data. Their performance, therefore, likely reflects both architectural properties and the biological characteristics of the datasets used during fine-tuning.

2.6 Peak calling

Protein binding and open chromatin data are highly imbalanced, with true signal often obscured by substantial noise. Peak-calling methods are computational approaches used to detect regions enriched for genuine epigenetic signal or peaks. In this study, these methods are applied both before and after training: initially to construct the training datasets, and subsequently to assess whether the model has effectively captured the underlying signal.

The choice of peak-calling strategy has important implications for downstream ML applications because peak callers effectively define the regulatory regions used as training labels or evaluation targets. Different algorithms vary in their sensitivity to weak enrichment, handling of background noise, treatment of broad versus narrow peaks, and susceptibility to sequencing or GC-content biases. Consequently, the same underlying epigenetic dataset may generate substantially different sets of candidate regulatory elements depending on the peak-calling framework employed. These differences can directly influence model training by altering the balance between positive and negative examples, modifying the apparent structure of regulatory signal profiles, and introducing systematic biases into the learned representations. In this context, peak calling should not be viewed solely as a preprocessing step, but also as a methodological component capable of shaping both predictive performance and biological interpretation in regulatory genomics models. This is particularly relevant when benchmarking architectures, as apparent differences in predictive performance may partially reflect differences in the construction of the underlying training labels.

2.6.1 LanceOtron

LanceOtron is a DL-based peak-calling framework designed for the analysis of genomic signal data derived from assays such as chromatin accessibility, protein–DNA binding, and histone modification profiling [147]. Compared to traditional methods such as MACS2, LanceOtron

has been shown to achieve improved performance in peak detection. Its workflow can be described as a multi-stage process integrating statistical enrichment and learned representations of signal shape.

The first stage focuses on the identification of candidate enriched regions through a 25-way enrichment testing scheme. This approach combines five smoothing window sizes (100, 200, 400, 800, and 1,600 bp) with five threshold levels (2, 4, 8, 16, and 32), resulting in 25 distinct combinations. This strategy is designed to capture a broad spectrum of signal patterns, including narrow, broad, and weak enrichment regions, thereby generating a comprehensive set of candidate peaks.

In the second stage, local enrichment features are computed for each candidate region. Specifically, the maximum read coverage within each candidate peak is determined and compared against multiple background estimates. These include a global background defined by the mean coverage across the chromosome, as well as ten local background regions of increasing size (ranging from 10 kb to 100 kb) centred on the candidate peak. This multi-scale comparison provides a set of enrichment metrics that contextualise the signal relative to both local and global genomic backgrounds.

The third stage involves characterising the shape of the signal associated with each candidate region. A fixed window of approximately 2 kb surrounding the candidate peak is extracted, and the signal within this region is encoded into a representation capturing key structural features, such as peak height, width, and flanking patterns. This representation is then processed by a CNN, which outputs a score reflecting the likelihood that the observed signal shape corresponds to a true biological peak.

In the final stage, the enrichment features derived from the second step and the shape-based score from the third step are integrated. These inputs are fed into a multilayer perceptron, which produces a final classification score indicating whether the candidate region represents a true peak. Additionally, p-values can be used to filter peaks previously identified on the input tracks.

Ultimately, LanceOtron reports all peaks, or candidates to regulatory elements, exceeding a predefined threshold, along with associated metrics including enrichment statistics, shape scores, p-values, and gene annotations. This threshold influences the sensitivity and specificity of peak

detection; therefore, it is important to optimise and evaluate its selection. Benchmarking between LanceOtron and MACS2 demonstrated that LanceOtron achieves higher precision and F1-scores, while also exhibiting greater robustness to variability in signal shape and intensity.

2.6.2 MACS2 and MACS3

MACS2 (Model-based Analysis for ChIP-seq) is a widely used peak-calling algorithm and is often considered a standard baseline method for identifying enriched regions in genomic assays. Although originally developed for ChIP-seq data, it has also been extensively applied to related assays such as ATAC-seq and CUT&RUN.

At a high level, MACS2 operates on aligned sequencing reads to identify regions of significant enrichment [160]. The method first constructs a model of the fragment size distribution by analysing the spatial distribution of reads mapped to the forward and reverse strands. This allows estimation of the characteristic fragment length. Optionally, control (input) samples can be incorporated to improve background modelling and reduce false positives arising from sequencing biases or genomic artefacts.

Following this, MACS2 adjusts read positions by shifting them by approximately half the estimated fragment length and extending them to represent full DNA fragments. This transformation improves the localisation of enrichment signals by aligning reads more closely with the underlying binding sites.

The algorithm then scans the genome using a sliding window approach and evaluates enrichment at each position by comparing observed read counts to an expected background distribution, typically modelled using a Poisson process. The background can be estimated globally or locally, which could be improved by the previously mentioned incorporation of control data. Statistical significance is assessed by computing p-values for enrichment, which are subsequently corrected for multiple testing to yield q-values.

Significantly enriched regions are then merged into contiguous peaks. For each detected peak, MACS2 reports key statistics, including genomic coordinates, peak summits, fold enrichment, p-values, and adjusted q-values.

MACS3 represents an updated implementation of the same underlying peak-calling framework, retaining the core `callpeak` algorithm while introducing improvements in software engineer-

ing and functionality. The codebase has been modernised, including migration to more recent Python versions, resulting in enhanced computational efficiency, improved memory management, and more robust error handling. Additional features have also been introduced, such as variant calling capabilities from ChIP-seq data.

Overall, the primary advantages of MACS3 over MACS2 lie not in fundamental algorithmic changes but in improved performance, maintainability, and integration within modern bioinformatics workflows.

2.6.3 Hybrid log-normal peak calling

DL-based epigenetic predictions differ fundamentally from experimental sequencing data. In particular, they do not consist of reads, fragments, or count-based coverage profiles. Consequently, conventional peak-calling algorithms developed for sequencing experiments cannot be directly applied to these outputs. To address this limitation, I developed a custom peak-calling strategy specifically designed for the continuous prediction scores produced by our DL models. The approach is based on statistical modelling of the distribution of predicted signal intensities, or chromatin accessibility and protein binding signal, in the context of this thesis.

Genome-wide inspection of the prediction values revealed a strongly right-skewed unimodal distribution: most genomic bins exhibit low predicted intensity, whereas only a small fraction display high values. The distribution rises steeply on the left side toward the mode and then gradually tapers off into a long right tail. This pattern is consistent with a log-normal distribution, which is frequently used to describe positive-valued background processes with occasional extreme events. Accordingly, the predicted signal was modelled under the assumption that background values follow a log-normal distribution. The peak-calling procedure consists of four main steps.

The first step involves fitting the bulk distribution of prediction values. Here, the outputs from HYPER are modelled as log-normally distributed, with the long upper tail corresponding to genuine regulatory activity. Although a single log-normal distribution provides a reasonable approximation of the data, the predicted values are not fully captured by the tail region of its corresponding fitted distribution. In particular, the upper range of the observed data extends beyond what would be expected under a single-component log-normal model. To better represent

the empirical distribution, the data were therefore modelled using a mixture of two log-normal distributions.

The parameters of these two log-normal components are estimated using maximum likelihood. The first model, referred to as the bulk fit, is estimated using the lower 95th percentile of the prediction values. This excludes candidate peaks so that the fitted distribution reflects the background signal without being influenced by extreme values. The second model, referred to as the full fit, is estimated using the complete set of predicted values and therefore better captures the behaviour of the right tail of the distribution. In both models, the location parameter is fixed to zero (`loc = 0`), as prediction scores are strictly non-negative.

A log-normal random variable X has a probability density function (PDF):

$$f(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0$$

Where σ represents the shape parameter and $\exp \mu$ the scale parameter. The parameters ($\hat{\sigma}$,$\hat{\mu}$) are estimated using maximum likelihood estimation (MLE) separately for the bulk and full datasets.

The second step involves identifying the crossover point between the two fitted distributions. Because the corresponding PDFs overlap, there exists a point at which the full fit begins to describe the observed data more accurately than the bulk fit. This crossover point, denoted x^*, is defined as the value at which the two PDFs intersect:

$$f_{\text{bulk}}(x^*) = f_{\text{full}}(x^*)$$

For this, SciPy's `fsolve` function was employed. This function solves nonlinear equation systems by iteratively linearising them with a Jacobian matrix, computing a correction step, and updating the solution guess until convergence.

The third step involves constructing the hybrid model. For this, both distributions are assembled based on the crossover point x^*.

$$f_{\text{hybrid}}(x) = \begin{cases} f_{\text{bulk}}(x) & x < x^* \\ f_{\text{full}}(x) & x \geq x^* \end{cases}$$

The rationale for this approach is that the bulk fit, having been estimated without the influence of extreme values, more accurately captures the shape of the background signal within the central region of the distribution. In contrast, the full fit incorporates the complete dataset, including the right tail, and therefore provides a better characterisation of the tail behaviour relevant for peak, or enriched signal regions, detection. Cubic spline interpolation is applied to the resulting hybrid PDF, which employs cubic polynomials to connect the data points and ensures a smooth transition between these two fits.

The fourth step involves the peak-calling procedure itself. Two commonly used criteria are applied: statistical significance and fold change. Statistical significance is assessed using the survival function (i.e., the right-tail p-value) derived from the full log-normal fit.

$$p(x) = P(X > x) = 1 - F_{\text{full}}(x) = \text{SF}_{\text{full}}(x)$$

Here, the full fit p-value was used instead of the hybrid model because its tail best models the null distribution for assessing how extreme a value is relative to the genome-wide signal.

On the contrary, fold-change was computed relative to the mean of sub-threshold signal (genomic regions with no signal) values.

$$\text{FC}(x) = \frac{x}{\bar{x}_{\text{bulk}}}, \quad \text{where } \bar{x}_{\text{bulk}} = \mathbb{E}[X \mid X \leq x^*]$$

A genomic bin is called a peak if and only if its p-value is below the p-value threshold and its fold change exceeds the fold change threshold.

$$p(x) < p_{\text{thresh}} \quad \text{and} \quad \text{FC}(x) > \text{FC}_{\text{thresh}}$$

Default thresholds were set to $p \leq 0.001$ and fold change ≥ 5, defining regions with a considerably high change in signal and strong statistical significance. This dual criterion prevents calling statistically rare but biologically weak signals (low fold change) or high-amplitude artefacts that are not unusual in the distribution tail.

Higher thresholds were also tested; the selected values served as a sensitive yet stringent starting point, given the 128-bp bin resolution and potential signal dilution. These thresholds yielded peak counts comparable to those called directly from the bp-resolution raw data.

This approach facilitates the identification of enriched signal regions by fitting the data distribution in a flexible manner to the predicted signal, thereby improving sensitivity to subtle features. These steps were necessary to enable a careful, high-resolution analysis of individual predictions when identifying peak regions in DL-reconstructed tracks. However, for analyses involving large numbers of predictions, such as genome-wide applications, it remains useful to employ established peak-calling tools such as LanceOtron or MACS2/MACS3, albeit with reduced or no stringency to accommodate the characteristics of the predicted data.

2.7 Mappability stress test: Fake Fastqs

ML models trained on epigenetic data, such as Enformer, are sequence-to-signal models. They use raw DNA sequence as input and predict quantitative signal tracks. Importantly, they do not have access to sequencing reads, insert sizes, or other experimental artefacts. Therefore, when these models predict a peak, the prediction is based entirely on sequence patterns learned during training.

In contrast, experimental epigenomic data are affected by several sources of bias and variability, including sequencing noise, limited read depth, GC-content bias, mappability issues, and fragment length distributions. As a result, differences between model predictions and experimental observations may arise either because the model is incorrect or because the experiment failed to capture the signal accurately. It is therefore necessary to distinguish sequence-driven predictions from effects caused by experimental or alignment biases.

This distinction is especially important when the model predicts a peak that is not observed in the experimental data. Such cases are usually labelled as false positives. However, the absence of an experimental peak does not necessarily mean that the region lacks regulatory activity. The model may be detecting genuine sequence-encoded regulatory features that were missed due to limited experimental sensitivity or alignment bias. To assess this issue, a mappability stress test was needed.

To achieve this, synthetic “fake FASTQ” datasets were generated to assess genomic mappability in a controlled way. These datasets systematically reconstruct paired-end fragments using all insert sizes observed in the experimental data and reposition them across the genome.

A pipeline comprising two main steps was developed in collaboration with Dr Edward Sanders. First, each chromosome is tiled so that a centre position is defined every 5 base pairs. This step size was chosen to keep the dataset computationally manageable. Second, at each centre position, synthetic paired-end fragments are generated using a set of 60 insert sizes spanning the range observed in the experimental dataset. Because the synthetic reads are 40 bp long, fragments generated from neighbouring centre positions overlap substantially. As a result, a given genomic position is covered by fragments originating from approximately eight adjacent centre positions.

Under ideal alignment conditions, all synthetic fragments should map uniquely back to their original genomic location. Since approximately eight centre positions contribute coverage to each base and 60 insert sizes are generated at every centre position, a perfectly mappable region is expected to have a coverage value of $(60 \times 8 = 480)$. Thus, a coverage value of 480 represents the expected number of synthetic fragments overlapping a genomic position when all fragments align correctly. Regions with coverage lower than 480 indicate reduced mappability, meaning that some fragment lengths fail to align uniquely, which could lead to under-detection of peaks. In contrast, regions with coverage higher than 480 suggest multi-mapping or repetitive sequences, where reads align to multiple locations and experimental coverage may be inflated.

The resulting insert-size count track provides a genome-wide measure of alignment reliability. By comparing predicted peaks with regions showing abnormal insert-size counts, it becomes possible to determine whether apparent false positives occur in areas affected by mappability artefacts. Importantly, this approach does not prove that a false positive represents true biological activity. Instead, it helps identify cases where differences between the model's predictions and the experimental data may be explained by alignment bias rather than inaccuracies in the model itself. Predictions located in regions with stable insert-size counts are therefore more likely to reflect genuine sequence-driven regulatory potential, while predictions in problematic regions can be treated with caution, increasing overall confidence in the model's results.

2.8 Feature attribution methods

DL models are becoming increasingly large and complex, often resembling black boxes more than traditional statistical methods or conventional computational programs. As their ability to capture intricate patterns improves, so do questions about how they arrive at specific decisions and which intermediate representations or inputs drive those outcomes. This growing complexity highlights the need for methods that can interpret and explain model behaviour.

The field of interpretable ML addresses this need by providing techniques to clarify how and why a model makes a particular prediction. Among these, feature attribution methods aim to quantify the contribution of each input data point to the final output. From a neural network perspective, this requires quantifying how the output of one neuron influences subsequent layers in the network. In molecular biology applications, this typically involves assigning a score to each nucleotide in a sequence, indicating how strongly changes at that position would influence the model's prediction.

From a biological perspective, feature attribution methods provide a means to link model predictions back to underlying regulatory mechanisms encoded in DNA sequence by assigning importance scores to individual nucleotides. High attribution scores highlight positions within a sequence that are likely to be functionally relevant, such as core binding motifs or regions contributing to cooperative binding. Moreover, analysing patterns of attribution across sequences can reveal higher-order regulatory features, including motif spacing, orientation, and interactions between distal elements, providing insight into the regulatory grammar governing gene expression. These interpretations allow model predictions to be compared with known biological knowledge.

All of the following techniques are implemented in Captum, a Python-based package for supporting interpretability of PyTorch models [161].

Integrated gradients

Integrated gradients (IG) is a feature attribution method [77]. It relies on defining a baseline input, usually a vector of zeros. In brief, the workflow proceeds as follows: for each data point in the input, the method constructs a straight-line path between the baseline and the actual

input. It then samples intermediate points along this path and computes the gradient at each point, which represents the local rate of change or slope of the model's output. These gradients are summed and averaged to approximate the integral along the path, yielding the integrated gradient for each input feature. Finally, this value is multiplied by the difference between the input and the baseline to obtain the final attribution score.

In addition to being computationally efficient, IG addresses the saturation problem in NNs. Many NNs use activation functions such as Sigmoid or Tanh, which are steep in their central region but become nearly flat at the ends. When a feature takes on very large or very small values, the model output may lie in one of these flat regions, where the gradient is close to zero. As a result, gradient-based attribution methods that rely only on the local gradient at the input can incorrectly assign a near-zero importance to a feature, even if that feature strongly influenced the prediction. IG solves this issue by accumulating gradients along a continuous path from a baseline input to the actual input. Instead of evaluating only the final gradient.

SHAP

Shapley values originate from cooperative game theory, where they are used to fairly distribute a total payoff among players based on their individual contributions to the overall outcome. In ML, SHAP (SHapley Additive exPlanations) adapts this concept to quantify the contribution of each feature to a specific model prediction [162]. A SHAP value for a feature is computed as the average marginal contribution of that feature across all possible subsets of features, comparing the model's prediction with and without the feature included.

SHAP is widely regarded as a standard method for feature attribution in model interpretability because it satisfies three key properties: local accuracy (the feature contributions sum to the model's prediction), consistency (if a feature's impact increases in a model, its SHAP value does not decrease), and missingness (features that do not affect the prediction receive zero contribution).

DeepLIFT

Deep Learning Important FeaTures (DeepLIFT) extracts feature attributions by assigning scores to input features based on how much the model's prediction changes relative to a baseline in-

put [78]. Similar to the IG approach, DeepLIFT requires a reference (baseline) input, which is typically an all-zeros matrix.

The core workflow of DeepLIFT consists of three main steps. First, the model generates predictions for both the original input and the baseline input. Second, for each neuron y, DeepLIFT computes the difference D_y between the activation produced by the original input and the activation of the baseline input. This difference must equal the sum of the contributions from all parent neurons (i.e., neurons in the previous layer) connected to neuron y. The contribution of a parent neuron x to neuron y is defined as the product of a multiplier m_{xy} and the activation difference D_x.

Since contributions in one layer depend on those in the previous layer, the multipliers are computed using a chain rule-like procedure, similar to the backpropagation of derivatives, to define multipliers m from any neuron x to a downstream neuron t.

$$m_{xt} = \sum_{y \in \text{path}} m_{xy} m_{yt}$$

Finally, DeepLIFT produces a per-feature importance score that indicates how much each input feature explains the difference in the model's output between the original input and the baseline reference. The backpropagation-like process is briefly summarised in Figure 2.6

DeepLIFTshap

DeepLIFTShap is an attribution method that assigns contribution scores to input features by combining ideas from DeepLIFT and SHAP. Unlike many other attribution methods, DeepLIFTShap requires more than one baseline reference. In genomic applications, in addition to an all-zeros matrix, another baseline may be based on randomly selected non-blacklisted regions with similar GC content and low signal.

DeepLIFTShap consists of three main steps. First, a distribution of baselines is selected. Second, the DeepLIFT procedure is applied separately for each baseline. This produces one attribution matrix per baseline, where the scores reflect the differences between each feature and the corresponding baseline. Finally, the attributions are averaged across all baselines. This average is treated as an estimate of the SHAP values, as it approximates the expected contribution of

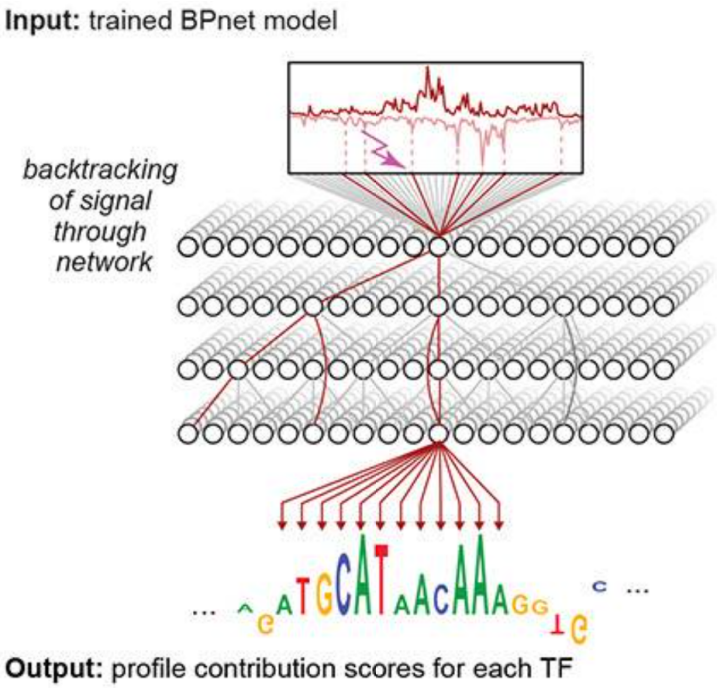


Figure 2.6: Diagram of the backpropagation-like process of DeepLIFT. DeepLIFT estimates the contribution of each base pair in an input DNA sequence by propagating the prediction backwards through the network, recursively attributing the output to individual inputs across successive layers. *Source:* Reproduced from Figure 2A of Avsec, Nature, 2022 [68] under the Creative Commons Attribution 4.0 License.

each feature by comparing model predictions with and without that feature.

In addition to its implementation in Captum, DeepLIFTShap was originally implemented as part of the BPNet-refactor [163] prediction module.

2.9 Motif identification tools

Feature attribution methods provide insight into which regions of the input data contribute most strongly to a model's predictions. This information can be leveraged to identify sequence motifs, which can then be visualised as sequence logos and compared with known motif databases to evaluate their biological relevance. These tools enable the discovery, comparison, and genome-wide mapping of predicted motifs. From a biological perspective, this allows the identification of TFBSs and the validation of whether the model has learned known regulatory patterns or potentially novel ones. These approaches therefore bridge model interpretation and biological understanding, linking predictive features to underlying regulatory mechanisms. Several such methods are publicly available, and those used in this study are based on distinct

underlying principles, as described below.

2.9.1 TF-MoDISco

TF-MoDISco is an attribution-based framework for discovering sequence motifs and interpreting genomic sequence models [164]. It takes one-hot encoded sequences along with a per-nucleotide attribution score matrix as input. The method is compatible with feature attribution approaches such as DeepLIFT, DeepLIFTshap, IG, and others.

The workflow consists of four primary steps: (1) identifying short, highly important subsequences known as seqlets; (2) computing pairwise similarities or differences between motifs; (3) clustering similar motifs to form consensus motifs; and (4) generating position weight matrices (PWMs) for each consensus motif. The outputs include PWMs, absolute contribution weight matrices (CWMs) that summarise contribution patterns and highlight the most influential nucleotide at each position, and hypothetical CWMs, which represent the importance of all four nucleotides at every position.

Importantly, when comparing predicted motifs to known motif PWMs, the hypothetical CWMs should be used rather than the absolute contribution matrices. This is because the importance of all four nucleotides at each position is required for comparison to standard PWMs, which represents nucleotide probabilities or weights across positions. By default, TF-MoDISco uses DeepLIFTshap to generate these contribution scores.

TF-MoDISco incorporates downstream analysis with Tomtom to enable comparison of its discovered motifs against known motif databases.

2.9.2 TomTom

Tomtom is a tool within the MEME Suite designed for motif–motif comparison [165]. It takes one or more query motifs and searches them against one or more target motif databases, ranking the targets by similarity and providing optimal alignments.

Tomtom requires a query motif file containing PWMs in MEME format, along with a target motif database for comparison. It systematically compares each query motif to every motif in the target database and calculates a similarity score, by default using the Pearson correlation

between aligned columns, although alternative distance metrics are available, such as Kullback-Leibler, etc.

In addition, Tomtom estimates the statistical significance of each match, reporting a q-value by default. Only matches below a user-defined significance threshold (default ≤ 0.5) are considered. These significantly enriched epigenetic regions are included in the output. The final report summarises all significant matches, including their alignments and associated significance scores.

2.9.3 FIMO

FIMO (Find Individual Motif Occurrences) is a tool within the MEME Suite that scans DNA sequences to identify matches to individual motifs. It evaluates each motif independently, sliding it across every position in the input sequences and scoring potential matches using a log-odds ratio relative to a background model.

The default background model in FIMO is predefined and represents expected nucleotide frequencies under a null scenario. It was generated by estimating single-nucleotide frequencies across a large collection of human genomic sequences sampled from NCBI databases and then averaging these frequencies. This background model serves as the baseline for assessing the statistical significance of motif matches.

FIMO requires a motif file containing PWMs in MEME format and a FASTA file with the sequences to be scanned. The search process involves four main steps. First, for each motif at every sequence position, FIMO computes a log-likelihood ratio score. Second, it assigns a p-value to each score based on a null model defined by background nucleotide frequencies. Third, it applies Benjamini–Hochberg false discovery rate (FDR) correction across all tested motif–position combinations. Ultimately, it reports only those matches that pass a user-defined significance threshold, by default corresponding to a p-value lower than 0.0001.

FIMO can be used to identify genome-wide occurrences of a known motif. In this case, the input consists of a known motif file including its different flavours extracted from a public motif database in MEME format, and a reference genome in FASTA format. FIMO can be used to locate these motifs' genomic instances. When combined with Tomtom, these predicted binding sites can subsequently be intersected with open chromatin regions or protein binding

sites by integrating nucleotide-level importance scores derived from epigenetic models.

2.9.4 Motif databases

Motif databases compile and curate TF binding models, typically represented as position frequency matrices (PFMs) or PWMs. PFMs store the observed nucleotide counts at each position within aligned binding sites, whereas PWMs are derived from PFMs by converting nucleotide frequencies into log-likelihood scores, thereby quantifying binding preferences at each position.

These databases are derived from both *in vivo* and *in vitro* experimental assays. *In vivo* techniques such as ChIP-seq identify protein–DNA interactions within their native chromatin context in living cells or tissues. In contrast, *in vitro* methods such as SELEX [166] and HT-SELEX [167] characterise intrinsic DNA-binding preferences under controlled experimental conditions, independent of chromatin context.

Over time, several publicly available databases of experimentally validated TFBSs have been established and widely adopted by the research community. Prominent examples include JASPAR [168], HOCOMOCO [169], and TRANSFAC [170]. These resources are commonly used to annotate de novo discovery of motifs through tools like the MEME Suite or TF-MoDISco. Among these databases, JASPAR is the largest and most widely used open-access repository, containing binding profiles for over one thousand TFs across multiple species. HOCOMOCO is another extensively used resource, providing high-quality, ChIP-seq–derived binding motifs for approximately 700 human and mouse TFs.

2.9.5 Nucleotide importances to find motifs

Nucleotide importance scores represent how much each input base pair contributes to the prediction of a given signal peak. These scores are typically distributed in a “carpet-like” pattern, where stretches of consecutive bases receive consistently high or low values, which can later be interpreted as potential TF binding motifs. This representation is particularly useful for studying the functional impact of genetic variants, as it highlights not only where motifs occur but also how strongly they influence model predictions.

A combination of the approaches described above forms a conventional bioinformatics pipeline for motif discovery and interpretation. In this workflow, four main methods are used: FIMO, TomTom, HYPER, and a feature-attribution approach such as DeepLIFTShap.

First, FIMO is applied to the DNA sequence using a TF database such as JASPAR or HOCOMOCO. This step identifies candidate motif occurrences by scanning the sequence and reporting genomic coordinates of matches based on statistical significance (p-values). It provides a comprehensive map of potential TF binding sites, although it is only based on DNA sequences rather than accounting for their functional relevance in a specific epigenetic context.

Second, HYPER is run on the same DNA sequence to generate predicted signal tracks, or reconstructed epigenetic profiles, and nucleotide importance scores are computed using DeepLIFTShap (or a similar feature-attribution method). These scores are derived with respect to specific outputs of interest (e.g., presence or absence of a peak), enabling the identification of sequence regions that are most influential for the prediction. This step introduces a functional layer, linking sequence patterns to model-derived regulatory activity.

Third, for each motif identified by FIMO in the first step, the corresponding nucleotide importance scores are extracted from the importance tracks. These scores are then aggregated across motif instances and normalised (e.g., so that each row sums to 1), producing PWMs that reflect importance rather than raw sequence frequency. The resulting matrices are stored in MEME format, effectively converting importance profiles into interpretable motif representations.

Fourth, TomTom is run on the generated MEME files containing the importance-derived PWMs, using the same TF database as in the first step. This operation compares the derived motifs against known TF motifs and returns a list of statistically significant matches based on similarity metrics and p-values. In this way, motifs inferred from model importance scores can be annotated with putative TF identities.

Fifth, nucleotide importance score coverage is extracted from the HYPER-predicted importance bigWig tracks for each matched motif instance.

Finally, all results are integrated into a ranking framework that prioritises motifs based on multiple complementary criteria, including FIMO p-values (sequence match confidence), TomTom p-values (motif similarity), the structure and sharpness of the PWMs, and the magnitude of predicted importance or enrichment. These factors are combined through an averaged summation,

enabling an integrated approach that identifies the most biologically relevant and functionally active motifs.

However, because this framework integrates multiple dimensions, it may also prioritise motifs that are not directly or exclusively associated with the regulatory function under study. In particular, motifs may be selected if they occur at overlapping genomic locations with true binding sites or if their PWMs share similarity with canonical motifs of interest. Rather than representing a limitation, this behaviour is intentional and beneficial. Certain TF motifs, such as NF-E2, may be underrepresented or incompletely characterised in databases like JASPAR, leading to weaker scores in some criteria (e.g., motif similarity) while still showing a strong signal in others (e.g., nucleotide importance).

By combining positional, statistical, and importance-based features, the framework increases the likelihood of capturing these partially represented or variant motifs, allowing them to appear in the final ranking even if they do not score highly across all individual metrics. This approach provides a more flexible and inclusive representation of the regulatory landscape, particularly in regions with overlapping or combinatorial binding, where multiple related motifs may contribute to the observed signal.

This flexibility enables a more comprehensive representation of the regulatory landscape, capturing both primary drivers and closely related motif variants, and allowing for a nuanced interpretation of the sequence features underlying the observed signal patterns and their modulation by genetic variation.

2.10 Distribution similarity metrics

Performance metrics such as MSE, BCE, sensitivity, recall, and others quantify how close the DL model's predictions are to the true values. However, assessing biological relevance requires a broader perspective that captures the overall shape of the predictions and their similarity to real signal patterns. To achieve this, distribution-based comparisons are used to evaluate how closely the predicted data distribution matches the experimental data. From a biological perspective, this is essential for ensuring that models accurately reproduce meaningful regulatory features, such as peak shapes, signal intensity, and variability associated with protein binding or chromatin accessibility. Capturing these patterns is critical for reflecting true regulatory

activity, as it enables the identification of functional elements and preserves the structure of biological signals beyond point-wise accuracy.

Pearson correlation coefficient

In the context of DL-based models, Pearson correlation coefficient (PCC) is the most common method to measure the linear relationship between two distributions. This makes it very useful when identifying peaks since it is scale-invariant.

For two genomic tracks, where X is a real value, and Y is a predicted value, the PCC formula is

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Here, the numerator is equivalent to the covariance between the distributions, and the denominator is the standard deviation, which normalises the results so it always falls between -1 and 1, where 1 is a perfect match, 0 indicates no linear relationship, and -1 reflects a perfect inverse relationship.

PCC is often considered a gold standard for evaluating DL models in epigenetics because it is invariant to read counts and emphasises the shape of the signal rather than its absolute scale. However, it has notable limitations. It assumes a linear relationship between variables, which can lead to underestimation of non-linear patterns such as exponential trends or those arising from non-normally distributed data, an important limitation given that epigenetics datasets are often not normally distributed. Additionally, its performance can be compromised in highly imbalanced datasets.

Cosine similarity

Cosine similarity measures how closely two vectors align in direction, disregarding their magnitude. It is computed as the cosine of the angle between the vectors, effectively the dot product divided by the product of their norms. A value of 1 indicates perfect alignment, while -1 reflects vectors pointing in opposite directions.

In the context of epigenetic data analysis, cosine similarity is particularly useful for sparse datasets, as it focuses on the relative pattern of non-zero signal across bp bins rather than

absolute values. However, it has limitations when dealing with outliers, since extreme values can still yield high similarity scores if the overall directional pattern is preserved.

Kullback-Leibler divergence

In information theory, relative entropy, also known as Kullback–Leibler (KL) divergence, is used to quantify how much information is lost or gained when one probability distribution is used to approximate a reference distribution. It provides a measure of how different the predicted data distribution is from the true distribution, making it a useful tool for evaluating how well a model captures the underlying structure of the data.

Considering P as the reference distribution and Q as the approximated distribution, then the KL divergence is expressed as

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

where the logarithm represents the difference in information between the two distributions. As a general guideline, KL divergence values that are less than or equal to 0.2 are often considered indicative of a good match between distributions.

There are several advantages to using this metric for validating DL model outputs. It is scale-invariant, as it normalises both distributions to sum to one, allowing comparisons that focus on the relative placement of the signal rather than the absolute magnitude. This makes it particularly useful for identifying whether predicted peaks align with true signal regions. Furthermore, it emphasises informative regions of the data, helping to highlight signal peaks even in datasets that are heavily imbalanced toward background noise, and providing a more meaningful assessment of biological relevance.

Wasserstein distance

Wasserstein distance (also known as Earth Mover's Distance) quantifies the cost of transforming one distribution into another by measuring the "effort" required to move probability mass across a geometric space. Different orders of the Wasserstein distance determine how strongly long-distance transport is penalised: the first order assigns a linear cost proportional to mass

times distance, the second order increases this penalty to mass times distance squared, and higher orders (e.g., cubic) penalise long-distance movements even more heavily, making them particularly sensitive to outliers.

Considering P as the reference distribution and Q as the approximated distribution, the first-order Wasserstein Distance is

$$W_1(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| \, d\gamma(x,y)$$

This formulation can be understood as finding an optimal transport plan (γ) that redistributes all the mass from distribution P to distribution Q while minimising the total transport cost or penalisation.

In this study, the Wasserstein distance will be computed without explicitly accounting for genomic coordinates, as it will be applied exclusively to peak regions, which are already considered biologically meaningful. As a result, the metric will compare the distributions of signal within these enriched regions between the predicted and experimental data. Consequently, lower Wasserstein distance values indicate greater similarity between the two distributions.

One of the key advantages of this metric for validating epigenetic DL model outputs is its sensitivity to both the shape and spatial distribution of the signal. It penalises predictions that are overly smooth or diffuse, for example, a "flat" prediction in place of a sharp peak, even when they are correctly located. In this way, it provides a more informative assessment by addressing not just whether the signal is in the right place, but also whether its shape and structure are accurately captured.

Kolmogorov-Smirnov test

The Kolmogorov–Smirnov (KS) test compares the empirical cumulative distribution functions (ECDFs) of two datasets, focusing on differences in their value distributions. Let $F_{1,n}$ represent the ECDF of the reference data and $F_{2,m}$ that of the predicted data, where n and m denote the respective sample sizes. Here, $F_{1,n}$ can be described as

$$F_{1,n}(x) = \frac{\text{number of elements in Sample 1} \leq x}{n}$$

In other words, it can be viewed as a step function that increases by $\frac{1}{n}$ at each data point, representing the proportion of values in the dataset that are less than or equal to that point.

Then the KS statistic D is calculated as

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

In other words, both datasets are first sorted and represented as step functions, after which the vertical differences between them are evaluated. The KS statistic, D, is then defined as the maximum vertical distance between these two step functions. The KS statistic ranges from 0 to 1, with values closer to 0 indicating greater similarity between the two empirical ECDFs.

In the context of DL models applied to epigenetic data, this metric offers several advantages. It penalises diffuse or "blurry" predictions, thereby favouring sharp, well-defined signals over background noise. It is also scale-invariant, enabling comparisons that focus on the overall structure of the distribution rather than absolute values. As a result, it helps assess whether both the range and intensity of the predicted signal are realistic, providing a more biologically meaningful evaluation of model performance.

Jensen-Shannon divergence

The Jensen–Shannon divergence (JSD) is another measure of similarity between two probability distributions. It is derived from KL divergence but introduces two key improvements that make it more suitable for validation tasks: (i) it is symmetric, meaning the divergence between the reference and predicted distributions is bidirectional, and (ii) its values are bounded between 0 and 1, where 0 indicates identical distributions and 1 indicates no overlap between them.

Instead of comparing both distributions, P and Q, directly, JSD generates an average distribution M.

$$M = \frac{1}{2}(P + Q)$$

Then it calculates how much P and Q differ from the average M.

$$\text{JSD}(P \parallel Q) = \frac{1}{2}D_{KL}(P \parallel M) + \frac{1}{2}D_{KL}(Q \parallel M)$$

where D_{KL} is the KL divergence.

JSD is particularly well suited for comparing epigenetic datasets because of its robustness to zeros and noise. Rather than directly contrasting zero-valued predictions with non-zero reference values, which can lead to inflated divergence, it introduces an intermediate average distribution M, stabilising the comparison and preventing extreme values. Additionally, it is effective at capturing differences in peak structure, penalising overly smooth or flat predictions and favouring those that accurately reproduce sharp, well-defined signal patterns.

3 | Comparative evaluation of deep learning architectures for predicting CTCF binding

3.1 Introduction

As described in Chapter 1, ML methods are capable of predicting regulatory activities such as open chromatin or TF binding from DNA sequence. This provides a direct framework for studying how regulatory information is encoded and interpreted within the genome. Many regulatory assays, such as chromatin accessibility, reflect the combined activity of multiple TFs and regulatory processes, resulting in signals that are inherently heterogeneous and difficult to interpret. This makes it challenging to understand how individual sequence features contribute to the signal and to interpret what the model has learned. In this context, predicting the binding of a single TF provides a more defined and interpretable problem. CTCF is particularly well suited for this purpose, as it has a well-characterised binding motif and a relatively consistent sequence-driven binding behaviour across the genome. This allows a clearer link between DNA sequence and observed binding signal compared to more complex regulatory assays. CTCF provides a simplified setting in which differences in model performance can be more directly attributed to model architecture rather than underlying biological complexity. This makes it an appropriate system for systematically comparing different DL approaches before extending the analysis to more complex regulatory signals.

We hypothesised that profile-based and long-range transformer architectures would outperform conventional CNN-based classification and regression models in predicting CTCF binding patterns. To test this, this chapter presents a comparative evaluation of three DL models representing two widely used architectural paradigms, CNNs and transformers, focusing on their ability to predict CTCF protein binding. Specifically, the models evaluated include a DeepSEA-based CNN framework (Section 2.5.1), the profile-based BPNet architecture (Section 2.5.2), and the transformer-based REnformer framework (Section 2.5.3).

Because each model was originally developed with a distinct preprocessing strategy tailored to its biological objectives and architectural design (Section 2.3), separate preprocessing pipelines are required for each approach to preserve the relevant regulatory information required for

model training. To ensure a fair comparison between these architectures, model validation is structured across three levels, each designed to assess the models' ability to reproduce cis-regulatory activity at the alpha-globin locus. The model that performs best across most or all validation levels will be selected for further evaluation in an extended validation stage, which examines genome-wide signal and compares predictions with experimental data.

In addition to predictive accuracy, feature importance analyses will be performed to identify the sequence features, motifs, and regulatory regions that most strongly influence model predictions. This enables assessment of whether the sequence features influencing model predictions are consistent with known functional elements, such as TF binding motifs, enhancers, or chromatin architectural features, rather than spurious patterns. If the selected model demonstrates strong agreement with experimental signal and biologically meaningful feature attribution patterns, it will be used for downstream analysis.

Chapter aims

The specific objectives of this chapter are as follows:

- To compare the performance of DeepSEA (CNN-based), BPNet (profile-based CNN), and HYPER (transformer-based) models.
- To evaluate the predictive accuracy of these models for CTCF binding sites.
- To assess the best-performing model's ability to predict genome-wide signal profiles and generalise to previously unseen sequences.
- To establish a robust and reliable validation pipeline for DL models in epigenetic signal prediction.

This chapter is organised as follows. The first two sections provide an overview of the datasets used and the model architectures evaluated. The third section presents a comprehensive comparison of model performance, structured across three levels of validation: (i) reconstruction of protein-binding tracks, (ii) identification of nucleotide-level importance for motif discovery, and (iii) prediction of variant effects based on SNPs. Subsequently, the best-performing model is subjected to extended validation, including genome-wide signal prediction and genome-wide SNP effect assessment. The chapter concludes with a discussion of the results, key insights,

and overall conclusions.

This chapter evaluates whether DL architectures can accurately predict CTCF binding directly from DNA sequence while learning biologically meaningful regulatory representations. In addition to assessing overall predictive accuracy, the chapter examines the ability of the best-performing model to reconstruct genome-wide signal profiles and generalise to genomic regions not encountered during training. Successful recovery of experimentally observed binding patterns across unseen sequences would support the interpretation that the model captures sequence features strongly associated with regulatory activity. Conversely, failure to accurately reconstruct signal profiles or generalise beyond the training data would suggest that important aspects of regulatory organisation are not adequately represented within the sequence representations learnt by the analysed ML architecture.

Together, these analyses aim to assess not only comparative model performance, but also the extent to which learned sequence representations reflect biologically interpretable features of genome regulation.

3.2 Description of the DL architecture

Although all three models are designed to learn regulatory patterns from functional genomics experiments measuring protein-DNA interactions, they differ in the biological scale and signal resolution they model.

DeepSEA-based CNN

The DeepSEA-based CNN is a classification model originally developed to predict CTCF protein-binding-enriched regions from short DNA sequences. The model takes 1 kb one-hot encoded DNA sequences as input and processes them through multiple convolutional and pooling layers to identify sequence motifs associated with CTCF protein binding. Its primary advantage is its relatively simple and computationally efficient architecture, which is effective at learning local sequence features and predicting whether a genomic region contains regulatory activity corresponding to CTCF binding. However, because it operates on short genomic windows and produces probabilistic outputs, it cannot model precise binding profiles or long-range chromatin

interactions as effectively as BPNet or HYPER. More details about this model's architecture are available in Section 2.5.1.

BPNet

BPNet is a CNN-based regression model designed to predict protein-binding signal at base-pair resolution from short DNA sequences. BPNet takes 1 kb one-hot encoded DNA sequences as input and quantifies the protein-binding signal within it. Unlike DeepSEA, BPNet was trained directly on strand-specific ChIP-seq or ChIP-nexus signal profiles, enabling it to learn precise TF footprints and motif interactions from nucleotide-resolution data. Its architecture uses dilated convolutional layers without pooling, allowing the receptive field to expand while preserving positional resolution across the sequence. A major advantage of BPNet is its ability to generate base-resolution attribution profiles that are consistent with local motif syntax, including features related to motif spacing, orientation, and potential cooperative binding interactions, thereby supporting biological interpretation. However, compared with transformer-based models such as HYPER, BPNet remains limited to relatively local genomic contexts and cannot efficiently capture long-range enhancer–promoter or chromatin interactions spanning hundreds of kilobases. More details about this model's architecture are available in Section 2.5.2.

HYPER

HYPER is a transformer-based architecture derived from the pretrained Enformer and REnformer frameworks, which were originally trained on thousands of functional genomics tracks including ChIP-seq, DNase-seq, ATAC-seq, and gene expression data. Unlike the CNN-based approaches, which take 1 kb DNA sequences, HYPER quantifies the epigenetic signal from approximately 200 kb DNA sequences, enabling the model to capture long-range regulatory dependencies and chromatin interactions across large genomic regions. Architecturally, HYPER combines convolutional layers for local feature extraction with multi-head self-attention transformer layers that model interactions between distant regulatory elements. This provides a major biological advantage for studying CTCF, whose function is strongly linked to chromatin looping and higher-order genome organisation. Compared with DeepSEA and BPNet, HYPER can therefore learn broader regulatory context and distal interactions; however, this increased

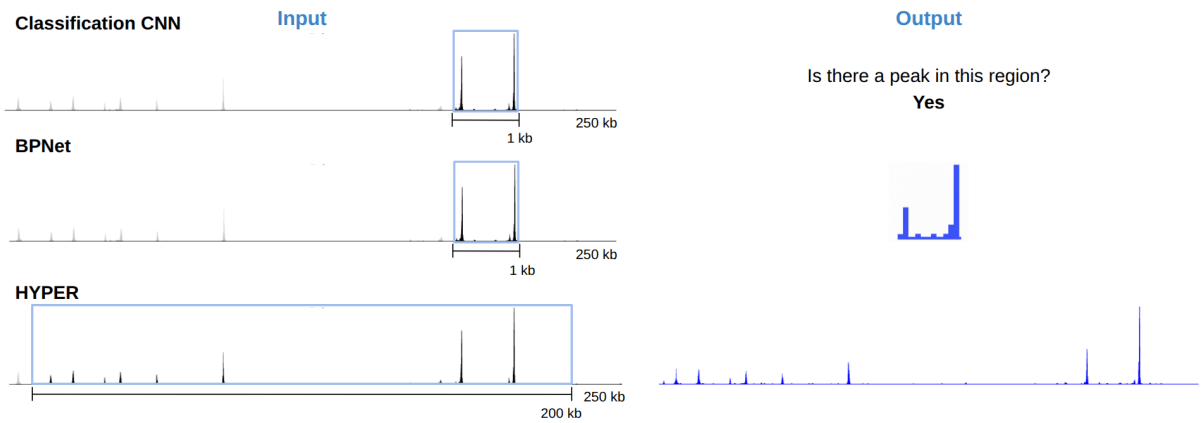


Figure 3.1: Diagram illustrating the inputs and outputs of the three models. Light blue squares on the left indicate the length of the DNA sequence provided to each model: the CNN-based models receive 1 kb sequences, whereas HYPER processes 196,608 bp sequences. On the right, representative visualisations show the different outputs: the classification CNN produces a binary prediction indicating whether epigenetic activity is present in the input sequence, BPNet generates a 1-kb-resolution reconstruction of the epigenetic signal, and HYPER produces a 128-bp-resolution reconstruction of the signal.

complexity requires substantially greater computational resources and reduces nucleotide-level precision relative to BPNet's base-resolution predictions. More details about this model's architecture are available in Section 2.5.4.

Despite differences in the models' architecture and pretraining approaches, all three approaches aim to learn biologically meaningful regulatory patterns while minimising technical noise and sequencing bias. Figure 3.1 presents a representative scheme of each of their inputs and outputs. Although they operate at different biological scales, from local motif recognition in DeepSEA, to nucleotide-resolution binding profiles in BPNet, and long-range chromatin interactions in HYPER, each model seeks to identify the sequence features underlying TF binding and genome regulation.

3.3 Available datasets

In-house 50 donors CTCF binding

The primary analyses in this chapter are based on an in-house dataset comprising 50 donors from the Oxford Biobank, generated in the Hughes lab. This dataset, in addition to both polyA+

(stable fraction) and polyA- (non-polyadenylated ribosomal depleted fraction) RNA-seq and chromatin marks such as H3K27ac, also contains high-resolution epigenetic profiles of erythroid cells, including CTCF ChIP-seq and ATAC-seq across all 50 individuals. All donors were randomly selected and include representative male and female donors, resulting in a cohort with inherent inter-individual genetic variation.

Erythroid cells were derived from peripheral blood mononuclear cells (PBMCs) isolated from the 50 inviduals' blood samples. CD34+ haematopoietic progenitors were isolated from PBMCs by antibody labelling and magnetic bead sorting (Sachith Mettananda protocol, modified by Damien Downes). Cells underwent seven days of expansion followed by fourteen days of erythroid differentiation. CD34+ purity and erythroid maturation were routinely confirmed by flow cytometry. Genomic assays were performed on all donors at the late erythroblast stage, just before enucleation, ensuring consistent maturation across the cohort. Long-read whole-genome sequencing using the $10\times$ linked-read platforms was also performed for each donor at approximately 30X coverage to allow for high-quality variant calling and extensive genome phasing.

Figure 3.2 shows the aggregated epigenetic profiles from the 50 individuals, including CTCF

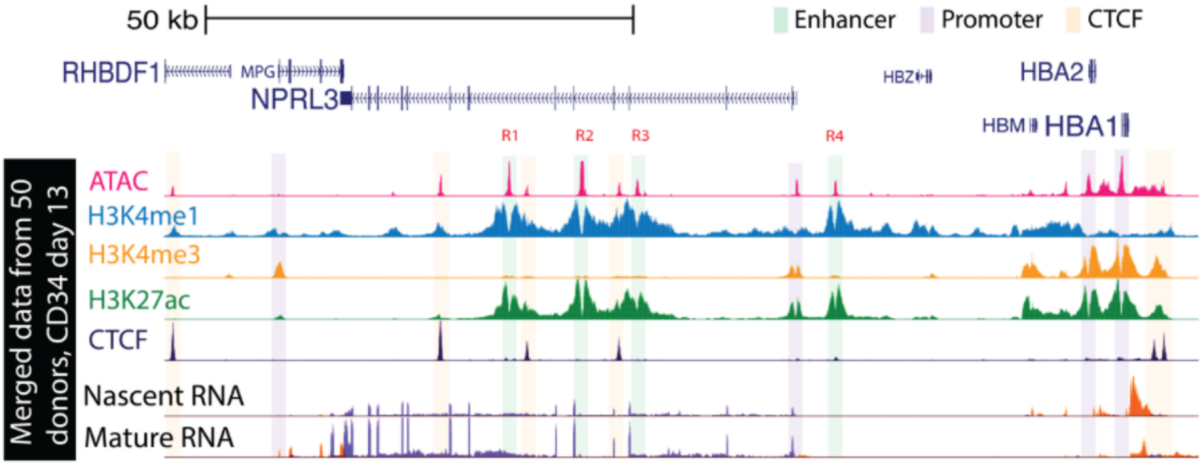


Figure 3.2: Combined epigenetic data from the 50 individuals. Epigenetic profile at the HBA locus. ATAC (pink track), H3K27ac (green track) and CTCF (purple track) ChIP-seq combined data from 50 individuals. H3K4me1 (blue track) and H3K4me3 (yellow track) ChIP-seq from the 5 available sets. Nascent RNA (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. *Source:* Reproduced from Figure 2.2 in Gosden, 2024 (doctoral thesis) with permission of the authors [171].

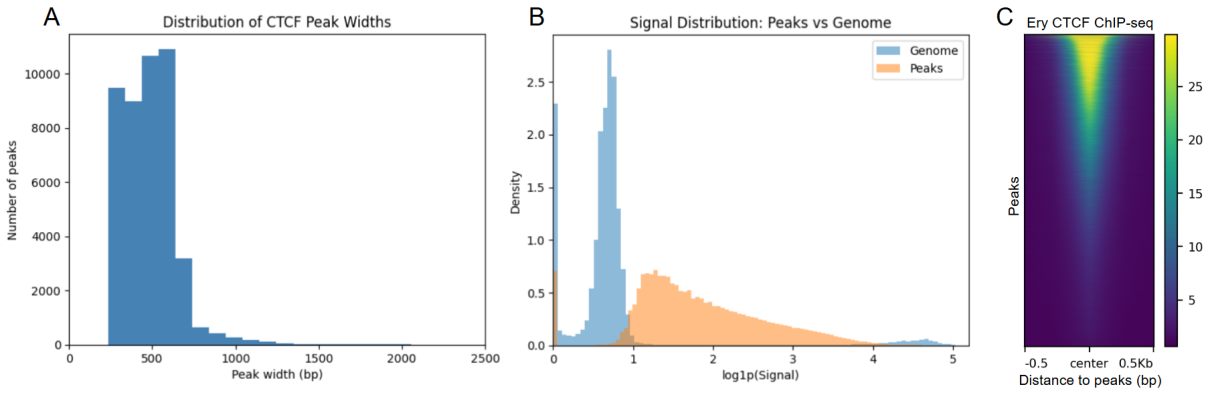


Figure 3.3: CTCF peaks and signal properties. A) Distribution of CTCF peak widths, showing that the majority of peaks are shorter than 750 bp. B) Distribution of genome-wide binding signal (blue), which is generally below 1, whereas signal intensity within called peaks (orange) predominantly ranges between 1 and 4. Peak-associated signals may also include zero values, as peak regions can contain positions with no measurable coverage after normalisation or thresholding. In addition, a small proportion of genome-wide signal values exceed 4; these likely represent background noise, technical artefacts, or high-coverage regions unrelated to specific CTCF binding events. C) Erythrocyte CTCF binding signal on peaks.

binding, at the alpha-globin locus. The signal captures known regulatory elements within the locus, including CTCF-associated domain boundaries and regions overlapping enhancer elements (R1, R2, R3 and R4).

CTCF binding in this dataset exhibits characteristic signal properties. Peak widths range from narrow regions ($\sim$ 200 bp) to broader domains exceeding 1 kb (Figure 3.3A,C), with most peaks shorter than 750 bp. Furthermore, the genome-wide signal distribution is left-skewed, with low background signal across most regions, while the distribution of peak coverage is right-skewed and typically ranges between 1 and 4 (Figure 3.3B).

ENCODE's cell-type-specific CTCF binding

In addition to the in-house dataset used for benchmarking, both the DeepSEA and HYPER models were trained on a large compendium of CTCF binding profiles derived from the ENCODE project, which was also used to train the original DeepSEA-based CNN model. This compendium comprises 33 primary human cell types derived from diverse tissues, including immune cells, neural support cells, epithelial cells, endothelial cells, fibroblasts from multiple organs, and muscle-related cells. Together, these cell types span many human organ systems. All represent differentiated somatic cell types with well-established lineage identities and have

been highly characterised in the literature.

Raw sequencing data were obtained from ENCODE and processed using the CATCH-UP pipeline, followed by signal track generation in BigWig format. For this work, normalisation was applied using ZEN to reduce technical variability, ensuring that differences in signal more accurately reflected underlying biological regulatory activity rather than experimental artefacts. More details about ZEN and the preprocessing process are available in Section 2.2. Peak calling was performed using LanceOtron to identify enriched genomic regions corresponding to putative CTCF protein binding sites.

The processed dataset provides a diverse representation of CTCF binding across cell types and serves as the training foundation for models evaluated in this chapter. Supplementary Figure S1 illustrates the distribution of the datasets used by HYPER to associate CTCF protein binding (train data) and for validating its learning (test data). Together, these datasets provide both a controlled erythroid-specific benchmark and a broader multi-cell-type context for evaluating model generalisation.

3.4 Models comparison framework

After training the three models, a comparative evaluation was conducted under the working hypothesis that HYPER would outperform the alternative approaches and therefore be employed for subsequent downstream analyses. For this comparison, the alpha-globin locus was chosen as a benchmark region due to its well-established erythrocyte-specific activity and high evolutionary conservation, as described in Section 1.7. The evaluation framework is structured into three complementary validation levels, as summarised in Figure 3.4.

At the first validation level, I compare the experimental CTCF protein-binding track at the alpha-globin locus with the predictions generated by each of the three models. This analysis comprises two components: (i) a global comparison across the entire genomic interval, including background regions, and (ii) a focused comparison restricted to high-signal regions (i.e., peak regions), thereby assessing performance both in signal detection and in background discrimination.

The second validation level assesses the ability of each model to recover known CTCF bind-

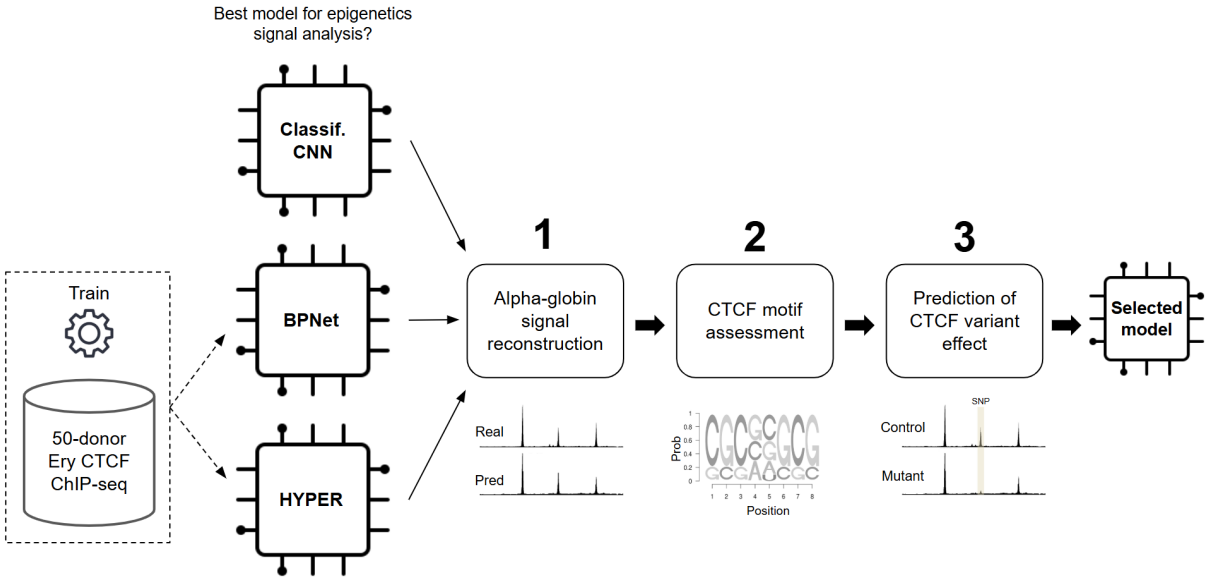


Figure 3.4: Validation workflow for comparative evaluation of DeepSEA (classification CNN), BPNet, and HYPER models. BPNet and HYPER were trained using the in-house erythroid CTCF dataset from 50 donors, while the pretrained DeepSEA-based model was used as a baseline. Model evaluation was performed across three validation stages: (1) reconstruction of alpha-globin CTCF binding profiles, (2) recovery of the CTCF_C2 motif using attribution-based analyses, and (3) prediction of CTCF variant effects. The best-performing model was subsequently selected for genome-wide evaluation.

ing motifs. Nucleotide-level importance scores derived from feature attribution methods were compared with canonical motif logos from public databases, providing a measure of biological interpretability.

Finally, model generalisation was evaluated using previously characterised CTCF variants from the in-house 50-donor dataset [171].

The best-performing model across these validation stages was selected for further validation at a genome-wide scale, using the `hg38` genome assembly. This analysis assesses the model's capacity to reproduce genome-wide binding signal distributions and predict the direction and magnitude of variant effects using a curated set of gain-, loss-, and neutral-function SNPs.

3.4.1 Local evaluation/assessment of CTCF binding predictions

The three erythrocyte-specific CTCF models were applied to predict protein-binding signals across the genomic region chr16:40,961–155,649 (114, 688 bp), encompassing the alpha-globin locus. Because the CNN-based models operate on fixed 1 kb input windows, predictions were

generated across overlapping sequence segments and merged to reconstruct continuous signal profiles across the locus. The region contains four major CTCF binding sites within the RHBDF1 and NPRL3 loci.

As shown in the top panel of Figure 3.5, the DeepSEA-based model captures the major CTCF binding peaks at the expected genomic locations. However, the predicted track also exhibits substantial variability in background regions, characterised by pronounced fluctuations ("valleys") across non-peak regions. This behaviour is expected to some extent because the model was originally trained as a classification model, predicting the probability of binding rather than directly regressing the magnitude of the epigenetic signal. Consequently, variation in the predicted probabilities across non-peak regions can produce a heterogeneous background when these predictions are represented as a continuous track. Nevertheless, the pronounced local fluctuations may also indicate that the model is sensitive to local sequence patterns that do not correspond to strong experimental binding signal. Because the original model outputs probabilities rather than signal intensities, predictions were rescaled for comparison with the experimental data.

In contrast, BPNet exhibits limited separation between peak and background regions (Figure 3.5, middle panel). Although BPNet's predicted values are inherently on the same scale as the experimental signal, peak regions remain only weakly distinguishable from background signal. Despite BPNet's ability to model broader sequence context through dilated convolutions, this does not translate into improved peak resolution in this setting.

HYPER most accurately recapitulates the experimental binding signal profile (Figure 3.5, bottom panel). The predicted profile closely mimics the shape and localisation of the observed CTCF binding peaks, with clear separation between signal and background distributions. Although the absolute scale of the predicted signal exceeds that of the experimental data, the relative structure and peak positioning are preserved with high fidelity. The strong performance of HYPER is likely facilitated by its pretraining on large-scale epigenetic datasets that include CTCF binding data, a regulatory signal that is highly conserved across cell types. Moreover, because it is not known whether the alpha-globin region was excluded from the pretraining data, the model may have encountered this genomic locus during training. HYPER's ability to model long-range dependencies may also contribute to its accurate reconstruction of CTCF binding patterns, particularly where binding is influenced by distal regulatory elements. Im-

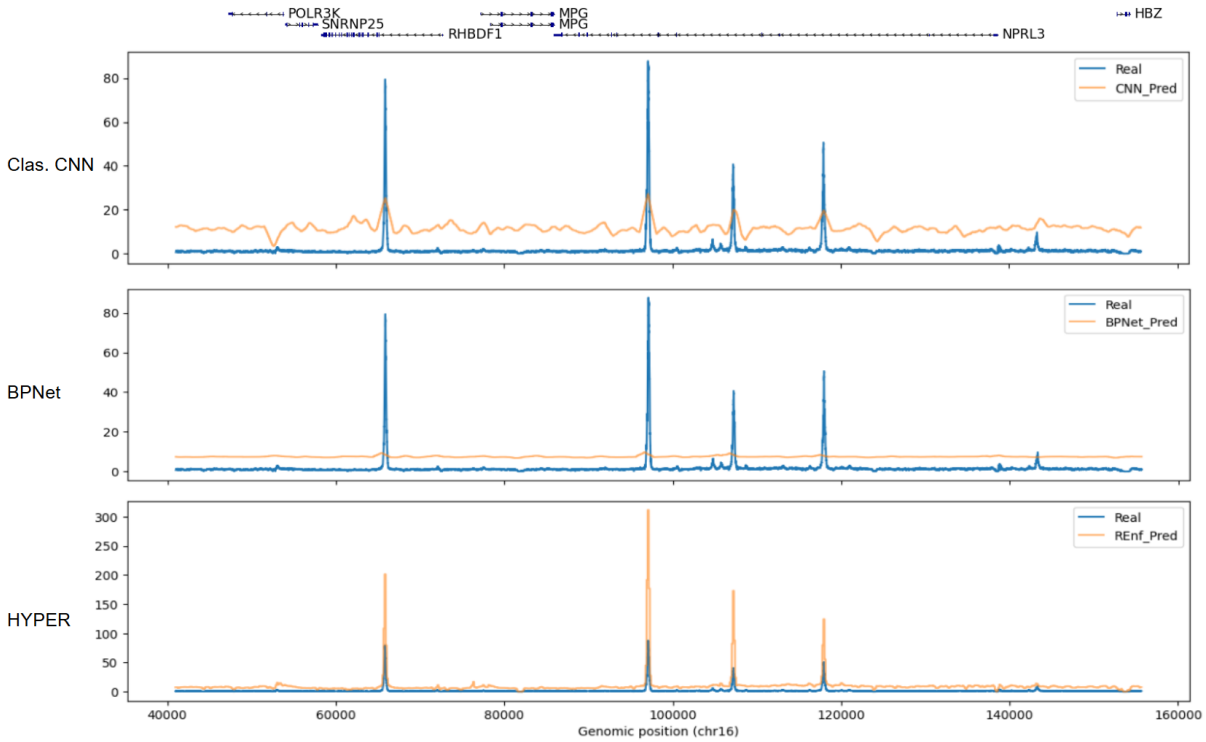


Figure 3.5: Predictions of the three CTCF models across the genomic region chr16:40,961–155,649. In all panels, the experimental CTCF ChIP-seq signal at the alpha-globin locus is displayed in blue, while the model predictions are in orange. The top panel shows the prediction of the DeepSEA-based model. The middle panel presents the BPNet protein-binding prediction, and the bottom panel illustrates the corresponding protein-binding prediction generated by HYPER.

portantly, the background distribution remains stable and well separated from peak regions, indicating improved signal-to-noise discrimination compared to the other two models.

To quantify similarity between predicted and experimental tracks, two complementary metrics were used: the NRMSE and the cosine similarity. The NRMSE assesses similarity in signal magnitude and scale, with lower values indicating greater similarity. Furthermore, cosine similarity measures the similarity in signal shape independently of absolute magnitude, with values closer to 1 indicating greater similarity. Together, these metrics assess both quantitative and structural agreement between predicted and observed profiles.

As shown in Table 3.1, all three models achieve relatively good scale agreement with the experimental data, indicating that their predicted values fall within a comparable numerical range. Notably, BPNet achieves the lowest and best NRMSE among the three models. However, this favourable scale similarity does not translate into superior predictive performance when considering the actual signal profiles, as BPNet's predictions exhibit limited concordance with the

observed peak structure.

The discrepancy between BPNet's low NRMSE and poor visual reconstruction likely reflects class imbalance within the genomic signal, where background regions substantially outnumber peak regions. The proportion of the true binding signal is small relative to the extensive background regions. BPNet performs well in predicting the background signal, thereby reducing the overall NRMSE. However, this masks its comparatively weaker performance in capturing peak regions. The imbalance is further evident when comparing the signal-region scales in the HYPER predictions: these are substantially larger than those in the experimental data, yet the NRMSE is not significantly affected by this difference.

To address this limitation, cosine similarity was introduced as a complementary metric. Cosine similarity better captures agreement in signal shape and therefore provides a more informative measure of peak reconstruction performance. According to this metric, HYPER demonstrates the highest agreement with the experimental track, while the DeepSEA-based model and BPNet show substantially lower shape similarity scores. These results reinforce the hypothesis that HYPER provides the most biologically faithful reconstruction of CTCF binding profiles, specifically at the alpha-globin locus.

Model	**Scale**	**Shape**
	NRMSE	**Cosine Similarity**
Classification CNN	0.12	0.44
BPNet	0.08	0.36
HYPER	0.18	0.95

Table 3.1: Similarity metrics between predicted and experimental whole-tracks. The first column reports the normalised root mean squared error (NRMSE), which quantifies differences in scale between predicted and observed signals; lower NRMSE values indicate greater scale similarity. The second column presents the cosine similarity score, a shape-based metric that evaluates similarity in signal profile independently of magnitude; values closer to 1 indicate greater concordance in signal shape.

3.4.2 Fine-scale modelling of CTCF binding sites

TFs recognise and bind to specific DNA sequences to regulate gene expression. As previously mentioned (Section 2.9.4), known TF binding sites are commonly represented using PWM,

which describe how likely each nucleotide (A, C, G, or T) is to occur at every position within a binding site. Importantly, a single TF may exhibit multiple variants of its binding motif; therefore, all relevant motif variants must be considered when identifying potential binding sites.

In this subsection, I assess the ability of the three models to reproduce a known CTCF binding motif. To provide a more stringent evaluation, I selected a CTCF peak that is not highly predominant in the human alpha-globin locus, thereby testing the models' capacity to identify binding sites with lower coverage in comparison to other CTCF binding sites in the region. The selected peak spans chr16:106688–107569 and is located between enhancers R1 and R2 (highlighted by a red rectangle in Figure 3.6A). The specific CTCF binding site analysed is 20 bp in length and is situated at chr16:107200–107221.

To extract nucleotide-level importance scores from the model predictions, I used Captum's implementation of DeepLIFTShap alongside BPNet's internal DeepLIFTShap function. In both cases, hypothetical importance scores were computed. This feature attribution approach was selected instead of standard DeepLIFT because it enables evaluation of the predicted contribution of all possible nucleotides at each sequence position, rather than only the nucleotide observed in the input sequence. As a result, hypothetical importance scores provide a more informative representation of potential regulatory motifs and sequence features influencing the model predictions.

Following extraction, the nucleotide importance profiles were visually inspected and compared against established CTCF motif variants from Jaspar and HOCOMOCO databases. Based on the site length and the structural characteristics of the motif, specifically the presence of a "CC" dinucleotide at the 5' end, a "GC" at the 3' end, and a central cluster of guanines, the CTCF_C2 variant from the JASPAR database was identified as the closest match. The nucleotide probability distribution, or sequence logo, for this variant is shown at the top of Figure 3.6B.

The three lower panels of Figure 3.6B display the predicted nucleotide importance logos generated by the DeepSEA-based CNN, BPNet, and HYPER, respectively, from top to bottom. The importance logo derived from the DeepSEA-based CNN model emphasises the terminal guanine at position $107,217$. Still, it fails to capture the flanking "CC" at positions $107,204$–$107,205$ and the central guanine cluster. In contrast, BPNet accurately identifies most of the key nucleotides; however, it differs from the other two models at positions with lower nucleotide oc-

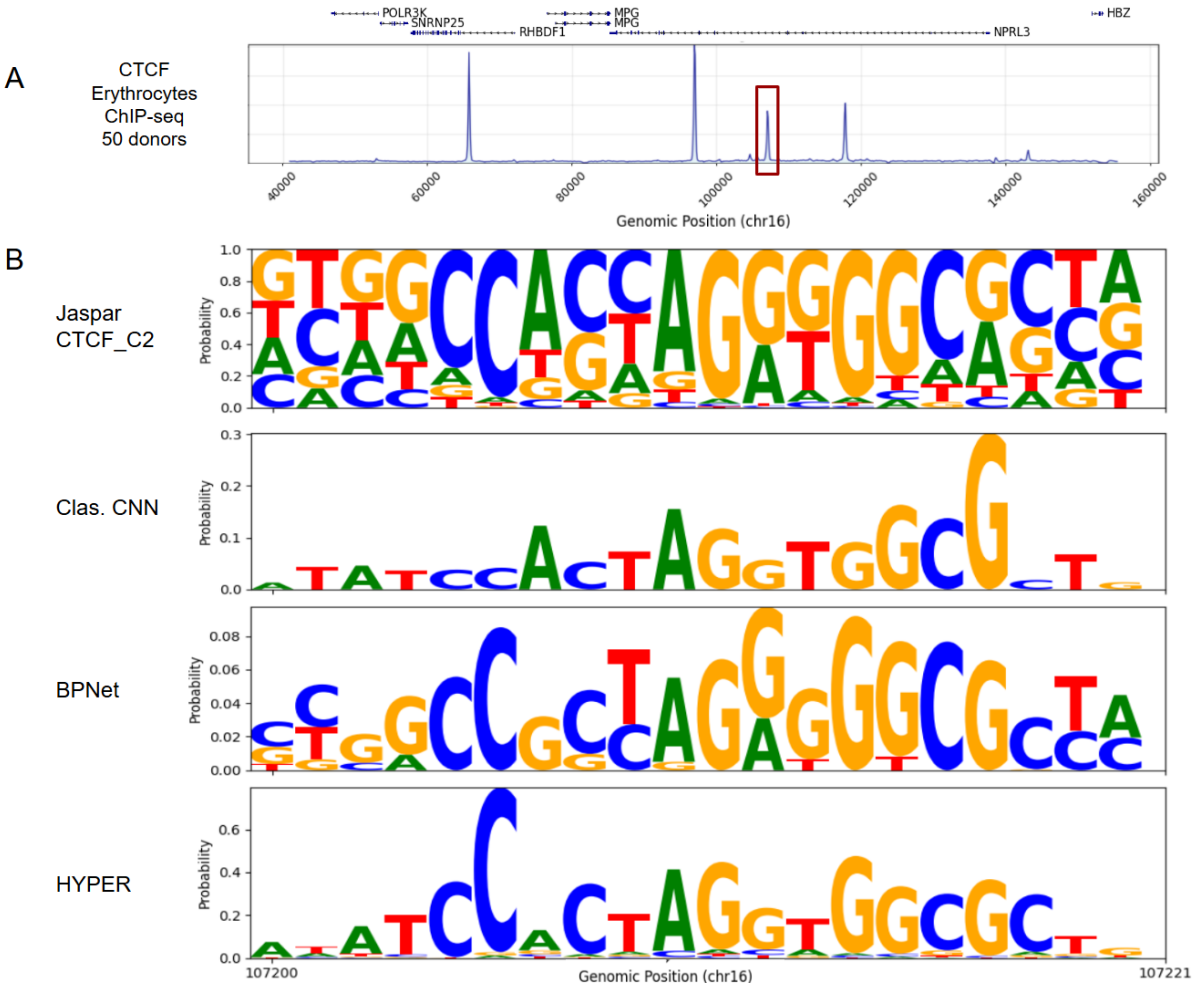


Figure 3.6: Nucleotide importance of the motif CTCF_C2. A) An instance of the CTCF_C2 binding motif is identified within the third CTCF binding site of the human alpha-globin locus. B) The CTCF_C2 sequence logo from JASPAR spans 20 bp and displays strong nucleotide preferences, notably a “CC” at positions 4 and 5 and a cytosine at position 16. The predicted nucleotide attribution profiles generated by the DeepSEA-based CNN, BPNet, and HYPER are shown below and collectively recapitulate the characteristic structural features of this CTCF_C2 motif instance.

currence probabilities. These discrepancies may stem from differences in the DeepLIFTShap implementation, as BPNet uses a custom version of the method.

Unlike BPNet, HYPER’s importance profiles tend to emphasise a single dominant nucleotide per position. In summary, both BPNet and HYPER demonstrate strong performance in recapitulating the CTCF_C2 nucleotide motif based on their predicted importance scores. Importantly, recovery of the CTCF_C2 motif should be interpreted as evidence that the models learn sequence features consistent with known CTCF binding determinants, rather than as direct proof that they have inferred the full mechanistic grammar of CTCF binding. In contrast, the DeepSEA-based CNN model appears less capable of fully capturing the motif structure.

The next step was to formalise the visual assessment through quantitative analysis. To this end, I compared the nucleotide importance distributions produced by the three models (query motifs) with the CTCF_C2 PWM (target motif) using two complementary metrics: a significance metric, reflecting how likely it is that a given query motif corresponds to an instance of CTCF_C2, and a shape similarity metric, quantifying the resemblance between the distribution profiles of the query and target motifs.

Statistical significance was evaluated using TomTom's p-value, while shape similarity was assessed using the PCC. The resulting values are reported in Table 3.2. Because PCC is scale-invariant, the feature attribution scores could be directly compared to the PWM without additional normalisation. In contrast, TomTom required prior normalisation of the importance scores.

Although HYPER yielded the lowest p-value, all three models demonstrated strong statistical significance in matching the query motif to the CTCF_C2 instance. In contrast, the shape similarity metrics exhibited greater variability. BPNet showed the lowest PCC, whereas HYPER achieved the highest shape similarity. This may be attributable to HYPER's stronger positional nucleotide preferences (i.e., more pronounced peaks), which render its profile more similar to the target motif.

Model	**Significance**	**Shape**
	p-value	**PCC**
Classification CNN	0.00016	0.62
BPNet	0.0025	0.53
HYPER	0.000021	0.8

Table 3.2: Models' performance metrics for predicting the nucleotide importance of a CTCF_C2 instance.

3.4.3 Generalisation to an unseen CTCF genetic variation

The final level of validation involved applying the three models to predict the binding effects of a SNP previously shown to result in loss of CTCF binding at a specific site. This SNP was identified by Dr Matthew Gosden [171] in Donor 7 from the cohort of 50 erythrocyte donors. The variant is present in both alleles and is located at chr16:8,655,299, where an adenine sub-

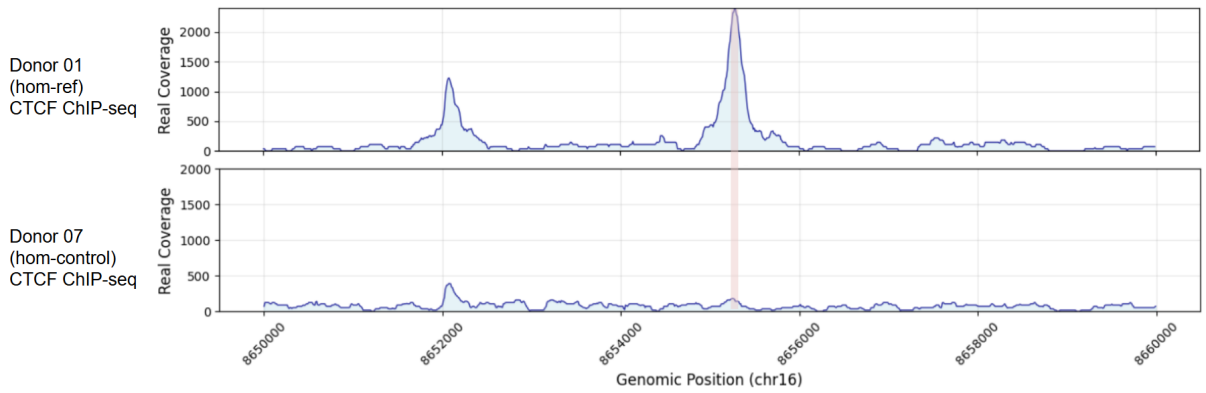


Figure 3.7: Effect of the SNP at chr16:8,655,299 on CTCF protein binding. Donor 1, who does not carry the mutation, exhibits the expected CTCF binding activity at this locus (highlighted in light red). In contrast, Donor 7 carries the SNP, resulting in loss of CTCF binding at the site.

stitutes for a cytosine. To predict the effect of this variant, the reference genome (`hg38`) was used as the model input, with the alternate allele substituted at the variant position, rather than Donor 7's personal genome. This approach isolates the effect of the single SNP from the surrounding genetic background, so that the resulting difference between reference and alternate predictions can be attributed exclusively to this variant, consistent with the standard variant-effect-prediction protocol used to benchmark these models [68]. The functional consequences of this CNN are illustrated in Figure 3.7, which shows the loss of a CTCF binding site at the mutated locus. For comparative analysis, Donor 1, who does not carry this mutation, was used as the reference individual.

To evaluate the generalisation capacity of the three models, both the reference CTCF sequence and the corresponding SNP-containing sequence were used as input for prediction. Preliminary inspection of the reconstructed binding tracks indicates that all three models capture the functional impact of the variant, correctly predicting a loss of CTCF binding at the mutated locus.

The DeepSEA-based CNN predicts a peak located slightly distal to the core functional site, which is absent in the mutated sequence, as shown in the top blue panel of Figure 3.8. However, the overall predicted signal exhibits relatively high noise. BPNet, illustrated in the middle red panel of Figure 3.8, more clearly captures the loss-of-function event at the variant locus. However, the predicted peaks are broader than expected, and the coverage intensities differ from those observed in the experimental data. In contrast, as shown in the bottom green panel of Figure 3.8, HYPER most faithfully reproduces the variant's effect, accurately reflecting the

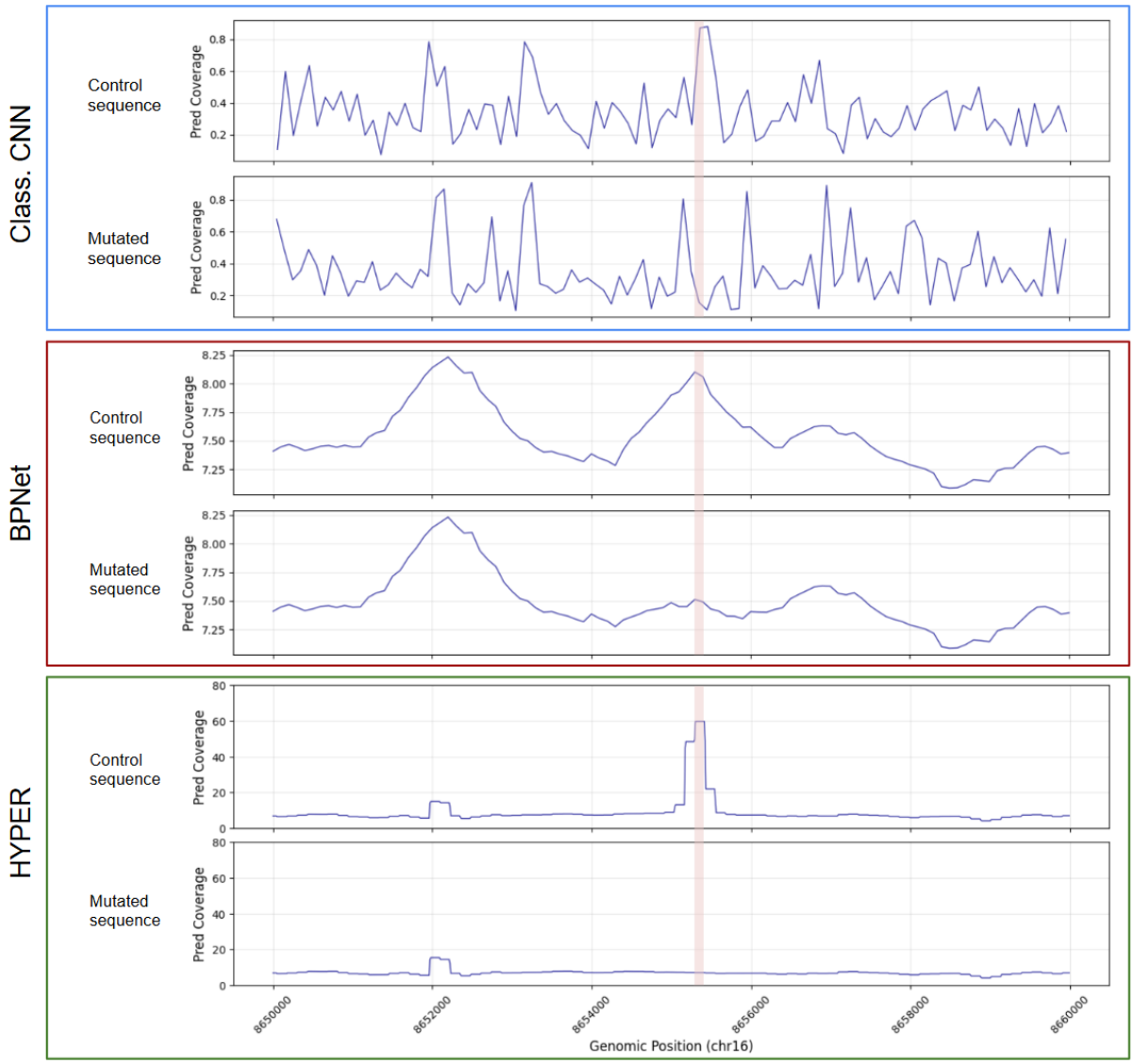


Figure 3.8: Predicted effects of the SNP at chr16:8,655,299 on CTCF binding. The top panel (blue) shows the predictions generated by the DeepSEA-based CNN, with the reference (control) sequence displayed above and the mutated sequence below. This model identifies a relatively narrow and subtle loss-of-function signal near the functional site (highlighted in light red). The middle panel (red) presents the BPNet predictions, again with the reference sequence on top and the mutated sequence below, and indicates a broader loss-of-function effect affecting the second peak. The bottom panel (green) displays the predictions from HYPER, which capture a sharp and well-defined loss-of-function signal at high resolution, closely localised to the functional site.

shape, intensity, and resolution of the experimentally derived binding profiles.

In summary, this section systematically benchmarked the models across three complementary validation levels. At the first level, binding track reconstruction was most accurately performed by HYPER, which outperformed the other two architectures. At the second level, focused on CTCF motif discovery through nucleotide importance scoring, both HYPER and BPNet

demonstrated comparable performance and surpassed the DeepSEA-based model. At the third level, variant-effect identification via *in silico* mutagenesis was successfully achieved by all three models; however, HYPER predicted tracks that more closely resembled the observed peaks in shape and intensity. Based on this comprehensive comparative evaluation, HYPER was selected for all subsequent downstream analyses.

3.4.4 Comparisons summary

In summary, the three models, the DeepSEA-based CNN, BPNet, and HYPER, were evaluated across three validation levels to compare their performance in predicting CTCF protein binding. The first level focuses on reconstructing the alpha-globin CTCF binding signal in human erythrocyte cells. All three models accurately reproduce the signal intensity, with NRMSE values below 0.2; however, only HYPER closely matches the shape of the experimental signal, achieving a cosine similarity of 0.95.

The second validation level examines CTCF_C2 motif discovery based on nucleotide importance scores. While all three models identify the motif with strong statistical significance (p-values below 0.01), only HYPER attains a high shape accuracy, with a PCC of 0.8.

The final validation level evaluates each model's ability to predict the loss of protein-binding function caused by a SNP involving a single-nucleotide substitution on chromosome 16. Although the DeepSEA-based CNN model suggests a loss of function at the site, the effect is not clearly pronounced. In contrast, both BPNet and HYPER show a clear loss of the CTCF binding site. These results are summarised in Table 3.3.

Model	**Track reconstruction**		**Motif assessment**		**Variant effect**	**Selected**
	Scale	**Shape**	**Significance**	**Shape**		
Classif. CNN	✓	-	✓	-	-	-
BPNet	✓	-	✓	-	✓	-
HYPER	✓	✓	✓	✓	✓	✓

Table 3.3: Summary of the comparison results across the three validation levels. Track comparison is evaluated in terms of scale and shape, using NRMSE and cosine similarity, respectively. Motif assessment includes statistical significance (p-value) and shape similarity (PCC). The variant effect is evaluated qualitatively through visual inspection.

3.5 Extended genome-wide validation

As an extension of the comparative model evaluation, additional validation analyses were performed specifically for the HYPER model at the genome-wide level. Genome-wide predictions were generated for the CTCF protein binding class by applying the model across the entire genome in $114,688$ bp sliding windows with a $40,960$ bp overlap. The resulting tracks were then merged, and coverage values within overlapping regions were averaged.

Due to HYPER padding during prediction, the model processes a $196,608$ bp input window but produces coverage predictions only for the central $114,688$ bp. Consequently, each chromosome's predicted track contains gaps of $40,960$ bp at both the beginning and the end. To ensure a fair comparison between the experimental data and the reconstructed tracks, the first and last $40,960$ bp of each chromosome were excluded from the experimental data.

Two further validation steps were implemented: (i) comparison of predicted peak signal distributions across the entire genome, and (ii) assessment of SNP generalisation. These analyses were designed to derive quantitative performance metrics based on confusion matrix–like statistics. Peak-based analysis of the predicted output from models such as HYPER presents an immediate technical difficulty, as no peak calling approach has been developed for them. Approaches used on conventional genomics data such as the MAC peak caller depend on the fragment data within the aligned BAM file, which is missing in ML output as only the genome profile is predicted. Some aligners, such as LanceOtron, can work on genome profiles; however, they themselves are neural-network-trained on primary data to recognise peak shape distributions. Due to the windows used in HYPER predictions, peak shape appears blocky at high resolution; see HYPER panel Figure 3.8, which could confound this approach.

Therefore, for the first analysis, I used a custom pipeline conceptually analogous to a peak-calling procedure. This framework is described in Section 2.6.3. Also, due to the structural characteristics of predictions generated by the CNN-based baseline models, which, as shown previously, did not accurately reproduce signal peaks, it was not feasible to call this data meaningfully, and so this genome-wide validation framework was not applied to those architectures.

3.5.1 CTCF signal peaks similarity

Because epigenetic datasets are typically highly imbalanced, where background regions vastly outnumber signal regions, it is important to assess how well the distribution of signal regions is preserved across the genome. To address this, a reliable set of signal peaks is first generated, whose coverage can then be compared between the experimental data and the predicted track. Figure 3.9 illustrates the workflow used in this section.

The tracks' resolutions need to be first adjusted to make the real and predicted tracks comparable. For this, the resolution of the experimental data was first reduced to match HYPER's output resolution of 128 bp bins. This was achieved using BedTools `makewindows` with a window size of 128 bp. The resulting BedGraph file was subsequently converted to BigWig format using the UCSC `bedGraphToBigWig` utility.

To obtain reference signal peaks from the original CTCF ChIP-seq dataset, and to later evaluate potential signal loss introduced by the binning process, peak calling was performed using MACS3 callpeak with the following parameters: a q-value threshold of 0.0001, retention of a single duplicate read, bypassing the BAM file dependent construction of the shifting model, extension of reads downstream to a fixed length of 150 bp, and a minimum peak length of 90 bp.

To identify biologically meaningful signal regions in both the binned experimental track and

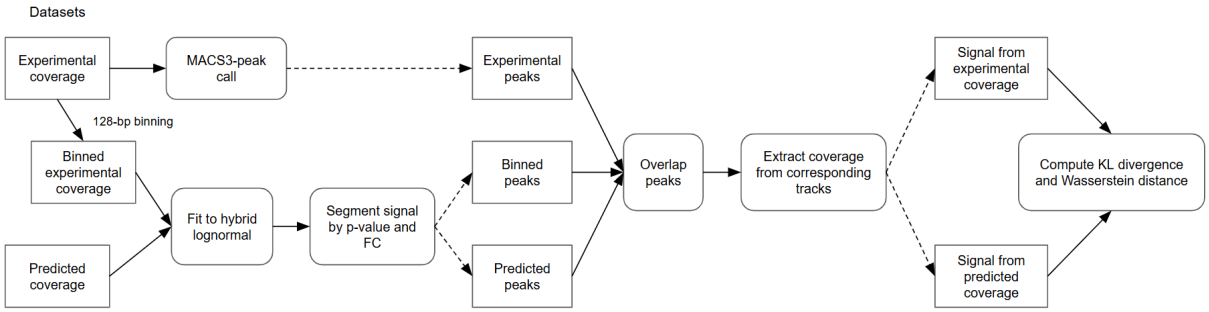


Figure 3.9: Workflow for signal segmentation on predicted tracks. Experimental data are first binned to a resolution of 128 bp. Both the predicted and binned tracks are then fitted to a hybrid log-normal distribution, from which signal peaks are identified based on p-values and fold-change thresholds. In parallel, peaks are called on the experimental track using MACS3. The resulting three peak sets are intersected to generate a consensus set of regions. This final set is used to extract signal coverage from both experimental and predicted tracks for subsequent comparison.

the predicted coverage track, an initial filtering step was applied to remove the large background peak centred around zero observed in the signal distribution plots. This near-zero peak primarily represents genomic regions with little or no regulatory activity and would otherwise dominate the distributions. Removing these low-signal regions enriched the dataset for informative signal values and produced distributions that more closely approximated a log-normal shape, facilitating subsequent modelling. Accordingly, the binned experimental track was filtered to retain values greater than or equal to 0.6, while the predicted coverage track was filtered to retain values greater than or equal to 0.002.

The peak-calling framework described in Section 2.6.3 was then applied to both tracks. Briefly, this method models the distribution of signal intensities using two log-normal distributions. The first distribution (the bulk fit) is estimated from the lower 95th percentile of the data and captures the main body of the distribution, which predominantly represents the background signal. The second distribution (the full fit) is estimated using the entire dataset and therefore better characterises the behaviour of the right tail, where the genuine signal is expected to be. Candidate peaks were identified using two thresholds: a p-value ≤ 0.001 and a fold-change ≥ 5. These thresholds balance statistical stringency and practical sensitivity, which is particularly important given the reduced resolution of the predicted data (128-bp bins) relative to the original ChIP-seq signal and the potential signal attenuation introduced during prediction.

Although this modelling framework was originally developed to describe both predicted and experimental coverage signals using a combination of multiple log-normal distributions, the real binned coverage data showed a much simpler and more symmetric shape than the predicted data. To maintain consistency and allow direct comparison between datasets, the same overall modelling strategy was still applied to both. JSD was also calculated to measure how closely the fitted distribution matched the observed data, as described in Section 2.10.

Typically, the hybrid model uses two separate fits: one to describe the main bulk of the signal distribution, which mostly represents background and moderate signal levels and is usually referred to as the bulk fit, and another to capture the smaller high-signal tail corresponding to strongly enriched regions, which I refer to as the full fit. However, in the real binned coverage data, the main bulk distribution already described the overall shape of the data reasonably well because the high-signal tail was much less prominent. Therefore, only the bulk fit based on the lower 95% of the data was used for the final model. The same peak-calling thresholds were

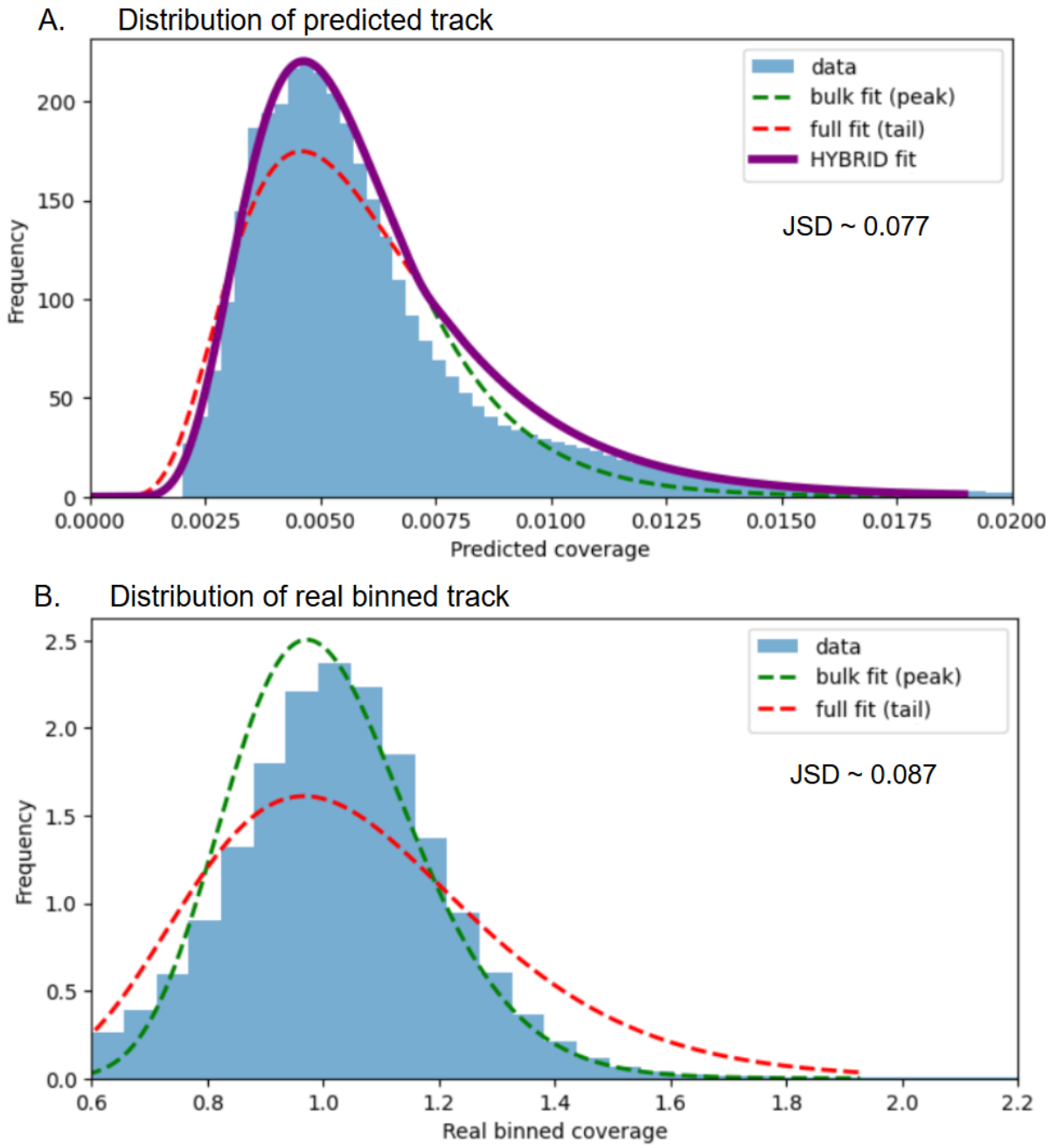


Figure 3.10: Distributions of predicted and real binned coverage values. A) Histogram of predicted protein-binding coverage fitted with a hybrid log-normal model (purple). The central peak of the distribution is modelled using a bulk fit (green) estimated from the lower 95th percentile of the data. At the same time, the right tail is captured by a full fit (red) estimated from the entire dataset. B) Histogram of real binned protein-binding coverage. In this case, the distribution is best described by the bulk log-normal fit (green) based on the lower 95th percentile of the data. The full fit (red), used for the predicted distribution, is not included as it does not improve the representation of the empirical distribution.

retained (p-value ≤ 0.001 and fold-change ≥ 5).

Both datasets showed a high concentration of values close to zero. Since log-normal models

require strictly positive inputs, additional thresholds were applied to remove near-zero coverage values before modelling. Specifically, predicted coverage values were filtered using a threshold of coverage ≥ 0.002, while real binned coverage values were filtered at coverage ≥ 0.6. The higher threshold for the binned data was chosen to account for symmetry and to enable the hybrid distribution to better capture the right tail.

As shown in Figure 3.10A, the predicted coverage distribution is best described by a hybrid model consisting of a mixture of the bulk fit (green dotted line), which captures the peak region of the distribution, and the full fit (red dotted line), capturing the behaviour of the right tail and achieving a JSD score of approximately 0.077. The purple curve shows the resulting combined distribution.

In contrast, the real binned coverage distribution is better described by the bulk fit alone, as illustrated in Figure 3.10B, which achieves a JSD value of 0.087. Given that an ideal JSD value is 0, indicating a perfect match between the distributions, both distributions demonstrate a very good fit.

After identifying signal peaks for each distribution, peaks were sorted and merged using BedTools. Peaks overlapping low-mappability or multi-mapping regions were also identified and counted, yielding 130 peaks in the real base-pair resolution dataset, 5 peaks in the binned real binding coverage dataset, and 220 peaks in the predicted signal set. These peaks were subsequently excluded from further analysis.

The final peak sets consisted of $44,947$ peaks obtained with MACS3 from the base-pair resolution experimental CTCF ChIP-seq track, $28,021$ peaks identified using the bulk log-normal model from the binned real dataset, and $33,209$ peaks detected in the predicted CTCF binding coverage. These numbers are obtained following filtering.

A visual inspection of a randomly selected 700 kb locus on chromosome 16 (Figure 3.11) illustrates the relative sensitivity of each peak set in capturing signal. Peaks extracted with MACS3 from the full CTCF ChIP-seq track appear highly sensitive, identifying numerous regions, including some where the signal is not clearly distinguishable from background noise. In contrast, the binned representation is more conservative, primarily capturing regions with a strong, evident signal, as binning smooths the track and emphasises dominant peaks. Consequently, smaller peaks within the region chr16:1,450,000–1,700,000 are largely missed.

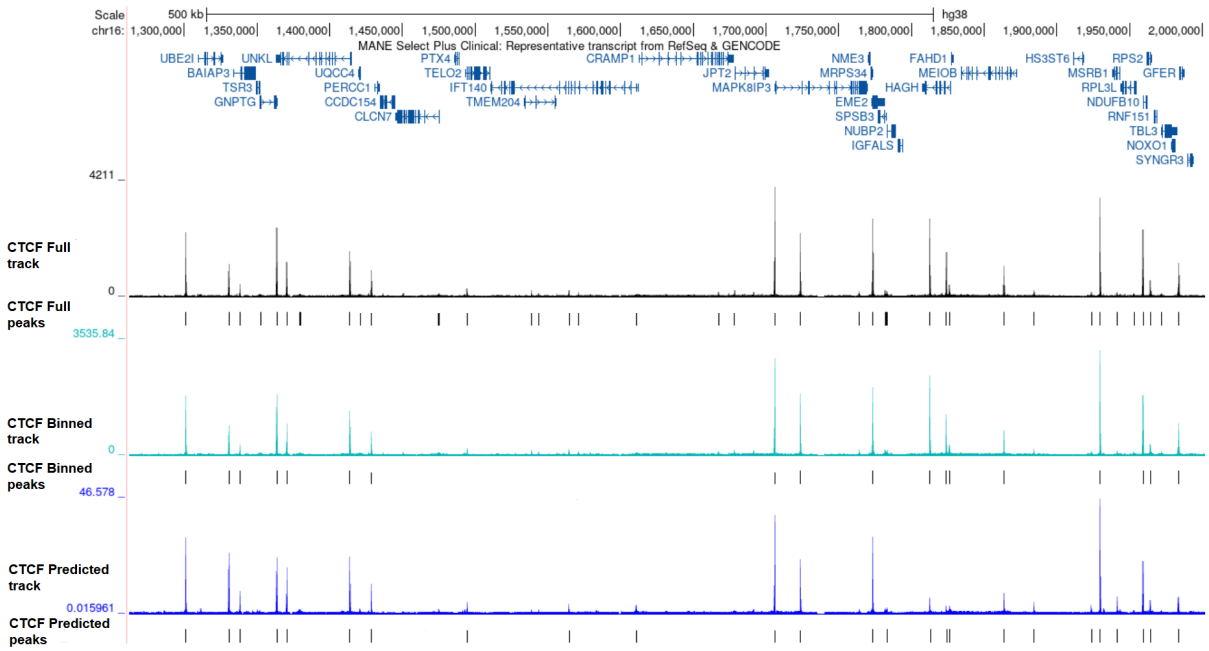


Figure 3.11: Genome browser view of the region chr16:1,300,000–2,000,000 displaying CTCF tracks. The full CTCF ChIP-seq signal and corresponding MACS3-called peaks are shown in black. The binned CTCF ChIP-seq signal with its bulk log-normally distributed peaks is shown in light blue. The HYPER-predicted CTCF peaks, along with their hybrid log-normally distributed peaks, are shown in blue.

By comparison, the predicted peaks achieve a more balanced performance, retaining sensitivity to smaller peaks while avoiding excessive noise. Notably, these smaller peaks are well represented in the chr16:1,450,000–1,700,000 interval, alongside the major signal peaks across the locus. Overall, this suggests that the predictive model can recover biologically relevant signals with improved precision–recall trade-offs compared with both MACS3-derived and binned peak sets.

As shown in Figure 3.12A, the peaks identified in the binned dataset overlap 97.7% with the peaks from the full-resolution experimental dataset. The predicted peaks overlap 70.3% with the intersection between the binned and real peak sets. However, 84.2% of the predicted peaks overlap with the full-resolution experimental peak set. This suggests that the predicted signal may capture CTCF binding regions more effectively than the binned experimental data, or that binning reduces signal resolution and results in the loss of detectable peaks.

Those $9,853$ predicted peaks that are not present in the 128 bp–binned overlapping experimental dataset show a distribution largely concentrated around 0.025, which corresponds to the lower bound of the distribution, as shown in Figure 3.12B. These peaks may represent false

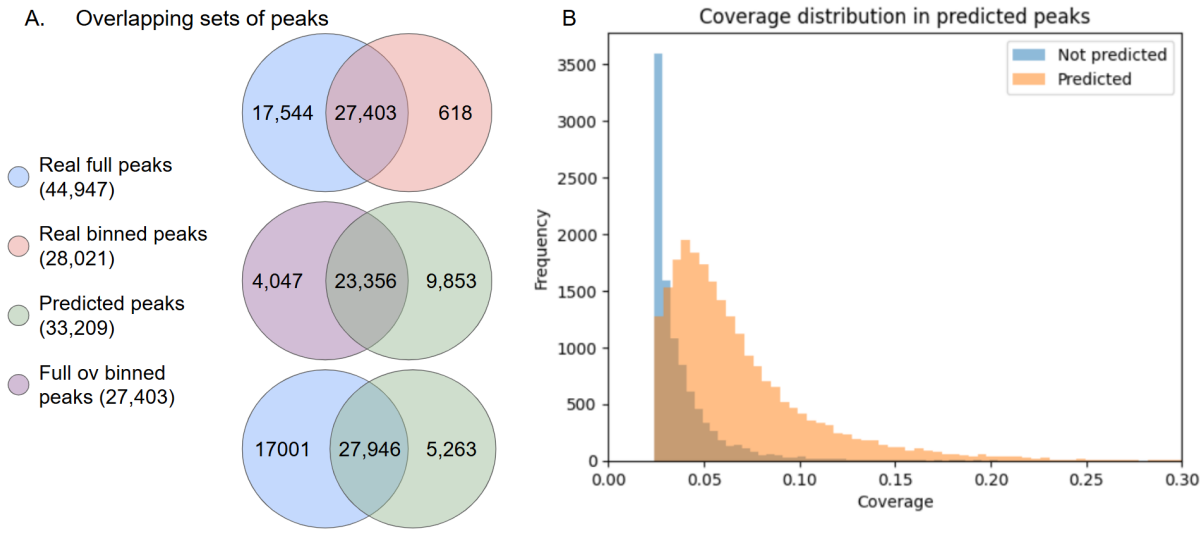


Figure 3.12: Called peak sets comparison and statistics. A) Peaks called from the experimental base-pair-resolution CTCF protein binding track are in blue. Peaks called from the 128-bp binned real data are shown in red. Overlaps between the binned and real protein-binding tracks are shown in violet. Predicted binding signal peaks are in green. B) Distribution of predicted peaks overlapping with binned and full real peak sets (in orange) and non-overlapping peaks classified as false positives (in blue).

positives, potentially arising from the model learning signals from overrepresented or repetitive sequences. Alternatively, they could correspond to genuine signal regions that were not detected in the experimental track due to signal loss introduced during the binning process.

When extracting the signal from the set of $23,356$ highly reliable peaks, the coverage values were obtained from both the experimental protein-binding track and the predicted CTCF track. The resulting signal distributions were then compared using two complementary similarity metrics: KL divergence and Wasserstein distance. Score values close to zero are interpreted as high similarity between two distributions for both metrics.

The KL divergence, which quantifies the difference between two probability distributions by measuring the information lost when one distribution is used to approximate another (more details in Section 2.10) yielded a value of 0.16. Additionally, the Wasserstein distance (also known as the Earth Mover's Distance), which measures the minimal cost of transforming one distribution into the other by accounting for the distance between their values, resulted in 2.77×10^{-6}.

Both metrics indicate a high degree of similarity between the experimental and predicted signal distributions across these high-confidence peak regions, suggesting that the predicted track accurately captures the overall structure of the experimental CTCF binding signal.

Once the similarity between predicted and observed peaks has been established, a further level of validation is to assess HYPER's ability to predict gain or loss of CTCF protein binding signal resulting from SNVs. This involves evaluating whether the model can accurately capture both the emergence and disappearance of peaks, as well as the direction and magnitude of the underlying variant effect.

Despite the generally strong reconstruction of CTCF binding profiles, model performance was not uniform across all genomic contexts. Reduced predictive accuracy was more frequently observed in regions characterised by weak or diffuse enrichment signals, low signal-to-noise ratios, repetitive sequence composition, or complex local chromatin organisation. In particular, regions with broad low-intensity binding patterns or ambiguous peak boundaries were often reconstructed less accurately than sharply defined CTCF binding sites. Certain discrepancies may also reflect technical biases present in the underlying experimental data, including GC-content bias, mappability limitations, and variability introduced during peak calling or sequencing coverage normalisation. These observations suggest that predictive performance remains partially dependent on both data quality and the extent to which CTCF-associated sequence patterns resemble those represented in the training set. Consequently, failure regions may provide insight into limitations of current sequence-based modelling approaches as well as genomic contexts in which CTCF binding is less strongly encoded by local sequence composition alone. Systematic analysis of such failure modes may therefore represent an important avenue for improving both model robustness and biological interpretability in future regulatory genomics frameworks.

3.5.2 HYPER for genome-wide SNP generalisation

As described previously, the Hughes lab has generated a SNP collection from the erythrocyte dataset, derived from the variants called across the cohort of 50 donors [171]. These SNPs were subsequently categorised based on their genomic and functional contexts, including their association with CTCF-binding sites, enhancers, or promoter regions.

We further analysed this dataset to identify SNPs associated with CTCF binding, including variants overlapping known enhancer or promoter regions. For this study, only SNPs located directly within functional CTCF binding sites (inside called peaks) were retained for downstream analysis, while SNPs located outside these regions, often referred to as distal or putatively causal variants, were excluded to reduce potential confounding effects. More than

700,000 SNPs were identified at this step.

Ideally, reads from the two alleles should map to their respective positions in approximately equal (50/50) proportions. Deviations from this expected balance indicate potential allele-specific mapping bias. To minimise such biases, the WASP allele-specific analysis framework was applied [172]. WASP detects potential mapping bias by testing whether sequencing reads containing alternative alleles still align correctly to the genome after remapping. Because phased genomes were available for each donor, all sequencing reads were instead aligned to personalised donor genomes, improving the accuracy of allele-specific mapping. This analysis identifies allelic imbalance, where one allele is supported by significantly more sequencing reads than the other, suggesting that the SNP may influence regulatory activity or TF binding. Such SNPs represent candidate molecular quantitative trait loci (QTLs) with potential causal regulatory effects. To ensure that downstream analyses focused on robust and unbiased CTCF binding signals, variants exhibiting allelic-specific mapping bias were excluded from the final dataset.

Following this filtering step, a CTCF binding site p-value was computed to quantify allele-specific binding bias. For each SNP, counts of CTCF ChIP-seq reads supporting the maternal and paternal alleles were aggregated across the 50 donors. A beta-binomial distribution was first fitted to account for overdispersion in the data. Subsequently, a beta-binomial test was applied to assess whether the observed allelic counts significantly deviated from the expected 1:1 ratio. The resulting p-value reflects the strength of allele-specific binding: highly significant values ($p \leq 0.001$) indicate strong preferential binding to one parental allele, whereas non-significant values ($p \geq 0.05$) suggest insufficient evidence to reject the null hypothesis of no allelic bias.

The filtered SNP set was subsequently stratified by regulatory impact: whether each variant resulted in a gain of binding, a loss of binding, or no measurable change in binding affinity. To determine functional effects, peak calling was performed with LanceOtron using a threshold of peak score equal to 0.5. A SNP was considered to exert a regulatory effect if the presence of the alternative allele was associated with either the emergence of a novel peak or the disappearance of an existing peak at the variant locus.

To mitigate the influence of experimental variability and technical noise inherent to high-throughput sequencing data, only SNPs associated with substantial changes in signal coverage

were classified as functional. Variants producing minimal or negligible differences in coverage between alleles were categorised as neutral.

A fourfold threshold (equivalent to a log_2 fold change ≥ 2) was selected as a stringent and biologically meaningful cutoff [173]. In high-throughput sequencing assays such as ChIP-seq, moderate fold changes may arise from technical variation. By adopting a $\geq 4\times$ threshold, the analysis ensures that only robust and substantial changes in TF occupancy are classified as functional. This threshold also serves a second purpose: it defines a set of variants for which the true regulatory label (gain, loss, or neutral) can be assigned with confidence from the allelic ChIP-seq data itself. Variants with intermediate coverage differences, falling between the neutral and four-fold cutoffs, cannot be reliably distinguished from technical noise, and so their true functional status remains ambiguous. Restricting downstream benchmarking to variants passing this threshold therefore ensures that predicted variant effects are evaluated against a high-confidence ground truth, rather than against labels that may themselves be uncertain.

To formally define the damage function, let ALT denote the binding coverage at the functional site in the sequence carrying the alternative allele, and REF denote the binding coverage at the same site in the reference (control) sequence. A SNP was classified as a gain-of-function variant if: ALT/REF ≥ 4, indicating that the alternative allele exhibits at least a four-fold increase in binding coverage relative to the reference allele. Conversely, a SNP was classified as a loss-of-function variant if: REF/ALT ≥ 4, indicating that the reference allele exhibits at least a fourfold higher binding coverage than the alternative allele.

However, the fold-change formulation described above does not accommodate cases in which one of the coverage values equals zero, such as when a peak emerges de novo in the presence of the alternative allele or is eliminated. In such instances, the ratio-based metric becomes undefined (division by zero) or tends toward infinity, complicating interpretation.

To address this limitation, a normalised formulation was adopted. Specifically, a SNP was classified as a gain-of-function variant if: ALT/(ALT+REF) ≥ 0.8, as a loss-of-function variant if: REF/(REF+ALT) ≥ 0.8, and as a neutral variant if: ALT/(ALT+REF) ~ 0.5. This proportional framework preserves the stringency of the original fourfold criterion while remaining mathematically stable when one of the coverage values is zero, in which case the proportion approaches 1 rather than becoming undefined, allowing for robust classification of extreme regulatory effects.

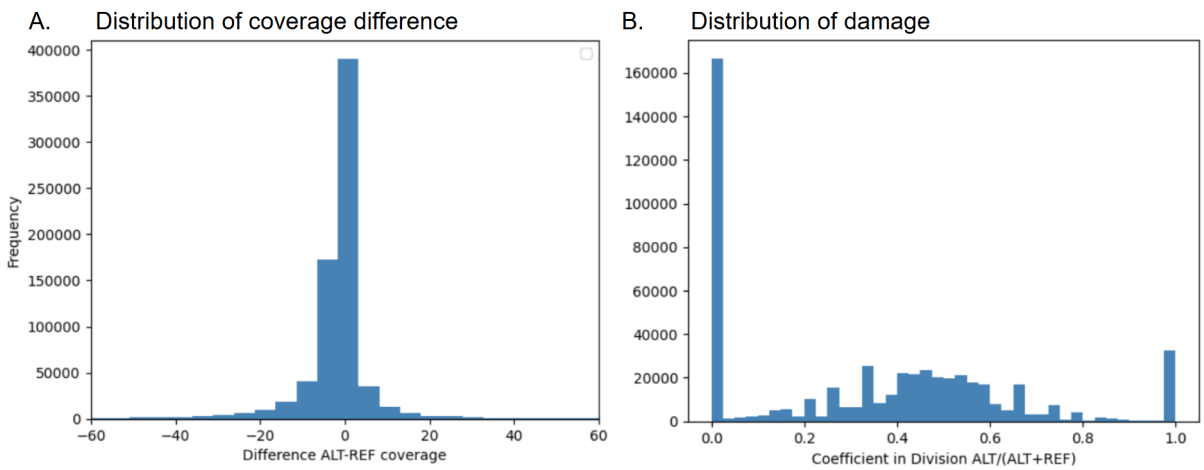


Figure 3.13: Divergence in binding coverage between reference (control) and alternative (mutant) alleles across all SNPs. A) Histogram showing the distribution of the difference between mutation coverage and control coverage for all SNPs. B) Histogram showing the distribution of the proportion of mutation coverage relative to the total coverage (mutation + control) across all SNPs, excluding SNPs for which mutation and control coverages are equal to zero.

As the majority of coverage differences are centred near zero, a high proportion of SNPs are expected to be classified as neutral. This pattern is reflected in the two highest-frequency bins of the coverage difference histogram shown in Figure 3.13A. In contrast, among SNPs exhibiting substantial coverage differences, the changes are predominantly biased toward higher reference coverage relative to the alternative allele.

This indicates that most pronounced regulatory effects correspond to reductions in binding associated with the mutation, and are therefore expected to represent loss-of-function variants. This trend is consistent with the tallest bin observed in the distribution of the normalised coverage proportion in Figure 3.13B. This figure also illustrates an increasing accumulation of coefficient coverage values toward the neutral function.

After applying the previous steps, the dataset consisted of 394 gain-of-function CTCF SNPs, $1,145$ loss-of-function SNPs, and 20,325 neutral SNPs. Additional postprocessing steps included filtering the dataset to retain only CTCF sites, excluding those that overlap promoters or enhancers, to ensure consistent, context-specific functional interpretation across all analysed CTCFs. I further restricted the dataset to heterozygous SNPs, thereby enabling comparison of the two alleles within the same cellular environment.

I also examined which mutations were most frequently classified as gain-of-function, loss-of-function, or neutral. To do this, I counted all single-mutation combinations for each functional

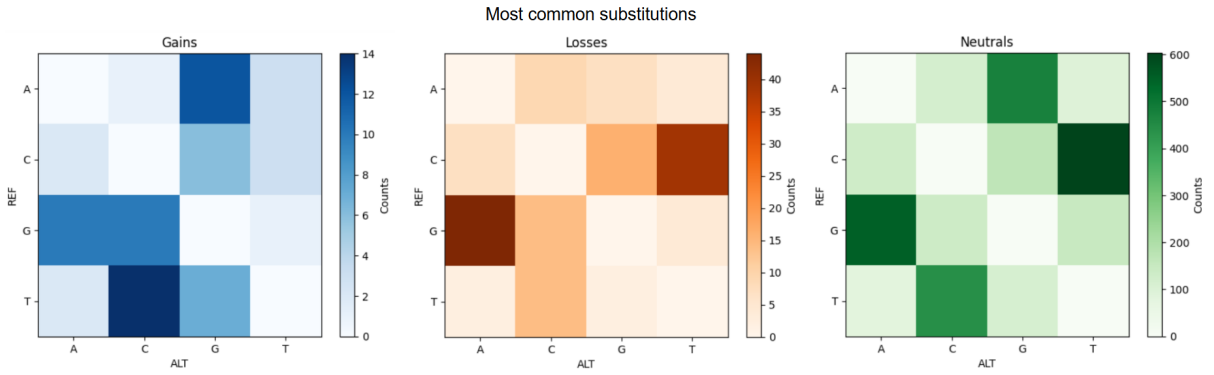


Figure 3.14: Heatmaps with the frequency of mutations across the three variant effect classes. Gain-of-function SNPs are shown in blue, loss-of-function variants in orange, and neutral variants in green.

class. The resulting counts are displayed as a heatmap in Figure 3.14. The data indicate that the most common mutation associated with gain-of-function is a thymine-to-cytosine substitution. In contrast, loss-of-function and neutral outcomes are most frequently represented by guanine-to-adenine and cytosine-to-thymine replacements.

Further analysis involves the linkage disequilibrium (LD) measure (r^2), which had previously been calculated for each SNP using TOPMed, which integrates data from over $100,000$ genomes across diverse ancestries and variants [174]. This metric quantifies the correlation between pairs of SNPs and supports the identification of potentially causal variants. When multiple mutations occurred on the same haplotype, the r-squared value helped quantify the degree of correlation among them.

By incorporating LD information alongside the damage score calculations (ALT-REF and ALT/(ALT + REF)), I aimed to identify independent SNPs with stronger predicted effects on CTCF binding. Here, the damage score represents the proportional change in predicted binding coverage between the mutated sequence (ALT allele) and the reference sequence (REF allele), providing a measure of how strongly a variant is predicted (by HYPER) to disrupt regulatory activity. The subtraction metric |ALT - REF| captures the absolute change in predicted signal, whereas the ratio metric (ALT / (ALT + REF)) normalises this effect relative to the total signal intensity.

In cases where two SNPs were highly correlated (high r^2), I retained only the SNP with the largest absolute damage score for gain-of-function SNPs or the smallest score for loss-of-function SNPs. Additionally, to further reduce the false-positive rate, I excluded SNPs located

in low-mappability or multi-mapping genomic regions, as described in Section 2.7. However, this final filtering step had little impact on the overall counts: only 2 gain-of-function SNPs, 2 loss-of-function SNPs, and 207 neutral SNPs overlapped with these regions. The larger number of neutral SNPs mainly reflects the substantially larger pool of neutral variants in the dataset. Moreover, neutral SNPs located in regions without a clear signal peak that also fail the mappability test are considered unreliable, as it is unclear whether they represent a true biological signal or mapping artefacts, which could ultimately lead to misleading predictions.

After establishing a high-confidence SNP set consisting of 71 gain-of-function, 172 loss-of-function, and $3,051$ neutral variants, I evaluated the performance of the HYPER model in classifying these variants based on their predicted damage scores. To do this, I used HYPER to predict CTCF binding coverage for each *in silico*–mutated genomic region containing a single SNP, with each SNP evaluated independently.

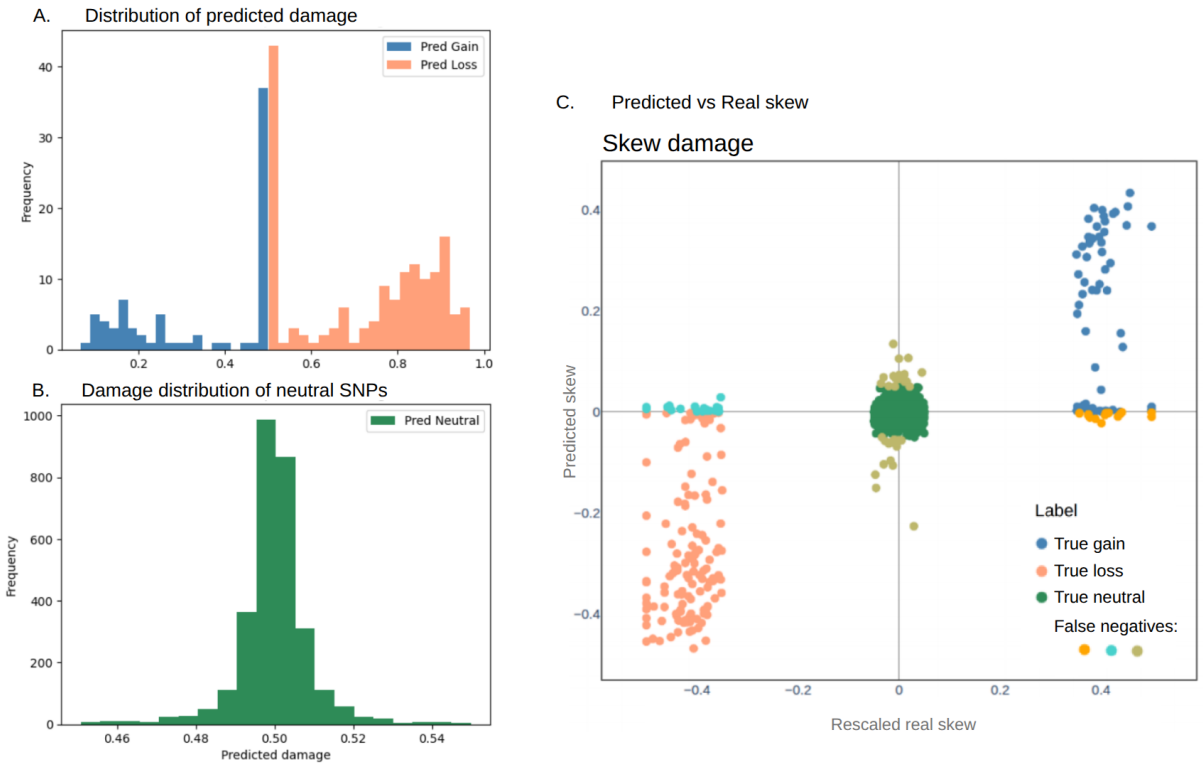


Figure 3.15: Distribution of HYPER-predicted variant damage for CTCF SNPs. A) Distribution of predicted damage scores for gain-of-function and loss-of-function variants. B) Distribution of predicted damage scores for neutral variants. C) Predicted versus real skew damage plot illustrating correctly predicted variant effects, true gains (blue), true losses (salmon), and true neutrals (green), and incorrectly predicted effects (light blue, light green, and orange). The skew, or damage, is rescaled by a shift of $+0.5$, so false positives and true positives are separated by zero on the Y axis, and neutrals are concentrated in the middle.

The predicted binding coverage for these SNPs shows a clear separation in damage score distributions between gain-of-function and loss-of-function variants, as illustrated in Figure 3.15A. In contrast, neutral variants display a Gaussian distribution of damage scores, similar to what is observed in the experimental data (Figure 3.15B).

To quantitatively assess model performance, I generated a skew-damage plot based on the confusion matrix that summarises the classification outcomes (Figure 3.15C). Of the 71 gain-of-function SNPs, 56 were correctly predicted. Among the 172 loss-of-function SNPs, 141 were correctly identified. Finally, out of $3,051$ neutral variants, $3,018$ were correctly classified.

The resulting performance metrics for each class are presented in Table 3.4. Overall, the model performs well across categories. These numbers show strong performance in identifying neutral variants and good performance in detecting loss-of-function variants, while gain-of-function variants are predicted with slightly lower reliability. This difference likely reflects the class imbalance and smaller sample size of gain-of-function SNPs, which can make learning their patterns more difficult. The similarity between the F1 score and the corresponding precision and recall values indicates a balanced performance across these metrics, suggesting that the model achieves a consistent trade-off between correctly identifying variants and minimising false positives.

Although these results indicate strong model performance, I investigated whether additional factors could be optimised to improve variant effect prediction further.

One potential confounding factor is genomic imprinting, a phenomenon in which only one parental copy of a gene is expressed while the other is epigenetically silenced, depending on whether it is inherited from the maternal or paternal allele. At imprinted loci, allele-specific binding patterns may arise independently of the variant itself, potentially biasing the interpretation of variant effects [175]. One approach to address imprinting is to detect occurrences where

Metric \Label	**Gain**	**Loss**	**Neutral**
Precision	0.79	0.87	0.99
Recall	0.73	0.9	-
Accuracy	0.85	0.85	-
F1	0.76	0.89	-

Table 3.4: Performance metrics of HYPER for predicting variant effects.

one allele is systematically underrepresented in sequencing reads. This was already addressed during the WASP allele-specific analysis, which identifies and filters loci showing strong allelic imbalance.

To further control for potential imprinting-related effects, I assessed whether the number of donors in which a given SNP is observed influences the quality of variant effect predictions. Variants detected across multiple donors are less likely to reflect donor-specific or imprinting-related artefacts [176], and therefore may provide more robust signals for evaluating model performance. However, it is important to note that genomic imprinting affects only a small fraction of the genome (approximately 0.1%) [177]. Therefore, this analysis is not expected to alter the overall results substantially.

To evaluate the model's ability to identify the functional effects of SNPs correctly, I analysed SNPs present in 3, 5, 8, 10, 14, 18, 22, and 25 donor sets separately. The analysis was limited to SNPs observed in up to 25 donors because the number of true positives became zero beyond this point, resulting in division by zero in the performance metrics. As shown in Figure 3.16, increasing the number of donors leads to a gradual, nearly linear decline in performance metrics for both gain-of-function and loss-of-function SNPs, driven by decreases in both true positives and false positives. However, this effect is more pronounced for loss-of-function SNPs.

Performance remains relatively stable up to 14 donors, after which the metrics decline more

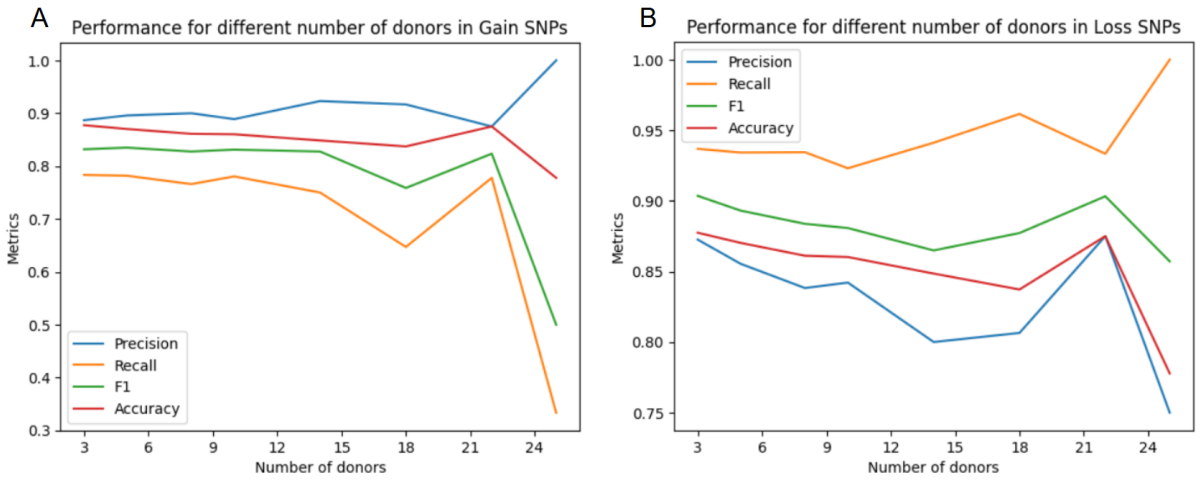


Figure 3.16: Evaluation of model performance in detecting variant effects of SNPs present in varying numbers of donors. A) Performance for identifying gain-of-function SNP variant effects when increasing the number of donors containing the SNPs. B) Performance for identifying loss-of-function SNP variant effects when the number of donors containing the SNPs is large.

rapidly due to the sharp reduction in true positives. A temporary increase in performance is observed at 22 donors: for gain-of-function SNPs, all metrics improve except precision, while for loss-of-function SNPs, all metrics improve except recall. This occurs because of a substantial reduction in true gains, true losses, and false losses. However, this apparent improvement is misleading, as performance drops sharply beyond 22 donors when the number of false gains reaches zero.

Overall, these results suggest that the model performs best when SNPs are present in a small number of donors. However, because the aim is to investigate imprinting, having more donors carrying the SNP strengthens the evidence. Nonetheless, restricting the analysis to SNPs observed in up to 10 donors still yields reasonably strong performance for both gain- and loss-of-function predictions.

Another factor considered for further optimisation of variant effect prediction was the allele-specific bias at CTCF binding sites, which had been previously quantified following the TOPMed analysis. In this case, the degree of allelic skewness, measured through the CTCF binding site p-value, reflects how strongly CTCF binding favours one allele over the other. I therefore evaluated how different thresholds of allele bias affect the model's ability to predict variant effects and whether restricting the analysis to sites with stronger allelic bias improves predictive performance. To assess this, SNPs were filtered according to their p-values using the following thresholds: 10^{-11}, 10^{-9}, 10^{-8}, 5×10^{-8}, 10^{-6}, 5×10^{-6}, 10^{-5}, $5 \times 10{-}5$, 10^{-4}, $5 \times 10{-}4$, 10^{-3}, 5×10^{-3}, 0.01, and 0.05.

As shown in Figure 3.17A, neutral variants display very low statistical significance, with most p-values exceeding 0.4. Variants that were incorrectly predicted as neutral (shown in olive in Figure 3.17A) also exhibit relatively low statistical significance; however, their skew values deviate from zero, which explains why their true classification is not neutral. In contrast, correctly predicted gain-of-function and loss-of-function variants show strong statistical significance. False losses and false gains tend to have predicted skew values closer to zero, suggesting that they may correspond to variants with weak or negligible allelic bias and may therefore resemble neutral variants. Additionally, their p-value distribution, particularly in the false gains set, shows a small concentration of values toward reduced statistical significance, a pattern not observed in the other variant groups.

When predictions were performed on SNPs filtered by p-value, the predictive performance

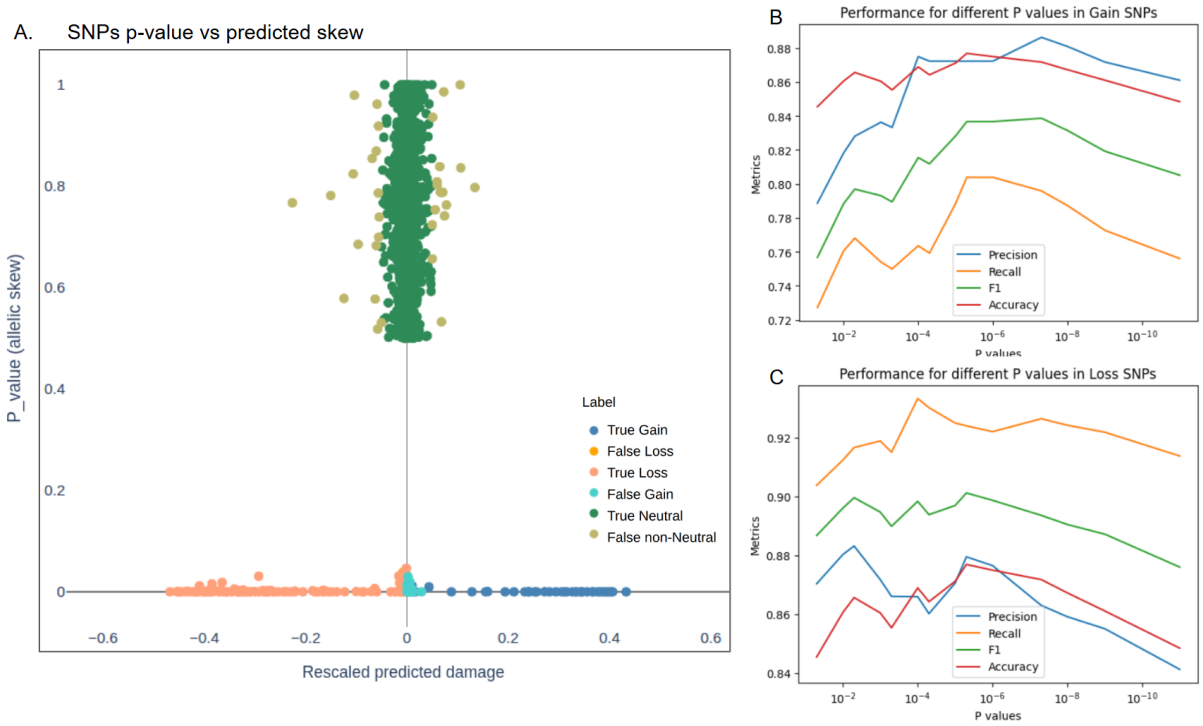


Figure 3.17: Evaluation of SNPs' p-value influence in the prediction of variant effects. A) SNPs p-value versus predicted skew (damage score). B) Performance scores for identifying gain-of-function SNP variant effects when decreasing the SNP's p-value. C) Performance scores for identifying loss-of-function SNP variant effects when decreasing the SNP's p-value.

of HYPER improved substantially up to a threshold of 10^{-5} for gain-of-function SNPs (Figure 3.17B) and 10^{-4} for loss-of-function SNPs (Figure 3.17C). At these thresholds, the model achieved F1 scores of approximately 83% and 90%, respectively. However, applying more stringent p-value thresholds beyond these points resulted in a roughly linear decline in performance. This decrease is likely attributable to the progressive reduction in the number of true-positive variants available for evaluation as the filtering becomes more restrictive.

Overall, these results suggest that the statistical significance of allele-specific binding at CTCF sites indirectly influences the model's predictive performance. Moderate filtering for allele-specific bias appears to enhance variant effect prediction, whereas overly stringent filtering reduces the available signal and negatively impacts model performance.

When optimising these two metrics simultaneously, it is not sufficient to simply select the highest number of donors or the most stringent p-value. The SNP dataset is already limited in size, and applying filters based on multiple parameters would further reduce it. Therefore, I implemented a small script that performed a grid search to identify the optimal combination of parameters by trying each and all of the combinations: the largest number of donors (between 1

and 10) and the smallest p-value threshold (between 0.01 and 0.00001) that maximises the mean F1 score across gain-of-function and loss-of-function SNPs. Neutral variants were excluded from this optimisation because their p-values are not statistically meaningful.

The results indicate that requiring at least 4 donors and a p-value threshold of 0.001 as the maximum yields the best performance, achieving F1 scores of 0.846 and 0.91 for gain-of-function and loss-of-function SNPs, respectively, and an overall accuracy of 88.7%. Under these conditions, the model correctly predicts the variant effect for 44 out of 49 gain-of-function SNPs and 81 out of 92 loss-of-function SNPs. As shown in Figure 3.18, the number of false positives is substantially reduced compared with that observed in Figure 3.15C.

A final analysis examined whether the mutation frequencies observed for each class in Figure

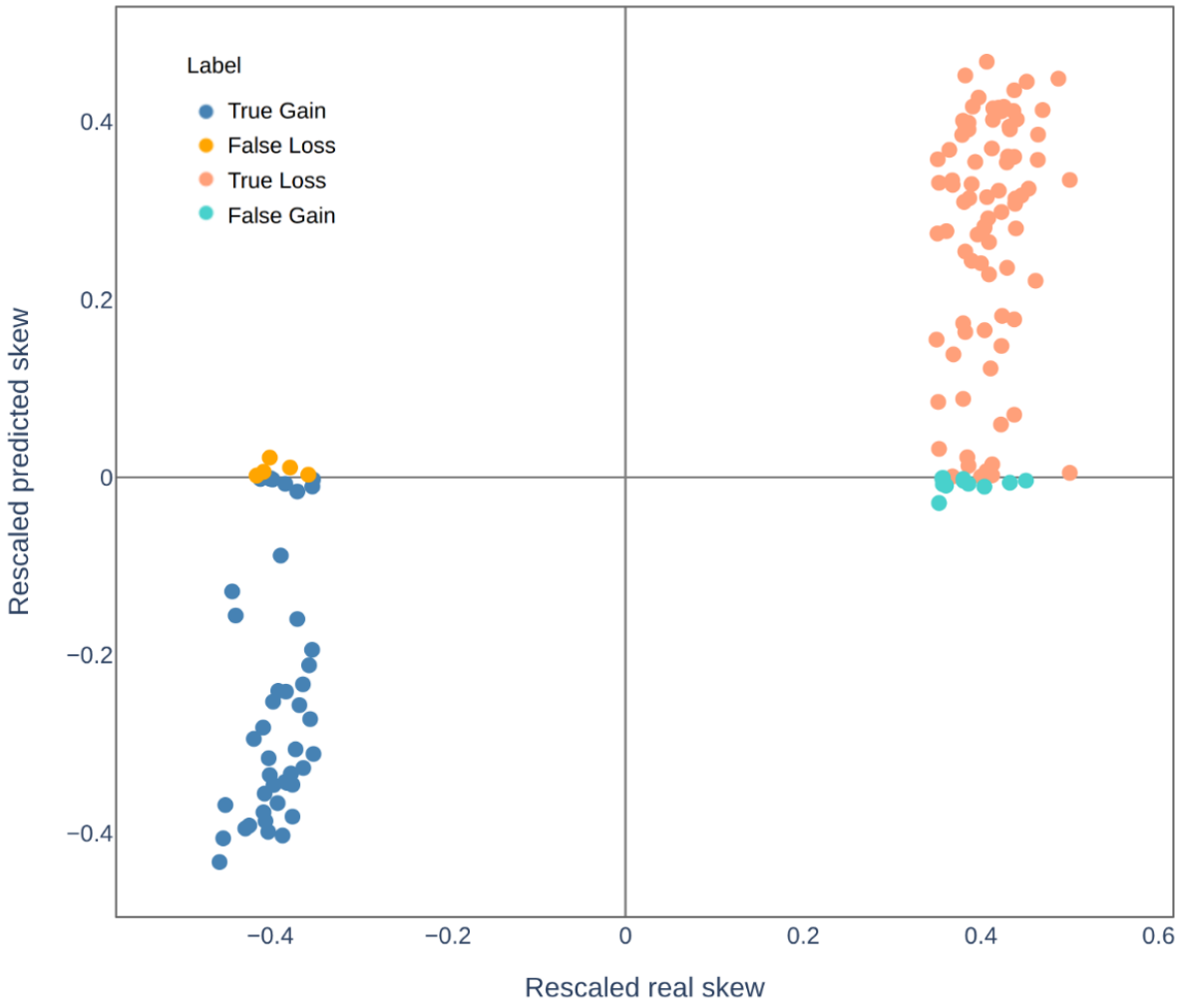


Figure 3.18: Predicted versus real skew of variant effects under optimised parameters. 2-dimension map illustrating correctly predicted variant effects, true gains and true losses, and incorrectly predicted effects, false gains and false losses, considering a minimum of 4 donors and a maximum of 0.001 for p-value. The skew is rescaled just as in Figure 3.15C.

3.14 were also captured by the model. To investigate this, I counted how often each type of single-nucleotide substitution occurred within the three variant effect classes and whether the model predicted them correctly or incorrectly.

As shown in Figure 3.19, adenine-to-guanine substitutions remain among the most frequent mutations associated with gain-of-function, whereas thymine-to-cytosine substitutions are more often incorrectly predicted. Loss-of-function and neutral SNPs again display similar patterns: cytosine-to-thymine substitutions are typically predicted accurately, while guanine-to-cytosine is among the most frequently misclassified mutations in both classes. In addition, thymine-to-cytosine substitutions are commonly misclassified in the loss-of-function class. Although the number of false positives is relatively low, certain mutation types still appear more frequently among the errors. This likely reflects their overrepresentation in the dataset: the more often a mutation type occurs and is predicted, the greater the chance that the model will misclassify proportionally.

In summary, we began with a curated set of SNVs located within well-defined CTCF binding sites, along with their allelic skew. This dataset was processed to accurately classify each

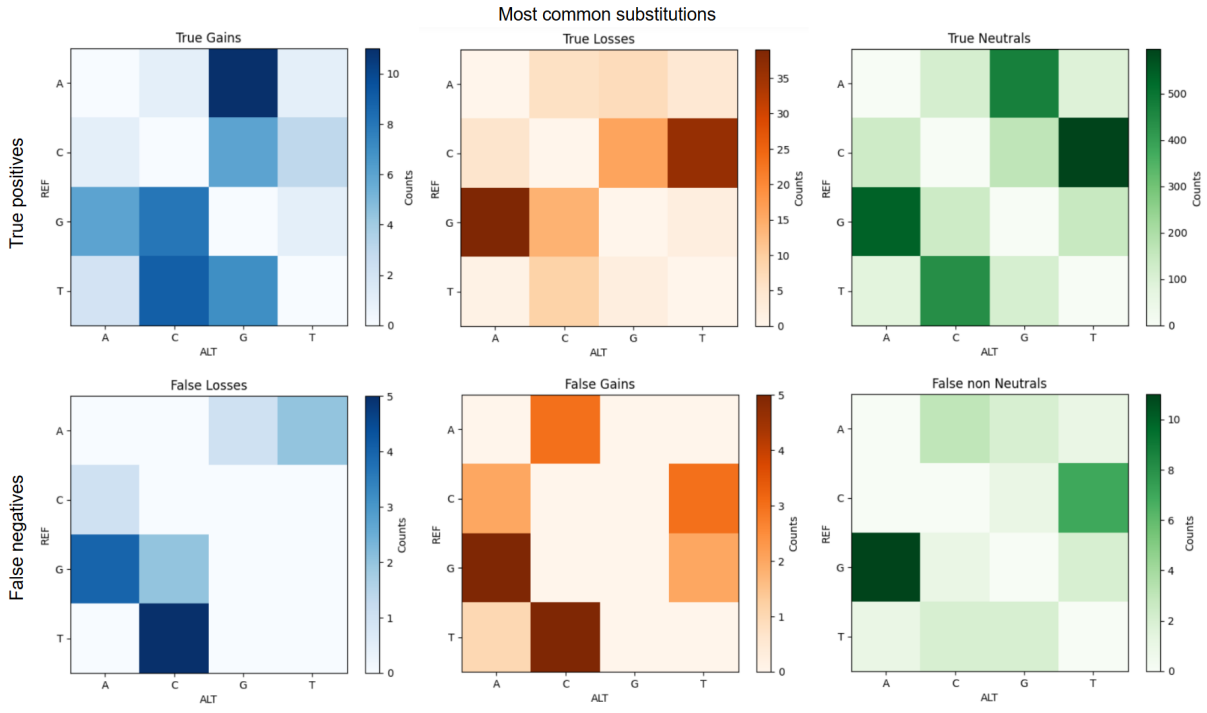


Figure 3.19: Heatmaps with the frequency with which variant effect classes are correctly and incorrectly predicted. The first row represents true positives, while the second row shows the corresponding false negatives for each class, displayed in the same colour. Gain-of-function variants are shown in blue, loss-of-function variants in orange, and neutral variants in green.

variant as gain-of-function, loss-of-function, or neutral. The resulting set was then used as a validation benchmark to evaluate HYPER's ability to capture CTCF binding changes for each SNP at a genome-wide scale. With optimised parameters, such as the allelic p-value threshold and the number of donors, the model achieved an accuracy of 88.7%.

Although these results demonstrate strong predictive performance for variant-effect classification, several limitations should be considered when interpreting the outputs of *in silico* mutagenesis approaches. First, the predicted effects remain dependent on the quality and representativeness of the underlying training data, including potential biases introduced during sequencing, peak calling, and coverage normalisation. In addition, the predicted damage scores are not explicitly calibrated probabilities and therefore should be interpreted primarily as relative measures of predicted regulatory disruption rather than direct quantitative estimates of biological effect size. False-positive predictions may also arise in repetitive or low-mappability regions, or in sequence contexts under-represented in the training data.

Furthermore, the current framework evaluates variants independently and therefore does not capture potential epistatic interactions between neighbouring regulatory variants or broader haplotype-dependent effects that may influence CTCF binding *in vivo*. Regulatory activity may also be modulated by chromatin state, three-dimensional genome organisation, or cell-state-specific cofactors not fully represented within sequence-only models. Consequently, while the observed predictions are consistent with biologically meaningful regulatory effects, they should be interpreted as computational estimates requiring experimental validation rather than definitive evidence of causal mechanism.

3.6 Discussion and conclusion

This chapter aimed to benchmark three DL architectures, two CNN-based models (DeepSEA-based and BPNet) and one transformer-based model (HYPER), for predicting CTCF protein binding across three validation levels: binding track reconstruction, nucleotide importance estimation for motif discovery, and generalisation to unseen sequences. The evaluation was performed on a large, high-resolution in-house erythrocyte CTCF-binding dataset comprising profiles from 50 healthy individuals.

At the binding track reconstruction level, the profiles generated by the two CNN-based models

capture a substantial amount of local noise. BPNet, however, incorporates dilated convolutional layers that expand its receptive field relative to standard convolutional architectures, leading to improved signal reconstruction compared with the DeepSEA-based model. Despite this, the output resolution of both CNN models remains insufficient to fully capture key shape characteristics of CTCF peaks, such as their sharpness and narrowness.

In contrast, HYPER accurately reconstructs both signal and background across the entire prediction window, capturing subtle variations at the appropriate scale. Similarity between predicted and observed tracks was quantified using cosine similarity and NRMSE. While all models exhibited low reconstruction error, HYPER achieved the highest cosine similarity, indicating that although all models approximate the overall signal scale, only HYPER effectively reproduces the most similar peak shapes.

The second validation level focused on computing nucleotide importance scores using DeepLIFT-Shap, which were then compared against a well-characterised CTCF motif (CTCF_C2) from the JASPAR database. The DeepSEA-based model was unable to recover the motif's sequence logo accurately, whereas both BPNet and HYPER performed well, yielding highly significant TomTom p-values. Although BPNet's sequence logo more closely resembles the JASPAR PWM, highlighting its strong ability to capture motif grammar, likely due to its use of dilated convolutions, it assigns overly high nucleotide probabilities at each position. This introduces ambiguity when identifying precise motif instances. In contrast, HYPER produces a more balanced and realistic representation, ultimately outperforming BPNet in recapitulating the shape of a true CTCF_C2 motif instance.

The third validation level leveraged an in-house curated set of CTCF SNPs to evaluate each model's ability to capture variant effects. In the example considered, the SNP induces a loss of protein binding, reflected by the disappearance of a CTCF peak at the functional site. Both HYPER and BPNet clearly captured this loss, while the DeepSEA-based model also reflected the effect, albeit less visibly. Consistent with earlier observations, BPNet reconstructed broader peaks with similar binding intensity, whereas HYPER more accurately reproduced the signal with appropriate intensity, width, and shape.

Given that HYPER outperformed the other models across all three validation levels, further analyses focused on its genome-wide performance. This extended validation included two additional components: a genome-wide assessment of signal reconstruction and a comprehensive

evaluation of CTCF SNP effects.

In the previous track-level comparison, both background and noise were considered. However, because the true binding signal is relatively underrepresented in these datasets, it was necessary to verify whether the model accurately captures signal regions explicitly. To address this, I constructed a curated set of CTCF peaks and extracted coverage from both the original and predicted binding profiles.

First, the real data were downsampled to match the 128 bp resolution of HYPER's output. Peak calling on the full-resolution CTCF track was performed using MACS3. At the same time, log-normal distribution models were applied to both the binned real data and the predicted profiles to identify peaks. This resulted in three peak sets, which were intersected to generate a high-confidence set of $23,356$ CTCF peaks used as ground truth for signal comparison.

This overlap analysis also revealed $9,853$ predicted peaks that did not intersect with the curated set and were therefore considered false positives in this context. However, because the predicted peaks show greater concordance with the full-resolution data than with the binned set, these discrepancies may arise from signal loss during binning or from HYPER capturing signal in underrepresented or ambiguous regions.

The KL divergence and Wasserstein distance between the real and predicted signals were 0.16 and 2.77×10^{-6}, respectively, indicating a high degree of similarity between the two tracks.

Based on these results, several key observations can be made:

- BPNet leverages dilated convolutions to expand its receptive field relative to the DeepSEA-based model, although this expansion remains limited in capturing long-range dependencies.
- The inclusion of a bias model in BPNet reduces its ability to learn local background signals effectively.
- HYPER is capable of modelling interactions between regulatory elements across kilobase-scale distances.
- Standard CNN-based architectures struggle to accurately capture CTCF motifs, suggesting limited capacity to identify more complex regulatory patterns.
- Because BPNet was originally trained on base-pair-resolution ChIP-nexus data, it can

align motifs more precisely to the underlying sequence than traditional CNN models.

- HYPER benefits not only from its architecture and ability to model long-range interactions, but also from pretraining on a large epigenetic dataset, which may provide a richer prior over regulatory sequence patterns and contribute to its improved performance.
- HYPER produces higher-resolution predictions and demonstrates improved performance in identifying CTCF binding influenced by distal regulatory elements.

These results indicate that HYPER can accurately model the CTCF protein-binding signal, capturing both the underlying signal and variant effects at high resolution. In this context, and given the available computational resources and data, HYPER represents the most suitable model for downstream analyses. An open question, however, is whether this level of performance can be maintained when applied to more complex regulatory patterns.

4 | Evaluation of HYPER for predicting chromatin accessibility and transcription factor binding signals

4.1 Introduction

CTCF binding sites are characterised by long, highly conserved DNA motifs, yielding high binding specificity and sharp peak profiles. In contrast, its regulatory role depends on interactions with other elements such as enhancers, which are formed by clusters of TFs. In ChIP-seq these tend to produce broader, less defined peaks due to shorter, more degenerate motifs, greater sequence variability, and cell-type-specific binding. For example, within the beta-globin locus in erythroid cells, the TF GATA1 facilitates chromatin accessibility at CTCF binding sites and promotes the recruitment of both CTCF and the cohesin subunit RAD21 to these regions [178, 179]. This highlights the importance of cooperative interactions between TFs in establishing and maintaining chromatin architecture.

CTCF's distinct binding behaviour reflects its role as an architectural protein. It has been described as a pioneer-like factor, capable of binding nucleosome-occupied or otherwise inaccessible DNA, whereas many TFs preferentially bind pre-accessible chromatin [180]. In addition, its interaction with the cohesin complex generates strong, directional binding patterns associated with loop extrusion and domain boundary formation [181]. By contrast, TF binding is often combinatorial, occurs in clusters, and exhibits positional flexibility [182], making it inherently more variable and difficult to model. Consistent with its role in three-dimensional genome organisation, CTCF binding is relatively stable across cell types and often conserved across species, and disruption of the protein can lead to large-scale chromatin reorganisation such as altered TADs [183], whereas TF perturbations typically result in more localised, context-specific regulatory effects.

In contrast, open chromatin signals represent an even more complex and heterogeneous regulatory landscape than those associated with individual TFs or CTCF. Regions of open chromatin correspond to nucleosome-depleted DNA that can accommodate a wide range of regulatory elements, including promoters, enhancers, insulators, and clusters of TF binding sites. Rather than reflecting the binding of a single factor, chromatin accessibility captures the aggregate effect of

multiple interacting proteins, effectively representing a combinatorial regulatory signal [184]. Moreover, chromatin accessibility is highly dynamic and can change rapidly in response to external stimuli or genetic perturbations [185]. This temporal variability, combined with strong cell-type specificity, further complicates its interpretation and predictive modelling.

Given this context, and considering that the previous chapter demonstrated that DL models are capable of learning CTCF binding signals from DNA sequence, an important open question is whether these models can also capture more complex regulatory relationships. In particular, it remains to be determined whether DL approaches can learn the combinatorial and cell-type-specific interactions between regulatory elements that give rise to nonlinear patterns of gene regulation.

CNNs have been widely used to predict TF binding and chromatin accessibility directly from DNA sequence, typically trained on ChIP-seq, ATAC-seq, or DNase-seq data. A range of architectures have been proposed, including models that leverage dense connectivity to improve feature propagation [186], as well as earlier foundational approaches such as DeepBind [66, 187] and BPNet [68]. These methods differ in their formulation, ranging from classification of binding events to base-resolution signal prediction, but collectively demonstrate strong performance across regulatory genomics tasks.

For chromatin accessibility, CNN-based models have also been widely applied to infer cell-type-specific open chromatin landscapes and identify underlying sequence motifs [188]. More recent approaches further improve performance by explicitly modelling experimental biases, enabling a clearer separation of true regulatory signal from assay-specific artefacts [189].

Although BPNet remains one of the most widely used approaches for modelling TF binding at base resolution, transformer-based models have recently demonstrated clear advantages. Self-attention mechanisms allow them to model interactions between distal regulatory elements separated by tens to hundreds of kilobases, which are difficult to capture with convolutional architectures limited by receptive field size. In practice, this leads to improved modelling of enhancer–promoter interactions and combinatorial TF binding. In addition, transformers can be pre-trained on large-scale genomic or epigenomic datasets and fine-tuned on sparse inputs such as scATAC-seq, improving robustness to noise and enabling more accurate inference of cell-type-specific regulatory activity.

Recent architectures have focused on extending sequence context and improving scalability for regulatory modelling. Transformer-based models pretrained on large-scale epigenomic datasets enable joint learning across multiple assays and cell types [190, 191], supporting tasks such as epigenomic signal prediction, enhancer-gene interaction modelling, and variant effect prediction. In parallel, alternative approaches have been proposed to address the computational limitations of self-attention, using more efficient long-range operators to model genomic sequences spanning tens of kilobases while maintaining tractability, such as HyenaDNA modules [192].

Collectively, these studies highlight a complementary relationship between model classes. CNNs excel at detecting local sequence motifs and short-range patterns, whereas transformer-based architectures are better suited for capturing long-range dependencies and generalising across cell types. However, these benefits come at a cost: transformer models typically require substantially greater computational resources for training and inference, whereas CNN-based models remain more accessible, particularly in resource-constrained settings.

Furthermore, there is consistent evidence that integrating multiple epigenomic modalities improves the prediction of TF binding and chromatin accessibility. Despite these advances, important gaps remain. In particular, most existing models have not been specifically developed or evaluated in the context of erythroid biology, nor have they systematically addressed regulatory modelling at single-cell resolution within blood cell lineages.

Chapter aims

Given that the previous chapter demonstrated strong performance of HYPER in predicting CTCF binding, signals that are relatively well-defined and sequence-driven, the next step is to assess its generalisability to broader regulatory tasks. This includes TF binding, which is often highly cell-type-specific and characterised by short, degenerate motifs, as well as chromatin accessibility, which reflects a composite signal arising from the combined activity of multiple regulatory elements within open chromatin regions.

In the context of erythropoiesis, processes such as differentiation, lineage commitment, and maintenance are regulated by a well-characterised network of TFs, including GATA1, TAL1 (SCL), and NF-E2. GATA1 acts as a master regulator of erythroid differentiation, TAL1 is critical for early lineage commitment, and NF-E2 plays a key role in late-stage maturation. The

binding motifs and regulatory interactions among these factors are well established, and the availability of high-quality, high-resolution epigenomic datasets profiling their activity provides a robust benchmark for evaluating models of TF binding and chromatin accessibility.

Accordingly, these datasets constitute a strong validation framework for assessing the ability of HYPER to generalise beyond CTCF and accurately model complex, cell-type-specific regulatory landscapes. In particular, they enable the evaluation of the model's capacity to reconstruct open chromatin profiles and to predict TF binding signals.

The specific objectives of this chapter are as follows:

- To train the HYPER model on TF binding and chromatin accessibility datasets derived from erythroid cells, extending its application beyond CTCF prediction.
- To evaluate the model's ability to accurately reconstruct open chromatin landscapes in RBCs.
- To assess the model's performance in predicting cell-type-specific TF binding profiles, focusing on key erythroid regulators such as GATA1, TAL1, and NFE2.
- To test the model's generalisation capability by evaluating its prediction of genetic variation effects, including known mutations and SNPs.

This chapter begins by outlining the datasets and model configuration used for training and evaluation. It then presents a comprehensive assessment of model performance in predicting TF binding profiles for TAL1, NFE2, and GATA1, alongside chromatin accessibility in erythroid cells. The model's generalisation capacity is subsequently evaluated through phenotypic perturbation analyses, including the assessment of known variants such as the Vanuatu mutation and additional SNPs.

To further validate model performance, a motif enrichment analysis is performed at the alpha-globin locus, where regulatory motifs are well characterised, enabling comparison between model-inferred and experimentally established binding patterns. Extending beyond validation, the model is also applied in an exploratory setting to identify candidate de novo SNPs predicted to have high functional impact. The chapter concludes with a discussion of these results and their broader implications.

This chapter tests the hypothesis that transformer-based architectures can reconstruct erythroid

TF binding and chromatin accessibility profiles directly from DNA sequence while retaining biologically meaningful regulatory representations. More specifically, it investigates the application of HYPER to erythroid regulatory prediction by training the model on erythroid TF binding and chromatin accessibility datasets. The analysis focuses on the model's ability to reconstruct open chromatin landscapes in RBCs and predict binding profiles for key erythroid regulators, including GATA1, TAL1, and NFE2. In addition, the chapter evaluates whether the model can generalise beyond the training data by assessing its predictions of genetic variant effects, including known mutations and SNPs. Accurate recovery of erythroid-specific regulatory patterns and variant-associated changes would suggest that the model captures biologically relevant sequence determinants of erythroid gene regulation.

4.2 Available datasets

The pretrained Enformer model has been trained on a wide range of tissues, including several blood-related datasets such as hematopoietic multipotent progenitor DNase-seq, GATA1 erythroblast ChIP-seq, erythroid-stage and whole blood CAGE, as well as histone modification profiles from peripheral blood mononuclear cells. This suggests that the model has prior exposure to the erythroid lineage and to cis-regulatory activity during differentiation. However, it does not include erythrocyte-specific tracks or open chromatin data from late-stage erythroid cells.

This chapter therefore focuses on analysing binding profiles of three key TFs involved in erythropoiesis: GATA1, NFE2, and TAL1. In addition, chromatin accessibility in erythrocytes is incorporated as an extra regulatory layer, enabling a more comprehensive evaluation of HYPER's ability to learn and predict diverse epigenetic signals.

Open chromatin track

The open chromatin profile was obtained from the previously described high-throughput ATAC-seq in-house dataset, consisting of erythrocyte ATAC-seq tracks generated from 50 healthy donors. To obtain a representative and robust signal, individual profiles were merged into a single aggregated track, reducing donor-specific noise and variability while enhancing shared

accessibility patterns across individuals. This aggregation improves signal-to-noise ratio and provides a more stable estimate of regulatory element activity. The top track in Figure 4.1 depicts chromatin accessibility across the alpha-globin locus, capturing the major regulatory elements within the region, including promoters and boundary elements.

GATA1 protein binding track

In addition, GATA1 binding profiles were derived from ChIP-seq experiments performed on samples from two donors within the same in-house epigenetic collection. These data provide genome-wide information on the binding of GATA1. The second track in Figure 4.1 shows the GATA1 binding profile, with strong enrichment observed at all four MCS-R elements, as well as within intron 2 of the NPRL3 gene and upstream of the HBM gene.

NF-E2 protein binding track

NF-E2 protein binding profiles were obtained from Olijnik and others, 2020 (accession number GSE125753 [193]). These ChIP-seq experiments were performed in human erythroid cultures derived from peripheral blood mononuclear cells of a healthy male donor. Erythroid differentiation was induced in two steps. In the initial phase, cells were expanded, and in the second phase, specific growth factors were added to drive erythroid maturation. NFE2 binding sites across the genome were identified using the antibody NF-E2 (sc-477, Santa Cruz).

The second-to-last track in Figure 4.1 presents the NFE2 binding profile, showing enrichment at R1 and R2, as well as at the NPRL3 promoter. In contrast, little to no signal is observed at the facilitator elements R3 and R4.

TAL1 protein binding track

TAL1 protein binding profiles were obtained from Palii et all, 2011 (accession number GSE25000 [194]), generated by ChIP-seq in human pro-erythroblasts derived from $CD34^+$ hematopoietic progenitor cells, representing early erythroid differentiation. This dataset provides a well-characterised view of TAL1 regulatory activity in a relevant cellular context. The bottom

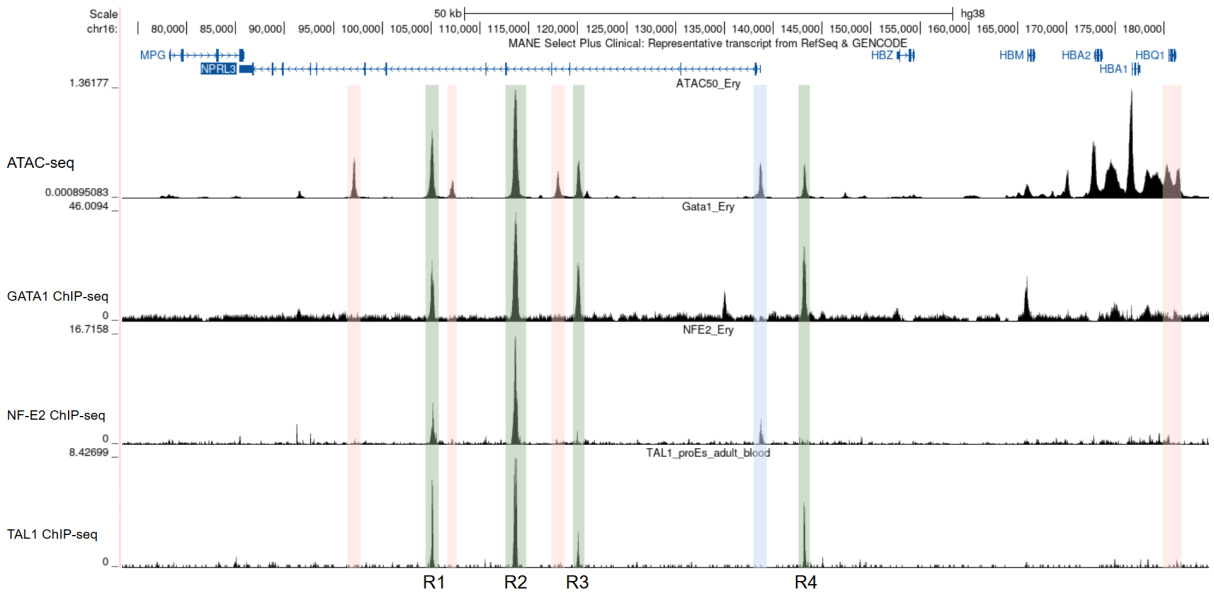


Figure 4.1: From top to bottom, the tracks display the merged open chromatin profile derived from the 50-donor dataset, followed by GATA1, NFE2, and TAL1 binding profiles. Enhancer-like elements (R1 to R4) within the human alpha-globin locus are highlighted in green, while the promoter region is indicated in blue. CTCF binding sites are shown in light red.

track in Figure 4.1 shows the TAL1 binding signal, with clear enrichment across all MCS-R enhancer-like elements.

These datasets were preprocessed and normalised following the pipeline described in Section 2.2. Peaks were called for the analyses in Sections 4.3.2 and 4.3.3 using the approach outlined in Section 2.6.3. As shown in Figure 4.2, the chromatin accessibility track displays a broad and symmetric signal across the genome, whereas TF binding profiles are more variable. GATA1 exhibits a narrower signal with strong enrichment at peak centres, while NFE2 binding is more sparsely distributed, with a lower maximum signal (1.6) and enrichment confined to a smaller fraction of peaks, consistent with either more selective binding or comparatively lower ChIP enrichment for this factor. In contrast, TAL1 shows a lower maximum signal intensity still ($\sim$0.8) and a pronounced discontinuous, striped pattern across the full range of peaks, rather than continuous enrichment. This pattern suggests substantially lower signal-to-noise in the TAL1 ChIP-seq track. A likely explanation is weaker ChIP enrichment or lower sequencing depth for this dataset.

This chapter also evaluates HYPER's ability to identify nucleotide importance sites that correspond to known PWMs for GATA1, NF-E2, and TAL1. This provides an additional layer of validation, as accurate recovery of established motif sequences indicates that the model

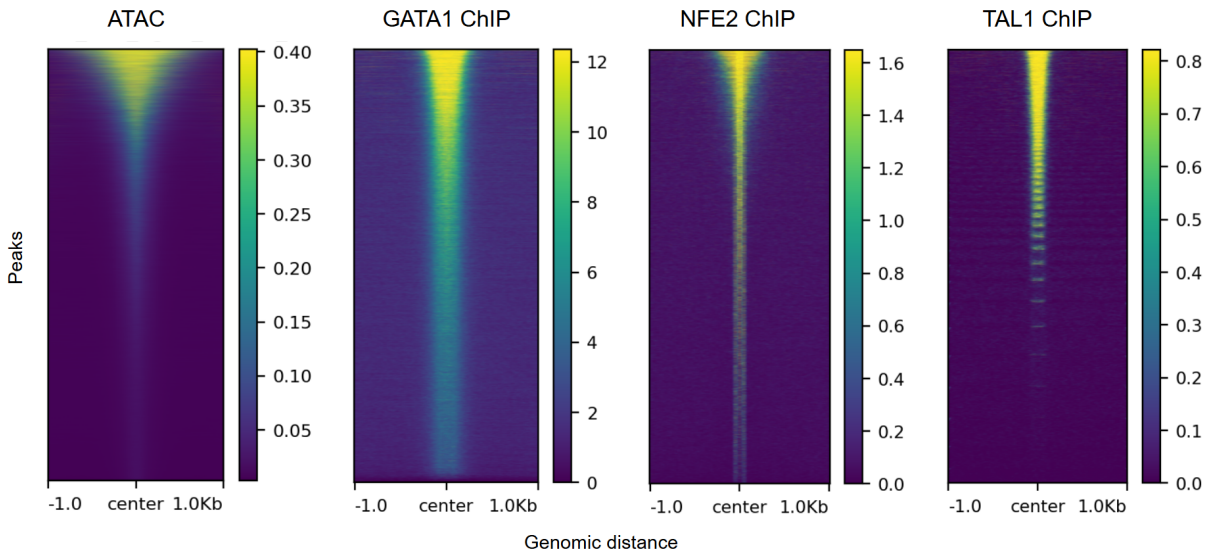


Figure 4.2: Heatmaps showing ATAC-seq and ChIP-seq signal intensity for GATA1, NFE2, and TAL1 across genomic regions centered on peak summits (±1 kb). Each row represents an individual peak, ordered by signal intensity, with colour indicating normalised read density (yellow = high, purple = low).



Figure 4.3: Shown are JASPAR-derived motifs for GATA1, the GATA1::TAL1 co-binding motif, the TAL1::TCF3 co-binding motif, NF-E2, and the MAF::NFE2 co-binding motif. A brief description of this database format is available in Section 2.9.4.

is capturing biologically meaningful regulatory features rather than spurious patterns. The analysis includes both individual TF motifs and co-binding motifs, such as GATA1::TAL1, TAL1::TCF3, and MAF::NFE2 (Figure 4.3). Overall, this assessment tests whether HYPER not only predicts binding signals accurately but also learns the underlying sequence determinants, supporting its potential for motif interpretation and de novo discovery.

4.3 HYPER model for predicting epigenetic signal

Chromatin accessibility and TF binding represent heterogeneous and context-dependent signals arising from multiple interacting regulatory elements. Modelling these signals requires the integration of diverse sequence features to capture a shared regulatory code across assays that reflect different aspects of gene regulation. Evaluating HYPER in this context is therefore essential to determine whether the model can move beyond single-factor binding and generalise to more complex, combinatorial regulatory landscapes. This section assesses HYPER's ability to reconstruct chromatin accessibility and TF binding profiles in erythrocytes, using complementary analyses of signal reconstruction, peak detection, and motif-level interpretation to evaluate both predictive performance and biological relevance.

4.3.1 Model training

To develop a model capable of reconstructing epigenetic tracks from chromatin accessibility and TF binding profiles in human erythrocytes, the four previously described datasets were used. These tracks were employed to fine-tune the final two layers of the HYPER model. More details about HYPER's architecture are available in Section 2.5.4.

Before training, the dataset was split into 95% for training ($23,209$ genomic regions) and 5% for testing ($1,116$ genomic regions). More information about the creation of training datasets for HYPER is available in Section 2.3.3. The model was trained for 30 epochs, achieving a training Huber loss of 0.111. A $95/5$ split was selected to maximise the amount of sequence available for training, thereby increasing the diversity of examples and capturing rare or combinatorial motifs, which can improve the model's ability to generalise across sequence space.

Supplementary Figure S2 presents the distributions of the four epigenetic tracks. The ATAC-

seq track shows the least dispersion, with a strong concentration of values around the background level. In contrast, the other three tracks are more widely distributed, with TAL1 in particular displaying a broader and less pronounced peak. This difference is likely attributable to the much larger size of the 50-donor ATAC-seq dataset, making the other tracks comparatively smaller. The effect can be seen both in the density distributions in Supplementary Figure S2 and in the signal profiles in Figure 4.1, where GATA1, TAL1, and NF-E2 appear less smooth and exhibit higher background noise.

Overall, HYPER was successfully trained on heterogeneous epigenetic datasets, despite differences in signal quality and variability across tracks. These differences provide a basis for evaluating model performance, which is assessed in the following section through visual and quantitative comparisons with experimental TF binding data.

4.3.2 Evaluation of HYPER for prediction of TF protein binding

Whole-genome coverage track reconstruction

Similar to the validations presented in Section 3.4.1, this subsection focuses on both visual and quantitative comparisons between the experimental ChIP-seq data and the HYPER predictions. To this end, HYPER was used to generate genome-wide coverage predictions for the three TF tracks. As an initial step, the reconstructed and original coverage tracks at the alpha-globin locus were visually compared, as shown in Figure 4.4.

Overall, all three predicted tracks reproduce the enhancer-like elements of the alpha-globin locus, with strong positional agreement between predicted and experimental regions of enrichment. However, the GATA1 reconstructed track (Figure 4.4A) exhibits the highest level of background noise, although it still captures several small signal peaks.

Given that the pretrained Enformer model was originally trained on erythroid-lineage GATA1 data, the elevated background signal in the GATA1 predictions may reflect residual pretraining bias, whereby patterns learned during initial training are partially retained and predispose the model to predict GATA1-like signal even in low-signal regions. This is consistent with the higher baseline signal relative to NF-E2 and TAL1.

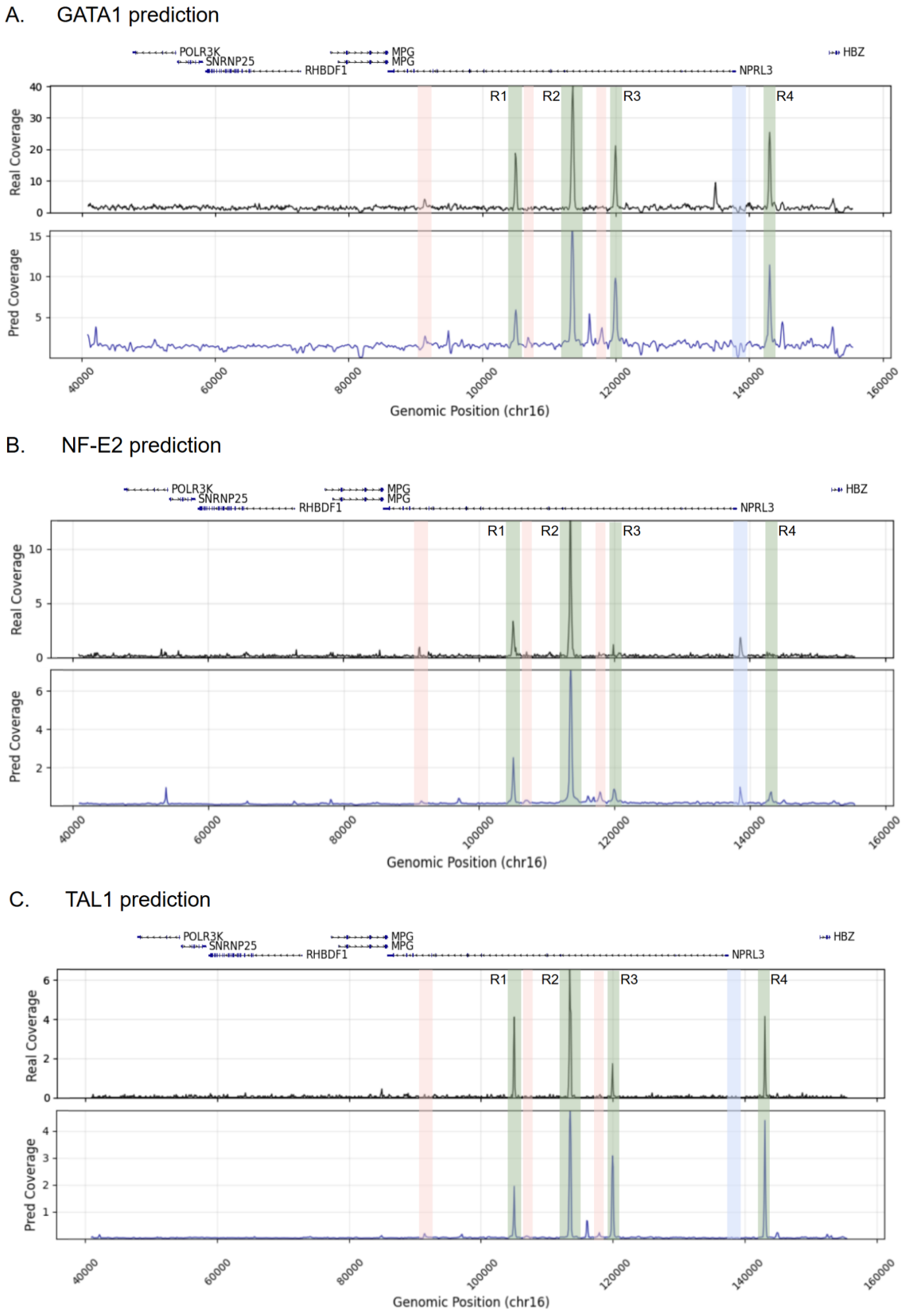


Figure 4.4: A) GATA1 binding profile from two merged donors in the 50-donor dataset (black) alongside the HYPER prediction (blue). B) NF-E2 binding profile from the GSE125753 dataset (black) with the corresponding HYPER prediction (blue). C) TAL1 binding profile from the GSE25000 dataset (black) compared to the HYPER prediction (blue).

However, data-specific factors may also contribute, as the GATA1 dataset is derived from a limited number of donors, where inter-individual variability may further increase background noise in the aggregated track. In contrast, the reconstructed tracks for NF-E2 (Figure 4.4B) and TAL1 (Figure 4.4C) display much lower background levels while preserving clear signal enrichment.

Two similarity metrics were computed to evaluate the agreement between the genome-wide predicted TF binding tracks and their corresponding reference tracks (Table 4.1). The NRMSE quantifies differences in signal magnitude, whereas cosine similarity assesses the similarity in signal shape between experimental and predicted tracks as described in Section 2.10. More information about the MSE is available in Section 2.4. All three TFs showed strong agreement in scale, as indicated by low NRMSE values. However, differences were observed in shape similarity: TAL1 and NF-E2, both derived from lower-resolution datasets, exhibited reduced concordance with their reference profiles, while GATA1, based on a high-quality 50-donor dataset, showed substantially better agreement. Nevertheless, due to the strong imbalance between signal and background, further evaluation restricted to signal-enriched regions is required.

TF	**NRMSE**	**Cosine similarity**
GATA1	0.014	0.87
NF-E2	0.0014	0.44
TAL1	0.006	0.41

Table 4.1: NRMSE quantifies differences in signal scale, whereas cosine similarity captures similarity in signal shape. NF-E2 exhibits the lowest error in scale, while GATA1 shows the highest agreement in signal shape.

Once the whole-genome signal had been reconstructed, I proceeded to identify regions of enrichment through peak calling, to enable comparison between the predicted coverage profiles and those obtained experimentally. To do this, peak calling was first performed on the experimental data using MACS3. A brief description of MACS3's pipeline is available in Section 2.6.2. As described earlier, the resolution of the experimental tracks was then reduced to 128 bp to ensure comparability with the predicted data.

Whole-genome GATA1 peaks analysis

To identify enriched regions genome-wide across the different signals, a custom approach based on statistical distribution fitting was employed. This method provides greater flexibility in defining the underlying distribution, allowing a closer fit to both predicted and binned tracks. Following the peak-calling workflow for HYPER-reconstructed tracks described in Section 3.5.1, and the methodology outlined in Section 2.6.3, the predicted and real binned tracks were modelled using a hybrid log-normal distribution.

The GATA1 reconstructed track exhibits a left-skewed distribution (Figure 4.5, left), rather than the expected right-skewed behaviour. This deviation may reflect elevated background signal and a compressed dynamic range in the predictions.

To address this, near-zero background values were excluded (values > 0.5), and the remaining data were fitted using a reflected hybrid log-normal distribution (Figure 4.5A, right). The bulk

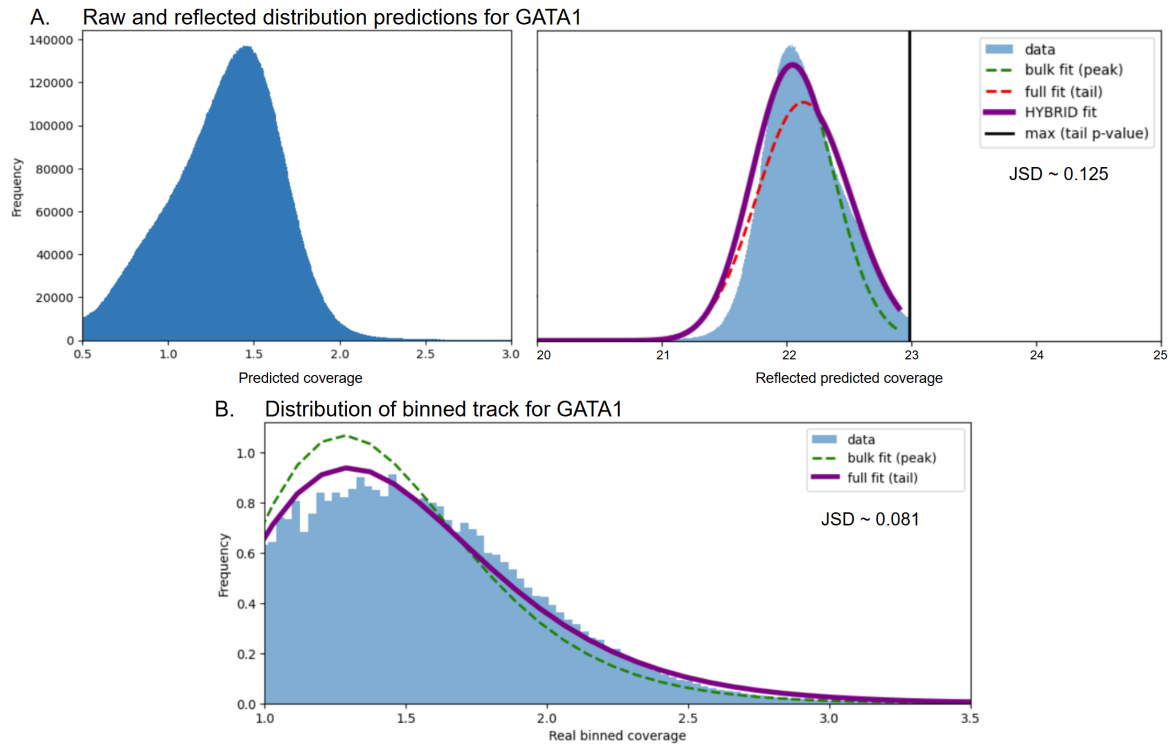


Figure 4.5: A) The left panel shows the raw predicted GATA1 binding values, which display a left-skewed distribution. The right panel presents the reflected predictions, producing a right-skewed distribution, along with the fitted hybrid lognormal model, where the bulk fit captures the central peak and the full fit models the right tail. Most of the signal is concentrated on the left. B) The real binned protein binding distribution exhibits a more symmetric shape and is fitted using only the full fit model.

component was defined up to the 0.85 quantile, yielding a JSD, measuring how different two probability distributions are, of 0.1254. More details about the JSD score are available in Section 2.10. Consequently, the left tail of the reflected distribution (more details in Section 2.10), corresponding to the signal, was used for downstream analysis.

The resulting fit emphasises the signal-dominated region, which is better captured than the background, although some asymmetry remains.

The binned real GATA1 protein binding coverage exhibits a more symmetric, Gaussian-like distribution, similar to the binned CTCF data (Figure 4.5B). To model this distribution, the hybrid log-normal framework was retained for consistency. Near-zero background values were removed using a cutoff (values > 1), and the data were fitted using only the full-fit component, as the bulk component was not suitable for a symmetric distribution. This approach resulted in a JSD score of 0.08.

The low JSD indicates a high degree of similarity between the fitted model and the empirical distribution, suggesting that the full-fit component adequately captures the underlying signal despite the departure from the typical right-skewed behaviour.

Given the moderate mean (1.25) and maximum coverage (113.9) values, together with a substantial and relatively constant background signal, peak calling was performed using stringent parameters. Peaks were identified with MACS3 using a strict significance threshold (q-value cutoff of 0.0001). A minimum peak length of 150 bp was applied to reduce false positives arising from background noise. Duplicate reads were handled by retaining only a single copy of reads mapping to the same genomic location, and peak calling was conducted without constructing a fragment size distribution model, instead relying on default shift and extension settings.

For peak identification in the predicted track, Section 2.6.3 defines thresholds based on both p-value and fold change, which are appropriate when the distribution contains a well-defined background mode and a pronounced right tail. However, this assumption does not hold for the GATA1 distribution. Because fold change is defined relative to the background, and the mode of a left-skewed distribution lies close to the maximum signal, most fold-change values fall within a narrow range (typically between 1 and 2). As a result, increasing the fold-change threshold would exclude only a minimal fraction of the signal and would not effectively discriminate

peaks. Therefore, peak selection for GATA1 was performed using p-value thresholds alone, which provide a more stringent and appropriate criterion under these conditions.

Accordingly, peak calling on the predicted track was performed using a stringent p-value threshold of 0.00001. In contrast, peaks from the 128 bp–binned experimental data were identified using the default parameters described in Section 2.6.3, with a maximum p-value of 0.001 and a minimum fold-change of 5.

After merging overlapping peaks (or those separated by a single base pair), trimming peaks at chromosomal boundaries to match the HYPER output format, and removing peaks overlapping low-mappability or multi-mapping regions, the final peak sets comprised $14,492$ peaks from the experimental GATA1 binding profile, $6,675$ peaks from the 128 bp–binned version of the real data, and $21,356$ peaks from the predicted track. Among these, $4,669$ peaks were shared across all three datasets.

The relatively low number of peaks identified in the experimental data may be attributed to the high background signal present in the dataset. Nevertheless, the main signal peak locations and intensity are properly recapitulated. Figure 4.6 illustrates the experimental, binned, and

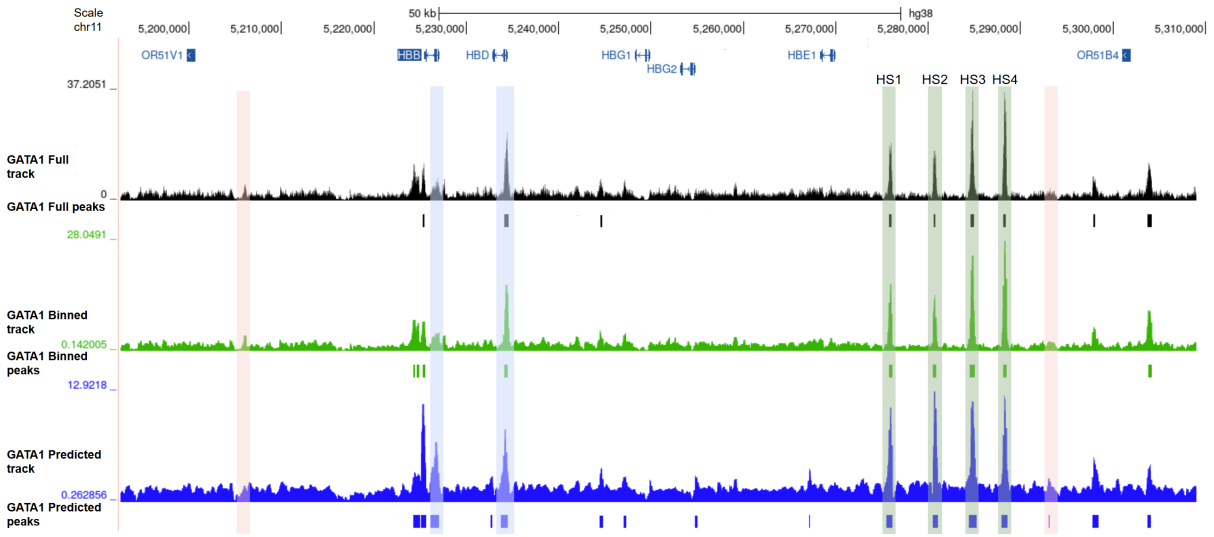


Figure 4.6: UCSC tracks of GATA1 protein binding and signal peaks across the beta-globin locus. The experimental base-pair resolution track is shown in black at the top, followed by its MACS3-called peaks. Below, the 128 bp–binned experimental track is displayed in green with its corresponding log-normal–derived peaks. The HYPER-predicted track is shown in blue, also accompanied by peaks identified using the log-normal approach. Vertically highlighted regions indicate cis-regulatory elements: HS1–4 enhancers (light green), CTCF binding sites (light red), and promoters (light blue). An interactive version of this UCSC Genome Browser session is available at: https://genome.ucsc.edu/s/arodrig/HYPER_hg38_WG.

predicted GATA1 ChIP-seq tracks across the beta-globin locus, along with the corresponding called peaks for each dataset. Most cis-regulatory elements are successfully detected; however, the elevated noise levels hinder complete peak identification in the experimental data, and some signal is lost during the binning process. Additionally, the model appears to have learned aspects of the local background signal, which may lead to the identification of potential false positives. Given the limited quality of the original data, it is difficult to determine the extent of this effect.

The signal corresponding to the overlapping set of peaks was extracted using the UCSC tool `bigWigAverageOverBed`. The similarity between the experimental and predicted data was then quantified using KL divergence and Wasserstein distance, yielding values of 0.102 and 1.22×10^{-5}, respectively, indicating a high degree of similarity between the two distributions. A brief description of these metrics can be found in Section 2.10.

Whole-genome NFE2 peaks analysis

Continuing the analysis, NFE2 protein binding coverage was filtered out by values > 0.05 to remove the biggest concentration of background values near zero. The following distribution exhibits a highly symmetric distribution of predicted protein binding coverage values, as shown in the top panel of Figure 4.7. This dataset is well described by the bulk fit alone, rather than the hybrid model. In this case, the bulk distribution is fitted up to the 0.98 quantile, achieving a JSD score of 0.085.

In contrast, the binned real coverage displays a more right-skewed distribution after filtering out values equal to zero, as seen in the bottom panel of Figure 4.7. Consequently, the hybrid lognormal model provides a better fit than either component alone. The central peak is best captured by a bulk fit up to the 0.95 quantile, followed by a full fit that is merged at their point of overlap. This results in a JSD similarity score of 0.176.

Peaks derived from the fitted hybrid distributions were identified using the default thresholds described in Section 2.6.3, namely a maximum p-value of 0.01 and a minimum fold-change of 5. In contrast, peak identification in the experimental signal proved more challenging than for the GATA1 dataset, likely due to the large disparity between the mean coverage (0.076) and the maximum value ($4,920$). This extreme dynamic range introduces numerous outliers, making

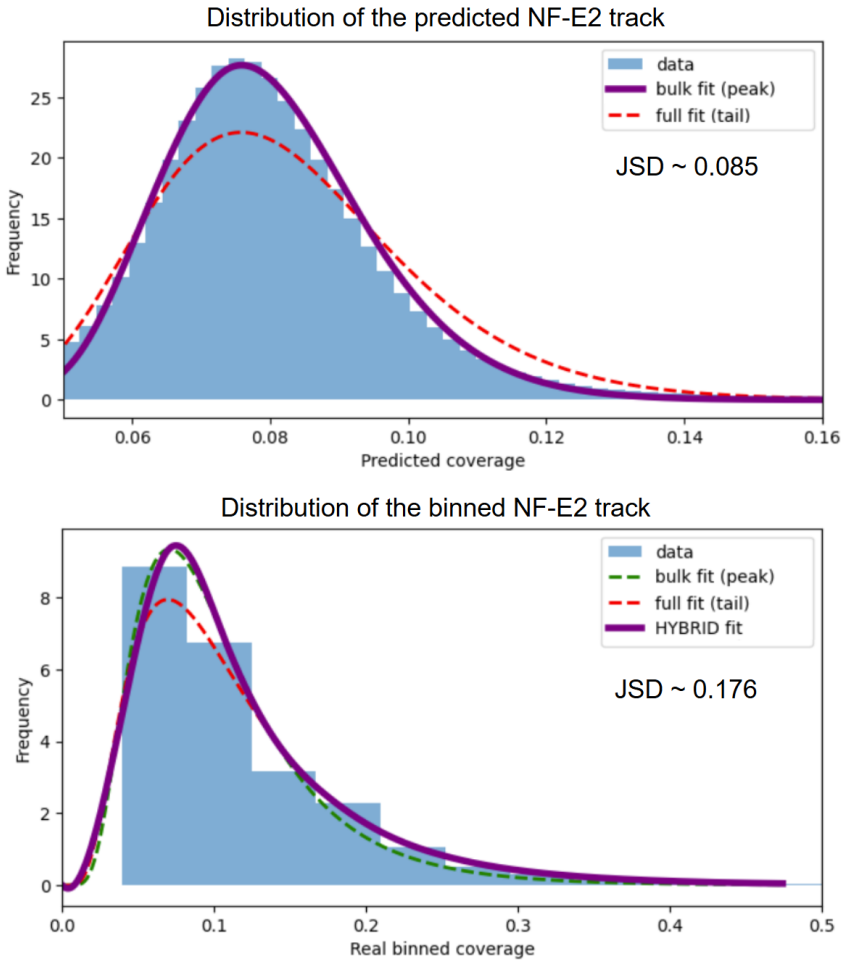


Figure 4.7: Distribution of reconstructed NFE2 protein binding coverage. The top panel shows the predicted binding coverage distributions alongside their bulk model fitting. The real binned protein binding distribution exhibits a more right-skewed shape and is fitted using the hybrid lognormal model.

peak calling from raw, unnormalised data less reliable and more susceptible to artefacts detected by MACS3. Consequently, slightly less stringent parameters were adopted: peak calling was performed with a q-value cutoff of 0.01, while all other settings were kept at their default values. As in previous analyses, no fragment size distribution model was estimated.

Following post-processing, merging overlapping peaks (or those separated by a single bp), trimming peaks at chromosomal boundaries, and removing peaks that did not pass the mappability filtering, the final peak sets consisted of $19,144$ peaks from the experimental base-pair resolution data, $72,317$ peaks from the 128 bp–binned version of the same data, and $17,061$ peaks from the predicted track. Of these, $3,900$ peaks were common to all three datasets.

The discrepancy in peak numbers between the base-pair resolution and 128 bp–binned profiles likely reflects the low coverage of the original data, which limits the ability of peak callers

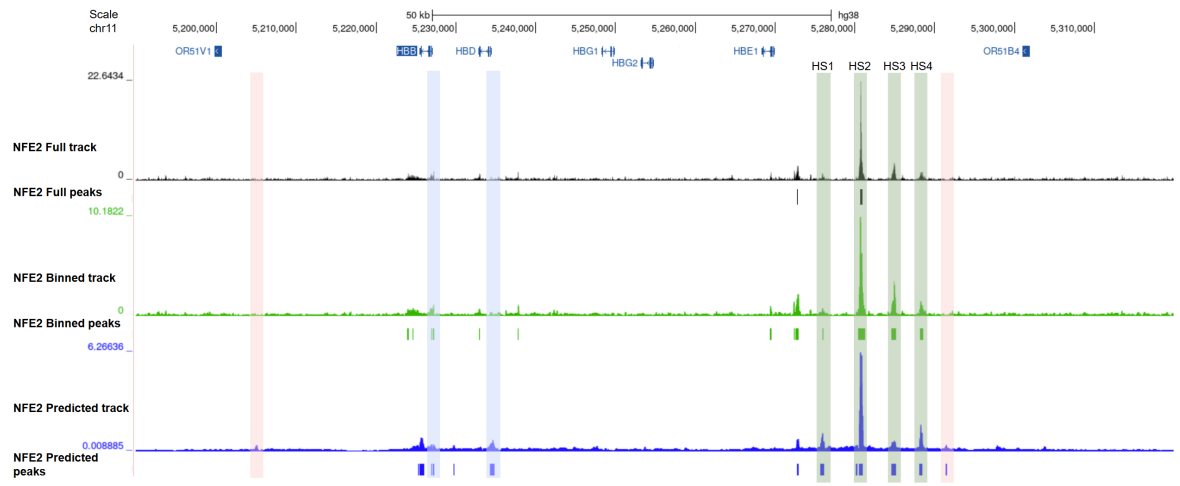


Figure 4.8: UCSC tracks of NF-E2 protein binding and signal peaks across the beta-globin locus. The experimental base-pair resolution track is shown in black at the top, followed by its called peaks. Below, the 128 bp–binned experimental track is displayed in green with its corresponding log-normal–derived peaks. The HYPER-predicted track is shown in blue, also accompanied by peaks identified using the log-normal approach. Vertically highlighted regions indicate cis-regulatory elements: HS1–4 enhancers (light green), CTCF binding sites (light red), and promoters (light blue). An interactive version of this UCSC Genome Browser session is available at: https://genome.ucsc.edu/s/arodrig/HYPER_hg38_WG.

to distinguish true signal from background. With relatively few reads available, reliable peak detection becomes more challenging. Nonetheless, in the beta-globin locus, both the binned experimental data and the predicted track show accurate peak classification, with cis-regulatory elements clearly represented, as illustrated in Figure 4.8.

The similarity between the experimental and predicted signal within this set of overlapping peaks was quantified with a KL divergence of 0.266 and a Wasserstein distance of 4.3×10^{-5}. These results indicate an overall good agreement between the two distributions. The comparatively higher KL divergence, relative to previous analyses, can be attributed in part to the low read coverage of the experimental data, which introduces noise and sparsity in the signal. In particular, regions with very low or zero counts, as well as sporadic fluctuations arising from sampling noise, can lead to local discrepancies that are strongly penalised by KL divergence. As a result, even if the main signal patterns are well captured, these local instabilities inflate the divergence, while the low Wasserstein distance suggests that the overall distributional shape remains similar.

Whole-genome TAL1 peaks analysis

Unlike GATA1 and NFE2, both TAL1 distributions, predicted coverage and real binned coverage, are better characterised by right-skewed behaviour; therefore, a hybrid lognormal model provides a better fit to their data points.

For the predicted coverage, the distribution exhibits a broad peak with a gradual increase and a slow decay, extending from frequencies of around $6,000,000$ and above (Figure 4.9, top plot). To improve the fit to a hybrid lognormal distribution, which typically shows a sharper decline around the peak, I filtered the data to retain values greater than 0.01. The bulk component was then fitted up to the 0.98 quantile. The resulting model achieves a JSD of 0.128 when compared to the predicted coverage distribution.

In contrast, the real 128-binned coverage appears considerably more irregular, with prominent

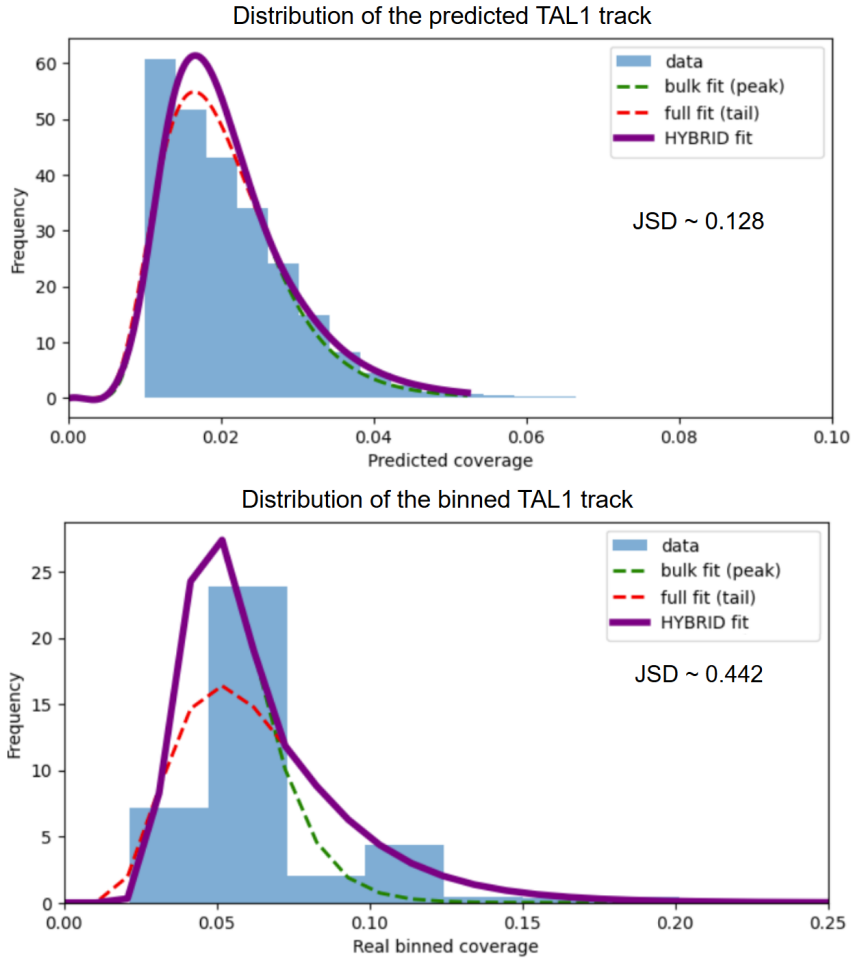


Figure 4.9: Distribution of reconstructed TAL1 protein binding coverage. The top panel shows a right-skewed distribution of predicted binding coverage, while the lower panel exhibits a more irregular distribution also fitted by a hybrid log-normal model.

peaks distributed between 0.03 and 0.15 that are often non-sequential (Figure 4.9, bottom). This irregularity makes it more challenging for the model to achieve a good fit. However, after filtering values > 0.02, the data shows improved conformity to a lognormal distribution. Fitting the bulk component up to the 0.8 quantile yields a JSD of 0.442 relative to the reference distribution.

Peaks derived from the fitted hybrid log-normal distributions were identified using the default thresholds described in Section 2.6.3. In parallel, peak calling from the TAL1 base-pair resolution experimental data was comparatively more straightforward than in the previous analyses of GATA1 and NF-E2. Although this dataset had the lowest resolution among the three protein binding profiles, its variance (0.07) was similar to its mean (0.016), resulting in a more homogeneous and predictable signal with fewer outliers and spurious peak amplifications. Consequently, high stringency was not required, and MACS3 peak calling was performed using the same parameter settings applied to the NF-E2 dataset.

Following post-processing of the called peak sets, the final datasets comprised $17,910$ peaks from the experimental base-pair resolution data, $10,509$ peaks from the 128 bp–binned version of the same data, and $25,812$ peaks from the predicted track. Among these, $5,275$ peaks were shared across all three datasets.

The limited coverage of the original raw data becomes apparent through the loss of signal during the binning process. In contrast, both the experimental base-pair resolution data and the predicted track more accurately capture the underlying signal. This is illustrated in the TAL1 binding profile at the beta-globin locus (Figure 4.10), where cis-regulatory elements are consistently identified across all three tracks; however, the binned track fails to capture smaller, more subtle peaks detected in the other two datasets.

The similarity between the experimental and predicted signal within this set of overlapping peaks was quantified using KL divergence and Wasserstein distance, yielding values of 0.37 and 2.44×10^{-5}, respectively. Although both metrics indicate that the distributions are broadly similar, the slightly higher KL value compared to previous analyses may reflect differences in the underlying signal structure. In particular, the irregular and discretised nature of the binned data likely makes it more difficult to accurately fit a lognormal distribution, introducing local discrepancies. Additionally, KL divergence is especially sensitive to small differences in low-signal regions and distribution shape, which can amplify the impact of these irregularities even

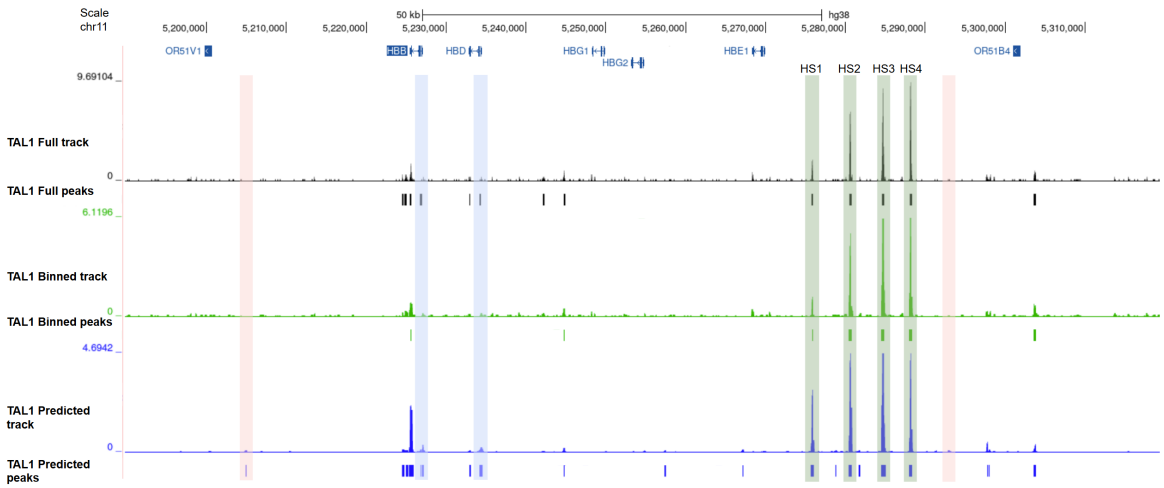


Figure 4.10: UCSC tracks of TAL1 protein binding and signal peaks across the beta-globin locus. The experimental base-pair resolution track is shown in black at the top, followed by its called peaks. Below, the 128 bp–binned experimental track is displayed in green with its corresponding log-normal–derived peaks. The HYPER-predicted track is shown in blue, also accompanied by peaks identified using the log-normal approach. Vertically highlighted regions indicate cis-regulatory elements: HS1–4 enhancers (light green), CTCF binding sites (light red), and promoters (light blue). An interactive version of this UCSC Genome Browser session is available at: https://genome.ucsc.edu/s/arodrig/HYPER_hg38_WG.

when the overall agreement between the signals is strong.

Summary of genome-wide TF binding peaks analyses

A summary of calculated metrics for the signal analysis of the three genome-wide TF protein binding datasets is shown in Table 4.2. Overall, the hybrid lognormal distributions provided a reasonable fit to the binned real data tracks across all TFs, including TAL1, although in this case the fit was less accurate, likely due to its more irregular distribution.

Peaks were then called from the three datasets using a p-value threshold of 0.001 and a fold-change threshold of $5\times$, except for GATA1, which used a more stringent p-value threshold of 0.00001. The number of peaks shared between the experimental data and its binned counterpart was consistently below $10,000$, suggesting a loss of signal during the binning process.

In contrast, the predicted datasets contained a substantially higher number of peaks, of which only 21-28% overlapped with those identified in the other two sets. Although this overlap represents a relatively small proportion, the fact that these peaks are consistently detected across all three datasets increases confidence that they correspond to true signal.

TF	Fitted distribution		JSD		Binned/Real peaks	Predicted peaks	Overlapping peaks	KL divergence	Wasserstein distance
	Predicted	Binned	Predicted	Binned					
GATA1	Reflected hybrid	Full only	0.125	0.081	6,044	21,356	4,669	0.102	1.22×10^{-5}
NF-E2	Bulk only	Hybrid	0.085	0.176	8,713	17,061	3,900	0.266	4.3×10^{-5}
TAL1	Hybrid	Hybrid	0.128	0.442	7,009	25,812	7,009	0.37	2.44×10^{-5}

Table 4.2: Summary of genome-wide peak analysis metrics. The similarity of the fitted hybrid lognormal distributions to the reference experimental data was quantified using the JSD score. Based on these distributions, peaks were called from the three tracks: original (base-pair resolution), binned, and predicted tracks. The number of peaks overlapping the real tracks (both bp resolution and 128-bp binned) is reported in the third and fourth columns, followed by the total number of peaks identified in the predicted dataset. The sixth column shows the number of peaks shared across all three peak sets. The overlapping peak coverages extracted from both the reference data and the predicted data were compared using KL divergence and Wasserstein distance, as presented in the final two columns of the table.

Coverage from these overlapping regions was subsequently extracted from both the experimental and predicted bigWig files. Comparison of these signals showed strong agreement for GATA1, while NF-E2 and TAL1 exhibited more moderate concordance, potentially reflecting issues such as peak overprediction and lower-quality or lower-coverage data.

4.3.3 Evaluation of HYPER for prediction of chromatin accessibility

After demonstrating that HYPER can accurately reproduce TF binding signals at a genome-wide scale, its performance was next evaluated on chromatin accessibility profiles in erythrocyte cells. The 50-donor erythrocyte dataset captures substantially greater inter-individual variability and provides a more representative population-level signal than the TF binding datasets. As a result, evaluation using distribution similarity metrics is more robust, as the signal is more consistent and less driven by dataset-specific noise.

HYPER was used to predict genome-wide chromatin accessibility, achieving low error in signal magnitude (NRMSE $= 0.0046$) and strong agreement in signal shape (cosine similarity $= 0.83$), indicating accurate reconstruction of the accessibility profiles. The reconstructed signal over the alpha-globin locus is shown in Figure 4.11. The model achieves high accuracy in recapitulating key cis-regulatory elements within this region. Enhancer-like elements are particularly well reconstructed, while CTCF binding sites and promoter regions are also detectable, albeit with reduced signal intensity compared to the experimental data.

This suggests that HYPER may exhibit a bias toward enhancer-associated features in chro-

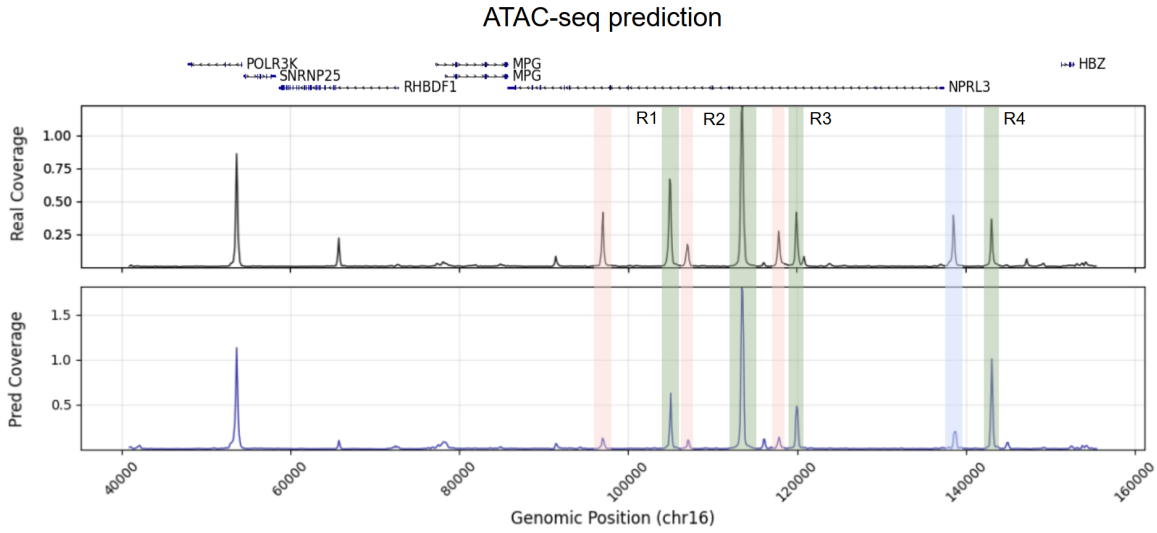


Figure 4.11: Predicted alpha-globin open chromatin signal using HYPER. The top track shows the chromatin accessibility profile obtained by merging all 50 donor tracks (black), with the corresponding HYPER prediction displayed below in blue.

matin accessibility profiles, reconstructing these regions with higher resolution and fidelity than CTCF-bound sites or promoters. This likely reflects underlying differences in chromatin architecture: enhancers such as R1–R4 are bound combinatorially by many erythroid TFs, which jointly displace nucleosomes across a broader region and produce a more pronounced accessibility signal. On the contrary, CTCF only binds to short DNA sequences where nucleosomes stay tightly in place. So CTCF only needs limited chromatin accessibility to bind. Promoters are different; their binding sites tend to stay open, even when the gene isn't actively transcribed. So their signal doesn't change much. Because of this, the model may have learned better to recognise the chromatin openness patterns at enhancers.

Signal extraction from the experimental data was straightforward and carried out using MACS3. Given the high coverage and quality of the dataset, and considering that the mean signal (0.0075) is substantially lower than the standard deviation (0.336) – an expected characteristic of tracks capturing diverse regulatory element activity – stringent parameters were necessary. Accordingly, a very strict significance threshold was applied (q-value $\leq 1 \times 10^{-9}$). Duplicate reads were handled by retaining only a single copy of reads mapping to the same genomic location, and peak calling was performed without estimating a fragment size distribution model, instead relying on default shift and extension settings.

Peak calling was performed on both the 128 bp–binned experimental coverage and the corresponding predicted track using the methodology described in Section 2.6.3. Both datasets

showed right-skewed distributions. To better capture the true signal peak, minimum thresholds were applied to each dataset. For the experimental data, values greater than 0.0025 were retained, while for the predicted chromatin accessibility track, values greater than 0 were included.

Different modelling approaches were then used to fit the distributions. For the predicted track, a bulk component model fitted to the 0.95 quantile provided the best fit. In contrast, the hybrid lognormal model more accurately captured the distribution of the experimental binned data. The fitted distributions are shown as purple curves in Figure 4.12.

The resulting fits achieved JSD scores of 0.091 for the predicted track and 0.11 for the binned

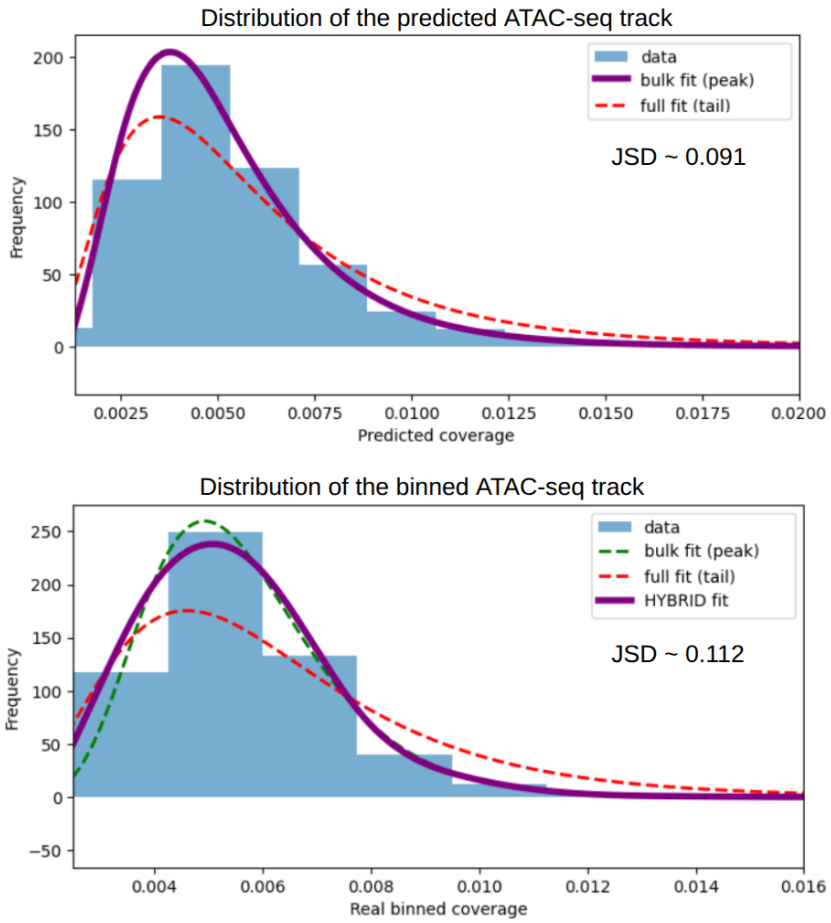


Figure 4.12: Distribution of the reconstructed open chromatin signal. The top panel shows a right-skewed distribution of predicted accessibility coverage fitted with a bulk component (purple). The lower panel displays a right-skewed distribution from the experimental binned data, modelled using an empirical hybrid lognormal fit (purple). Ideally, this hybrid approach captures the central peak through the bulk component and the right tail through the full distribution; however, in this case, the purple line represents an approximation of the empirical distribution, as the full reconstruction was limited by the large size of the original dataset.

experimental data, indicating a high degree of similarity to the experimental distribution.

Similar to the experimental track, peak identification required a high level of stringency. A maximum p-value threshold of 0.00001 was applied to both the binned and predicted tracks, with a minimum fold-change of 10 for the binned data and 5 for the predicted data. After merging overlapping peaks, trimming chromosomal boundaries, and filtering based on mappability, the final peak sets comprised $75,337$ peaks from the experimental base-pair resolution data, 38,010 peaks from the 128 bp–binned version of the same data, and $32,649$ peaks from the predicted track. Of these, $22,465$ peaks were shared across all three datasets.

The substantially larger overlap observed here, compared to the TF binding analyses, primarily reflects differences in data quality, resolution, and signal characteristics. In particular, the high number of shared peaks in the open chromatin profile indicates a more consistent and robust signal that is less affected by downsampling and noise. In contrast, this analysis of TF binding profiles is influenced by its lower coverage and higher variability, leading to reduced overlap between peak datasets.

Despite the use of stringent peak-calling parameters, the total number of detected peaks remains

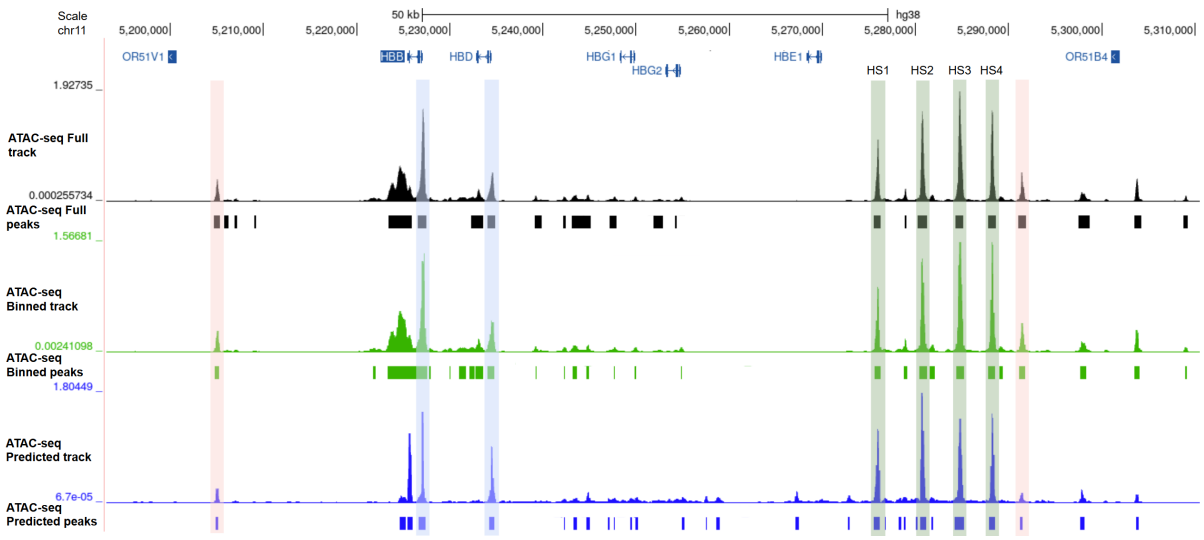


Figure 4.13: UCSC session with chromatin accessibility signal peaks across the beta-globin locus. The experimental base-pair resolution track is shown in black at the top, followed by its called peaks. Below, the 128 bp–binned experimental track is displayed in green with its corresponding log-normal-derived peaks. The HYPER-predicted track is shown in blue, also accompanied by peaks identified using the log-normal approach. Vertically highlighted regions indicate cis-regulatory elements: HS1–4 enhancers (light green), CTCF binding sites (light red), and promoters (light blue). An interactive version of this UCSC Genome Browser session is available at: https://genome.ucsc.edu/s/arodrig/HYPER_hg38_WG.

high, capturing not only well-defined regions but also smaller, subtle peaks, as illustrated in the beta-globin locus chromatin accessibility tracks in Figure 4.13. Nevertheless, cis-regulatory elements within this region are accurately recapitulated, including enhancer and promoter signals with intensities closely matching those observed in the experimental data. CTCF binding sites appear somewhat underrepresented, which may reflect the model's limited ability to capture CTCF-associated features in open chromatin profiles, as previously discussed.

The similarity between the experimental and predicted signal within this set of overlapping peaks was quantified using KL divergence and Wasserstein distance, yielding values of 0.15 and 3.69×10^{-6}, respectively. These results indicate strong overall agreement between the two distributions. The low Wasserstein distance reflects a high degree of similarity in the global signal shape, while the moderate KL divergence highlights minor local discrepancies, potentially arising from features such as CTCF binding sites and other chromatin-opening processes that are not fully captured by HYPER.

Summary of genome-wide open chromatin peaks analysis

In summary, although this section focuses on a dataset with more diverse signals arising from a variety of regulatory elements, the predicted signals show greater similarity to the original data than those observed in the TF binding analyses. This improvement is likely due to the higher-quality training data used for open chromatin prediction, which includes multiple samples from 50 healthy individuals, in contrast to the TF binding datasets that are based on lower coverage and fewer samples.

The similarity between predicted and experimental tracks, as measured by cosine similarity, is notably high for open chromatin data. Among the TFs, only GATA1 achieves a comparable level of agreement, which is consistent with the fact that its data originates from the same high-quality collection. A similar trend is observed in the fitted hybrid lognormal distributions, which show closer alignment for open chromatin than for TF binding predictions.

Furthermore, peak calling reveals a substantially higher overlap between predicted, binned, and experimental tracks in the open chromatin setting, with approximately $22,000$ shared peaks, significantly more than in the TF analyses. The signal distributions within these overlapping regions also demonstrate strong agreement between predicted and experimental data, compa-

rable to the performance observed for GATA1, while NF-E2 and TAL1 show somewhat greater discrepancies.

Overall, these results indicate that HYPER achieves its best predictive performance on chromatin accessibility data and on GATA1, highlighting the importance of high-quality, well-represented training data for accurate model predictions.

4.3.4 Model limitations

Chromatin accessibility and protein-binding assays rely on proteins that are targeted to identify DNA sequences in a sample that meet specific conditions. In chromatin accessibility assays, the Tn5 transposase preferentially inserts into regions of open or accessible chromatin. In contrast, protein-binding assays such as immunoprecipitation (e.g., ChIP-seq) use antibodies to identify DNA regions bound by a specific target protein.

Certain genomic regions have physical and molecular properties that promote protein binding even when the underlying DNA sequence does not satisfy the expected conditions of the assay. This phenomenon occurs in regions with extremely high transcriptional activity. However, because ChIP-seq assays are designed to detect specific protein–DNA interactions rather than general chromatin accessibility, as in ATAC-seq, they appear to be less influenced by the unusual conditions created by intense transcriptional activity.

The alpha-globin and beta-globin genes are among the most highly expressed genes in the genome [195]. As a result, transcriptional machinery continuously binds to their promoters to produce globin subunits, which are essential components of haemoglobin A. These loci are persistently maintained in an open state by remodelling complexes [196], allowing RNA polymerase II to efficiently transcribe the coding regions. Consequently, Tn5 is strongly enriched at these sites, leading to high signal intensity in sequencing data, as observed in the alpha-globin top track in Figure 4.1.

Although some TFs can recruit transcriptional machinery and initiate gene expression, transcriptional activity is influenced by many factors beyond DNA-encoded regulatory elements [197, 198]. Despite the strong enrichment of highly transcribed regions in chromatin accessibility assays, DL-based models primarily learn sequence-encoded features of open chromatin. As a result, they often underestimate signal intensity in these regions, producing lower pre-

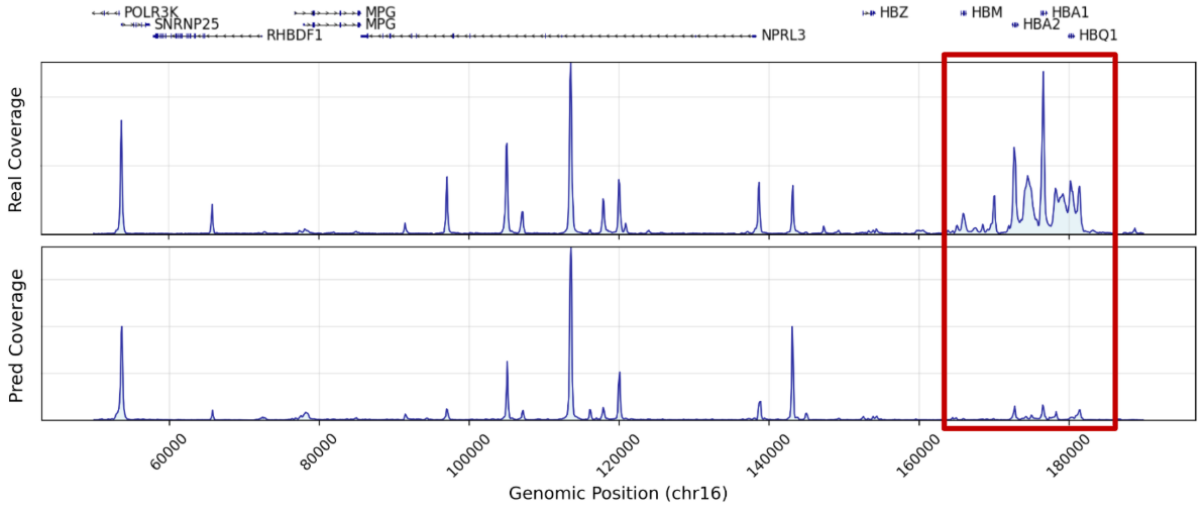


Figure 4.14: Chromatin accessibility tracks in erythrocytes at the alpha-globin locus. The experimental signal shows strong enrichment consistent with high transcriptional activity at the alpha-globin genes, whereas the HYPER-predicted accessibility track displays a markedly lower signal in the same region, marked in a red rectangle.

dicted signals compared to experimental data, as seen in the predicted alpha-globin track in Figure 4.14.

In summary, certain genomic regions have atypical chromatin environments that can influence epigenetic assays and lead to misleading signals. Importantly, these effects are not directly encoded in the DNA sequence, and therefore cannot be fully captured by DL models that rely mainly on sequence information. As a result, the model may underestimate or misrepresent signal in regions driven by high transcriptional activity or other non-sequence-related processes. Therefore, when comparing predictions with experimental data, it is important to account for regions such as highly transcribed loci, chromosome ends, and other atypical genomic contexts.

4.3.5 Reconstruction of silenced DNA regions

So far, the performance of HYPER has been evaluated in its ability to reconstruct signal-enriched regions. However, an equally important question is whether the model can accurately capture extended genomic regions characterised by the absence of signal.

The centromere is a crucial chromosomal region during cell division, as it serves as the attachment site for the molecular machinery responsible for segregating chromosomes. It is the point at which sister chromatids are held together and subsequently pulled apart during mitosis [199].

The pericentromeric region, which flanks the centromere on both sides, has long posed significant challenges for genomic research. This is largely due to its composition of satellite DNA, which is highly repetitive DNA sequences that can span extensive regions with near-identical patterns, making accurate assembly difficult [200]. As a result, these regions are often classified as multi-mapping and considered problematic in genomic analyses. Furthermore, the chromatin in these regions is typically compacted and inaccessible, adding another layer of complexity for experimental profiling.

Another class of genomic regions with low transcriptional activity are LADs, which are large chromosomal regions located at the nuclear periphery, where they interact with the nuclear lamina, a structural network connected to the inner nuclear membrane in eukaryotic cells. These domains are generally associated with gene repression [201] and play a key role in shaping epigenomic states and higher-order genome organisation [202].

Similar to pericentromeric regions, LADs are largely composed of heterochromatin. As such, they exhibit cell-type-specific weak or absent signals in NGS assays, including chromatin accessibility and protein-binding experiments. LADs are often conserved in position between some mammals like humans and mice [32]. They are also actively moving around throughout lineage development, which allows the release of repressed genes that need to be activated during cell differentiation and repressing other genes irrelevant to the lineage.

Based on this, I hypothesised that the regulatory features underlying gene silencing within LADs and pericentromeric regions may be partially encoded in the DNA sequence itself. If this is the case, HYPER might be able not only to reconstruct active signal regions but also to correctly predict these extended low-signal domains, thereby capturing both activating and repressive genomic contexts in a cell-type-specific manner.

To investigate this hypothesis, I examined the regions surrounding the centromere of chromosome 16, a chromosome known for having relatively high gene activity. The erythroid-lineage epigenetic tracks collected for this study reveal two clearly defined transcriptionally silent regions flanking the centromere. The first, spanning approximately chr16:32,200,000-46,500,000, corresponds to the pericentromeric region. The second, located around chr16:58,700,000-66,400,000, is consistent with a LAD, which appears to be conserved across mesoderm-derived cell types, as it is also observed in TIG-3 fibroblast cells.

HYPER's predictions across this extended region closely match the experimental data (Figure 4.15). Importantly, across all four tracks in both real and prediction contexts, the signal does not return to a true zero between peaks, but instead sits on a low, consistent, positive baseline.

This is a direct consequence of the ZEN normalisation procedure (Section 2.3), in which Z-

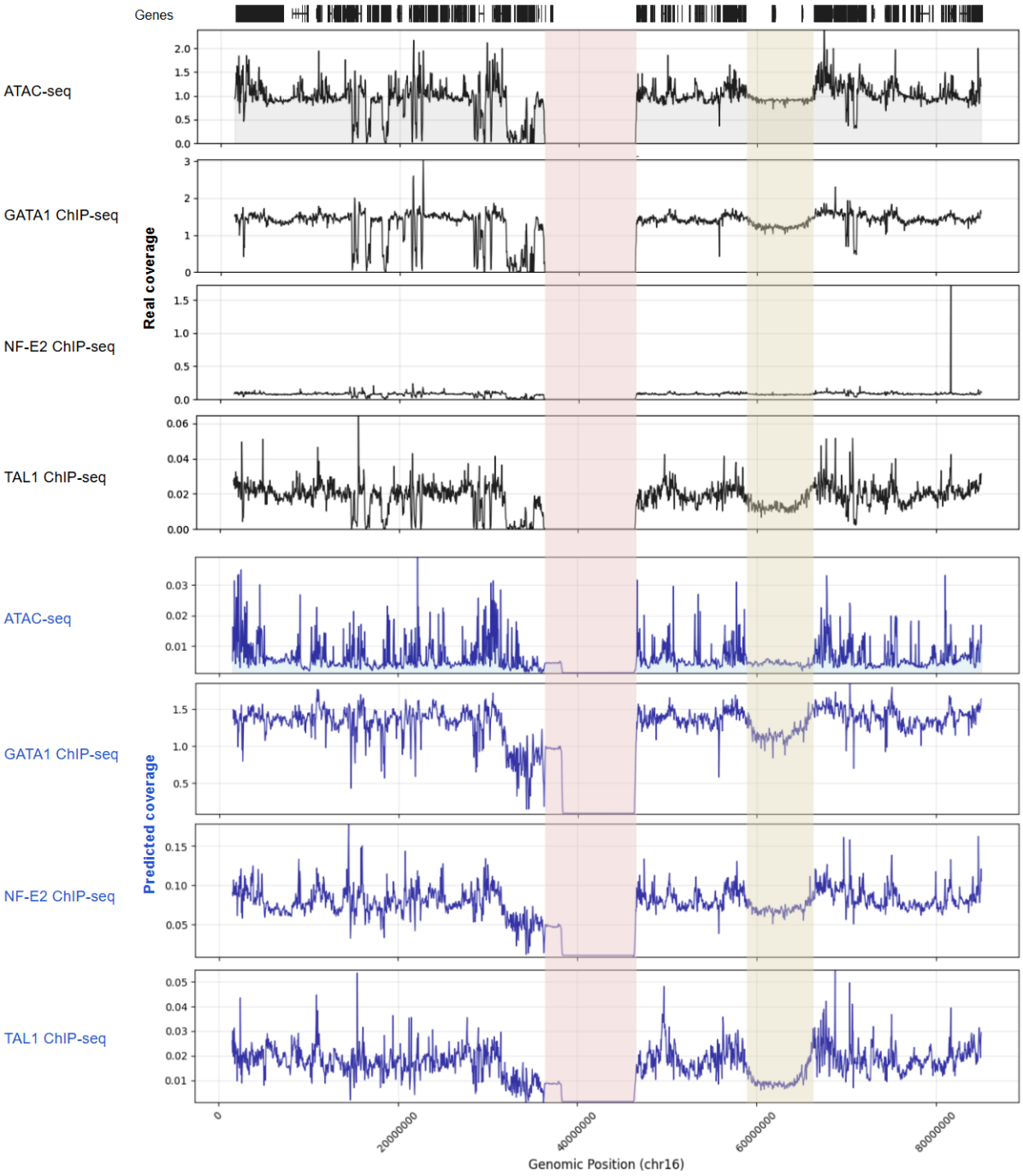


Figure 4.15: Real and predicted epigenetic profiles across the centromeric region of chromosome 16. Experimental data are shown in black and predicted tracks in blue for both TF binding and chromatin accessibility, spanning chr16:0–90,000,000. Two transcriptionally repressed regions are highlighted: the pericentromeric region in light red and a LAD in yellow.

scores are computed relative to the chromosome- or zone-specific background and then shifted upward by subtracting the minimum value, since the underlying signal is non-negative. As a result, only the single most extreme background bin in a given zone is mapped to exactly zero; all other background bins, which fluctuate around a central tendency rather than sitting at the true minimum, are shifted to small positive values above this floor. This produces the low, non-zero baseline visible throughout the real tracks. Because HYPER is trained to reproduce these shifted, non-negative targets, its predictions inherit the same floor, which is why a comparable non-zero baseline is also visible in the predicted tracks.

Regarding the silenced regions, the pericentromeric predicted signal is mostly absent, except for a ~ 4 Mb segment at the beginning of the pericentromere (chr16:32,050,000–35,950,000), where a flat, plateau-like signal with moderate intensity appears. In the experimental data, there are several reasons why pericentromeric regions still show some epigenetic activity despite being highly condensed. First, recent studies suggest that satellite DNA in these regions can be transcribed at very low levels during interphase, when chromatin becomes temporarily accessible, allowing Tn5 binding at low frequency [203].

Second, the highly repetitive nature of these regions can cause genome aligners to misassign reads, producing artificial peaks [204]. Additionally, immunoprecipitation-based assays may detect signal in these regions because the chromatin is "sticky" and can nonspecifically bind antibodies or beads with low affinity [205]. As a result, even background or artefactual signals present in the data may be learned by the DL model.

On the prediction side, several factors may explain why HYPER produces signals with inaccurate intensity. First, these regions are inherently difficult to sequence, which results in genome assemblies containing many undetermined nucleotides (e.g., stretches of "N"s). Consequently, the underlying sequence context is incomplete or ambiguous. The plateau-like signal described above is therefore likely a computational edge artefact, arising from the model's large receptive field encountering a gradual transition from high signal to uniformly low signal across repetitive elements, an effect commonly observed at the onset of pericentromeric regions.

In addition, when the model's sliding window spans predominantly low-signal regions, it is forced to learn from background noise rather than informative features. As a result, the predictions tend to resemble noise, albeit with slightly elevated intensity, reflecting the model's limited contextual guidance in such areas. Despite these limitations, the model remains capable

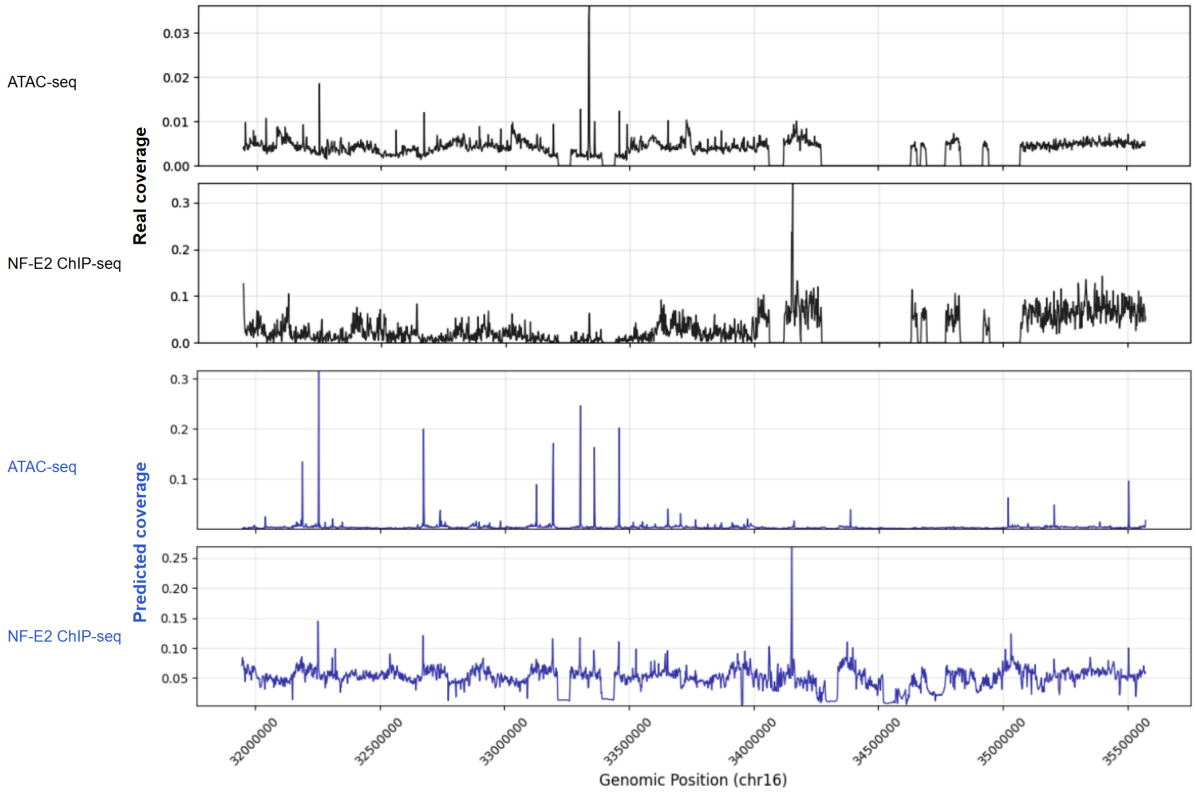


Figure 4.16: Real and predicted epigenetic profiles across the pericentromeric region of chromosome 16. Experimental data are shown in black and predicted tracks in blue for NF-E2 protein binding and open chromatin profiles, spanning chr16:31,900,000-35,580,000.

of accurately capturing narrow, high-intensity peaks. This is illustrated in Figure 4.16, which presents a zoomed-in view of the pericentromeric region shown in Figure 48 and highlights the preservation of these sharp signal features.

Similar to pericentromeric regions, LADs exhibit both chromatin "breathing" and stickiness. This means that chromatin is dynamic and can transiently open, allowing Tn5 access and facilitating nonspecific antibody binding. In addition, chromatin in LADs is less compact than in pericentromeric regions and is characterised by broad H3K9 methylation patterns. These features produce a continuous lamin-contact signal rather than sharp, isolated peaks [206], which explains why the signal in these regions appears flat but remains positive.

In summary, HYPER is able to reconstruct low-signal regions, but its outputs are strongly influenced by the chromatin's underlying molecular properties, as well as experimental biases, noise, and modelling artefacts. These factors do not prevent the correct identification of regions, but they are learned by the model and reflected in the predictions, sometimes affecting their appearance.

For example, pericentromeric regions typically show a flat, plateau-like near-zero signal at the boundary, followed by a near-zero, uniform profile. In contrast, LADs tend to exhibit a more plateau-like positive pattern, characterised by a noisy but broadly uniform signal with slightly higher intensity than pericentromeric regions. Despite these differences, both types of regions remain easily identifiable, either through post-processing pipelines or by visual inspection.

4.3.6 Generalisation to unseen sequences

HYPER's capacity to predict gain- and loss-of-function effects has previously been validated using a CTCF protein-binding model. This section extends that evaluation by examining HYPER's ability to predict the impact of variants in chromatin accessibility contexts, using a set of three known mutations. Furthermore, nucleotide importance scores derived from the model can be used to highlight key motifs underlying the predictions, which can then be compared to known TF binding sites using public databases. In this study, JASPAR is chosen for this comparison, as it is currently the most up-to-date publicly available resource.

Vanuatu mutation

The Vanuatu population often carries a single-nucleotide variation (SNV) on chromosome 16, located upstream of the HBM gene at chr16:159,709, where a T is replaced by a C [207]. This mutation creates a novel GATA-1 binding site and introduces a new promoter between the alpha-globin enhancer cluster and the alpha-globin genes, as shown in Figure 4.17. This ectopic regulatory element disrupts the interaction between these regions, decreasing alpha-globin gene expression and ultimately leading to alpha-thalassaemia.

To reproduce the effects of this variant using HYPER, the thymine at position chr16:159,709 was first replaced with a cytosine in the `hg38` genome assembly. Both the modified and the original (control) sequences were then used as input to HYPER to generate predicted epigenetic profiles. Visual inspection of the resulting tracks (Figure 4.18) indicates that HYPER captures the emergence of a novel chromatin accessibility peak, along with corresponding increases in predicted TF binding for GATA1, NF-E2, and TAL1 at the variant site. Although the GATA1 and NF-E2 signals retain a non-zero baseline signal outside peak regions, the gain of binding activity is consistently observed across the predicted profiles.

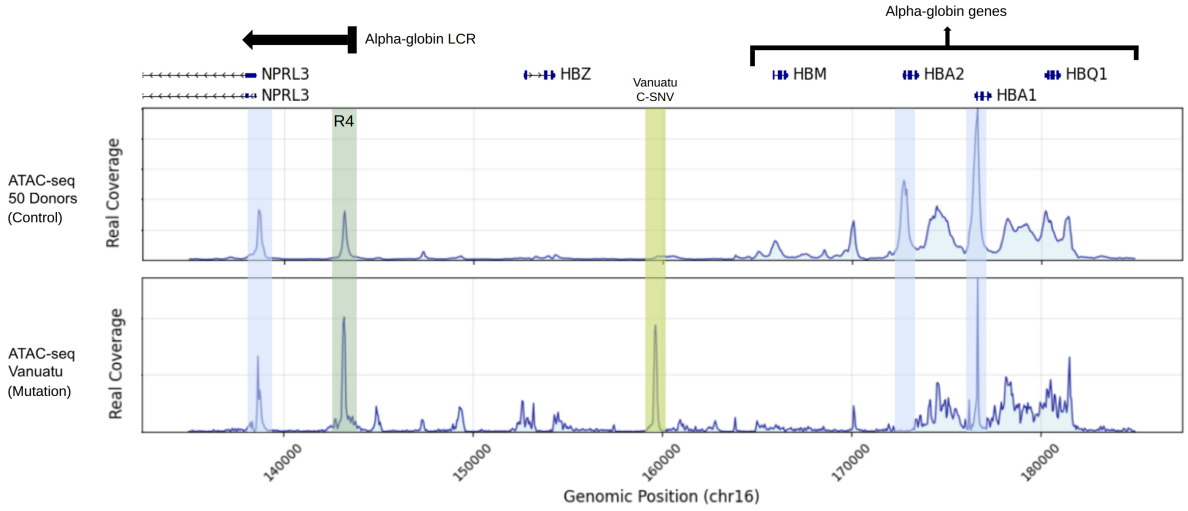


Figure 4.17: Effect of the Vanuatu SNV on chromatin accessibility. Control profiles were derived from the merged track of the 50 donor dataset, and the mutated profile from GSE159875. The Vanuatu mutation (chr16:159,709) is marked by a yellow-green bar. Enhancers and promoters are shown in green and blue, respectively.

As a next step, I computed nucleotide importance scores for this region based on the predicted peak at the Vanuatu mutation site across the three TF binding tracks and the chromatin accessibility track. This was done using DeepLIFTShap, followed by normalisation of the importance scores. More details on this process can be found in Section 2.8. Motif similarity analysis was then performed using TomTom against the JASPAR database. In parallel, TFMoDISco was applied to compare motif discovery between the two approaches (more details in Section 2.9.1).

As shown in Figure 4.20, visual inspection of the nucleotide importance logos suggests the presence of a GATA1 motif followed by a TAL1 binding site. TomTom consistently identified this motif across all four importance profiles as a unique and statistically significant match, with strong p-values. More details on TomTom's pipeline are available in Section 2.9.2. In contrast, TFMoDISco failed to recover any relevant TF motifs and returned results with low statistical significance.

Overall, HYPER successfully recapitulates the effects of the Vanuatu variant across all four epigenetic profiles, including the identification of a novel GATA1::TAL1 binding site. This demonstrates not only its ability to capture global changes in signal intensity but also its sensitivity to fine-scale regulatory alterations, such as the creation of new functional motifs.

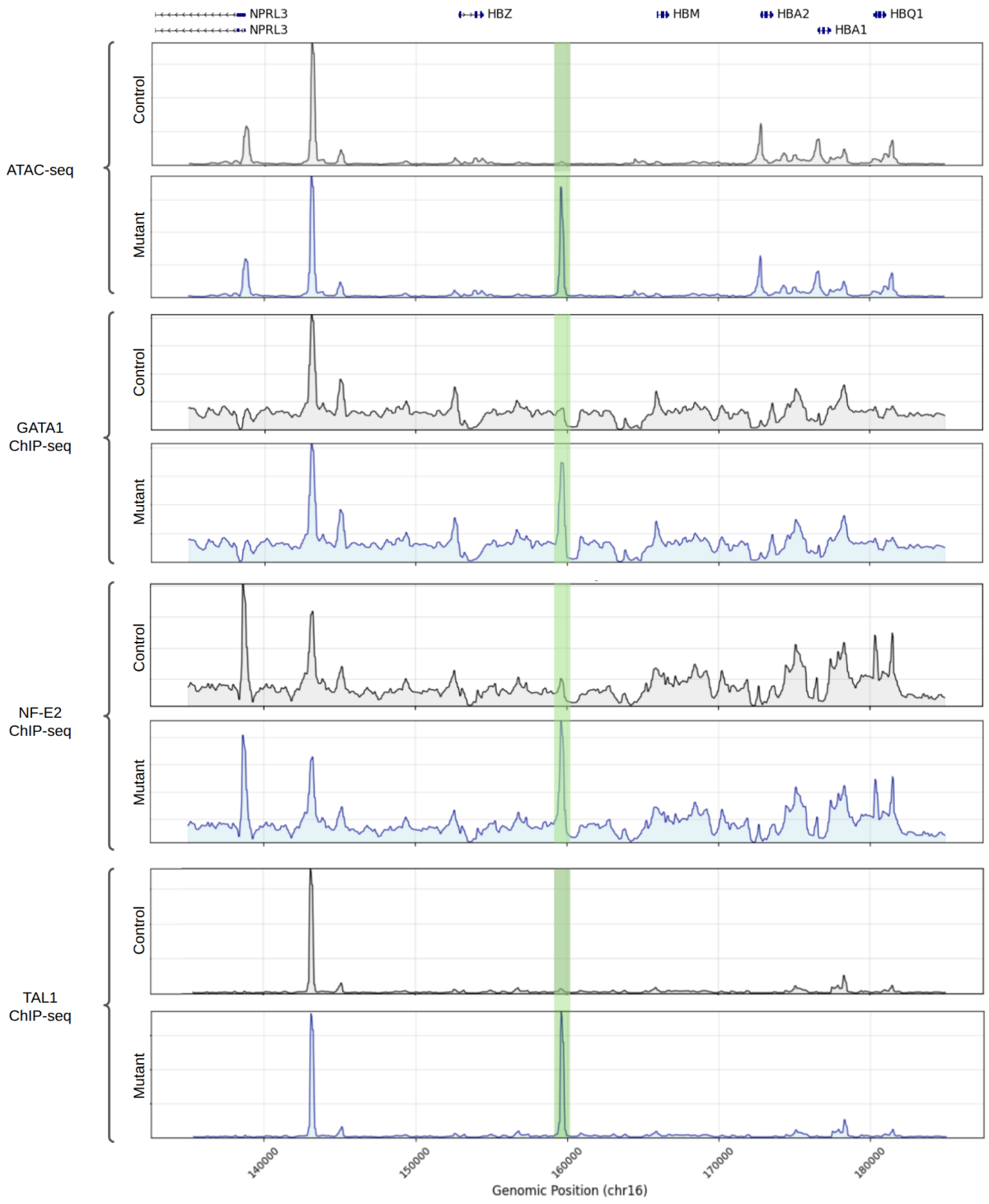


Figure 4.18: Predicted effect of the Vanuatu variant on erythrocyte epigenetic profiles. Tracks show chromatin accessibility and GATA-1, NF-E2, and TAL1 binding. Control and mutation profiles are shown in black and blue, respectively, with the SNP site highlighted in green. Blue tracks indicate variant-induced gain-of-function effects.

PPP1R3B double SNPs

To investigate HYPER's ability to capture more complex variant effects, a locus containing two known SNPs associated with both gain- and loss-of-function chromatin accessibility

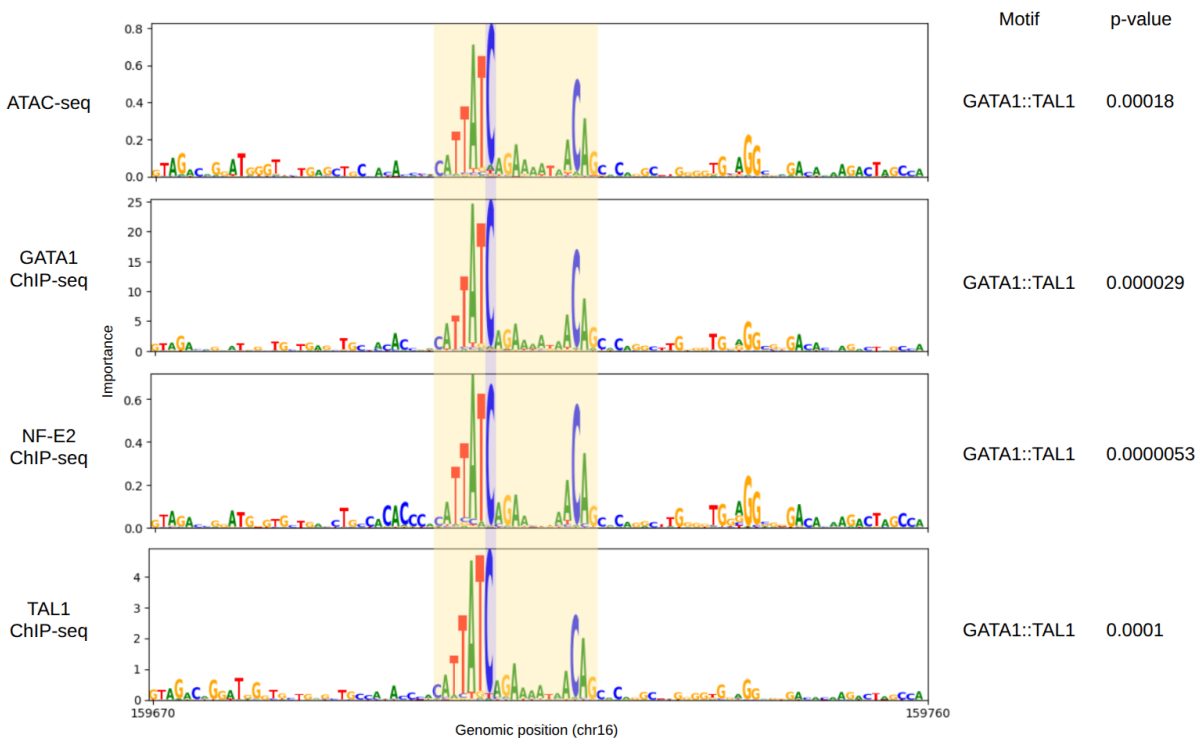


Figure 4.19: Nucleotide importance scores and motif discovery at the Vanuatu SNP functional site. DeepLIFTShap importance scores were computed across four epigenetic profiles for the newly formed regulatory element spanning chr16:159,500–159,830, resulting from the Vanuatu gain-of-function mutation, whose substitution site is highlighted in purple. The most important region of the peak, highlighted in light yellow, corresponds to a GATA1 binding site followed by a TAL1 motif. Both motifs were identified by TomTom with high statistical significance, as indicated by the p-values shown on the right.

changes was analysed. The PPP1R3B locus consists of a promoter and two downstream enhancer elements (Figure 4.20). The first enhancer (peak 1) contains a loss-of-function variant (chr8:9,164,667; A to C; rs330946), while the second enhancer (peak 2) contains a gain-of-function variant (chr8:9,172,650; G to A; rs3748136).

To evaluate the effects of the PPP1R3B regulatory variants, I used a combination of sequence-based prediction and interpretability analyses. Variant impact is first assessed through changes in predicted epigenetic signal, followed by nucleotide importance scoring to identify key regulatory regions, and finally motif enrichment analysis to link these signals to known TF binding motifs.

These variants were previously identified in a genome-wide association study (GWAS) study by Dr Matthew Gosden using the 50-donor dataset [171]. Although they have not been directly associated with disease phenotypes, they have been linked to blood-related traits, including platelet volume [208], reticulocyte count [209], and protein levels [210].

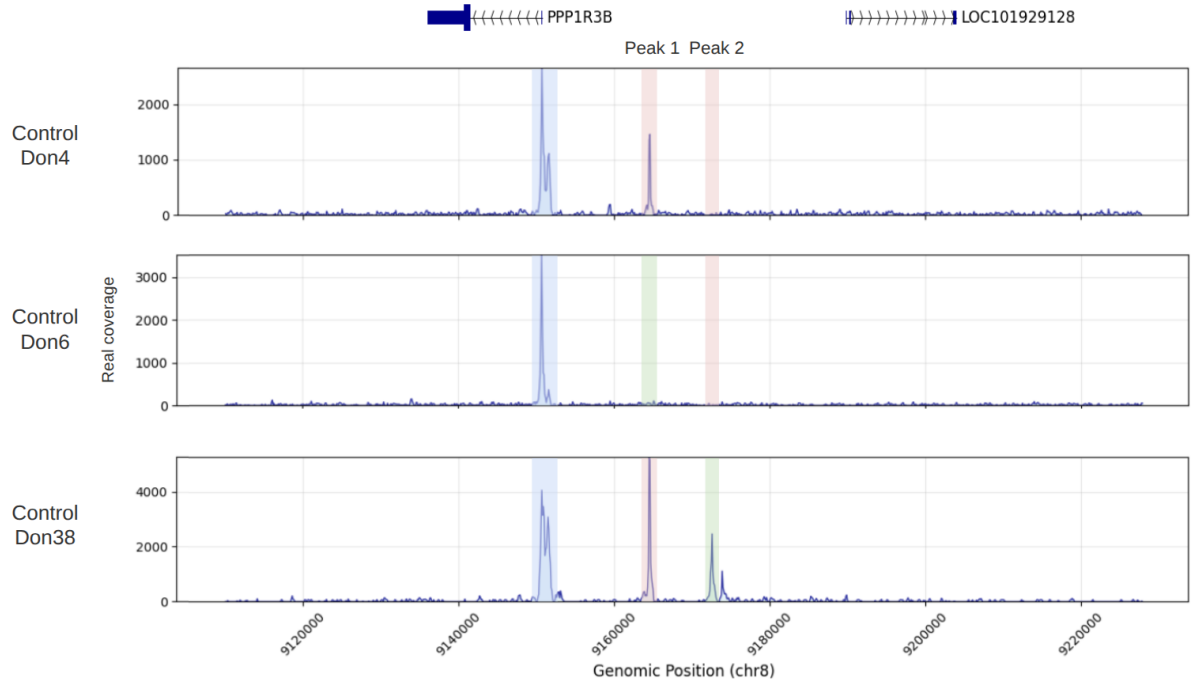


Figure 4.20: PPP1R3B locus showing combinations of SNP presence in chromatin accessibility tracks. In the absence of both SNPs, as observed in donor 4 (first plot), the locus contains a promoter (blue) and a single enhancer. Donor 6 (second plot) carries the loss-of-function SNP in peak 1 (green) with no mutation in peak 2 (red). In contrast, donor 38 (third plot) exhibits the gain-of-function SNP in peak 2 (green) while peak 1 (red) remains unmutated.

Three individuals representing distinct SNPs configurations were selected from the 50-donor dataset: one without mutations, one carrying the peak 1 variant, and one carrying the peak 2 variant. To reproduce these effects computationally, three versions of the PPP1R3B sequence were generated: a reference sequence and two mutated sequences corresponding to each SNP. These sequences were used as input to HYPER.

HYPER predictions indicate that the open chromatin model accurately captures both gain- and loss-of-function accessibility effects by respectively increasing or decreasing peak intensity (Figure 4.21). Specifically, the variant at peak 1 leads to a threefold reduction in predicted coverage, whereas the mutation at peak 2 results in a 50-fold increase in predicted coverage.

In contrast, predictions of TF binding coverage are more variable. NF-E2 is not predicted at peak 1 under either condition, but it appears to be induced by the variant at peak 2. Predictions for TAL1 and GATA1 are somewhat noisier; however, both TFs follow the expected pattern, with binding intensity decreasing in the presence of the peak 1 SNP and increasing with the peak 2 SNP.

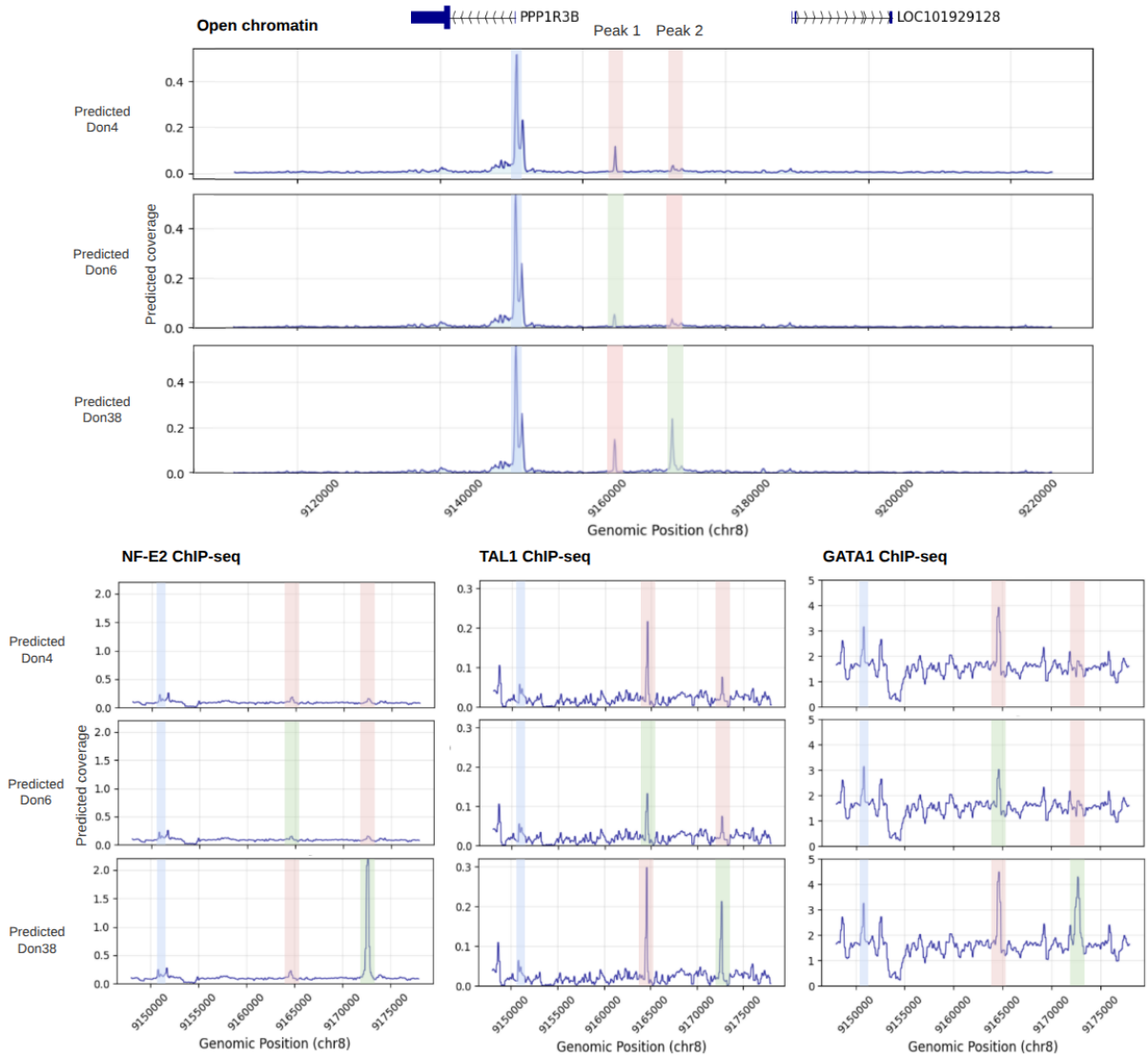


Figure 4.21: PPP1R3B locus with combinations of SNPs in predicted epigenetic profiles. In the absence of both SNPs, as observed in donor 4, the locus contains a promoter (blue) and a single enhancer. Donor 6 carries the loss-of-function SNP in peak 1 (green) with no mutation in peak 2 (red). In contrast, donor 38 exhibits the gain-of-function SNP in peak 2 (green) while peak 1 (red) remains unmutated. The main plot shows the open chromatin predictions, while the smaller plots below show the TF protein-binding predictions.

Overall, the predictions appear consistent and biologically plausible, as the SNP at peak 1 produces a weaker effect compared to the SNP at peak 2, an observation that is also supported by the experimental chromatin accessibility data. While the open chromatin predictions are highly accurate, the lack of control tracks for the ChIP-seq experiments limits direct comparison with TF binding predictions. A motif enrichment analysis based on nucleotide importance could further clarify whether these motifs are predicted to bind at peak 1 or peak 2.

Nucleotide importance scores, computed with DeepLIFTShap and derived from the control

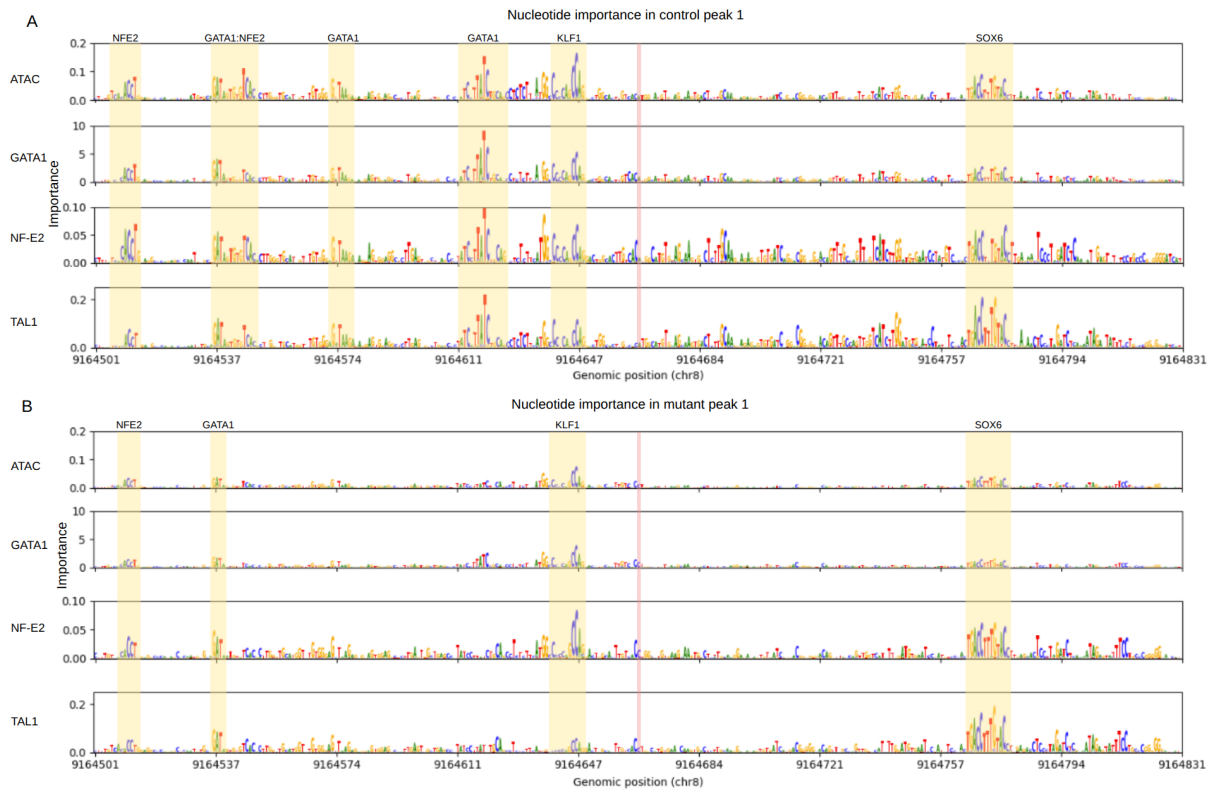


Figure 4.22: Nucleotide importance in peak 1 (loss-of-function SNP). A) Predicted nucleotide importance for the control version of peak 1, showing a clear signal peak across the four analysed tracks. B) Predicted nucleotide importance for the mutated version of peak 1, showing a markedly reduced (nearly absent) signal peak across the four analysed tracks. The mutation site is highlighted in light red. Yellow highlights indicate known motifs identified using TomTom.

(Figure 4.22A) and mutated (Figure 4.22B) tracks for peak 1, show an overall reduction in importance in the mutated condition. Approximately, ATAC-seq decreases from a maximum score of 0.2 to 0.07, GATA1 from 7 to 3, and TAL1 from 0.2 to 0.15. NF-E2 maintains a similar score range; however, it exhibits a clear loss of importance at the central motif (chr8:9164611–9164635), a trend that is also observed across the other three tracks. These findings suggest that the central motif is crucial for peak prediction and loses importance when the binding activity in the region decreases.

For peak 2, the gain-of-function mutation leads to a substantial increase in nucleotide importance scores from the control to the mutant tracks (Figure 4.23). This is particularly evident for GATA1 and NF-E2, where scores increase from approximately 1.1 and 0.2 to 12 and 3, respectively. In the control condition, importance is distributed across multiple motifs, typically two to four, depending on the track. Following the mutation, the signal becomes concentrated within a single dominant motif. For ATAC-seq, NF-E2, and TAL1, this dominant signal lo-

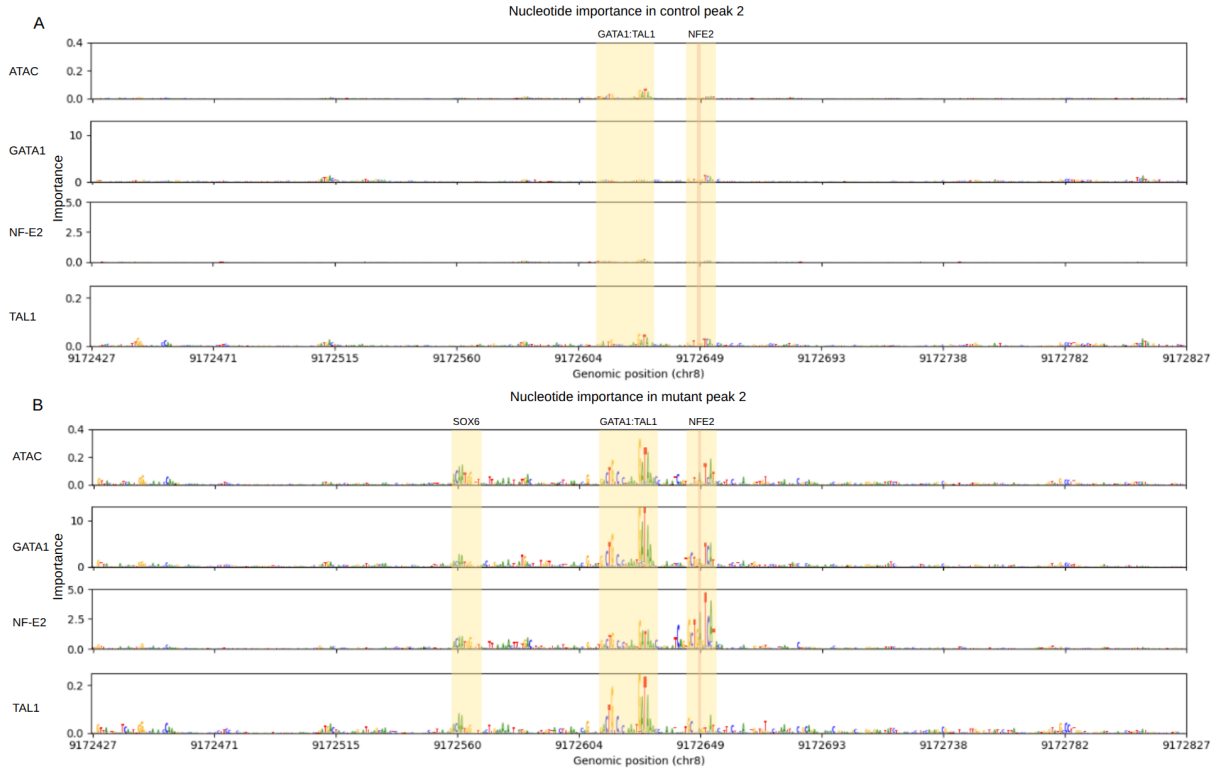


Figure 4.23: Nucleotide importance in peak 2 (gain-of-function SNP). A) Predicted nucleotide importance for the control version of peak 2, showing a distinct signal peak composed of multiple motifs in the central region. B) Predicted nucleotide importance for the mutated version of peak 2, showing a pronounced increase in signal across the four analysed tracks. The mutation site is highlighted in light red. Yellow highlights indicate known motifs identified using TomTom.

calises to chr8:9172614–9172635, while for the GATA1 track, the most influential motif is slightly shifted to chr8:9172639–9172654.

These results suggest that the mutation reorganises the regulatory signal, concentrating activity into a single dominant region rather than a distributed set of weaker motifs. This is consistent with the gain of a stronger binding site driving the increased signal.

Motif enrichment was performed to analyse peaks 1 and 2 and assess whether motif sequences and their associated importance scores can be linked to known TF binding motifs. To achieve this, I applied the pipeline described in Section 2.9.5, which relies on motif annotations from the JASPAR database to identify putative TF binding sites and their contribution scores based on sequence alignment, motif similarity, and enrichment (Figure 4.24). Importantly, this analysis is limited by the coverage and resolution of available motif databases, as similar TFs often share binding preferences and some motifs may be missing, leading to ambiguous or incomplete assignments.

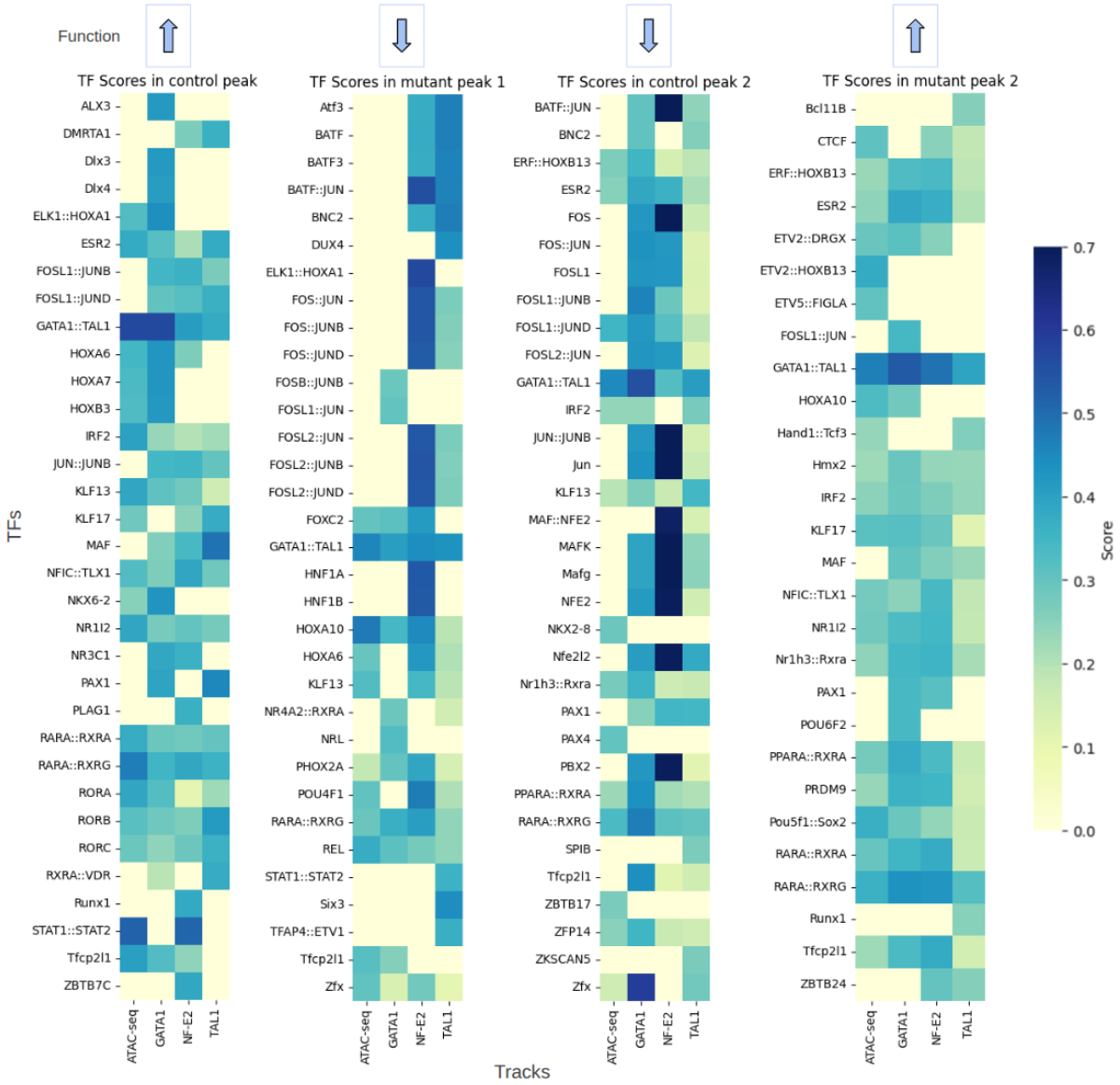


Figure 4.24: Heatmap of motif contribution scores for PPP1R3B SNPs. Each heatmap displays the contribution scores of the most significant matched motifs (rows) across the four epigenetic tracks (columns). The left panel corresponds to the control condition for peak 1, where the signal is strong. The second panel represents the mutant condition for peak 1, where the peak loses function. The third panel shows the control condition for peak 2, characterised by a low signal, while the right panel represents the mutant condition for peak 2, where the signal is markedly increased. All scores are normalised within a range of 0 to 1. The top row, indicated by arrows, represents the functional state of the peak: an upward arrow denotes a high signal peak, while a downward arrow indicates a low or absent signal peak.

For control peak 1, the results indicate that the most influential motifs in the ATAC-seq and GATA1 tracks correspond to the GATA1::TAL1 composite motif, consistent with its role as a key regulator of erythropoiesis. Additional relevant motifs include STAT1::STAT2 and MAF, which also contribute to the NF-E2 and TAL1 tracks.

In the mutant version of peak 1, the contribution of GATA1::TAL1 motifs is reduced, although they remain the dominant motifs in the ATAC-seq and GATA1 tracks. In contrast, motifs from the FOS and BATF families become more prominent in the NF-E2 and TAL1 tracks, respectively.

For control peak 2, the most important motifs across the ATAC-seq, GATA1, and TAL1 tracks again correspond to GATA1::TAL1, reinforcing its central role in driving the regulatory signal. In the NF-E2 track, however, important regions match a broader set of motifs, including NFE2 itself, its known co-binding partners, and members of the JUN family.

Consistent with the nucleotide importance analysis, the regulatory signal of peak 2 becomes highly concentrated within a single dominant region. This region strongly matches the GATA1::TAL1 motif across all four epigenetic tracks, suggesting that the mutation enhances or creates a high-affinity binding site for this complex. Overall, this points to a unifying mechanism in which increased binding of GATA1::TAL1 drives the elevated signal observed in the mutant peak, reinforcing its role as the primary determinant of regulatory activity in this context.

Summary of SNP generalization evaluation

This section aimed to evaluate the ability of the HYPER models for open chromatin and protein binding to generalise to unseen sequences containing SNP-induced mutations, based on known single-nucleotide substitutions.

Two genomic regions were analysed. The first contains a well-characterised SNP on chromosome 16, commonly observed in the Vanuatu population, which leads to a gain-of-function by creating a GATA1 binding site and activating a promoter. The second region includes two SNPs associated with PPP1R3B gene regulation: one corresponding to a loss-of-function and the other to a gain-of-function variant.

Overall, HYPER accurately recapitulated the effects of all three variants. Changes in functional activity, either loss or gain of signal, are clearly reflected in the reconstructed tracks across the four epigenetic profiles. Although the predicted TF binding tracks (e.g., GATA1, NF-E2, and TAL1) exhibit some noise, particularly in the Vanuatu and PPP1R3B analyses, the overall signal trends remain consistent with the expected biological outcomes.

In addition, nucleotide importance predictions successfully captured base-pair–level contri-

butions, revealing how certain motifs lose importance following disruptive mutations, while others retain relatively high contributions even in loss-of-function contexts. This highlights the nuanced regulatory landscape, where multiple motifs can contribute differently to the final signal.

Motif discovery analysis further supports these findings, showing that GATA1-related motifs remain among the most influential across predictions. In addition, motifs not typically associated with erythroid regulation were also identified; however, these are likely the result of motif ambiguity rather than true biological signals. In many cases, these motifs overlap with GATA1 or related binding sites, or share similar sequence logos (e.g., E-box-like motifs), leading to alternative assignments during motif identification. As a result, they appear among the top-ranked motifs despite not representing distinct regulatory drivers.

Although nucleotide importance and motif enrichment analyses provide useful insight into the sequence features associated with predicted regulatory activity, several limitations should be considered when interpreting these results across different genomic or evolutionary contexts. Attribution profiles may vary substantially depending on model initialisation, sequence context, or the specific attribution method employed, leading to instability in nucleotide-level importance estimates. In addition, cross-species comparisons introduce further ambiguity because homologous regulatory regions are often difficult to align precisely, particularly in repetitive or rapidly evolving non-coding regions.

Moreover, conservation of motif enrichment does not necessarily imply strict conservation of regulatory mechanism. Similar regulatory outputs may arise through motif turnover, compensatory evolution, or alternative combinations of TFBSs that preserve overall regulatory function despite substantial sequence divergence. Shared enrichment patterns may also reflect similarities in broader TF binding lexicons rather than direct conservation of individual regulatory elements. Consequently, motif enrichment analyses should be interpreted as evidence of potentially conserved regulatory features rather than definitive proof of mechanistically conserved regulatory grammar.

4.3.7 Binding motif enrichment

The alpha-globin locus is among the most transcriptionally active regions in the human genome, specifically in erythroid cells. Its regulatory landscape, including promoters, enhancers, and the TFs that bind to them, has been extensively characterised. Because of this detailed understanding, it serves as an ideal model system for testing and validating computational methods for motif discovery, allowing us to assess how accurately these approaches can identify known regulatory sequences and predict functional DNA elements.

In this section, following the approach used in the previous motif enrichment analysis based on predicted nucleotide importance, I perform a binding motif enrichment analysis to identify the most characteristic TF binding motifs within the alpha-globin MCS-R enhancer-like elements using HYPER's predicted epigenetic tracks and the pipeline described in Section 2.9.5. The aim is to assess whether the expected TFs are enriched at their corresponding sites and within the appropriate tracks.

R1 is one of the primary enhancers of the alpha-globin locus, where strong enrichment of GATA1 and TAL1 is expected, along with a moderate presence of NF-E2. The results of this analysis confirm that the most significant motifs in R1 consistently match the GATA1::TAL1 (Figure 4.25) composite motif across all four epigenetic profiles. Additionally, the GATA1 binding track reveals motifs that also correspond to other TF families, including VAX1, HOX, and POU factors.

R2 is the most potent enhancer within the alpha-globin cluster, contributing at least 50% of the regulatory activity of the alpha-globin genes [33]. Due to its exceptionally high transcriptional activity, R2 is one of the most densely occupied protein–DNA interaction regions in the genome, making it a key site for studying TF binding. Accordingly, GATA1, TAL1, and NF-E2 are all expected to be enriched in this region. However, the analysis shows that the most significant and enriched motifs predominantly correspond to the GATA1::TAL1 motif (Figure 4.25).

R3 acts to support and enhance the activity of R1 and R2. In this region, strong enrichment of GATA1 and TAL1 is expected, while NF-E2 is mostly absent. The results indicate that GATA1::TAL1 shows a very strong match to the most important motifs in R3, particularly in the ATAC-seq profile (Figure 4.25). Although its contribution is somewhat lower in protein-

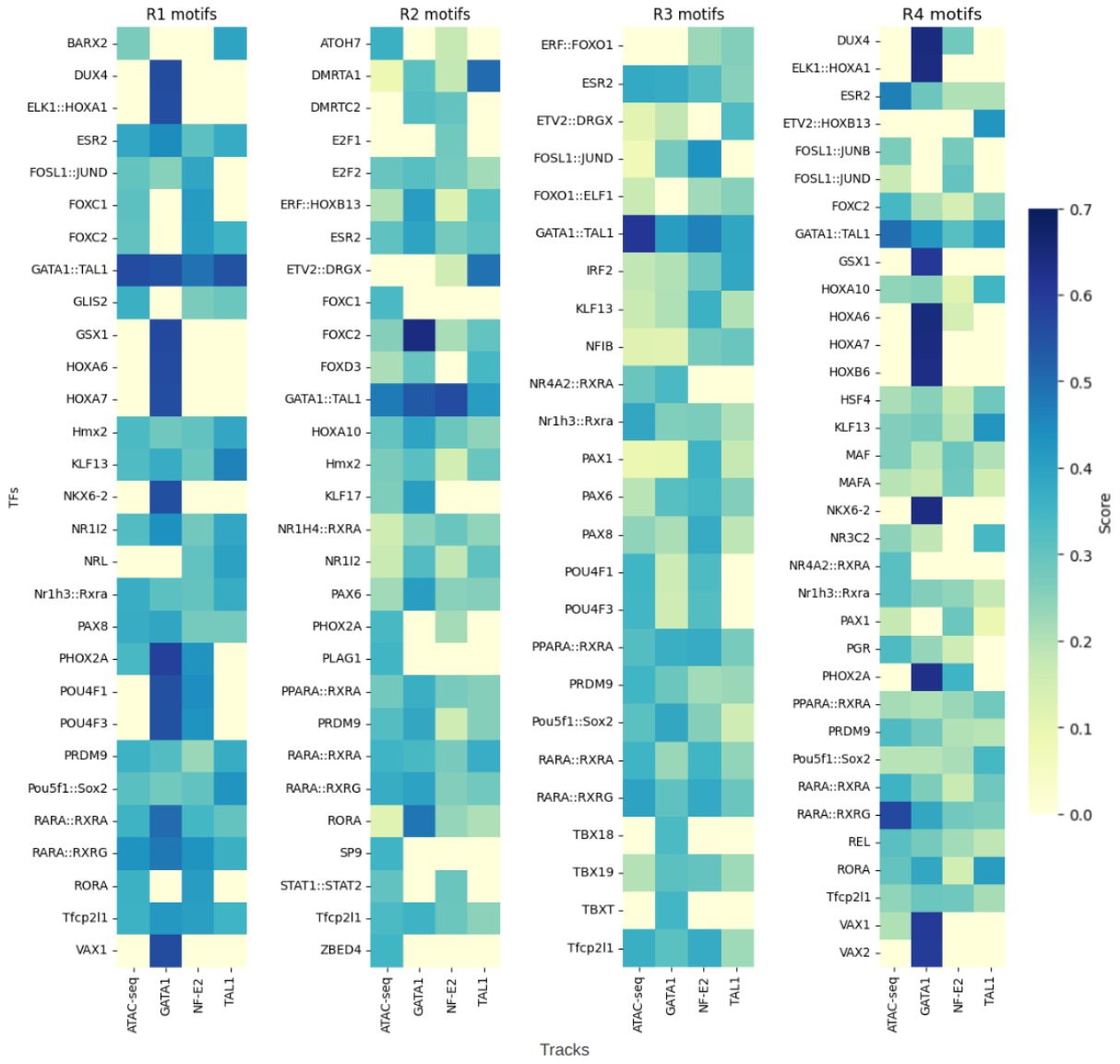


Figure 4.25: Heatmap of motif contribution scores for alpha-globin MCS-R elements. Each heatmap displays the contribution scores of the most significant matched motifs (rows) across the four epigenetic tracks (columns). From left to right, the heatmaps correspond to the R1, R2, R3, and R4 regions. All scores are normalised within a range of 0 to 1.

binding tracks, it still ranks among the highest-scoring motifs. Few other motifs reach a comparable level of contribution in the ATAC-seq data. Nonetheless, additional matches are observed, including POU5F1::SOX2 in the GATA1 track, FOSL1::JUND in the NF-E2 track, and IRF2 in the TAL1 track.

R4 is the final enhancer element and is located closest to the alpha-globin gene cluster. Similar to R3, it is expected to show strong enrichment of GATA1 and TAL1, with little to no NF-E2 presence. The results again identify GATA1::TAL1 as one of the top matches across all four epigenetic profiles (Figure 4.25). However, in this region, other TFs appear to play a more dom-

inant role. The open chromatin profile shows the strongest contributions from RARA::RXRG, while the GATA1 binding track highlights motifs associated with HOX and VAX family TFs. Additionally, NF-E2 shows strong contributions for PHOX2A, and the TAL1 track highlights KLF13 and RORA.

Overall, the methodology described in Section 2.9.5, combining HYPER for epigenetic profile prediction, DeepLIFTShap for extracting nucleotide-level importance, and FIMO with TomTom for motif discovery, performs well for two of the three most expected TFs, namely GATA1 and TAL1. These are consistently identified among the most significant motifs across all four enhancer-like elements of the alpha-globin locus, aligning with expectations for erythroid cells. However, NF-E2 was anticipated to show at least some enrichment in R1 and R2, yet it was not clearly detected even within its own protein-binding track.

Several factors could explain this discrepancy. One possibility is the limited representation of NF-E2 motif variants in public databases such as JASPAR; a broader range of motif "flavours" could improve detection by tools like FIMO and TomTom. Additionally, similarities between the NF-E2 motif and those of other TFs may lead to misclassification. In particular, enriched motifs from families such as AP-1, HOX, or nuclear receptors may reflect alternative assignments rather than true regulatory drivers, as they share sequence properties with canonical erythroid motifs, such as AP1-like binding for NF-E2, AT-richness for GATA1, or E-box similarity for TAL1. Overlapping binding sites and dense regulatory architecture can further contribute to this ambiguity. Another explanation is that the model may struggle to capture NF-E2-specific signals due to nucleotide-level biases or noise in the data. Despite these limitations, the overall approach still demonstrates good performance.

4.4 Identification of mutations modulating enhancer activity

The evaluation and validation of HYPER's predictive performance across multiple levels, particularly through the analysis of variant effects on four SNPs associated with either loss- or gain-of-function at CTCF, GATA1, TAL1, and NF-E2 binding sites, as well as regions of open chromatin previously identified in GWAS studies, provides a strong foundation for further exploration of the model's capabilities. Building on this, it becomes possible to investigate HYPER's potential to identify novel SNPs that may drive functional changes but have not yet been

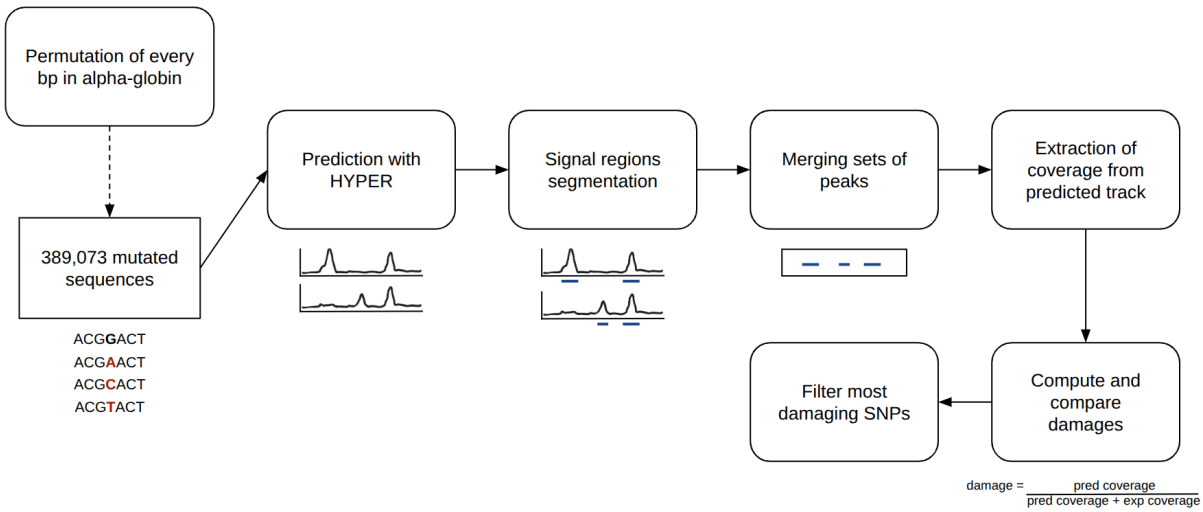


Figure 4.26: Workflow illustrating the identification of the most damaging SNPs. All base pairs within the alpha-globin locus are systematically permuted, and signal regions are subsequently segmented from the corresponding HYPER predictions. These regions are then merged and used to extract coverage from each predicted track. Finally, the impact of each variant on signal intensity is calculated, compared, and filtered to identify the most damaging SNPs.

reported in GWAS, or may not even have been observed in human populations to date.

Therefore, this section moves on to a more exploratory analysis aimed at uncovering potentially novel biological insights on variant effects from an in silico perspective. This section focuses on identifying the most impactful gain- and loss-of-function single-base-pair substitutions across the alpha-globin locus, specifically, variants that lead to pronounced increases or decreases in chromatin accessibility. The steps followed in this section are summarised in Figure 4.26.

To perform this analysis, each base pair within chr16:40959–170649 was systematically mutated to the three alternative nucleotides, resulting in a total of $389,073$ permutations. The mutated sequences were then used as input to HYPER to predict chromatin accessibility profiles in erythrocytes.

Once all predictions were generated, high-signal regions were identified from the predicted tracks. For this step, LanceOtron was used as a practical and efficient approach to segment signals across the $389,073$ HYPER-predicted short tracks. A brief description of LanceOtron's pipeline is available in Section 2.6.1. Since LanceOtron's peak score reflects experimental signal rather than DL–reconstructed signal, no peak score filtering was applied. Coverage was then extracted from these regions for each mutation-specific prediction.

The predicted coverage within each signal region was compared to that of the unmutated (con-

trol) sequence to quantify the effect of each variant. Let C_{pred} denote the predicted coverage for the mutated sequence and C_{ctrl} the coverage for the control sequence. To normalise these effects, the relative contribution of each condition was computed as

$$\text{Damage} = \begin{cases} \dfrac{C_{\text{pred}}}{C_{\text{pred}} + C_{\text{ctrl}}}, & \text{if } C_{\text{pred}} > C_{\text{ctrl}} \implies \text{Gain of function} \\ -\dfrac{C_{\text{ctrl}}}{C_{\text{pred}} + C_{\text{ctrl}}}, & \text{if } C_{\text{pred}} < C_{\text{ctrl}} \implies \text{Loss of function} \end{cases}$$

These quantities represent the proportion of the total signal attributable to the mutated or control sequence, respectively. Values close to 1 indicate dominance of one condition over the other (e.g., strong gain or loss of signal), while values near 0.5 indicate little or no change between conditions.

All possible nucleotide substitutions were evaluated at each base pair position across the $125,240$ bp region. Of these, $49,575$ positions (approximately 40%) showed some change in accessibility when mutated. However, most of these effects were minimal and closely associated with a neutral effect (~ 0.5 of damage rate), making them imperceptible.

To focus on meaningful changes, a stringent threshold was applied, restricting the analysis to SNPs that produced at least a fourfold change in coverage relative to the control sequence. This corresponds to absolute damage scores greater than approximately 0.8, indicating that the predicted signal is strongly dominated by either the mutated or control sequence. Biologically, this reflects variants that induce substantial changes in regulatory activity, such as the creation or disruption of high-affinity binding sites, leading to marked increases or decreases in chromatin accessibility.

A damage value of 1 indicates that a variant creates a peak de novo, while a value of -1 reflects complete loss of the peak signal. Nevertheless, values of ± 1 do not necessarily correspond to the largest absolute differences in coverage; in practice, the most substantial changes tend to approach 1 without fully reaching it.

4.4.1 Most damaging in situ mutations in the alpha-globin locus

Applying this threshold, only 11 SNPs exhibited the strongest gain-of-function effects, whereas $8,042$ showed the most pronounced loss-of-function changes (Figure 4.27).

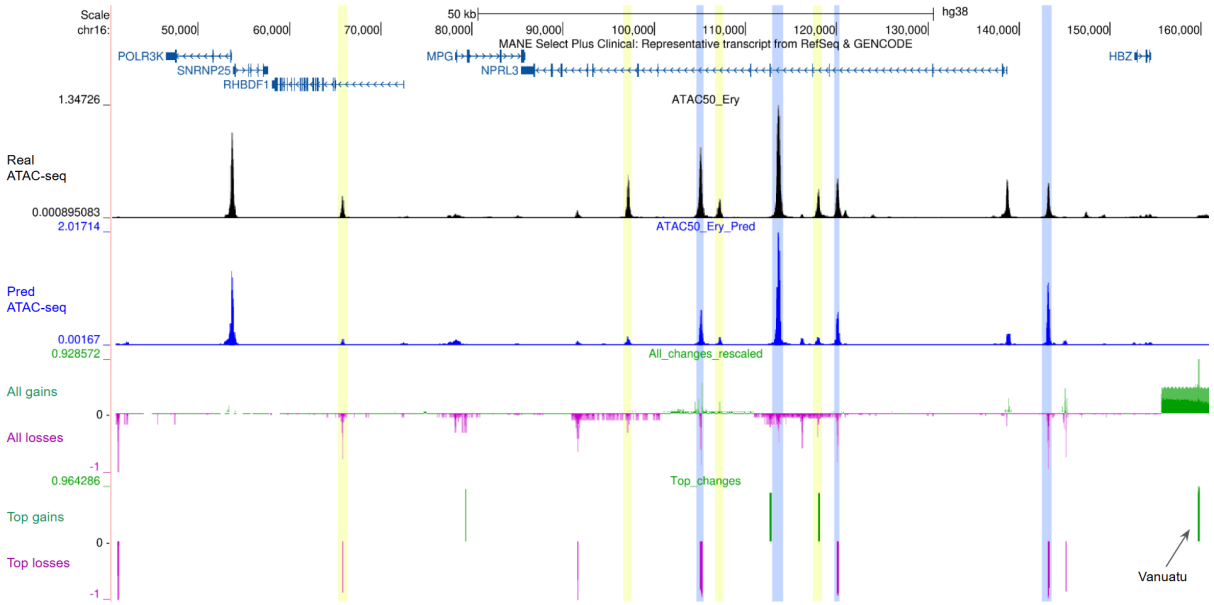


Figure 4.27: UCSC Genome Browser visualisation of alpha-globin variant effects. Observed open chromatin is shown in black, while predicted profiles are in blue. The second-to-last panel displays all variant effects based on their damage scores (rescaled to keep a baseline on zero instead of (+/-) 0.5, which is the neutral variant effect), and the bottom panel highlights the most significant effects, with gains shown in green and losses in pink. The MCS-R1 to R4 regions are marked in blue, and CTCF binding sites are indicated in yellow. For each variant, the x-axis corresponds to the position of the substitution, and the y-axis indicates the damage score.

Among these $8,053$ highly damaging variants, representing just 6.2% of the analysed region, none generated a peak de novo. However, $7,820$ reduced peak regions to zero signal. This does not necessarily imply that major peaks are lost; rather, it is more likely that smaller peaks are weakened or abolished. Consequently, only 222 of the $8,042$ SNPs reduced peak intensity without eliminating it.

Within this context, I quantified how often substitutions led to complete peak loss, substantial increases in coverage, or marked reductions in accessibility (Figure 4.28). Among variants causing enriched regions to disappear, substitutions of guanine to any other nucleotide were the most frequent. The greatest gains in accessibility were primarily associated with cytosine-to-adenine substitutions, as well as guanine replaced by cytosine or adenine. In contrast, reductions in signal (without losing the peak) were most commonly driven by adenine-to-cytosine substitutions.

Most of the 11 SNPs showing the largest increases in accessibility are located in regions not previously associated with regulation of the alpha-globin locus and have not been linked to

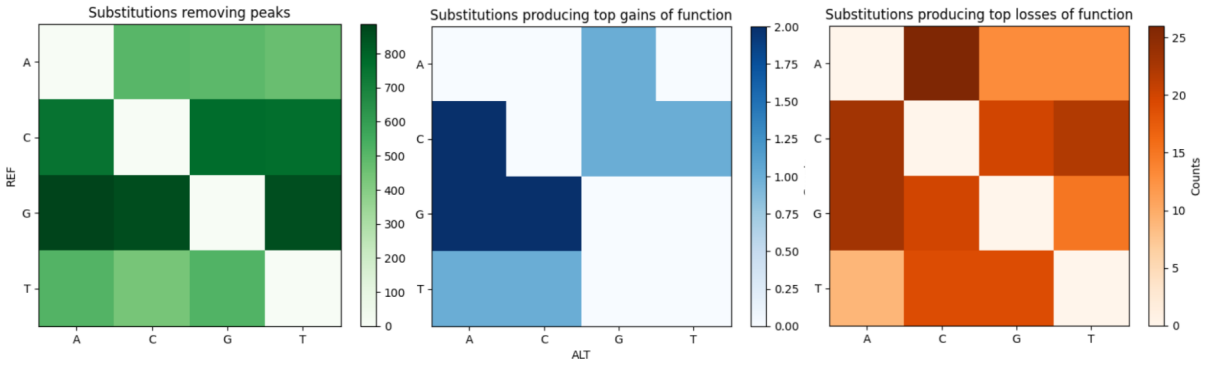


Figure 4.28: Heatmaps illustrating the most damaging variant substitutions. Substitutions that cause complete loss of a peak, likely those with initially low signal, are shown in green. Those leading to substantial increases in signal are represented in blue, while substitutions that reduce accessibility without fully eliminating the peak are displayed in orange.

ID	Position	Reference	Alternative	Previous coverage	New Coverage	Peak start	Peak end	Element	GWAS
G1	41,181	G	C	0.03	0.13	41,127	41,235	Unknown	Unknown
G2	41,191	C	A	0.03	0.16	41,127	41,235	Unknown	Unknown
G3	79,276	C	G	0.08	0.1	78,994	79,491	Unknown	rs754277305
G4	91,595	T	A	0.05	0.23	91,368	91,819	Unknown	Unknown
G5	112,725	C	A	0.03	0.16	112,653	112,886	R2	Unknown
G6	118,051	G	A	0.07	0.36	117,711	118,369	CTCF3	rs1900123061
G7	118,086	G	A	0.07	0.38	117,711	118,369	CTCF3	Unknown
G8	159,662	G	C	0.01	0.05	159,418	159,956	Unknown	Unknown
G9	159,675	C	T	0.01	0.05	159,418	159,956	Unknown	Unknown
G10	159,709	T	C	0.01	0.27	159,418	159,956	New promoter	rs2142009890 (Vanuatu)
G11	159,714	A	G	0.01	0.07	159,418	159,956	Unknown	Unknown

Table 4.3: Summary of the 11 SNPs with the highest damage scores. All variants are located on chromosome 16. The reference allele indicates the original nucleotide, while the alternative allele represents the substituted nucleotide. Each variant exhibits a minimum fourfold change in damage. Notably, most of these SNPs occur in regions with no previous elements or mechanisms that have been associated with the alpha-globin regulation.

known GWAS phenotypes or blood traits, except for the Vanuatu mutation (rs2142009890), as summarised in Table 4.3. This motivated a closer examination of the underlying nucleotide-level changes. To investigate this, I used DeepLIFT to extract nucleotide importance scores for these variants and compared them with their corresponding control sequences, intending to identify sequence features that could explain the observed gains in accessibility.

Nucleotide importance profiles for the control and mutated sequences reveal the emergence of clear regulatory motifs from regions that previously showed low signal (Figure 4.29). In nearly all cases, the mutations appear to create or strengthen binding sites for GATA1 or the GATA1–TAL1 complex, suggesting a consistent mechanism underlying the observed increases in accessibility.

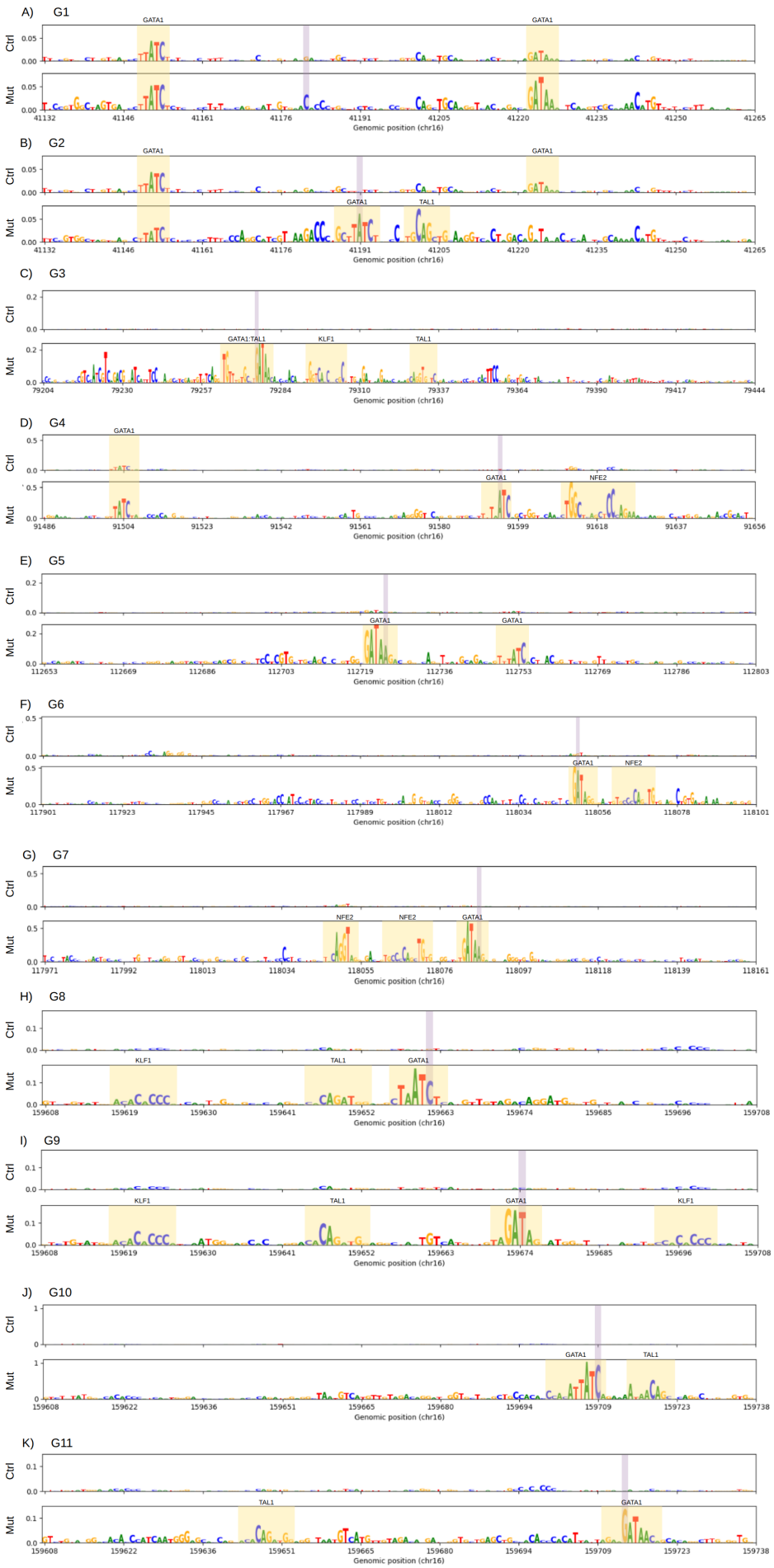

A) G1
B) G2
C) G3
D) G4
E) G5
F) G6
G) G7
H) G8
I) G9
J) G10
K) G11
Ctrl
Mut
GATA1
TAL1
KLF1
NFE2
GATA1:TAL1
Genomic position (chr16)

Figure 4.29: Nucleotide importance profiles for the top 11 highest gain-of-function variants. The substitution site is highlighted in red in both plots, showing nucleotide importance for the control sequence (top) and the mutated sequence (bottom). A) Variant G1 at position $41,181$ (G to C). B) Variant G2 at position $41,191$ (C to A). C) Variant G3 at position $79,276$ (C to G). D) Variant G4 at position $91,595$ (T to A). E) Variant G5 at position $112,725$ (C to A). F) Variant G6 at position $118,051$ (G to A). G) Variant G7 at position $118,086$ (G to A). H) Variant G8 at position $159,662$ (G to C). I) Variant G9 at position $159,675$ (C to T). J) Variant G10 at position $159,709$ (T to C). K) Variant G11 at position $159,714$ (A to G). Yellow highlights indicate known motifs identified using TomTom.

Interestingly, and in contrast to other high-ranking SNPs, variant G1 (Figure 4.29A) involves a substitution located approximately 40 bp away from the GATA1 functional site. This raises the possibility that the mutation is not acting directly on the core motif but instead generates or enhances a KLF-like motif that cooperates with GATA1 binding. Such an effect could reflect motif co-binding or spacing-dependent interactions between TFs.

Another notable case is G10 (Figure 4.29J), which corresponds to the Vanuatu mutation discussed in Section 4.3.5, further supporting the relevance of these findings in known regulatory contexts.

More broadly, the results suggest that the impact of these variants is not limited to the immediate motif itself. Changes in accessibility often extend to surrounding nucleotides, sometimes spanning dozens of base pairs. This pattern points toward a potential role for cofactors, cooperative binding, or local chromatin remodelling, where alterations in one site influence the binding landscape of neighbouring regions.

4.4.2 Most damaging in situ mutations in the alpha-globin MCS-R elements

I next examined SNP effects that impact enhancer-like elements within the alpha-globin locus. Aside from G5 (Figure 4.29E), which is associated with a gain of accessibility in the R2 region, most notable variant effects within the MCS-R elements correspond to losses of accessibility. No additional significant effects were detected for R2. In contrast, 23 variants were identified within the R1 enhancer, of which only 7 are linked to GWAS phenotypes. For the R3 and R4 enhancers, 6 and 7 significant variant effects were detected, respectively, with only one variant in each region associated with a known GWAS phenotype. These matches include

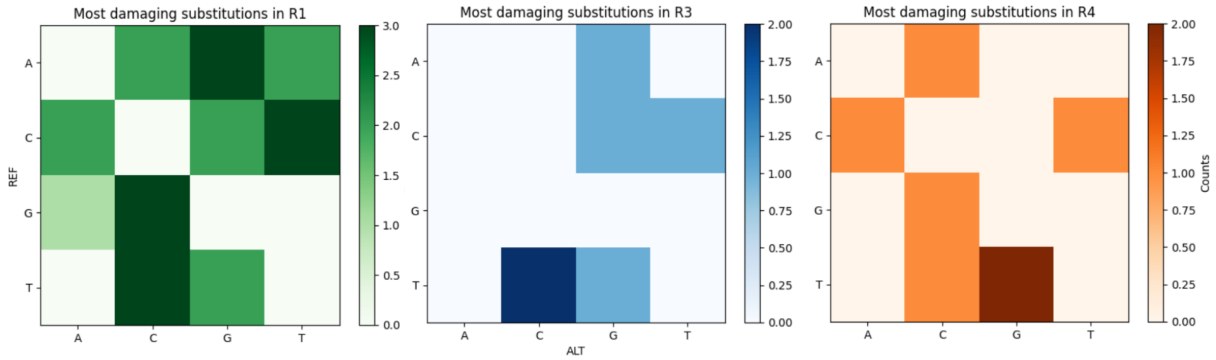


Figure 4.30: Heatmaps displaying the most damaging variant substitutions within MCS-R elements. Substitutions leading to loss of accessibility in R1 are shown in green, while those in R3 and R4 are represented in blue and orange, respectively. No additional variant effects were identified in R2 beyond G5.

rs1381570219, rs187174553, rs1448311315, rs1899473133, rs1596514460, rs372148821, and rs1180626767 in R1; rs983350401 in R3; and rs909089336 in R4.

To further characterise the effects of these variants, I analysed which nucleotide substitutions most frequently resulted in reduced chromatin accessibility (Figure 4.30). Thymine-to-cytosine substitutions were the most common changes associated with substantial accessibility loss in both the R1 and R3 enhancers, whereas thymine-to-guanine substitutions predominated in R4.

To better understand how these variants affect the accessibility profiles of MCS R1, R3, and

ID	Position	Reference	Alternative	Previous coverage	New Coverage	Peak start	Peak end	Element
L1	105,185	A	T	0.17	0.01	104,568	105,529	R1
L2	105,186	T	G	0.17	0.01	104,568	105,529	R1
L3	105,188	A	C	0.17	0.01	104,568	105,529	R1
L4	105,194	C	G	0.17	0.01	104,568	105,529	R1
L5	105,195	A	C	0.17	0.01	104,568	105,529	R1
L6	120,092	T	G	0.23	0.02	119,738	120,461	R3
L7	120,093	A	G	0.23	0.02	119,738	120,461	R3
L8	120,095	C	G	0.23	0.02	119,738	120,461	R3
L9	143,165	T	G	0.38	0.02	142,822	143,555	R4
L10	143,166	T	G	0.38	0.01	142,822	143,555	R4
L11	143,167	A	C	0.38	0.01	142,822	143,555	R4
L12	143,168	T	C	0.38	0.01	142,822	143,555	R4

Table 4.4: Summary of the most damaging loss-of-function SNPs in MCS-R enhancers. All variants are located on chromosome 16. The reference allele indicates the original nucleotide, while the alternative allele represents the substituted nucleotide. Each variant exhibits a minimum damage rate of 0.9. No loss-of-function variants were found for the R2 enhancer.

R4, I excluded those already identified in GWAS and retained only variants with a damage rate greater than 0.9. This filtering yielded five variants for R1, three for R3, and four for R4 (Table 4.4).

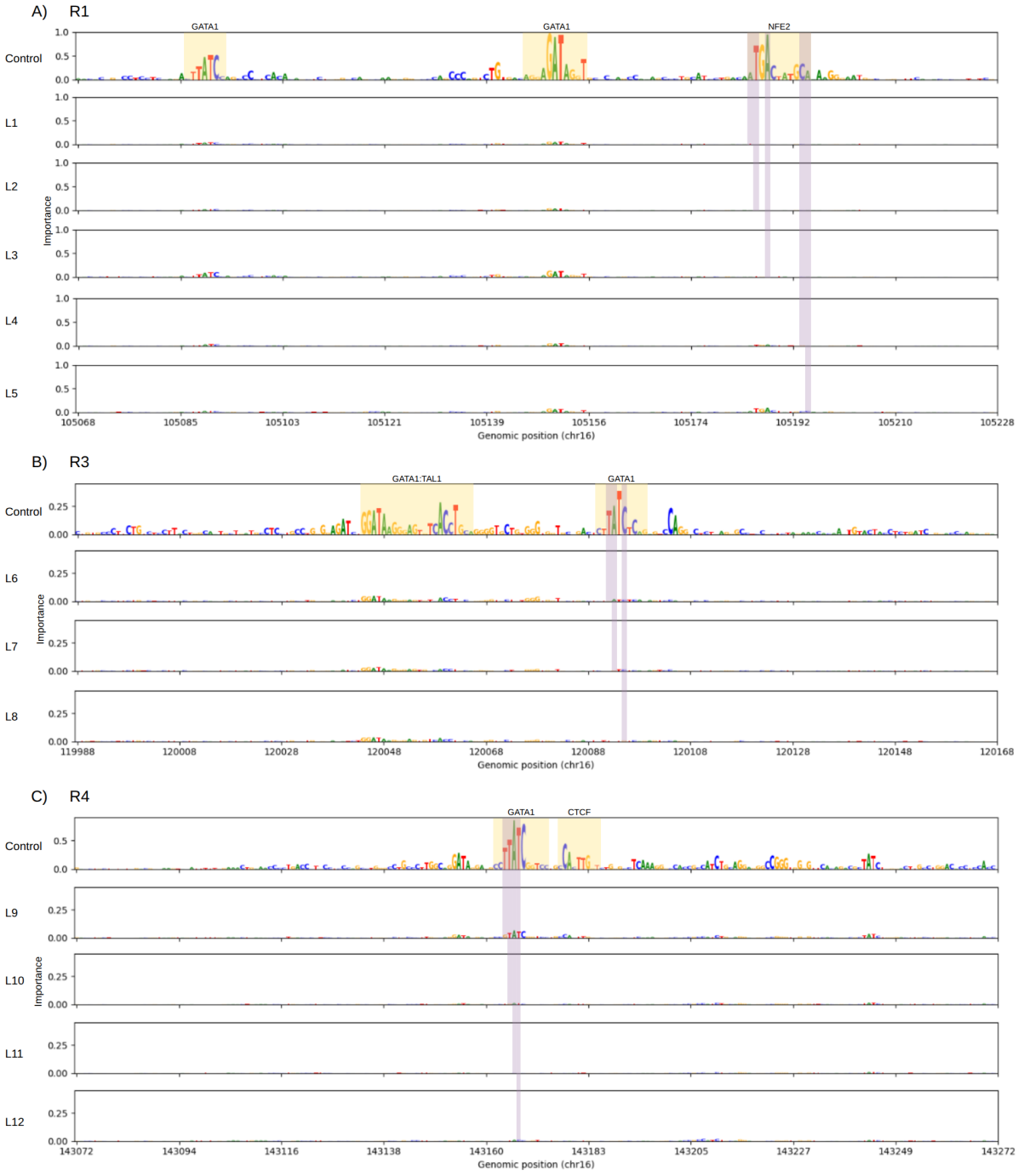


Figure 4.31: Nucleotide importance of the most damaging loss-of-function SNPs in the MCS-R elements. Substitution sites are highlighted in red, with the control sequence at the top and variants ordered by genomic position. A) R1 enhancer variants (L1–L5). B) R3 facilitator variants (L6–L8). C) R4 facilitator variants (L9–L12). Yellow highlights indicate known motifs identified with TomTom.

The differences in predicted coverage between the control and mutated sequences for each variant are often identical within the same enhancer-like element. Moreover, given that both the substitution positions and the nucleotide changes are frequently shared, this suggests that these variants likely disrupt the binding motif of the same TF. To investigate this further, I extracted nucleotide importance scores for each variant and compared them with their corresponding control sequences (Figure 4.31), enabling a more detailed assessment of how these substitutions alter the underlying regulatory signal and potential TF binding sites.

All three enhancer-like regions contain multiple well-defined motifs. Because this analysis focused on the most damaging loss-of-function variants, the excluded variants, with damage rates between 0.8 and 0.9, still showed similarly strong reductions in accessibility. Notably, mutations affecting the most critical base pairs tend to abolish TF binding at the primary site; however, they also disrupt binding at neighbouring motifs, even when those motifs exhibit high individual importance scores. This pattern supports the idea that these TFs bind cooperatively at these regions, as suggested in the previous nucleotide importance analysis.

Although it may initially appear that a single factor predominates in driving binding at these sites, variants that induce strong loss-of-function effects are also found within motifs corresponding to potential cofactors. Disruption at any of these sites is sufficient to markedly reduce or completely abolish binding, reinforcing the notion that coordinated co-binding is required to maintain chromatin accessibility in these enhancer elements.

Summary and conclusions of in-situ mutagenesis analysis

Summarising, this analysis assessed HYPER's ability to predict the functional impact of SNPs across the alpha-globin locus by systematically mutating each base pair and measuring changes in chromatin accessibility. While most mutations had minimal effects, a small subset caused substantial changes, with loss-of-function variants being far more common. Gain-of-function variants were rare and typically associated with the creation or strengthening of TF binding motifs, particularly involving GATA1 and the GATA1–TAL1 complex.

In enhancer-like regions (MCS-R elements), damaging variants clustered at specific positions and shared similar substitutions, indicating disruption of common TF binding motifs. Nucleotide importance analyses showed that mutations affect not only individual sites but also

neighbouring motifs, supporting cooperative TF binding.

Overall, these findings highlight both known and novel regulatory variants and demonstrate that chromatin accessibility is highly sensitive to specific nucleotide changes. Although experimental validation is required, this approach provides a robust framework for prioritising candidate variants and improving our understanding of how genetic variation influences gene regulation and disease.

4.5 Discussion and conclusions

This chapter extends HYPER from predicting well-defined regulatory features, such as CTCF binding, to more complex and cell-type-specific signals, including TF binding and chromatin accessibility in erythroid cells. The model was fine-tuned on high-resolution datasets for key erythroid regulators (GATA1, TAL1, NF-E2) and open chromatin profiles, enabling sequence-based reconstruction of multiple epigenetic modalities.

Model performance was evaluated using both locus-specific visualisation and quantitative comparisons, based on a hybrid log-normal signal extraction framework. This approach enabled consistent identification of enriched regions across predicted and experimental tracks. Overall, the model reproduced key features of the experimental data, although performance varied with signal quality. In particular, TAL1 showed weaker similarity in signal shape due to its irregular distribution, whereas GATA1 more closely matched the reference signal in both magnitude and shape. Chromatin accessibility predictions performed best, with substantially higher peak overlap ($\sim 22,000$ shared peaks) and strong similarity metrics, highlighting the importance of high-quality training data.

Beyond signal-rich regions, the model was also evaluated on repressive genomic contexts, including LADs and pericentromeric regions. HYPER successfully reproduced these low-signal domains, capturing both near-zero and diffuse patterns characteristic of inactive chromatin. Although predictions reflect some experimental noise and bias, these regions remain clearly identifiable, indicating that the model captures both activating and repressive regulatory states. However, in highly transcribed regions, the model tends to underestimate signal intensity, reflecting its reliance on DNA-encoded features and its limited ability to capture effects driven by transcriptional machinery and other non-sequence-dependent processes.

To assess generalisation, variant effect prediction was performed using known mutations, including the Vanuatu variant and PPP1R3B-associated SNPs. HYPER accurately recapitulated gain- and loss-of-function effects, with predicted changes in accessibility and TF binding reflecting the expected direction and magnitude of signal change. Nucleotide importance analysis further revealed base-level contributions, showing how mutations disrupt or enhance regulatory elements.

Motif enrichment analysis linked these predictions to underlying regulatory mechanisms. Using the pipeline described in Section 2.9.5, GATA1::TAL1 emerged as the dominant motif across enhancer elements, consistent with erythroid biology. While additional motifs were detected, these likely reflect motif similarity and database limitations, particularly for NF-E2. Despite these ambiguities, the framework effectively identifies key regulatory drivers.

Finally, an exploratory large-scale *in silico* mutagenesis analysis was performed across the alpha-globin locus. Most mutations had negligible effects, while a small subset produced substantial changes in accessibility, predominantly loss-of-function. These effects were frequently associated with disruption or creation of TF binding motifs, particularly involving GATA1 and its cofactors, and often reflected cooperative interactions within enhancer elements. Notably, many candidate variants fell outside known GWAS loci, highlighting the potential of this approach to identify novel regulatory elements. Although experimental validation is required, this framework provides a systematic strategy for linking sequence variation to regulatory function.

An additional factor complicating the interpretation of regulatory conservation is enhancer turnover, whereby regulatory function can be preserved despite substantial sequence divergence through the gain, loss, or rearrangement of individual TFBS. In this context, equivalent regulatory outputs may arise from different underlying sequence architectures, reflecting distributed redundancy across enhancer elements and compensatory changes in motif composition. Consequently, conservation of regulatory activity does not necessarily require strict conservation of nucleotide sequence or motif organisation. Instead, regulatory systems may retain functional robustness through flexible combinations of binding sites that preserve overall transcriptional behaviour despite evolutionary drift. This evolutionary plasticity may partially explain why predictive transferability can persist even when individual motifs or local sequence contexts differ substantially.

Based on these results, my conclusions can be summarised in the following lines:

- HYPER accurately reconstructs epigenetic signals, with the strongest performance in chromatin accessibility and GATA1, highlighting the importance of high-quality training data.
- The model captures both active and repressive chromatin states, successfully identifying LADs and pericentromeric regions despite noise and signal sparsity.
- HYPER reliably predicts gain- and loss-of-function effects from SNPs, with attribution methods revealing base-level regulatory mechanisms underlying these changes.
- Core erythroid motifs (GATA1::TAL1) are consistently recovered, while additional motifs reflect ambiguity due to overlapping binding sites and motif similarity.
- *In silico* mutagenesis reveals that chromatin accessibility is highly sensitive to nucleotide changes affecting cooperative TF binding.
- This framework enables the identification of novel regulatory variants beyond known GWAS signals.

Together, these results demonstrate that HYPER predicts epigenetic signals with high accuracy while capturing sequence features associated with regulatory activity at nucleotide resolution.

The model's ability to generalise across sequence perturbations and recover biologically meaningful motifs is consistent with the presence of partially transferable regulatory principles encoded within genomic sequence. However, these results do not necessarily imply that the underlying regulatory logic is fully conserved across contexts. Similar predictive behaviour may also arise from shared motif composition, convergent regulatory syntax, representation biases learned during pretraining, similarities in GC-content or sequence composition, or the reuse of common TF lexicons across related regulatory systems. Consequently, the observed transferability should be interpreted as evidence that aspects of regulatory organisation are sufficiently similar to support predictive generalisation, rather than definitive proof of conserved mechanistic regulatory grammar.

This naturally motivates the next step: evaluating how well regulatory grammar can be transferred across evolutionary distances and whether conserved regulatory mechanisms can be identified in divergent genomes.

5 | Transference of regulatory grammar across species

5.1 Introduction

Understanding how gene regulatory mechanisms are conserved and diverge across species is a central challenge in regulatory genomics. While DNA sequence provides the primary substrate for regulatory activity, it remains unclear to what extent sequence-based regulatory logic can be transferred across evolutionary distances, particularly in species where experimental epigenomic data are limited. This raises a fundamental question: can ML models trained on sequence-to-function relationships in one species generalise to predict regulatory activity in another? Addressing this requires disentangling which components of regulatory grammar -understood as the identities, combinations, and arrangement of TFBSs and regulatory elements- are evolutionarily conserved and which are species-specific, and quantifying how predictive performance varies with phylogenetic divergence. To this end, this chapter systematically investigates cross-species generalisation of sequence-based regulatory models, using conserved genomic regions to assess how regulatory activity and underlying sequence features evolve across multiple mammalian species.

The epigenome presents a complex hierarchy of conservation that is central to understanding how regulatory programs are maintained and modified across evolution. This hierarchy comprises five distinct tiers: sequence conservation, functional conservation, structural conservation, positional conservation, and network conservation. Among these, sequence and functional conservation are the most commonly studied in comparative genomics, where sequence conservation refers to the preservation of nucleotide identity, and functional conservation reflects the maintenance of regulatory activity [211]. While sequence conservation is often assumed to imply functional conservation, this relationship is not one-to-one: regulatory activity can be preserved despite substantial sequence divergence [212]. This decoupling presents a fundamental challenge for interpreting genome function across species, as it limits the extent to which sequence similarity alone can be used to infer regulatory roles.

The remaining layers of conservation, structural, positional, and network, capture increasingly abstract aspects of genome organisation. Structural conservation refers to the preservation of higher-order chromatin folding [213]; positional conservation reflects syntenic relationships be-

tween regulatory elements across genomes [214]; and network conservation describes the maintenance of gene regulatory interactions despite changes in underlying sequence features [215]. These forms of conservation are more difficult to measure directly, typically requiring large-scale, high-resolution datasets that remain sparse across species. As a result, our current understanding of regulatory conservation is largely constrained to sequence and functional levels, leaving open questions about how these relate to higher-order regulatory organisations.

This limitation is particularly important in the context of cross-species inference. A central goal in regulatory genomics is to determine whether regulatory logic learned in one species can be transferred to another, especially in cases where experimental data are limited or unavailable. However, the incomplete and uneven availability of epigenomic datasets across species makes direct comparison challenging, restricting our ability to systematically assess conservation and divergence of regulatory mechanisms. Consequently, there is a need for computational approaches that can leverage sequence information to infer regulatory activity across species and provide a scalable framework for comparative analysis [216]

Sequence-to-epigenome models are well-suited to address this challenge. By learning predictive relationships between DNA sequence and regulatory activity, these models offer a means to interrogate functional conservation directly from genomic sequence. In principle, such models can identify shared regulatory features that underpin conserved cell-type-specific programs, while also revealing species-specific differences in regulatory grammar. Moreover, by comparing predicted regulatory activity across species, this enables the quantification of how regulatory function changes with evolutionary distance, providing a bridge between sequence-level variation and functional divergence.

Early efforts to apply ML to cross-species regulatory prediction emerged between 2008 and 2012. These studies employed traditional approaches such as SVMs and random forests to predict epigenetic features across species, typically training on human data and evaluating in mouse or jointly modelling multiple species [217–219] While these approaches demonstrated the feasibility of transferring regulatory annotations, they were limited by reliance on bulk tissue data, restricted species coverage, and dependence on sequence alignment, which biased analyses toward highly conserved regions.

The field has since undergone a substantial shift with the advent of large-scale epigenomic consortia such as ENCODE and Roadmap, and the increasing availability of single-cell datasets

[220–222]. These resources have enabled the application of DL models, particularly CNNs, which can capture complex regulatory patterns directly from sequence. Recent studies have shown that such models can generalise across species to a certain extent, successfully predicting regulatory activity in held-out genomes [155, 223], although most work remains focused on closely related species such as human and mouse. Efforts to extend these approaches to a broader phylogenetic range are ongoing [224, 225].

Despite these advances, several important limitations remain. First, predictive performance tends to decline with increasing phylogenetic distance, suggesting that regulatory grammar is only partially conserved. Second, many studies rely on heterogeneous or bulk datasets, limiting resolution and introducing confounding effects due to mixed cell populations. Third, differences in experimental protocols across datasets introduce batch effects that complicate cross-species comparisons. Finally, relatively few studies incorporate interpretability methods, making it difficult to determine whether models are capturing biologically meaningful regulatory features or exploiting superficial sequence similarities.

Taken together, these challenges highlight a key gap in the current understanding of cross-species epigenetics analyses: it remains unclear which aspects of regulatory logic are conserved across species, how predictive performance degrades with evolutionary distance, and whether sequence-based models can reliably infer regulatory activity in species lacking experimental data. Addressing these questions is essential not only for understanding the evolution of gene regulation but also for enabling the transfer of regulatory annotations to less well-characterised organisms.

Building on this motivation, this chapter aims to systematically investigate the extent to which regulatory patterns learned from sequence generalise across species and cell types. In particular, it focuses on conserved genomic regions as a basis for comparative analysis, enabling controlled assessment of predicted regulatory activity across multiple mammalian species. By explicitly modelling performance as a function of phylogenetic distance, this work seeks to quantify the limits of cross-species generalisation and identify both conserved and species-specific regulatory elements.

To address the limitations of previous studies, emphasis is placed on the use of matched cell types, high-quality datasets, and independent validation strategies. In addition, attribution methods are employed to provide nucleotide-level insight into model predictions, enabling di-

rect comparison of regulatory syntax across species. This allows not only the identification of conserved functional elements, but also the dissection of sequence features underlying regulatory divergence. Through this integrated approach, the chapter aims to provide a more rigorous and interpretable framework for studying the conservation and evolution of gene regulatory mechanisms.

Chapter aims

Building on the deep-level validation performed in the previous two chapters, where the model's performance was rigorously tested by predicting distinct regulatory profiles and generalising across simple single-nucleotide substitutions, achieving high accuracy in both tasks, the next step is to assess the model's proficiency in capturing and transferring the learned regulatory rules across species. This can be done by training a model on regulatory data from one species and then applying it to the genome of another, previously unseen species. In this scenario, the model is forced to base its predictions on general sequence principles that generalise across species, rather than on patterns that are specific to the training organism.

The literature shows that many regulatory elements are remarkably conserved between mouse and human [226], despite roughly 100 million years of evolutionary divergence. This conservation provides a strong foundation for cross-species transfer of regulatory models and motivates the use of the mouse as a training system whose learned regulatory grammar can be decoded onto the human genome. At the same time, it is also compelling to zoom in on human-specific regulatory evolution over a much shorter time scale. Ancient hominins offer a unique opportunity to study regulatory evolution in the absence of experimental data: although humans and ancient homo subspecies diverged only around 400–800 thousand years ago, high-quality genomes have been reconstructed from ancient samples, often from finger bones excavated across Europe and Asia [227, 228]. These genomes allow the reconstruction of putative transcriptional programs purely from DNA sequence, enabling the observation of how regulatory elements have potentially changed in sequence and inferred function during this period of human evolutionary history.

Extending beyond humans and primates, the analysis will also include additional mammal and bird species whose genomes have been sequenced and improved over the last decade. This broader cross-species perspective will help distinguish deeply conserved regulatory prin-

ciples from lineage-specific innovations. For this, the alpha-globin locus will again serve as the primary validation locus, due to its exceptionally high sequence and regulatory conservation across vertebrates and the well-characterised nature of its enhancers, promoters, and TF motifs. All previously described and validated computational pipelines, motif enrichment analyses, nucleotide importance analyses, and model-based CRE-prediction workflows will be reused in this chapter to ensure methodological continuity and facilitate direct comparison across species.

Within this framework, the chapter's aims are:

- To train HYPER models on mouse and chicken erythroid regulatory data, capturing regulatory grammar in both a mammalian and a distantly related vertebrate system.
- To characterise regulatory conservation and divergence across mammals, using model predictions and motif-level analyses to identify shared and lineage-specific regulatory features.
- To analyse evolutionary changes at the alpha-globin locus, comparing nucleotide-level importance and motif composition across human, chimpanzee, ancient hominins, and related species.
- To investigate recent regulatory evolution in the human lineage, identifying sequence and functional changes between human and ancient human genomes using model-based predictions.

This chapter begins by outlining the datasets and model configurations used to train sequence-to-epigenome models on human, mouse, and chicken erythroid-lineage data. It then presents a cross-species reconstruction of the alpha-globin locus across these species, including a detailed analysis of motif enrichment and nucleotide-level importance within the MCS-R regulatory elements. Building on this, the study extends to additional mammalian and avian species, applying the trained models to predict regulatory activity at the alpha-globin locus in genomes lacking matched epigenomic data. This includes comparative analyses across closely related hominins, such as Neanderthal, Denisovan, and early Homo Sapiens genomes. The chapter concludes with a discussion of the results in the context of existing literature, highlighting key insights and methodological limitations.

5.2 Available datasets

To identify and compare a set of cell-type-specific regulatory elements across two organisms, it is necessary to use datasets derived from analogous cell types. Although the systems under study may originate from different species, and therefore are not directly equivalent, they can still be meaningfully compared at a regulatory and epigenomic level, provided they capture the same core features of interest, in this case, erythroid-lineage-defining key transcriptional regulators and chromatin accessibility landscapes.

At the end of this section, multiple high-coverage genome sequences from different species are described for use in the analysis, despite the absence of experimental validation data.

5.2.1 Mouse erythroid-lineage

Five publicly available datasets were used, comprising one chromatin accessibility dataset and four TF binding profiles from erythroid-lineage cells.

Open chromatin track

This dataset (GSM5588920) comprises multi-omics profiling of erythroid cells derived from mouse embryonic stem (mES) cell differentiation of embryoid body (EBs) in an ex vivo system that recapitulates primitive erythropoiesis. It was originally generated to identify and map accessible regulatory elements within the alpha-globin locus and surrounding genomic regions [34], providing a genome-wide catalogue of candidate enhancers and regulatory domains that serve as a foundational resource for the present study.

CTCF protein binding track

The CTCF ChIP-seq dataset (GSE174593) was generated in mouse erythroid cells as part of a study investigating the cis-regulatory landscape and 3D chromatin architecture underlying embryonic globin gene expression across developmental stages, with particular focus on its silencing in adult erythroid cells and potential reactivation. Chromatin immunoprecipitation followed by high-throughput sequencing was used to map genome-wide CTCF binding sites,

enabling the identification of insulator elements and higher-order chromatin domain organisation [229].

GATA1 protein binding track

This ChIP-seq dataset (GSE47492) was generated from Ter119^+ mouse erythroid cells, representing a differentiated erythroid population. It was originally produced to support the characterisation of interactions between promoters and distal regulatory elements at loci such as the alpha-globin region [230], using Capture-C, a high-resolution method for mapping cis-regulatory interactions. The accompanying GATA1 ChIP-seq data provides complementary information on TF binding, facilitating interpretation of the regulatory landscape and chromatin context.

NF-E2 protein binding track

The NF-E2 ChIP-seq dataset, also obtained from GSE47492, was generated from Ter119^+ mouse erythroid cells to profile genome-wide binding of a key erythroid TF using an NF-E2–specific antibody.

TAL1 protein binding track

The TAL1 ChIP-seq dataset (GSE36029) was generated in mouse erythroid progenitor cells (G1E-ER4) as part of the Mouse ENCODE project [231]. ChIP-seq was performed using standard protocols. This dataset provides genome-wide maps of TAL1 occupancy, enabling the identification of erythroid regulatory elements and supporting integrative and comparative analyses of TF binding and gene regulation.

All tracks were preprocessed and normalised using the pipeline described in Section 2.2, and peaks were identified with LanceOtron using peak score thresholds of 0.9 for ATAC-seq, 0.8 for CTCF, GATA1, and NF-E2, and 0.6 for TAL1. The latter lower threshold was applied to account for its lower coverage.

Although these tracks do not achieve the same resolution and depth of coverage as those from

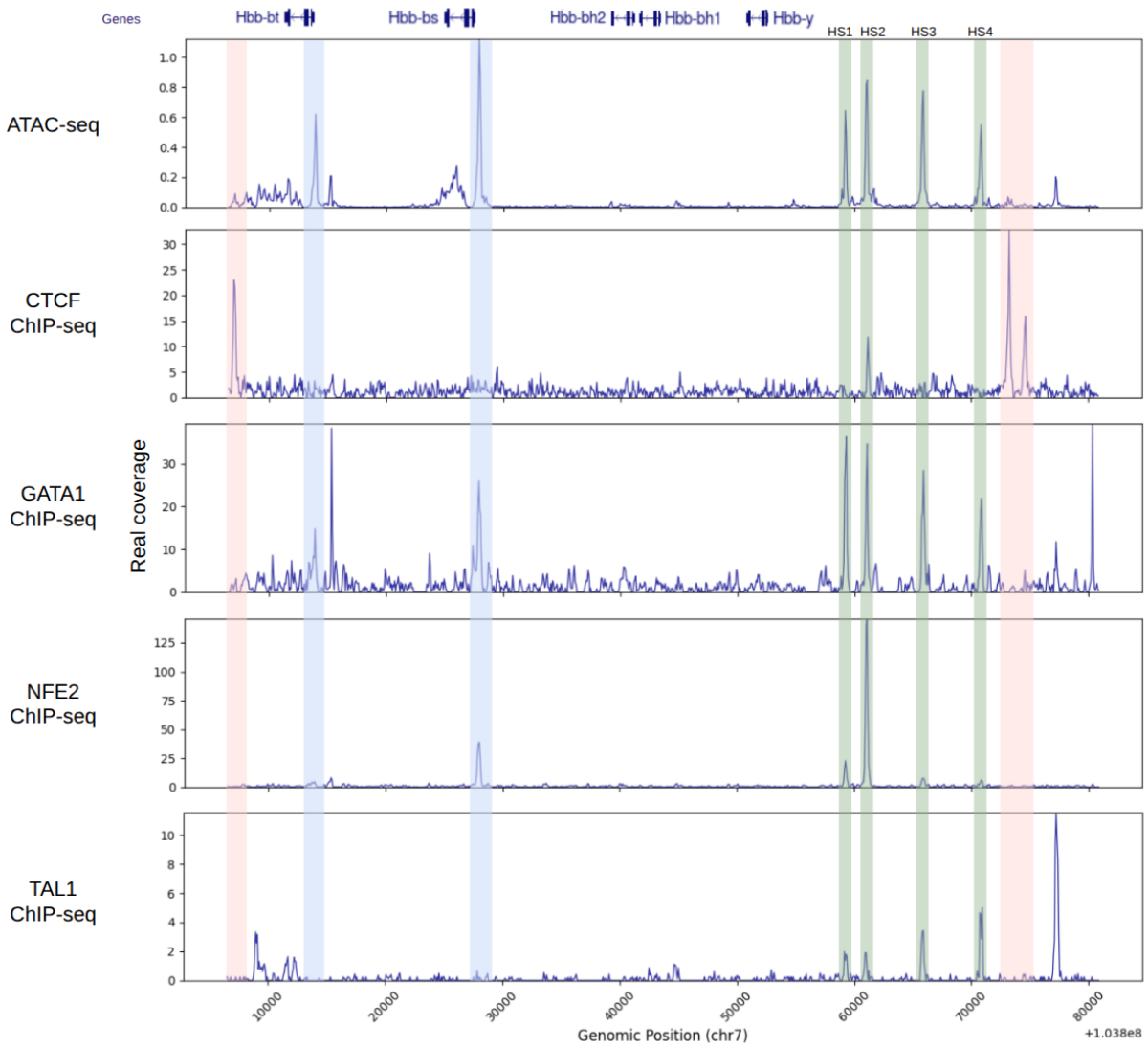


Figure 5.1: Open chromatin and TF binding at the murine β-globin locus. From top to bottom, the normalised tracks represent ATAC-seq, followed by CTCF, GATA1, NF-E2, and TAL1 binding profiles. Enhancer-like elements (HS1–HS4) are highlighted in green, while the promoter region is shown in blue. CTCF binding sites are indicated in light red. Notably, a CTCF-bound region overlapping HS2 coinciding with TFBS is visible but not highlighted in red.

the 50-donor collection, they still capture regions of high chromatin accessibility and protein binding. Figure 5.1 shows the beta-globin locus across the five tracks, where all datasets accurately highlight enhancers, promoters, and protein-binding sites. In this case, ATAC-seq and CTCF ChIP-seq appear somewhat noisier than in the human analysis, with GATA1 also showing increased variability. In contrast, TAL1 and NF-E2 exhibit more consistent and controlled signal profiles.

Regarding the signal around peak centres (Figure 5.2), and consistent with the human epigenetic tracks, the chromatin accessibility profile displays a broad and symmetric pattern. The

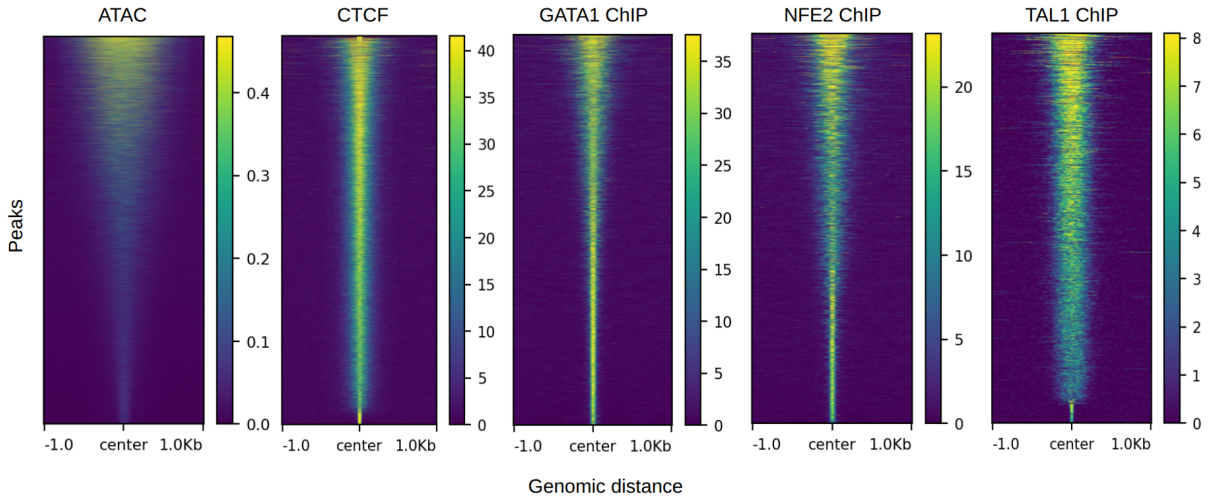


Figure 5.2: Mouse epigenetic profile heatmaps centred on peaks. Heatmaps showing ATAC-seq and ChIP-seq signal intensity for CTCF, GATA1, NFE2, and TAL1 across genomic regions centred on peak summits (± 1 kb). Each row represents an individual peak, ordered by signal intensity, with colour indicating normalised read density (yellow = high, purple = low).

CTCF binding profile is characterised by a very sharp, intense, and symmetric signal, indicating strong and highly localised occupancy across most peaks. Compared to CTCF, GATA1 exhibits slightly broader enrichment, while NF-E2 shows a similarly narrow profile but with lower intensity, suggesting less frequent or weaker binding events. Finally, as observed in the human data, TAL1 displays greater variability across peaks, evident from the striped pattern, which may reflect cooperative or context-dependent binding behaviour.

5.2.2 Chicken erythroid-lineage

Two publicly available datasets were used, comprising one chromatin accessibility dataset and one boundary element binding profile from erythroid-lineage cells.

Open chromatin track

The ATAC-seq dataset is an unpublished in-house dataset generated through a collaboration between the Sauka-Spengler and Higgs groups, derived from RBCs isolated from chick embryos collected before day 14 of development, representing an early embryonic stage. Blood was obtained from the yolk sac, and RBCs were subsequently enriched using a simple centrifugation-based separation protocol. This study aimed to profile chromatin accessibility in the erythroid lineage during chicken development, providing insight into the regulatory landscape underlying

erythropoiesis at early embryonic stages.

CTCF protein binding track

The CTCF ChIP-seq dataset (GSE51846) was generated from the HD3 chicken erythroblast cell line (non-induced state), representing an undifferentiated erythroid model. Chromatin immunoprecipitation was performed using a custom rabbit anti-CTCF antibody. This dataset was originally generated to map genome-wide CTCF binding sites in chicken erythroid and lymphoid cells, to investigate how CTCF contributes to 3D genome organisation, particularly the spatial clustering of CpG islands [232].

All tracks were preprocessed and normalised using the pipeline described in Section 2.2. Peak calling was performed with MACS3 (`bdgpeakcall`), using a signal cutoff of 4 for the ATAC-seq track, with both minimum peak length and maximum gap set to 100 bp. For the CTCF track, a lower signal cutoff of 1 was applied, with the same minimum length and maximum gap parameters (100 bp).

Similar to the mouse epigenetic profiles, the chicken datasets exhibit a noticeable level of noise. From a bioinformatics perspective, this can be partially mitigated by applying more stringent signal thresholds, reducing false positives at the expense of potentially excluding weaker true signals. Importantly, the key regulatory elements relevant to this study are well captured, show-

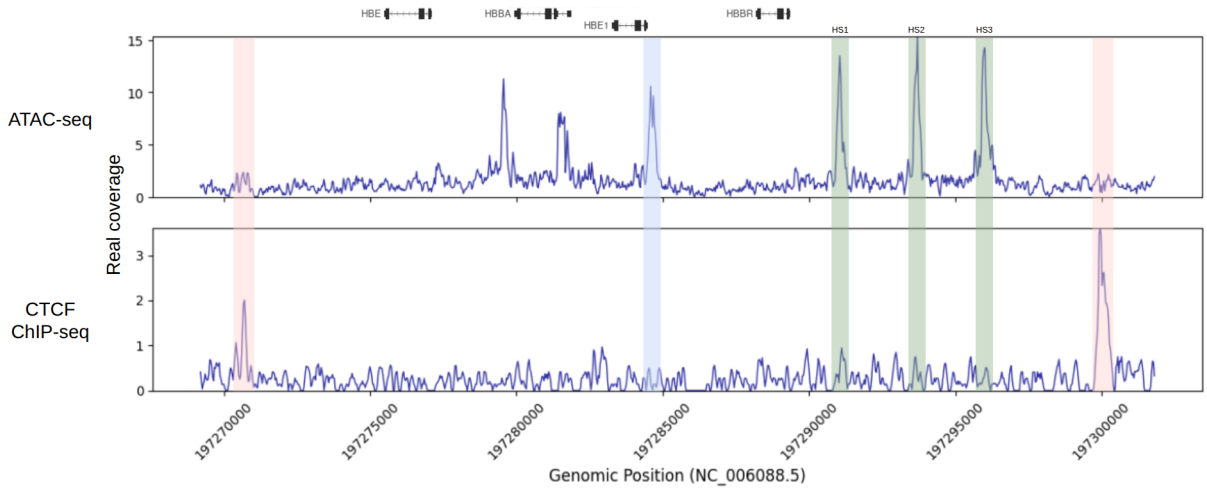


Figure 5.3: Open chromatin and CTCF binding at the chicken beta-globin locus. The normalised ATAC-seq track (at the top) is followed by the CTCF ChIP-seq track (at the bottom). Enhancer-like elements (HS1–HS3) are highlighted in green, while the promoter region is shown in blue. CTCF binding sites are indicated in light red.

ing strong and well-defined peaks that enable reliable identification, as illustrated in Figure 5.3, where the principal enhancer, promoter, and CTCF sites are clearly recapitulated.

With respect to signal enrichment around peak centres (Figure 5.4), the open chromatin track exhibits greater variability, likely reflecting the diversity of regulatory elements captured as well as potential noise in the data. In contrast, the CTCF profile displays sharper and more uniform peaks, indicating more precise localisation and consistent binding. This is expected, as ATAC-seq captures a broad range of accessible regulatory elements, whereas CTCF ChIP-seq targets a single, sequence-specific DNA-binding protein.

The subsequent exploratory analysis, focusing on sequence and functional conservation of regulatory elements in erythroid-lineage cells, involves two groups of genomes for which no experimental validation datasets are available. Nevertheless, this provides a valuable starting point for investigating the potential evolutionary dynamics of erythroid-like epigenetic regulation across millions of years.

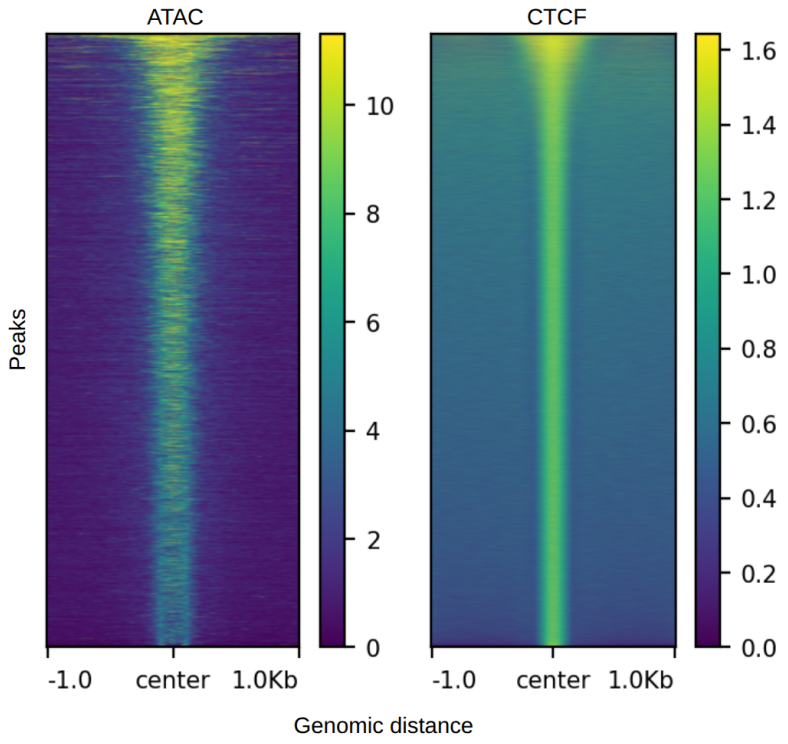


Figure 5.4: Chicken epigenetic profile heatmaps centred on peaks. Heatmaps showing ATAC-seq and ChIP-seq signal intensity for CTCF across genomic regions centred on peak summits (± 1 kb). Each row represents an individual peak, ordered by signal intensity, with colour indicating normalised read density (yellow = high, purple = low).

5.2.3 Ancient hominin genomes

To extend the comparative analysis beyond extant species, ancient genomes provide a unique opportunity to investigate regulatory evolution in closely related lineages. In this context, ancient hominin genomes offer valuable insight into recent human evolution, enabling the exploration of sequence variation within regulatory regions that may underlie differences in gene regulation between archaic and modern humans.

Although the number of deeply sequenced ancient hominin genomes remains limited, there are sufficient high-quality samples to reconstruct representative genomes and capture genetic variation within the lineage. Neanderthal remains are continuously being discovered, primarily across Europe and Asia, providing new material for genomic analysis. Importantly, Neanderthal populations are known to have been more genetically subdivided than present-day human populations [233], reflecting small, isolated groups with limited gene flow. This increased population structure introduces additional variability between individuals and must be considered in comparative analyses, as observed differences may reflect population-level divergence rather than lineage-specific regulatory changes.

Denisovan genomes, although fewer in number, provide complementary insight into archaic hominin diversity, representing a deeply divergent lineage with evidence of admixture with both Neanderthals and modern humans [234]. In parallel, early Homo sapiens genomes capture genetic variation from populations predating or coinciding with the initial dispersals of modern humans across Eurasia, offering a temporal bridge between archaic hominins and present-day populations. These genomes are particularly valuable for distinguishing ancestral from derived regulatory variants and for identifying changes specific to modern human lineages.

These ancient genomic data are obtained through a multi-step process combining archaeology, molecular biology, and computational analysis. Fossil remains, typically bones or teeth, are first identified and processed. DNA is then extracted, often from dense tissues such as the petrous bone, yielding highly fragmented and chemically damaged ancient DNA. Specialised library preparation methods are used to accommodate these characteristics, followed by high-throughput sequencing.

Computational pipelines are subsequently applied to process the data, including read alignment to a reference genome, authentication of ancient DNA based on characteristic damage patterns,

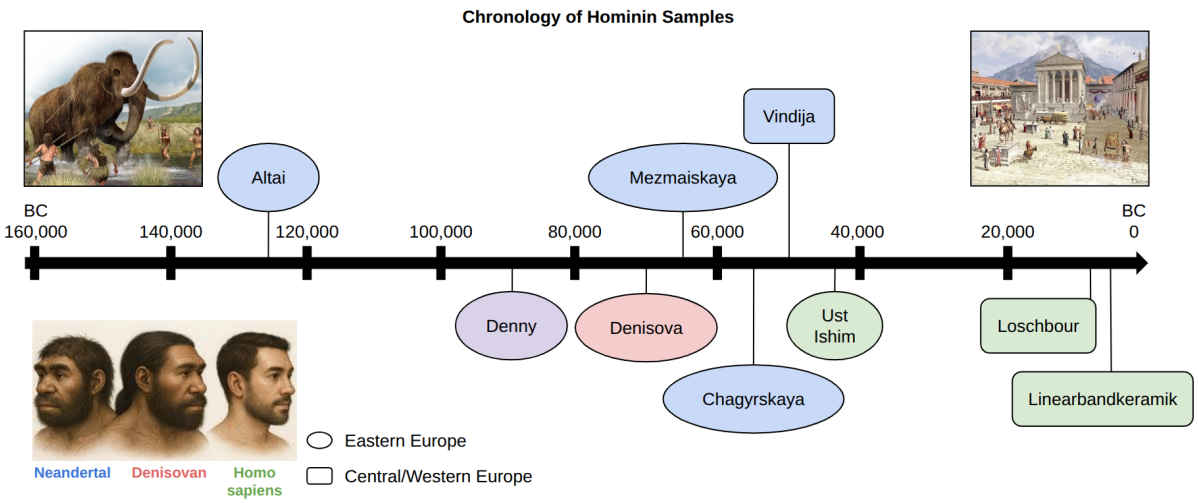


Figure 5.5: Timeline illustrating the estimated ages of the remains of the nine ancient hominins. Blue markers represent Neanderthals, red denote Denisovans, purple indicate Neanderthal–Denisovan hybrids, and green correspond to modern Homo sapiens. Marker shapes indicate geographic origin, with circles representing Eastern Europe and rectangles representing Central/Western Europe.

removal of contaminants, and genotype inference. The resulting reconstructed genomes enable downstream analyses of sequence variation. In this study, genetic variants from nine Neanderthal, Denisovan, and early Homo sapiens individuals were collected and processed. Variants were filtered based on quality, lifted over to the `hg38` reference genome, and incorporated using the `BCFtools` consensus pipeline to generate approximate whole-genome sequences for each ancient hominin.

Figure 5.5 depicts the temporal distribution, lineage classification, and geographic origin of the nine ancient hominin individuals included in this analysis. Supplementary Table S1 describes each sample.

5.2.4 Mammal and avian genomes

Building on the principle that evolutionary conservation can highlight functionally important cis-regulatory elements, this study utilises recently assembled genomes from 23 mammalian and avian species spanning a wide range of evolutionary divergence. These genomes provide a comparative context for investigating erythroid-lineage epigenetic activity, particularly at the globin loci. By including species with differing levels of relatedness, this dataset encompasses both conserved and lineage-specific genomic features, providing a valuable resource for ex-

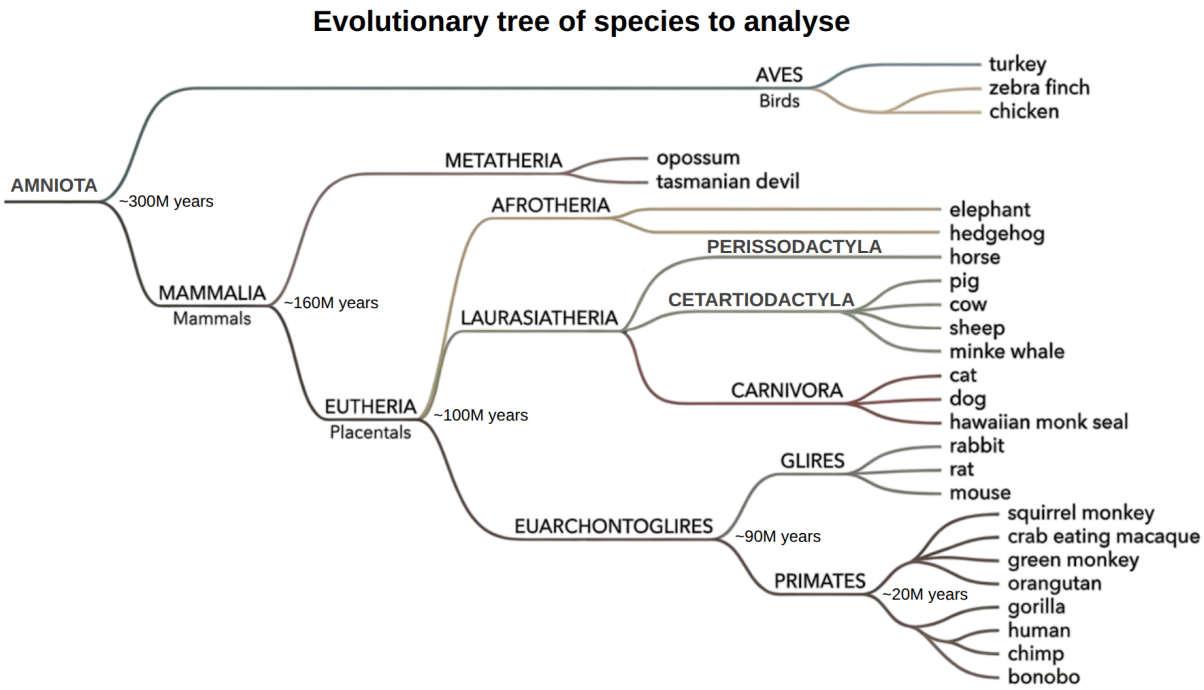


Figure 5.6: Evolutionary relationships between the 26 species. This simplified phylogenetic tree illustrates the evolutionary relationships among the species analysed, including primates, ungulates and cetaceans within the Cetartiodactyla clade, and other mammals and birds. The diagram reflects major eutherian mammal superorders and highlights divergence times in millions of years. The most recent split shown is between Hominoidea (e.g. human, chimpanzee) and other primates (monkeys), occurring approximately 20 million years ago, whereas the most distant divergence depicted is between primates and birds, at roughly 300 million years ago.

ploring regulatory signals associated with these regions.

Of the 23 species analysed, excluding human, mouse, and chicken, 9 belong to Euarchontoglires (7 primates and 2 glires), 8 to Laurasiatheria (3 carnivores, 1 cetacean, 3 ungulates, and 1 equid), 2 to Afrotheria, 2 to Metatheria (marsupials), and 2 to Aves (birds), as illustrated in Figure 5.6. The Supplementary Table S2 lists all 23 species along with their genome assemblies, publication dates, and approximate evolutionary distances from humans. These genomes were obtained from UCSC and Illumina, with minimal preprocessing limited to indexing using samtools and extracting chromosome size information.

5.3 HYPER model for predicting epigenetic signal in mouse

As an initial step in this cross-species analysis, I moved slightly further away from humans in evolutionary terms to examine another species within the same superorder of placental mam-

mals as humans, namely Euarchontoglires. This places the species at an approximate evolutionary distance of ~ 100 million years from humans. Despite this divergence, it is one of the most extensively studied model organisms, and in recent years, there has been a substantial increase in the availability of epigenetic profiling datasets for this species.

5.3.1 Model training

To enable the reconstruction of epigenetic tracks from chromatin accessibility and TF binding data in mouse erythroid-lineage cells, the final two layers of a pre-trained HYPER model were further optimised using the five datasets described in Section 5.2.1.

The dataset was divided into training and testing sets, with 95% of genomic regions $(17,868)$ used to train the model (train set) and the remaining 5% $(1,872)$ reserved for evaluation (test set). The model was trained for 70 epochs, reaching a final Huber loss of 0.136. The test set was not used during training and serves to assess how well the model generalises to new, unseen data.

Supplementary Figure S3 shows the distributions of the five epigenetic tracks in the training and test datasets. For each track, the two sets exhibit similar patterns and value ranges, indicating that the test data are representative of the training data. This ensures that model performance is evaluated on data drawn from the same underlying distribution as the training set, thereby reflecting the model's ability to generalise rather than differences between datasets.

The ATAC-seq, CTCF, and TAL1 tracks exhibit pronounced peaks at low signal values, reflecting a large proportion of background regions. TAL1 appears slightly left-skewed, indicating a concentration of values near zero with a gradual tail toward higher signals. In contrast, ATAC-seq and CTCF show mild right-skewness, suggesting that while most regions are near background levels, there is a subset of regions with elevated signal intensity corresponding to accessible or bound sites.

In comparison, the NF-E2 and GATA1 tracks display broader and more dispersed distributions, indicating greater variability in signal intensity across the genome. These wider distributions suggest a higher diversity of binding or activity states, with less dominance of background-level signal and more continuous variation between low and high signal regions.

5.3.2 Reconstruction of the murine alpha globin locus

Previous chapters have evaluated and validated HYPER's ability to predict epigenetic profiles with high accuracy, demonstrating a strong capacity to generalise to unseen sequences. Building on these results, this section focuses on the inspection of predicted profiles of open chromatin and TF binding in murine erythroid-lineage cells. To this end, we compare predictions generated using both the mouse-trained model and the human-trained model derived from erythroid-lineage cell data.

As previously noted, the alpha-globin locus is highly conserved across mammals. The primary difference between human and mouse lies in the presence of an additional regulatory element in the mouse locus. Specifically, the mouse locus contains a third facilitator element, termed Rm, located between the NPRL3 promoter and R3.

The four canonical enhancer-like elements (R1–R4) are otherwise conserved between human and mouse and occupy equivalent positions relative to the NPRL3 gene. Apparent differences in their reported locations arise from differences in gene structure and annotation between species, rather than true rearrangements of the regulatory elements. Overall, the regulatory architecture of the locus is preserved.

Both the human and mouse erythroid models were applied to predict epigenetic profiles across the murine alpha-globin locus. The resulting predictions, alongside the corresponding experimental tracks, are shown in Figure 5.7. The reconstructed open chromatin profiles successfully capture the key regulatory features of the locus, including all five enhancer-like elements and promoter regions (Figure 5.7A). As discussed in Section 4.5, epigenetics signal driven by transcriptional machinery and other chromatin-associated processes are not fully encoded in the DNA sequence alone, and therefore cannot be completely recovered by sequence-based models. This limitation is evident in the diminished predictive signal around the HBA and HBQ gene bodies, which is primarily driven by RNA polymerase II activity. Nevertheless, the primary focus of this analysis is on distal regulatory elements, which are accurately and consistently recapitulated by both models.

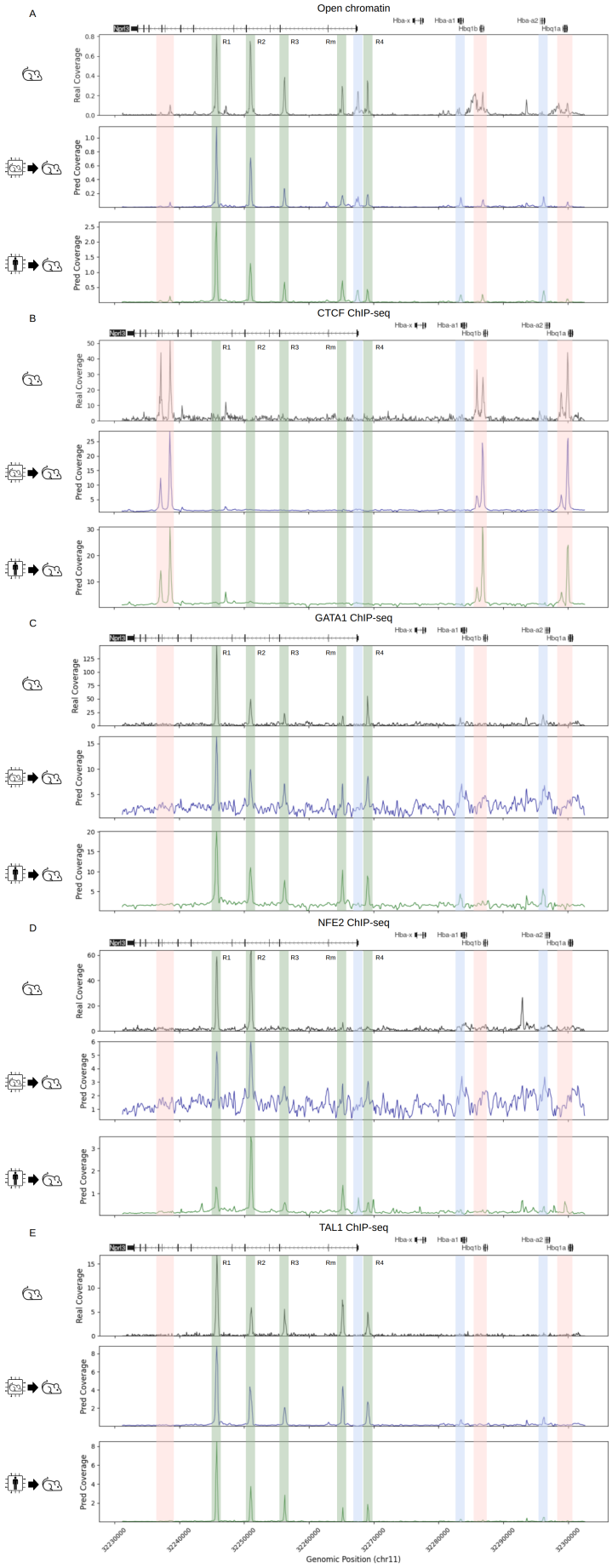

A
Open chromatin
B
CTCF ChIP-seq
C
GATA1 ChIP-seq
D
NFE2 ChIP-seq
E
TAL1 ChIP-seq
Hba-x
Hba-a1
Hbq1b
Hba-a2
Hbq1a
R1
R2
R3
Rm
R4
Real Coverage
Pred Coverage
Genomic Position (chr11)

Figure 5.7: Predicted epigenetic profiles across the mouse alpha-globin locus generated using both mouse and human models. The analysed region spans chr11:32,230,000–32,300,000, where erythroid-lineage cell tracks were predicted. Key regulatory elements are highlighted, with promoters shown in blue, enhancer-like elements in green, and CTCF sites in red. A) Open chromatin: experimental signal (black) compared with predictions from the mouse (blue) and human (green) models. B) CTCF binding: experimental signal (black) alongside mouse (blue) and human (green) predictions. C) GATA1 binding: experimental signal (black) with corresponding mouse (blue) and human (green) predictions. D) NF-E2 binding: experimental signal (black) compared to mouse (blue) and human (green) predictions. E) TAL1 binding: experimental signal (black) with predictions from the mouse (blue) and human (green) models. An interactive UCSC Genome Browser session of A and B, including predictions with the chicken model, is available at: https://genome.ucsc.edu/s/arodrig/HYPER_mm10_NPRL3_Pred.

Regarding the CTCF binding track (Figure 5.7B), the experimental signal is notably noisy, whereas the model predictions are considerably smoother and exhibit reduced background variability. Despite this denoising effect, both models successfully recapitulate the characteristic double CTCF peaks observed in the mouse locus.

In contrast, the reconstructed TF binding profiles for GATA1 (Figure 5.7C) and NF-E2 (Figure 5.7D) present a different behaviour. While the experimental tracks for these factors present similar noise levels to CTCF, the mouse model appears to partially learn and reproduce this local noise, particularly in the case of NF-E2, where true signal peaks become less distinguishable from the background. This effect is less pronounced for GATA1 but still noticeable. Interestingly, this issue is largely absent in the human model predictions, which remain smooth while still accurately capturing the main regulatory elements.

A likely explanation for this discrepancy lies in the statistical properties of the training data. Both GATA1 and NF-E2 mouse tracks exhibit substantially higher signal magnitude and variance compared to other factors, such as TAL1. This suggests a broader dynamic range and a heavier-tailed signal distribution, which may reflect technical noise or biological heterogeneity. As a result, the mouse model may overfit to these high-intensity and noisy patterns during training, leading to poorer generalisation and noisier predictions. In contrast, the human model, trained on a different distribution, may be less influenced by these noisy, high-intensity signals, allowing it to capture more consistent and conserved binding patterns.

On the other hand, TAL1 binding profiles are considerably cleaner in both experimental and predicted tracks (Figure 5.7E), with low background signal and well-defined peaks. Correspondingly, both models perform well in reconstructing TAL1 binding, highlighting how lower-

noise training data leads to more robust predictions.

Overall, both models demonstrate strong performance in reconstructing the epigenetic landscape of the murine alpha-globin locus. However, in some cases, particularly for TFs with noisy signal distributions, such as NF-E2, the mouse model tends to reproduce an elevated background signal. This behaviour is likely driven by the noisier, higher dynamic range of the mouse training data, especially considering that the human model was trained on higher-coverage data.

5.4 HYPER model for predicting epigenetic signal in chicken

Moving beyond mammals, I next evaluated HYPER's performance in predicting epigenetic activity in chicken, a species that diverged from humans approximately 300 million years ago and had not been previously encountered by its pretrained backbone, Enformer. Chicken is a well-established model organism. Although high-resolution epigenomic datasets remain relatively limited, the globin loci have been extensively characterised, and their regulatory elements are well defined.

5.4.1 Model training

For chicken erythroid-lineage cells, reconstruction of epigenetic tracks from chromatin accessibility and TF binding profiles was achieved by fine-tuning the final two layers of HYPER using the datasets outlined in Section 5.2.2.

The data were partitioned into 95% training ($7,987$ genomic regions) and 5% testing (827 genomic regions). The model was trained for 100 epochs, reaching a final training Huber loss of 0.034.

As shown in Supplementary Figure S4, the ATAC-seq and CTCF tracks both show unimodal distributions with strong overlap between the training and test sets, indicating consistent data partitioning and minimal distributional shift.

The ATAC-seq signal displays a broad peak centred around intermediate values with a slight right-skew, suggesting that a subset of regions exhibits higher signal intensity. A smaller sec-

ondary shoulder at lower values ($\sim$ 0.2–0.3) may reflect background or weakly accessible regions.

Similarly, the CTCF track exhibits a dominant peak with a narrower spread compared to ATAC-seq, indicating a more tightly distributed binding signal. The distribution shows mild right-skewness and a subtle low-signal component, such as the one in the accessibility distribution.

Overall, both tracks demonstrate clear signal enrichment above background with relatively well-defined distributions, with ATAC-seq showing greater variability and dynamic range than CTCF.

5.4.2 Reconstruction of the chicken alpha globin locus

Although the alpha-globin locus exhibits greater sequence divergence between mammals and avian species, its overall regulatory organisation remains broadly conserved. A key difference is the reduced number of alpha-globin genes in chicken, which contains three genes compared with five in humans. In contrast to the mammalian locus, where multiple enhancer-like elements (R1–R4) contribute to regulation, the chicken alpha-globin locus has been characterised as containing a single major functional enhancer, MCS-R2, with no additional enhancers identified to date. Despite these differences, the positional relationship between the enhancer and the NPRL3 gene is preserved across species.

As with the mouse analysis, I used the chicken erythroid-lineage model to predict epigenetic profiles across the chicken alpha-globin locus. In addition, I applied the human-trained model to the chicken genome, thereby evaluating model performance on a species not represented in either HYPER training or Enformer pretraining. Due to the limited availability of chicken experimental data, only open chromatin and CTCF binding tracks were available for direct comparison. Nevertheless, predictions for additional TF binding (e.g., GATA1, NF-E2, TAL1) from the human model were also included to assess whether conserved regulatory features, such as MCS-R2 and potential surrounding elements, could still be recovered.

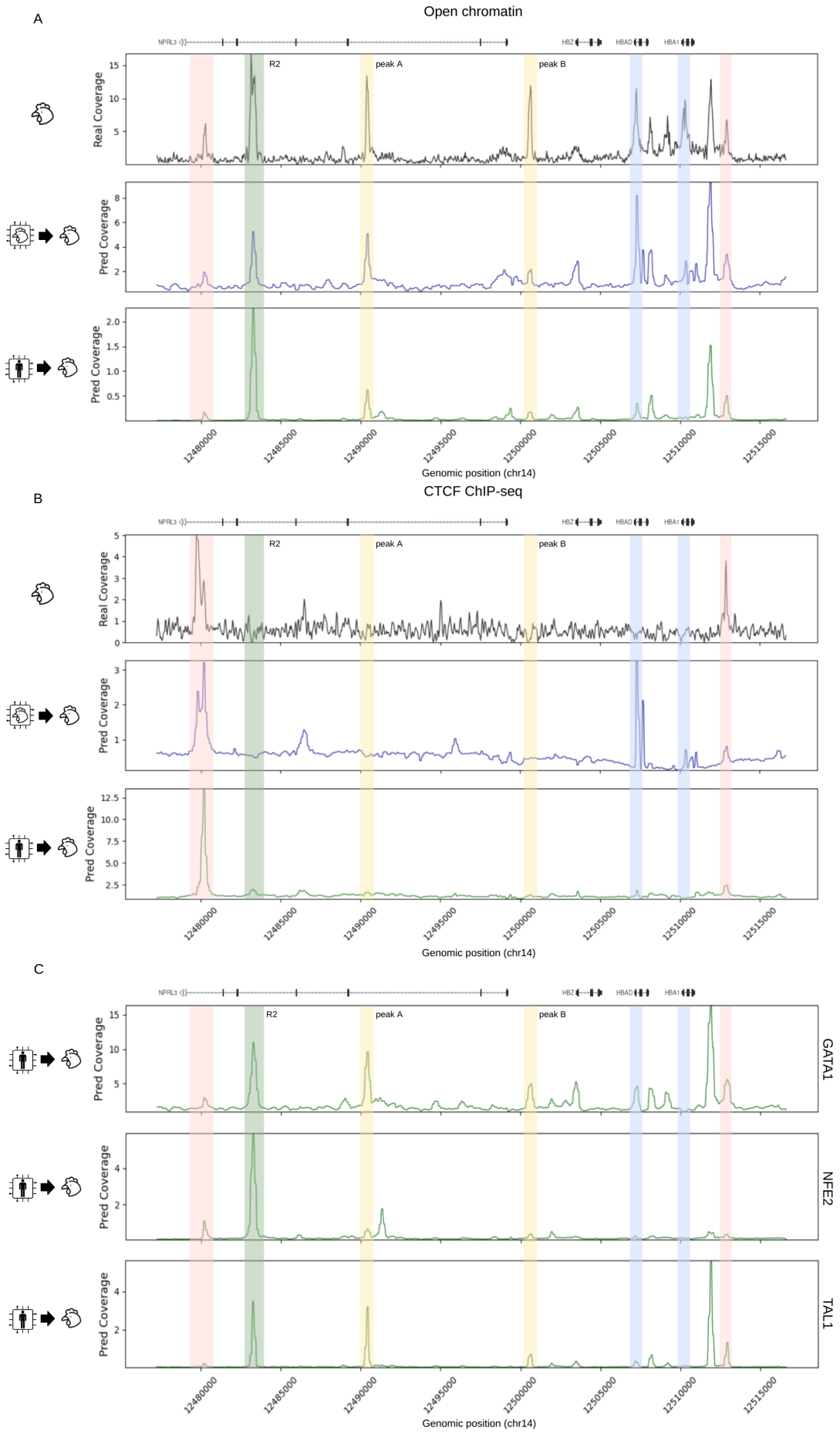

A
Open chromatin
NPRL3
HBZ
HBAD
HBA1
R2
peak A
peak B
Real Coverage
Pred Coverage
Pred Coverage
12480000
12485000
12490000
12495000
12500000
12505000
12510000
12515000
Genomic position (chr14)
B
CTCF ChIP-seq
Real Coverage
Pred Coverage
Pred Coverage
Genomic position (chr14)
C
GATA1
NFE2
TAL1
Genomic position (chr14)

Figure 5.8: Predicted epigenetic profiles across the chicken alpha-globin locus generated using both chicken and human models. The analysed region spans chr14:12,480,000–12,515,000, where erythroid-lineage cell tracks were predicted. Key regulatory elements are highlighted, with promoters in blue, enhancer-like elements in green, and CTCF sites in red. A) Open chromatin: experimental signal (black) compared with predictions from the chicken (blue) and human (green) models. B) CTCF binding: experimental signal (black) alongside predictions from the chicken (blue) and human (green) models. C) Predicted TF binding profiles generated using the human model, with GATA1 shown at the top, NF-E2 in the middle, and TAL1 at the bottom. Two signal peaks of interest are highlighted in yellow and named peak A and B. An interactive version of A and B, including predictions with the mouse model, is available at: https://genome.ucsc.edu/s/arodrig/HYPER_galGal6_NPRL3_Pred.

The resulting predictions, alongside the corresponding experimental tracks, are shown in Figure 5.8. Although the experimental open chromatin signal exhibits some degree of noise, particularly in regions associated with transcriptional activity around the alpha-globin genes, the reconstructed accessibility profiles successfully capture the key regulatory features of the locus (Figure 5.8A). Both models identify the MCS-R2 enhancer and promoter regions; however, the human model produces a smoother signal, whereas the chicken model tends to retain more local variability, especially near transcriptionally active regions. This behaviour is consistent with previous observations and likely reflects differences in the noise characteristics of the underlying training data.

The experimental CTCF binding track also exhibits a notable level of noise, slightly higher than that observed in the ATAC-seq signal. Consistent with the open chromatin predictions, the chicken model captures much of this local variability, reflecting the noisiness of the experimental data. In contrast, the human model produces a considerably smoother and cleaner signal (Figure 5.8B). Despite these differences, both models successfully recapitulate the major CTCF binding sites flanking the alpha-globin locus in chicken.

Given the strong performance of the human model in reconstructing the chicken erythroid epigenetic landscape, I further used it to predict binding profiles for the three principal regulators of alpha-globin expression: GATA1, NF-E2, and TAL1 (Figure 5.8C). As expected, these profiles show strong enrichment at the MCS-R2 enhancer. In addition, two other regions of elevated predicted binding emerge within the NPRL3 locus: one spanning chr14:12,490,219–12,490,573 (peak A), and another, weaker but consistently observed, at chr14:12,500,409–12,500,712 (peak B).

The predictions suggest that peak A, located within intron 2, is strongly enriched for GATA1 and TAL1 binding, with comparatively lower NF-E2 signal, which instead seems to bind adjacent to this region. In contrast, peak B, situated approximately 1 kb upstream of NPRL3, appears to be primarily enriched for GATA1, with minimal contribution from NF-E2 and TAL1.

Interestingly, peaks A and B display both positional and functional resemblance to the R3 and R4 regulatory elements described in mammals. Although sequence conservation for these elements can be detected between human and mouse, comparable conservation between human and chicken has not yet been demonstrated. Nevertheless, the absence of detectable sequence conservation does not preclude the existence of conserved regulatory function. Notably, even MCS-R2 exhibits only limited sequence conservation between distant species such as human and chicken (as assessed by a BLAT search in UCSC), supporting the idea that regulatory function can be maintained despite substantial sequence divergence. Experimental validation will, however, be required to more precisely characterise these regions and confirm their regulatory roles.

Overall, HYPER demonstrates strong cross-species generalisation on the chicken genome, accurately reconstructing key regulatory features of the alpha-globin locus. While the chicken model may capture more local variability, the human-trained model consistently identifies conserved regulatory elements with smoother and more robust predictions. These results highlight that regulatory function can be preserved despite sequence divergence, and that sequence-based models are capable of capturing this conservation across large evolutionary distances.

5.5 Generalising epigenetic signals in cross-species contexts

To further assess HYPER's ability to generalise epigenetic profiles across species, I next evaluated how well cross-species models can regress to the human genome. The motivation behind this analysis is to determine whether models trained on non-human species capture sufficiently conserved regulatory principles to accurately reconstruct human epigenetic landscapes, and to what extent species-specific biases influence this transfer. Importantly, the pretrained backbone of HYPER, Enformer, has already been trained on human and mouse genomes, meaning that any observed performance reflects both retained prior knowledge and the effects of subsequent species-specific fine-tuning.

As a first step, I tested the mouse- and chicken-trained models on the human genome, focusing on multiple levels of analysis. This includes the reconstruction of open chromatin and CTCF binding signal tracks across the alpha-globin locus, the assessment of nucleotide-level importance within regulatory elements for each model when applied to the human sequence and to the organisms they were trained on, and the identification of enriched TF motifs based on these importance scores.

Through this approach, I aim to evaluate not only the accuracy of cross-species signal prediction but also whether the different species models recover similar biologically meaningful regulatory features in human, thereby providing insight into the extent of functional conservation learned by HYPER across evolutionary distances.

5.5.1 Human alpha globin track

As a first step, I evaluated cross-species transfer back to the human genome by comparing the human-, mouse-, and chicken-trained erythroid models in reconstructing the epigenetic landscape of the human alpha-globin locus.

All three models accurately reconstruct both open chromatin (Figure 5.9A) and CTCF binding profiles (Figure 5.9B) across the locus. Since they are all built on the Enformer backbone, which has already captured key aspects of the human regulatory code, it is unsurprising that the human and mouse models achieve the highest fidelity. In the human model, HYPER fine-tuning further reinforces and specialises these patterns within an erythroid context, while in the mouse model, the closer evolutionary conservation of the locus relative to chicken likely contributes to its strong performance.

In contrast, the chicken model, fine-tuned on previously unseen and noisier data, predicts a higher baseline background signal. Nevertheless, it successfully captures all major regulatory elements within the locus. Notably, it is the only model that clearly recovers the HBM promoter, which is not detected by either the human or mouse models.

A similar trend is observed for CTCF binding: while the chicken model exhibits an elevated background signal, all three models accurately identify the key binding sites.

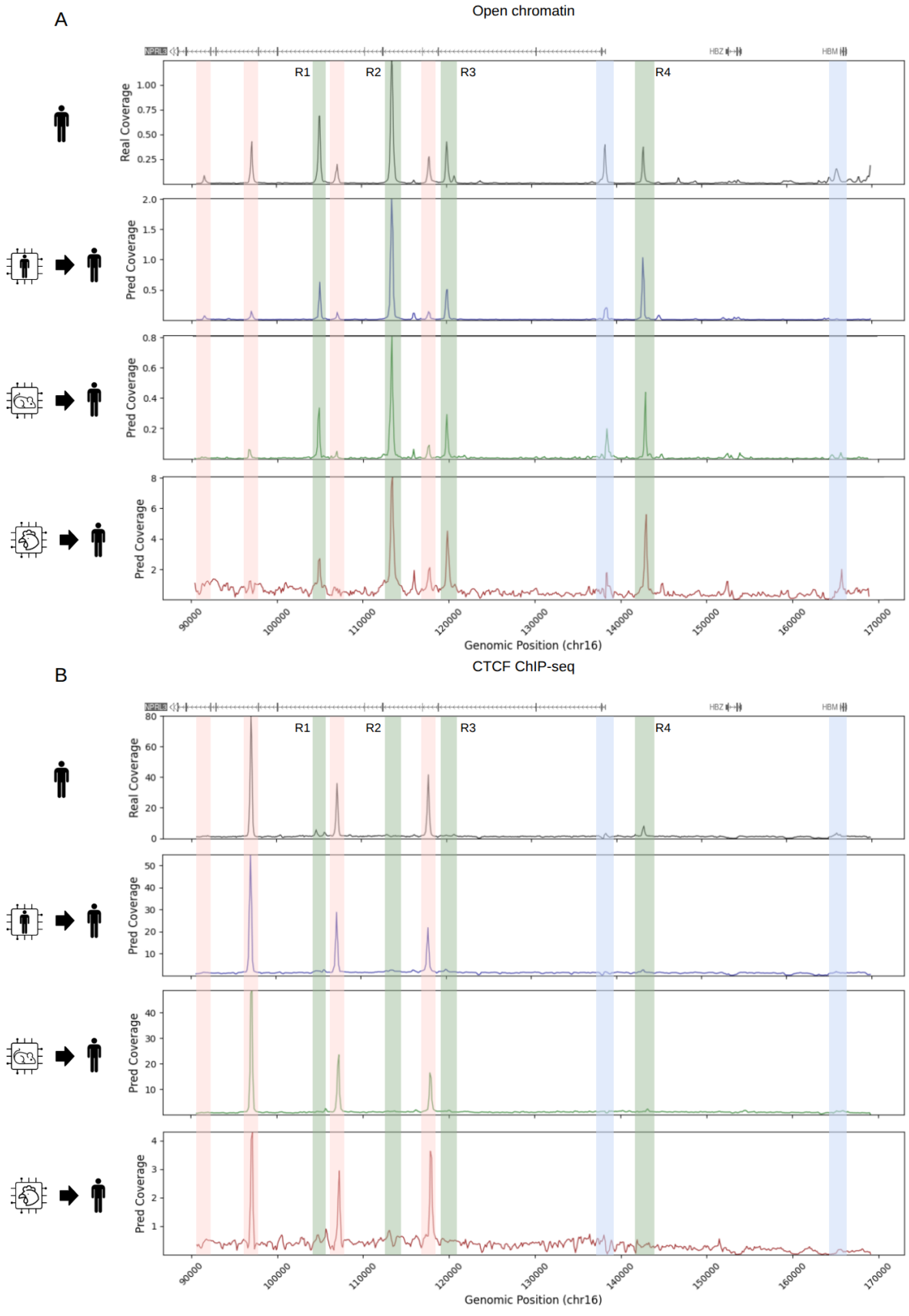

A
Open chromatin
NPRL3
HBZ
HBM
R1
R2
R3
R4
Real Coverage
Pred Coverage
Pred Coverage
Pred Coverage
90000
100000
110000
120000
130000
140000
150000
160000
170000
Genomic Position (chr16)
B
CTCF ChIP-seq
NPRL3
HBZ
HBM
R1
R2
R3
R4
Real Coverage
Pred Coverage
Pred Coverage
Pred Coverage
90000
100000
110000
120000
130000
140000
150000
160000
170000
Genomic Position (chr16)

Figure 5.9: Predicted epigenetic profiles across the human alpha-globin locus generated using both cross-species models. Key regulatory elements are highlighted, with promoters in blue, enhancer-like elements in green, and CTCF sites in red. A) Open chromatin: experimental signal (black) compared with predictions from the human (blue), mouse (green), and chicken (red) models. B) CTCF binding: experimental signal (black) alongside predictions from the human (blue), mouse (green), and chicken (red) models. An interactive version of these UCSC Genome Browser sessions is available at: https://genome.ucsc.edu/s/arodrig/HYPER_hg38_WG_XS.

Overall, these results indicate that, despite differences in signal quality introduced by species-specific training data, all models retain the ability to recover conserved regulatory features, with the chicken model revealing an unexpected sensitivity to certain elements missed by the others.

5.5.2 Cross-species nucleotide importance in R2

The next step in this cross-species analysis focuses on evaluating HYPER's ability to recover regulatory sequence features through nucleotide-level interpretation. Specifically, I assess whether the models can reconstruct sequence logos at a conserved regulatory element across species, the primary alpha-globin enhancer R2, using nucleotide importance scores derived from feature attribution methods, specifically DeepLIFTShap (Section 2.8). This approach provides insights into which bases within a sequence are most influential for the predicted regulatory signal in R2, enabling a direct link between model predictions and underlying cis-regulatory logic.

This step is important to assess whether models trained on different species identify the same functional bases when applied to the human genome, providing insight into the consistency and transferability of the learned regulatory code.

The nucleotide importance profiles for the R2 enhancers across the three species show strong agreement in the identification of key regulatory regions across all models. Each model captures the main clusters of high-importance bases, including complex motif structures where high- and low-importance positions are interleaved, such as the GATA1-like peaks observed to the right side of the plot in both human (Figure 5.10A) and mouse (Figure 5.10B), or the GATA1–TAL1-like pattern emerging across the left-central region in chicken (Figure 5.10C).



Figure 5.10: Nucleotide importance profiles for the R2 enhancer across species, predicted using all three models. A) Human R2 enhancer sequence, with importance profiles generated by the human model (top), mouse model (middle), and chicken model (bottom). B) Mouse R2 enhancer sequence, with predictions from the mouse model (top), human model (middle), and chicken model (bottom). C) Chicken R2 enhancer sequence, with predictions from the chicken model (top), human model (middle), and mouse model (bottom). Yellow highlights indicate known motifs identified using TomTom.

Across all predictions, peaks remain well-localised, with consistent motif positioning and clustering, indicating that the models preserve the underlying regulatory structure despite being trained on different species. This consistency suggests that the core functional elements of R2 are robustly recognised across evolutionary contexts, suggesting that important sequence features associated with erythroid regulatory activity are consistently represented across the models.

5.5.3 Species-specific nucleotide importance in the alpha-globin locus

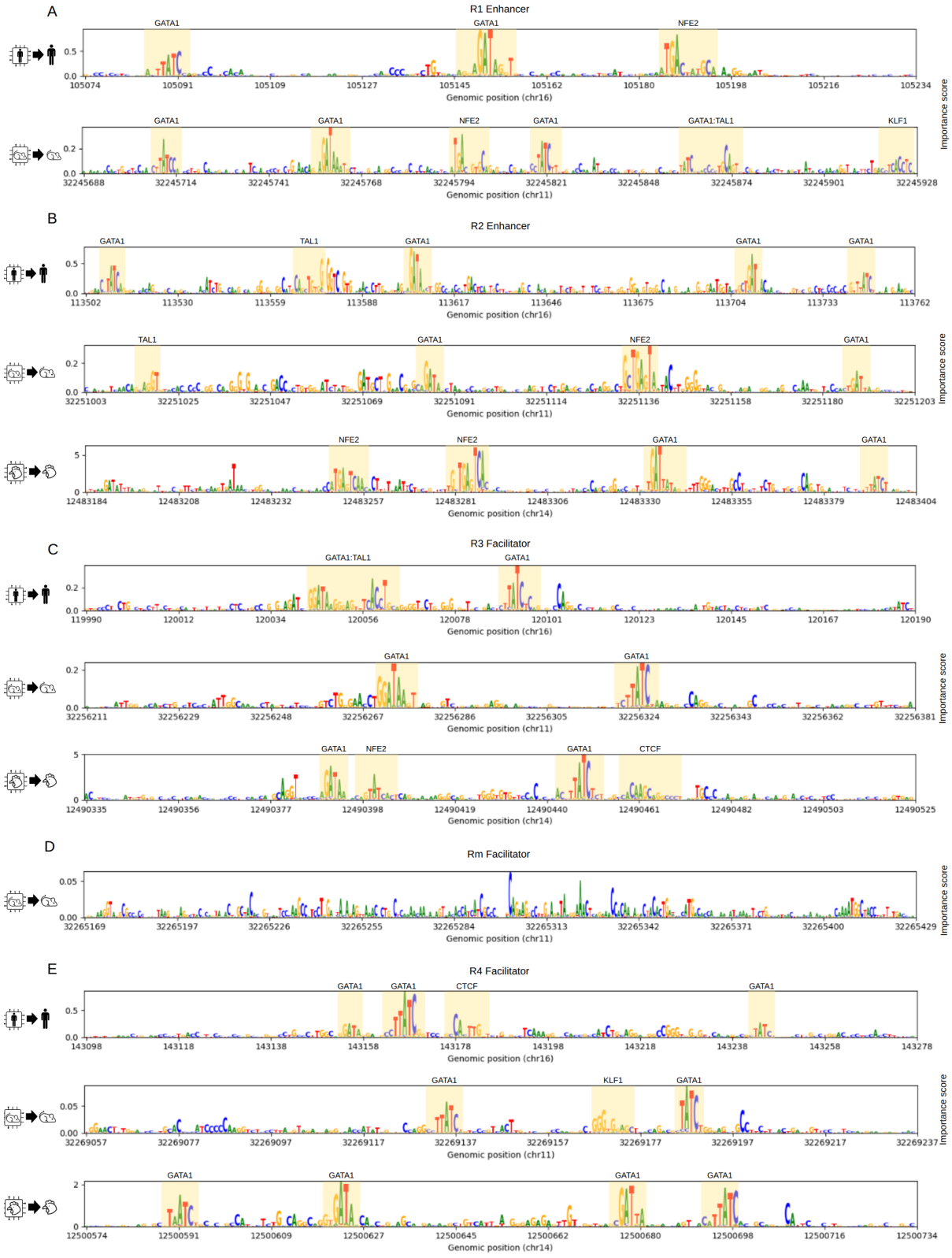

A
R1 Enhancer
GATA1
GATA1
NFE2
105074
105091
105109
105127
105145
105162
105180
105198
105216
105234
Genomic position (chr16)
GATA1
GATA1
NFE2
GATA1
GATA1:TAL1
KLF1
32245688
32245714
32245741
32245768
32245794
32245821
32245848
32245874
32245901
32245928
Genomic position (chr11)
Importance score
B
R2 Enhancer
GATA1
TAL1
GATA1
GATA1
GATA1
113502
113530
113559
113588
113617
113646
113675
113704
113733
113762
Genomic position (chr16)
TAL1
GATA1
NFE2
GATA1
32251003
32251025
32251047
32251069
32251091
32251114
32251136
32251158
32251180
32251203
Genomic position (chr11)
NFE2
NFE2
GATA1
GATA1
12483184
12483208
12483232
12483257
12483281
12483306
12483330
12483355
12483379
12483404
Genomic position (chr14)
C
R3 Facilitator
GATA1:TAL1
GATA1
119990
120012
120034
120056
120078
120101
120123
120145
120167
120190
Genomic position (chr16)
GATA1
GATA1
32256211
32256229
32256248
32256267
32256286
32256305
32256324
32256343
32256362
32256381
Genomic position (chr11)
GATA1
NFE2
GATA1
CTCF
12490335
12490356
12490377
12490398
12490419
12490440
12490461
12490482
12490503
12490525
Genomic position (chr14)
D
Rm Facilitator
32265169
32265197
32265226
32265255
32265284
32265313
32265342
32265371
32265400
32265429
Genomic position (chr11)
E
R4 Facilitator
GATA1
GATA1
CTCF
GATA1
143098
143118
143138
143158
143178
143198
143218
143238
143258
143278
Genomic position (chr16)
GATA1
KLF1
GATA1
32269057
32269077
32269097
32269117
32269137
32269157
32269177
32269197
32269217
32269237
Genomic position (chr11)
GATA1
GATA1
GATA1
GATA1
12500574
12500591
12500609
12500627
12500645
12500662
12500680
12500698
12500716
12500734
Genomic position (chr14)

Figure 5.11: Nucleotide importance profiles across alpha-globin MCS-R regulatory elements. A) Predicted sequence logos for the R1 enhancer in human and mouse. B) Predicted sequence logos for the R2 enhancer in human, mouse, and chicken. C) Predicted sequence logos for the R3 facilitator in human and mouse, alongside the putative R3 region in chicken. D) Predicted sequence logo for the Rm facilitator, specific to mouse. E) Predicted nucleotide importance profiles for the R4 facilitator in human and mouse, together with the putative R4 region in chicken. Yellow highlights indicate known motifs identified using TomTom.

To further investigate cross-species regulatory inference, I compared nucleotide importance profiles at the R1-R4 elements across species using nucleotide importances generated with the three models.

Given the high degree of conservation of the alpha-globin locus, it is reasonable to expect that its regulatory elements are governed by shared sequence features. In particular, conserved TF binding motifs should manifest as similar nucleotide importance patterns across species, even in cases where overall sequence conservation is limited. Therefore, this analysis aims to determine whether HYPER captures such conserved regulatory grammar and whether these signals are maintained across evolutionary distances.

To this end, I computed nucleotide importance scores for the R1 enhancer in human and the mouse equivalent using their respective species-specific models. Subsequently, I analysed R2, R3, and R4 across all three species, including the putative chicken regulatory regions corresponding to R3 (peak A) and R4 (peak B) analogues identified in the previous section.

The nucleotide importance profiles for the R1 enhancer (Figure 5.11A) in human and mouse show well-defined patterns, with prominent peaks marking key regulatory bases. Despite sequence divergence, the overall organisation is conserved, indicating that HYPER captures sequence features consistent with partially shared regulatory organisation. These peaks are consistent with canonical erythroid motifs, particularly GATA1 (WGATAA) and TAL1-associated E-boxes (CANNTG), suggesting cooperative binding. The human profile appears sharper, while the mouse signal is more distributed, likely reflecting differences in data quality or motif composition.

For the R2 enhancer (Figure 5.11B), strong motif conservation is evident across human, mouse, and chicken, with aligned clusters of high-importance bases. This dense structure is consistent with a combinatorial motif architecture involving GATA1, TAL1/E-box, and NF-E2 sites, in line with R2's role as the primary enhancer. Compared to R1, these patterns are more tightly

conserved, supporting stronger evolutionary constraints.

In contrast, the R3 facilitator (Figure 5.11C) shows a simpler structure, driven by a limited number of sequence importance peaks that are still well conserved. These are compatible with weaker or fewer GATA1 and E-box motifs and are also observed in the candidate chicken region, suggesting functional similarity. However, the absence of sequence conservation, together with the distinct regulatory architecture in chicken, makes it unclear whether this region represents a true R3 ortholog.

The mouse-specific Rm element (Figure 5.11D) displays a diffuse importance profile with no dominant peaks, indicating the absence of strong motif-driven binding. This suggests the presence of weak or degenerate GATA-like or E-box-like sequences, possibly reflecting its species specificity and a facilitator role mediated through cooperative interactions.

Finally, the R4 facilitator (Figure 5.11E) presents a small number of well-defined sequence peaks in human and mouse, also reflected in the corresponding chicken region. These are likely associated with GATA1, with possible contributions from E-box and CACCC-like (KLF1) motifs. While the chicken R4 region shows limited sequence conservation relative to its mammalian counterpart, its strong functional conservation supports its identification as an R4 ortholog. In particular, the presence of an increased number of GATA1 binding sites suggests that the regulatory architecture of this enhancer has been maintained or even reinforced in chicken.

Overall, despite increased variability in chicken, the key regulatory features are consistently recovered, indicating that HYPER captures conserved erythroid regulatory logic across species.

5.5.4 Cross-species motif enrichment in R2

A subsequent step involves using nucleotide importance scores derived from cross-species predictions, such as those examined in Section 5.5.2. This approach aims to determine whether HYPER can generate importance profiles that recover representative TF motifs when applied to sequences from other species, and whether these motifs can be reliably identified by the enrichment pipeline described in Section 2.9.5.

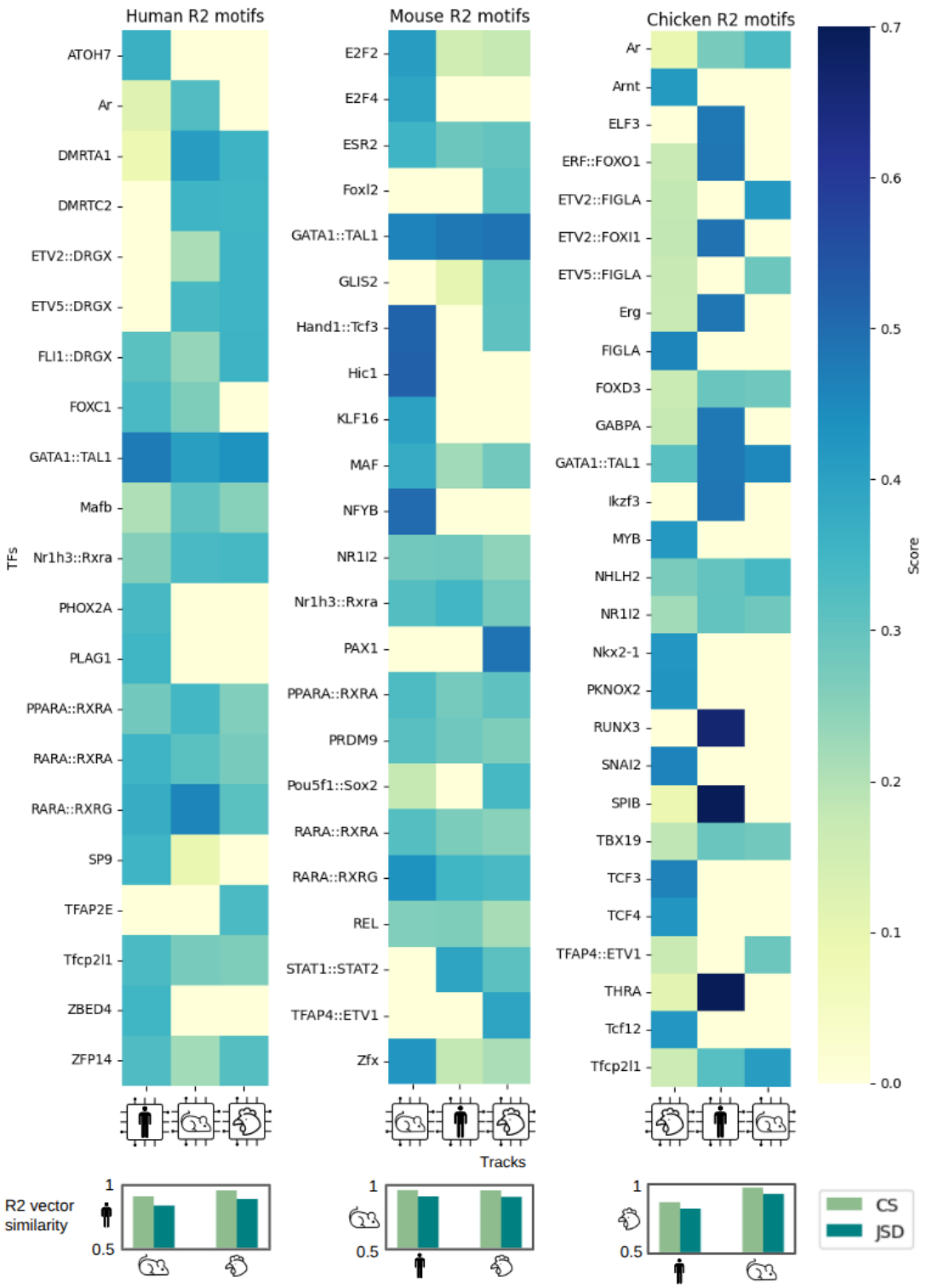


Figure 5.12: Motif enrichment of the R2 enhancer across species using cross-species predictions. Nucleotide importance scores were generated by applying all three models to each species-specific R2 sequence (Section 5.5.3) and analysed with FIMO and TomTom to identify enriched motifs against the JASPAR database. For visualisation, the top 10 contributing motifs were selected for each prediction instance and merged within each species, resulting in equally sized motif sets across species. At the bottom, cosine similarity (CS, shown in light green) and JSD (shown in teal) were computed between the R2 representations of the model's reference species and the species used to train the cross-species model.

As stated in Section 4.3.6, interpretation of these results should be made with caution, as motif-based analyses depend heavily on existing reference databases, which may not capture the full diversity of TF binding patterns. In contrast to the previous analysis, motif enrichment derived from cross-species predictions shows a more concentrated set of dominant motifs (Figure 5.12), particularly for the human and mouse R2 enhancers.

In both cases, the regulatory landscape is strongly driven by the GATA1::TAL1 complex, with additional contributions from motifs such as RARA::RXRG.

For the mouse R2 enhancer, while GATA1::TAL1 remains the primary contributor, additional motifs such as HIC1, NFYB, and Hand1::Tcf3 are prominently enriched when using the mouse model. However, predictions from the human model still recover a similar central motif structure.

The chicken R2 enhancer presents a more distinct pattern. When predicted using the human and mouse models, GATA1::TAL1 is again strongly enriched. In contrast, the chicken-trained model highlights a different set of motifs, including factors such as RUNX3, THRA, SPIB, and TCF family members.

To summarise the similarity between the predicted cross-species R2 representations and the predictions generated by the reference-species models, I compared the motif-contribution vectors (i.e., the columns of the heatmaps in Figure 4.3.6). Motifs that were not simultaneously identified by both the reference and cross-species models were excluded, thereby restricting the comparison to motifs consistently predicted by both models.

Two metrics were used to quantify these similarities. Cosine similarity measures how similar the sets of active motifs are, whereas JSD compares the relative contribution distributions of those motifs. For both metrics, values closer to one indicate greater similarity.

For human R2, the chicken model produced a motif contribution distribution more similar to the reference model than the mouse model. For mouse R2, both the human and chicken models generated highly similar contribution profiles across the shared motifs. In contrast, for chicken R2, the mouse model produced a representation more closely matching the reference model than the human model.

These differences may reflect variation in the sets of motifs identified across species. A clear example is observed in the chicken R2 predictions, where the human model assigned high

contributions to motifs such as RUNX3 and SPIB, while the reference chicken model and the mouse model either did not identify these motifs or assigned them negligible contributions.

In addition to the discrepancies discussed in Section 4.3.6, some of the identified TFs lack a clear association with erythroid regulation, including families such as RUNX, ETS, HIC1, and NFYB. However, other enriched motifs show strong similarity to canonical erythroid regulators and are more likely to reflect alternative or misassigned motif matches. Notably, motifs from the HAND1::TCF3 complex and the broader TCF family recognise E-box sequences, making them highly similar to TAL1 binding sites and therefore prone to overlap in motif identification.

Similarly, motifs associated with AT-rich regions, such as RARA::RXRG, or GC-rich regions, such as SP1 family motifs, may arise from sequence composition biases and could represent false negatives or indirect proxies of actual GATA1 binding.

Overall, these results suggest that cross-species predictions tend to emphasise conserved core motifs, particularly GATA1::TAL1, while species-specific models introduce additional variability and highlight alternative or ambiguous regulatory features.

5.5.5 Species-specific motif enrichment in the alpha-globin locus

A natural extension of the previous motif enrichment analysis is to perform a similar assessment using nucleotide importance scores derived from the previously computed species-specific nucleotide importance scores to infer motif enrichment across all MCS-R elements following the methodology described in Section 2.9.5. This approach is motivated by the need to translate base-level importance into interpretable TF binding signatures, and allows the identification of enriched motifs underlying the predicted regulatory activity, as well as the comparison of regulatory grammar across species.

The human enhancers (Figure 5.13, left) are strongly dominated by the GATA1::TAL1 complex, which shows consistently high contribution across all elements, reflecting the canonical erythroid regulatory program. Additional motifs such as RARA::RXRG contribute notably in R1 and R4, while ESR2 shows higher enrichment in R4. Overall, the human motif landscape appears relatively compact, with a high degree of motif overlap between elements.

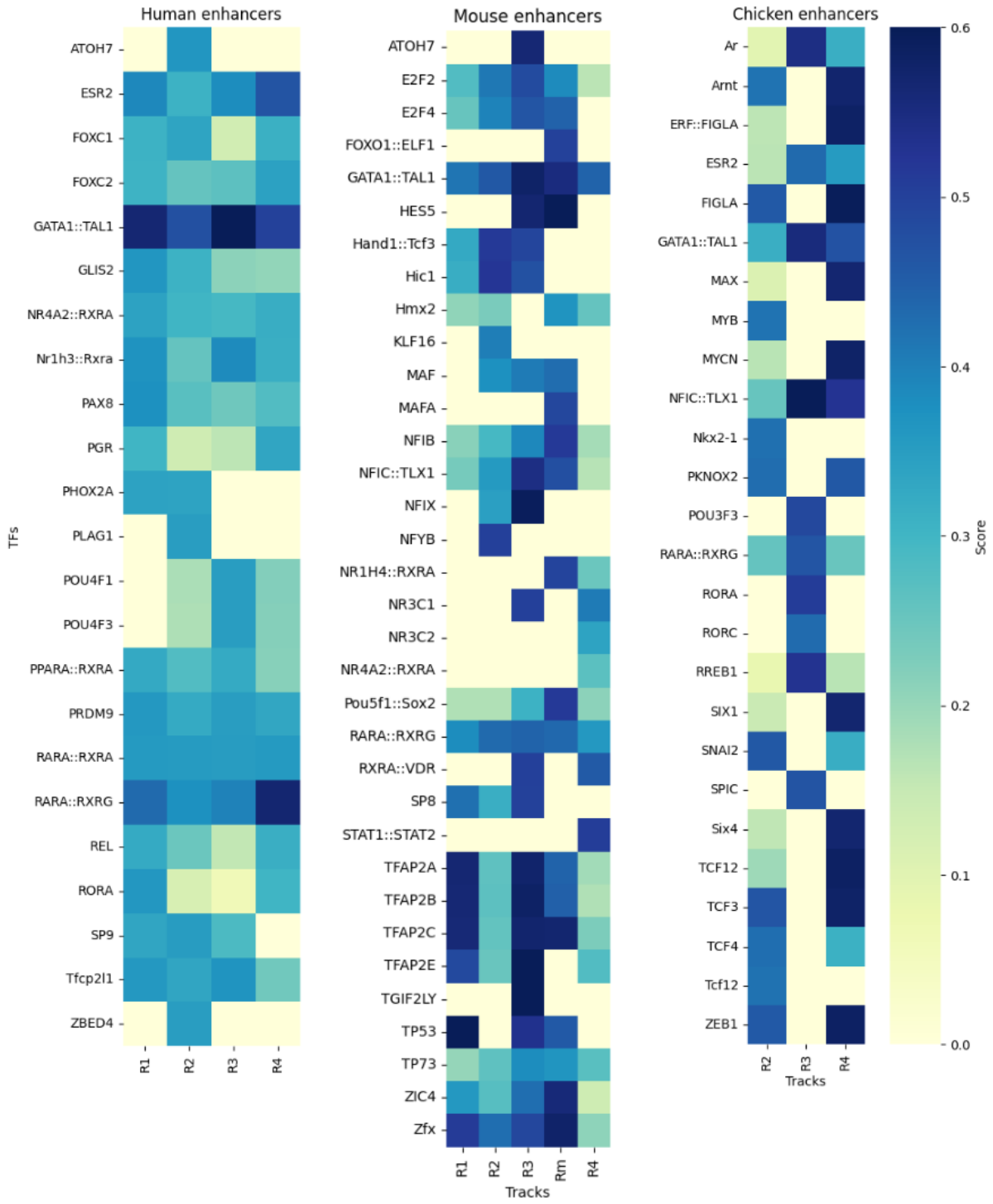


Figure 5.13: Motif enrichment across species-specific MCS-R elements. Nucleotide importance scores were obtained from the corresponding species-specific models (Section 5.5.3) and analysed using FIMO and TomTom to identify enriched motifs against the JASPAR database. A contribution score was calculated by combining motif significance with the average nucleotide importance across each motif. For the heatmap, the top 10 contributing motifs were selected for each enhancer-like element and then merged across elements within each species. The human set results in a smaller heatmap due to greater overlap of motifs between elements, while the mouse set shows a larger and more diverse set of motifs.

In contrast, the mouse enhancers (Figure 5.13, middle) exhibit a more diverse and heterogeneous motif composition. While GATA1::TAL1 remains an important contributor, several additional TFs display strong, element-specific enrichment, including TFAP family members (TFAP2A/B/C) in R1 and R3, HIC1 and E2F-related factors in R2, and HES5 and Hand1/Tcf3 in R4.

The chicken enhancers (Figure 5.13, right) show a distinct pattern compared to mammals. GATA1::TAL1 is primarily enriched in the putative R3 and R4 regions, but is less prominent in R2, where TCF family motifs (e.g., TCF3/4/12) dominate. Additional motifs such as NFIC::TLX1, MYCN, and RORA also contribute across elements.

Similar to the motif enrichment analysis described in Section 4.3.6, discrepancies in motif identification can arise from both methodological and biological factors. For example, limited representation of certain TF motifs, such as NF-E2, in databases like JASPAR may reduce their detectability by tools such as FIMO and TomTom. Additionally, similarities between motifs can lead to misclassification, particularly when TFs share related binding preferences.

Importantly, regulatory regions often contain overlapping or closely spaced binding sites, allowing multiple motifs to be assigned to the same sequence context. This can result in alternative motif interpretations depending on the scoring approach. In this pipeline, composite contribution scores were designed to accommodate such cases, enabling motifs with either high positional significance or high nucleotide importance to be retained among the top candidates. Finally, noise and biases in both the underlying data and the nucleotide importance estimates may further influence motif recovery.

Overall, motif enrichment highlights a conserved erythroid regulatory core centred on GATA1::TAL1 in mammals, alongside greater diversity in chicken. This divergence may reflect technical factors, such as experimental noise or model bias towards human sequence, as well as genuine biological differences arising from evolutionary distance and distinct regulatory architectures.

To further investigate the similarities between motif contribution representations of the same enhancers across different species, I computed motif contribution scores, as described in Section 2.9.5, for the super-enhancer elements of mouse and chicken using the human-trained model. This approach was chosen to maintain all prediction values within the same representational and regulatory framework within the human space, thereby enabling more consistent

cross-species comparisons. In addition, because the JASPAR database is predominantly composed of human TF motifs, projecting sequences from other species into the human regulatory space provides a more appropriate and biologically interpretable basis for motif analysis.

These motif contribution representations were normalised and subsequently compared using cosine similarity, as described in Section 2.10. As expected, enhancers belonging to the same species showed a considerable degree of association. For example, human R2 and R3 exhibited a cosine similarity of 0.55, while mouse R1 and R2 showed a similarity of 0.51 (Figure 5.14). This pattern may be explained by the presence of species-specific TF binding preferences and regulatory architectures that generate similar motif contribution landscapes within the same genomic and evolutionary context.

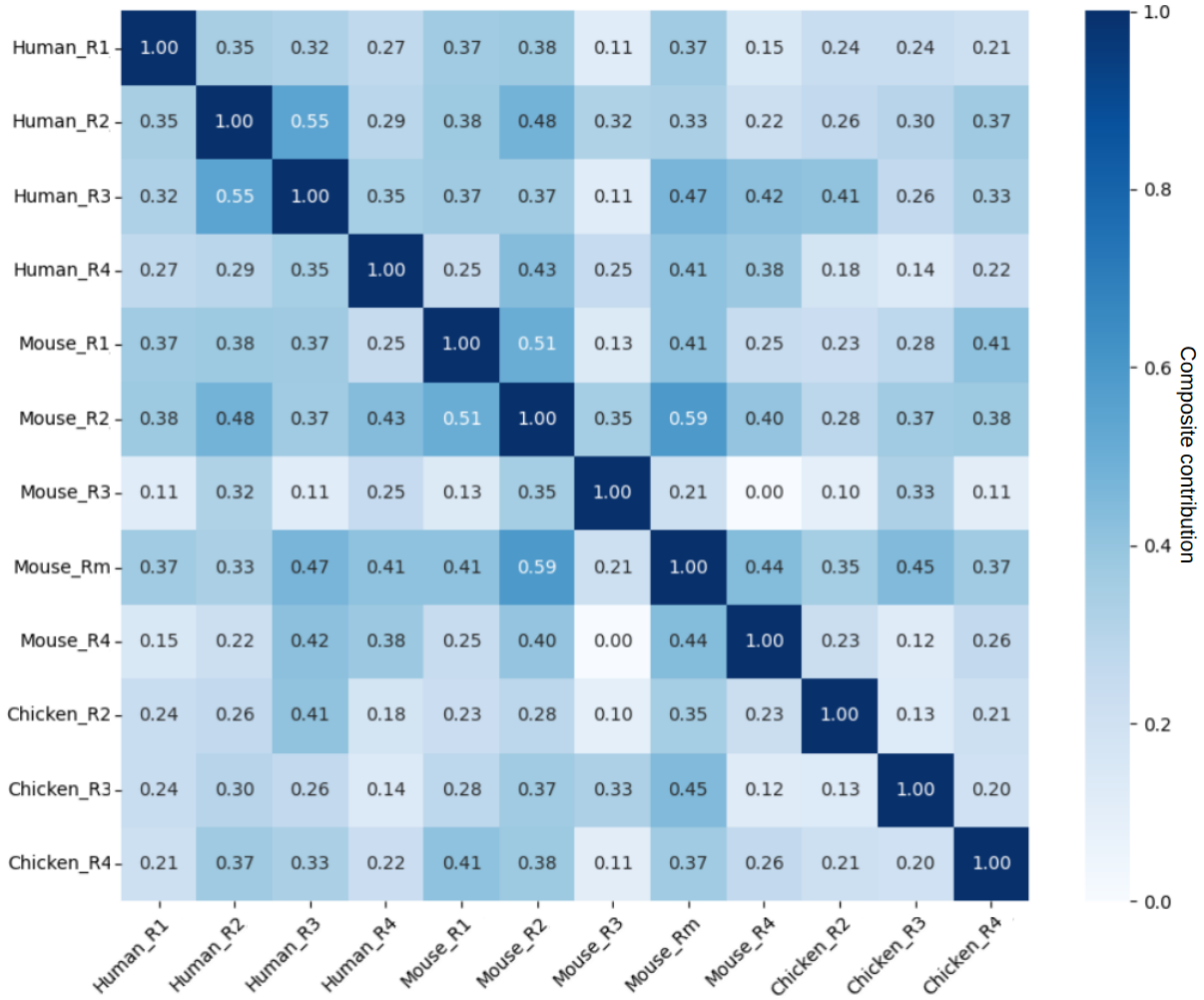


Figure 5.14: Motif contribution similarities for human, mouse, and chicken enhancers predicted with the human model. Cosine similarities were computed between the normalised representations of the super-enhancer elements for human, mouse, and chicken elements predicted with the human model.

Importantly, although these comparisons appear biologically meaningful for the R2 enhancer, it is essential to note that this type of analysis is strongly dependent on the underlying TF annotations. Current TF motif databases are not cell-type specific and still lack a comprehensive representation of regulatory "flavours" across different tissues, developmental stages, and species. Consequently, some inferred motif contributions may reflect limitations of the available annotations rather than true biological differences.

A more complete and accurately annotated TF database would substantially improve this type of comparative analysis. Under such conditions, motif contribution profiling could become a powerful framework for comparing orthologous regulatory elements across species, enabling the identification of conserved regulatory programs as well as species-specific motif usage and regulatory adaptations. This would provide deeper insight into how enhancer function evolves while preserving core transcriptional regulation.

5.5.6 Summary of predictions with cross-species models

Briefly, this section evaluated HYPER's ability to capture regulatory patterns using models applied to genomes outside their training domain. Epigenetic track reconstruction was first carried out with species-specific models (human, mouse, and chicken) to predict open chromatin and TF binding at the alpha-globin locus. These analyses were then extended to cross-species predictions, enabling the assessment of model generalisation across evolutionary distances.

To further interpret model predictions, nucleotide importance scores were computed using DeepLIFTShap for key regulatory elements (MCS-R regions). These importance profiles were used to generate sequence logos and identify functional bases driving predictions. Building on this, motif enrichment analysis was performed using FIMO and TomTom to map important sequence features to known TF motifs, both in species-specific and cross-species prediction contexts.

Overall, HYPER demonstrates strong performance in reconstructing epigenetic profiles across species, accurately capturing key regulatory elements even in genomes not represented during model training. While species-specific models tend to better reproduce signal characteristics (including noise), cross-species models consistently recover the core regulatory structure.

At the sequence level, nucleotide importance analyses revealed that the models learn a mostly

conserved regulatory grammar, primarily driven by canonical erythroid motifs such as GATA1 and TAL1. Motif enrichment further supports this, showing that GATA1::TAL1 remained the dominant regulatory signature, particularly in mammals, while increased variability in chicken reflects both evolutionary divergence and potential methodological limitations.

Together, these results indicate that HYPER captures sequence features consistent with partially conserved regulatory organisation across species, even when sequence conservation is limited, with cross-species predictions emphasising core motifs and species-specific models introducing additional regulatory variability.

5.6 Regulatory evolution of the alpha globin locus

This analysis is motivated by the observations from the previous chapter, where the human erythrocyte HYPER model demonstrated a strong ability to generalise across species, in some cases even outperforming models trained on those species-specific data. Two main factors likely contribute to this behaviour. First, the human model is trained on a large, high-resolution dataset, whereas equivalent epigenetic datasets for other species are more limited and often noisier. Second, this difference in data quality can lead to reduced performance or increased variability in species-specific models.

Building on the robust performance of the human model in capturing regulatory patterns across unseen sequences and species, I extended the analysis to explore its predictive capacity in genomes where erythroid epigenetic profiles are not available. This includes, first, ancient hominin genomes reconstructed by incorporating publicly available genetic variants onto the `hg38` reference assembly, and second, more evolutionarily distant species beyond the hominin lineage, including other mammals and birds.

An initial exploration of predicting epigenetic activity in the absence of experimental data was presented in Section 5.4.2, where the human model was applied to the mouse genome to infer TF binding patterns for GATA1, NF-E2, and TAL1. Given that these predictions were consistent with experimental ATAC-seq profiles, it is worth extending this approach further. This analysis, therefore, aims to assess whether HYPER can serve as a predictive tool for inferring regulatory landscapes without experimental data, and to evaluate the extent to which conserved regulatory logic can be recovered across deep evolutionary distances.

5.6.1 Recent evolutionary changes in the hominin alpha-globin locus

Before extending the analysis to more evolutionarily distant species, I first focused on a much shorter evolutionary timescale, within $\sim 600,000$ years, examining closely related human lineages, including Neanderthals, Denisovans, and early Homo sapiens. These groups belong to the genus Homo and represent closely related archaic human populations that coexisted and, in some cases, interbred.

Given the relatively short evolutionary distance between these groups, it is unlikely that structurally distinct regulatory differences have emerged, as such changes are more likely associated with speciation events [235]. This is even less expected in highly conserved regions such as the alpha-globin locus. Therefore, erythroid epigenetic profiles across these ancient genomes are expected to closely resemble those of modern humans, represented here by the aggregated signal from the 50-donor dataset.

However, investigating epigenetic profiles in Neanderthal erythroid cells remains highly informative. Even subtle sequence differences between these lineages could lead to changes in TF binding affinity or chromatin accessibility, potentially affecting gene regulation. In the context of erythropoiesis, such differences may relate to adaptations to distinct environments, such as hypoxia, climate, or metabolic demands, offering insight into how regulatory variation contributes to phenotypic diversity. Moreover, this analysis provides a stringent test of HYPER's sensitivity, assessing whether it can detect fine-scale regulatory variation within closely related genomes where large-scale differences are minimal.

The DNA sequences from the alpha-globin locus of each genome were used as input to HYPER to predict epigenetic activity. Although visual inspection of the reconstructed tracks for the nine ancient hominins does not reveal clear differences (Figure 5.15), a general slight reduction in activity is observed in Denisova, Mezmaiskaya, Vindija, Ust Ishim, and Linearbandkeramik.

Consequently, I conducted a more detailed analysis by computing damage scores to compare predicted coverage on ancient sequences with the control coverage (predicted on the ATAC-seq 50-donor sequence). Building on the approach used in Sections 4.4 and 3.5.2, and considering HYPER's strong performance in capturing variant effects from SNPs, peaks were called from the predicted tracks using LanceOtron without applying a peak score filter.

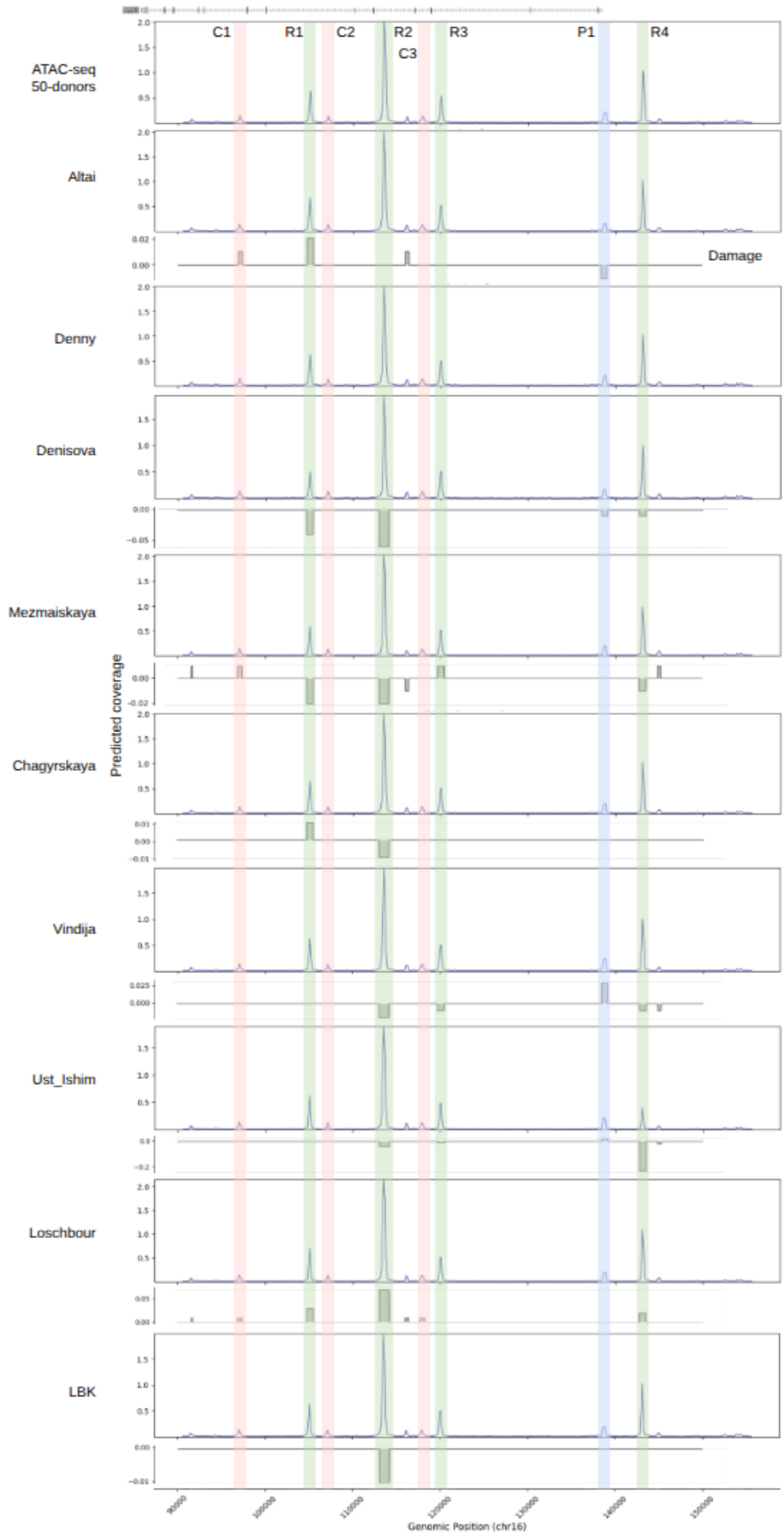

C1
R1
C2
R2
C3
R3
P1
R4
ATAC-seq
50-donors
Altai
Damage
Denny
Denisova
Mezmaiskaya
Chagyrskaya
Vindija
Ust_Ishim
Loschbour
LBK
Predicted coverage
Genomic Position (chr16)

Figure 5.15: Predicted alpha-globin chromatin accessibility in ancient hominin sequences. Erythroid open chromatin profiles were predicted for nine ancient hominin genomes reconstructed from variant data, with the control track (ATAC-seq from the 50-donor dataset) shown at the top. Where differences in signal were observed, a corresponding damage plot is displayed below each genome. In these plots, bars represent a damage score positioned along the x-axis according to the affected peak. Positive values indicate a gain of signal, while negative values indicate a loss of signal at the corresponding region. Key regulatory elements are highlighted, with promoters in blue, enhancer-like elements in green, and CTCF sites in red. An interactive UCSC Genome Browser session with the predicted signal tracks is available at: https://genome.ucsc.edu/s/arodrig/Hominins_NPRL3.

Coverage values were then extracted from a merged set of peaks across all genomes, and damage scores were calculated as the absolute difference between the predicted coverage in each ancient genome and the control at each peak.

Although this approach yields lower-magnitude values, it was chosen to better reflect the true scale of differences relative to the predicted signal. Given the minimal deviations observed, this formulation provides a clearer and more interpretable representation of the limited impact detected. Examining the elements in the locus from 5' to 3' relative to NPRL3 (as labelled in Figure 5.15), several subtle changes can be observed.

Upstream of the first CTCF site (C1), minor gains of accessibility are observed at two initial sites: a small peak with no known regulatory function and the first CTCF site (C1) in the Mezmaiskaya and Loschbour genomes. At the R1 enhancer, both gains and losses are observed: Altai, Chagyrskaya, and Loschbour display increased accessibility, with the largest gain in the Altai genome ($\sim 33\%$), whereas Mezmaiskaya and Denisova show reductions in accessibility, reaching up to $\sim 58\%$ loss in the Denisova genome.

R2 predominantly shows reduced accessibility across most genomes, with the largest decrease (~ 0.05 coverage) observed in Denisova. In contrast, Loschbour is the only genome displaying a comparable gain (~ 0.05, $\sim 3\%$). Other genomes, such as Altai and Denny, show minimal or no change. A nearby uncharacterised peak between R2 and C3 shows minor gains in Altai and Loschbour and a slight loss in Mezmaiskaya. The C3 site itself remains largely stable, with only a small gain in Loschbour.

At R3, only two notable changes are observed: a $\sim 2\%$ gain in Mezmaiskaya and a $\sim 3\%$ loss in Vindija. The NPRL3 promoter shows slightly higher variability, with a $\sim 4\%$ loss in Altai and a $\sim 10\%$ gain in Vindija. Finally, R4 predominantly exhibits losses of accessibility,

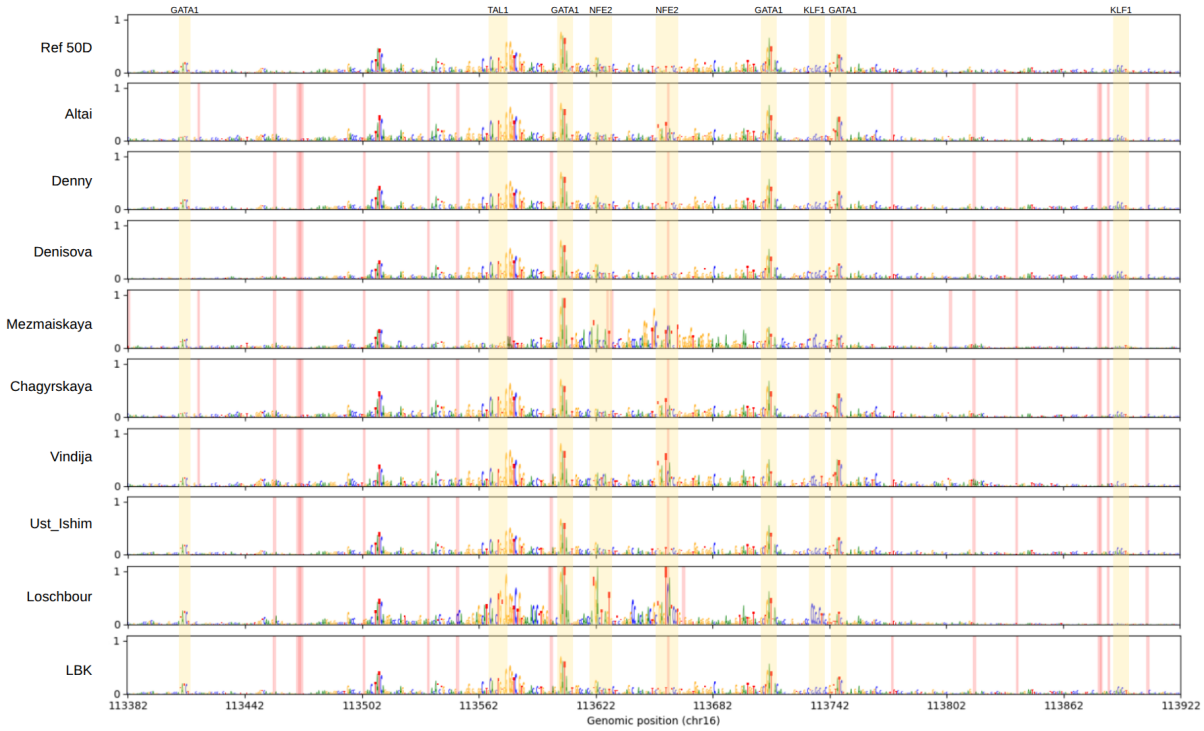


Figure 5.16: Nucleotide importance scores for ancient hominin R2 sequences with their variants. Feature attributions were derived from the predicted chromatin accessibility profiles for each hominin subject, specifically for the R2 peak prediction, with the control track (ATAC-seq from the 50-donor dataset) displayed at the top. Identified variants are highlighted in light red. Yellow highlights indicate known motifs identified using TomTom.

particularly in Denisova, Mezmaiskaya, Vindija, and Ust Ishim, with the largest effect reaching a $\sim 50\%$ reduction in Ust Ishim.

To obtain a more detailed view of how hominin genomic variants influence the predicted activity of regulatory elements, I extracted nucleotide importance scores using DeepLIFT from the chromatin accessibility predictions for the R2 enhancer (Figure 5.16). The majority of the identified SNVs are shared across the nine hominin genomes, which likely explains the high similarity observed among the reconstructed chromatin accessibility tracks. None of these variants were associated with known blood traits, except rs11248850 (chr16:113599, G to T substitution), which was shared by all subjects and has been linked to lower red blood cell counts and higher erythrocyte haemoglobin concentrations.

Although some of these importance changes coincide with SNVs positioned directly within functional motifs, and are therefore likely linked to local sequence perturbations, they may also arise indirectly from variants located outside the annotated functional sites. Such distal effects could alter the broader sequence context learned by the model, thereby modulating the

contribution of nearby motifs without directly disrupting them. A systematic dissection of these long-range or context-dependent effects, however, falls beyond the scope of this thesis.

Considering the genome-wide variant effect analysis in Section 3.5.2, where HYPER achieved $\sim 85\%$ accuracy in identifying variant effects, with strong precision for both gains (79%) and losses (87%), these results suggest that the observed differences are mostly reliable but remain limited in magnitude. Overall, the alpha-globin regulatory landscape appears highly conserved across ancient hominin genomes, with only minor, localised variations that are unlikely to substantially alter regulatory function.

5.6.2 Extended analysis throughout vertebrates

While the analysis of ancient hominin genomes offers insight into subtle regulatory variation across closely related lineages, it remains constrained by their high degree of conservation. To broaden this perspective, I next extend the analysis to more distantly related species, including other mammals and birds. This exploratory step aims to probe how far HYPER can generalise beyond familiar genomic contexts, and to investigate how regulatory patterns evolve as sequence divergence increases across larger evolutionary distances.

For this analysis, the human erythrocyte model was used due to its high accuracy and consistency in predicting epigenetic landscapes across species. Among the TFs examined, reconstructed TAL1 binding profiles showed the highest prediction accuracy compared to GATA1 and NF-E2. Consequently, this model was applied to predict TAL1 binding profiles and open chromatin across the alpha-globin locus in all 26 species. Supplementary Table S3 provides links to UCSC genome browsers displaying predictions for the five classes (open chromatin, TAL1, GATA1, NF-E2, and CTCF) across the 23 species.

To align the reconstructed TAL1 protein binding profiles to the human reference, I used two highly conserved exons flanking the R2 enhancers: exons 4 (ENSE00003538360) and 6 (ENSE00003466209). For this, I ran BLAT with a minimum score of 20 and a minimum identity of 60. The reconstructed TAL1 protein binding profiles (Figure 5.17) suggest that the R2 position is largely conserved across the alpha-globin locus in mammals. Overall, the predicted locus structure places these elements within introns in positions comparable to those in humans, indicating a high degree of potential functional conservation.

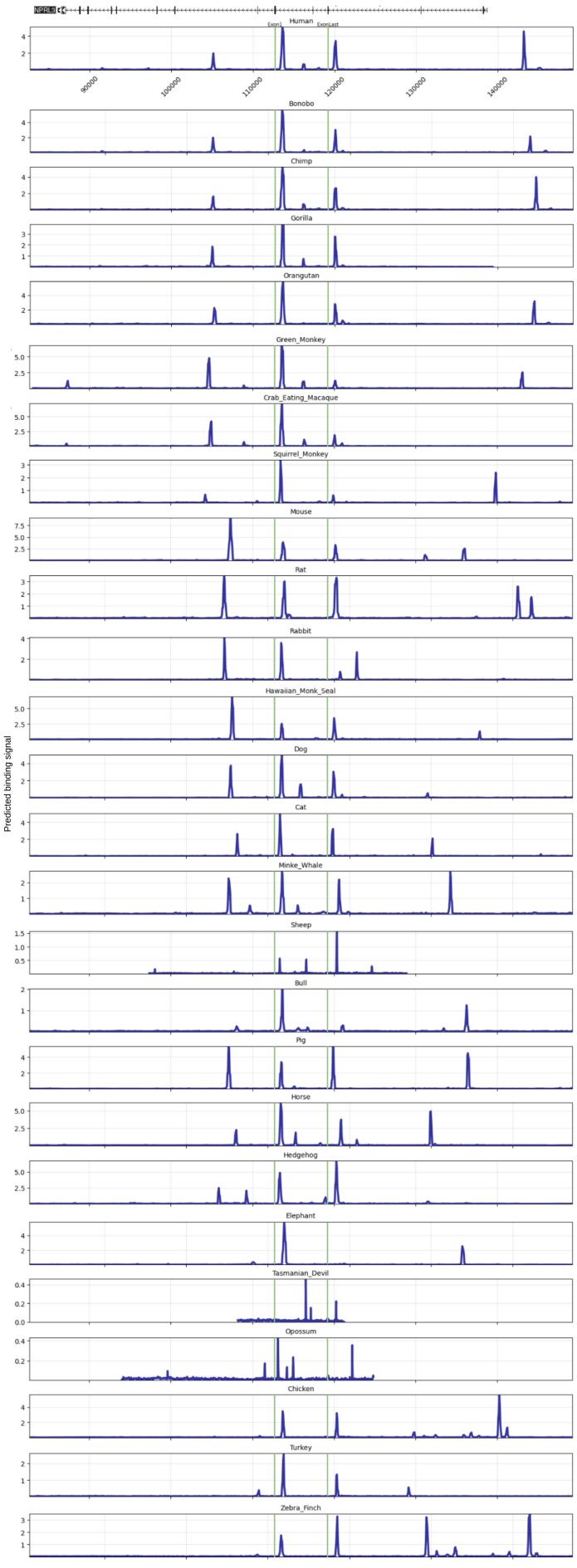

NPRL3
Human
Exon1
ExonLast
80000
100000
110000
120000
130000
140000
Bonobo
Chimp
Gorilla
Orangutan
Green_Monkey
Crab_Eating_Macaque
Squirrel_Monkey
Mouse
Rat
Rabbit
Hawaiian_Monk_Seal
Dog
Cat
Minke_Whale
Sheep
Bull
Pig
Horse
Hedgehog
Elephant
Tasmanian_Devil
Opossum
Chicken
Turkey
Zebra_Finch
Predicted binding signal

Figure 5.17: Predicted TAL1 binding profiles across the alpha-globin locus in multiple species aligned to two exons (green vertical lines) flanking the R2 enhancer in human. The ChIP-seq tracks were reconstructed using the human erythrocyte HYPER model for 26 species, including human.

Furthermore, HYPER captures lineage-specific variation by identifying regulatory elements that are present in some species but absent in others.

To complement the previous analysis, I next evaluated HYPER's ability to predict conserved regulatory elements based on sequence similarity within the alpha-globin locus. Candidate enhancer-like elements were identified by segmenting the predicted signal in each species into enriched and non-enriched regions using a threshold based on the top 10% of predicted coverage values, thereby allowing a more flexible level of stringency and the detection of a broader set of elements. Conventional peak-calling tools, such as LanceOtron, were not suitable in this context because the predicted coverage values and their standard deviations were very low, making it difficult to distinguish true signal from background. This likely reflects the fact that the predictions were inferred from models trained on other species.

Predicted enriched regions closely recapitulate regulatory elements within the reconstructed chromatin accessibility profiles of each species. To allow more flexible detection of residual regulatory features, even regions with relatively low enrichment were retained as peaks. Within this framework, and as illustrated in Figure 5.18, all enriched regions across the profiles were catalogued, enabling candidate functionally conserved elements to be mapped to specific genomic coordinates.

Building on this, the next step was to examine specific enhancer-like elements in each species, using known regulatory elements from human, mouse, and chicken as references, informed by their evolutionary relationships to the species analysed. This comparative approach highlights both conserved and lineage-specific features as reflected in the reconstructed open chromatin profiles. For example, turkey lacks an R1 enhancer, consistent with observations in chicken. Given their relatively short evolutionary divergence ($\sim$ 28 million years [236]), this supports the idea that the absence of R1 is a shared feature within this lineage. In contrast, the rat profile appears to contain a putative Rm facilitator, previously described as mouse-specific and absent in humans. Given that rats and mice are closely related rodents (diverging around 18 million years ago [237]), this observation is biologically plausible.

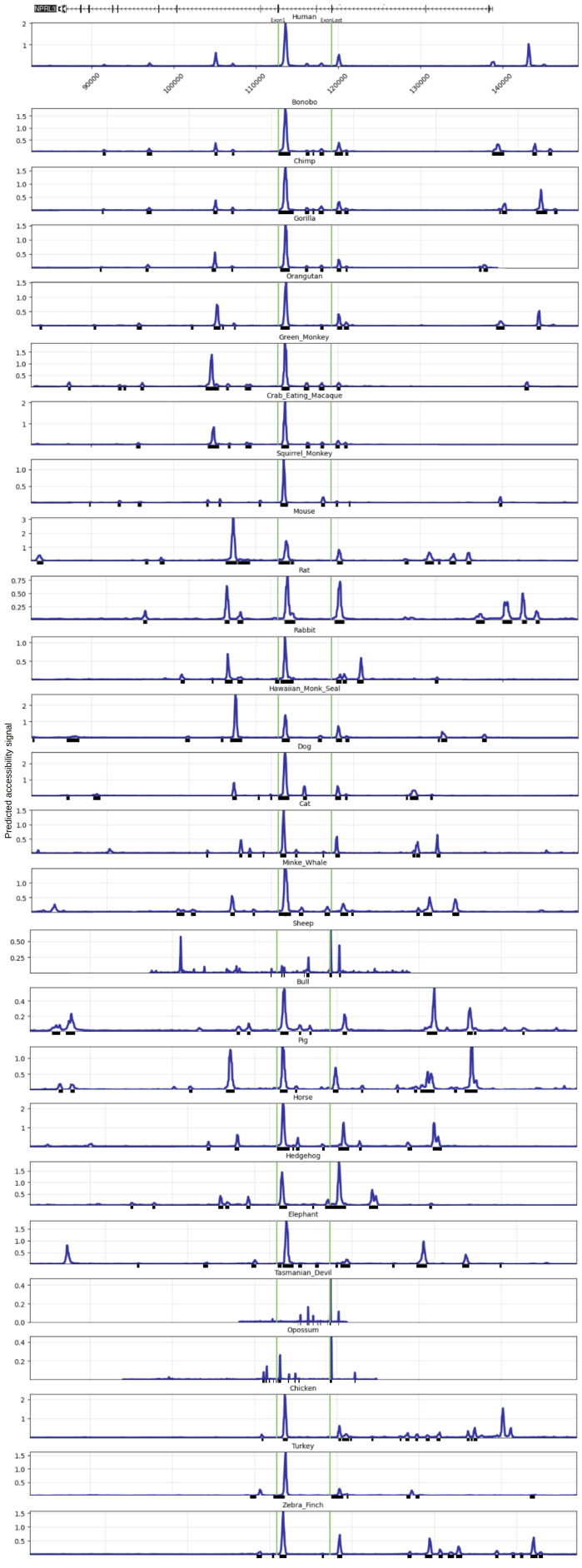

NPRL3
Human
Exon1
ExonLast
90000
100000
110000
120000
130000
140000
Bonobo
Chimp
Gorilla
Orangutan
Green_Monkey
Crab_Eating_Macaque
Squirrel_Monkey
Mouse
Rat
Rabbit
Hawaiian_Monk_Seal
Dog
Cat
Minke_Whale
Sheep
Bull
Pig
Horse
Hedgehog
Elephant
Tasmanian_Devil
Opossum
Chicken
Turkey
Zebra_Finch
Predicted accessibility signal

Figure 5.18: Predicted open chromatin profiles across the alpha-globin locus in multiple species aligned to two exons (green vertical lines) flanking the R2 enhancer in human. Chromatin accessibility tracks were reconstructed using the human erythrocyte HYPER model for 26 species, including human. Predicted enriched regions are shown as black boxes below the chromatin tracks.

Similarly, some elements appear to be retained in certain lineages while lost in others over larger evolutionary distances. For instance, an enriched region in the pig open chromatin profile aligns with the position of the human R4 element, suggesting potential functional conservation. This observation implies that, despite sequence divergence, key regulatory functions may be preserved in specific lineages while being lost or substantially altered in others; however, further verification through sequence conservation analysis would be needed to support this interpretation.

To measure sequence conservation of these elements, BLAT was applied, with parameters set to a minimum score of 20, minimum identity of 70, tile size of 11, and step size of 5, to identify regions within each species' alpha-globin locus that share sequence conservation with known human, mouse, and chicken enhancer-like elements. These conserved regions were then intersected with the segmented enriched regions to identify candidate elements that are both functionally active and sequence-conserved. The resulting elements are summarised in Table 5.1.

These results show that species with more recent divergence tend to exhibit higher sequence similarity in enhancer-like elements within the alpha-globin locus, as seen in comparisons such as mouse–rat, mouse–rabbit, and chicken–turkey or chicken-zebra finch. Although some species, such as cow and elephant, did not show elements that were both sequence- and functionally conserved, the identified candidate elements remain plausible, given their high sequence similarity and strong enrichment in predicted coverage.

In addition, the previously suggested presence of an Rm element in rat is further supported by these findings, with $\sim 87.5\%$ sequence similarity observed between the mouse Rm and the rat candidate region. This suggests that the element may be more broadly conserved within rodents than previously assumed. Similarly, the pig R4 element shows strong sequence conservation, with a BLAT similarity score of 91.49%. Considering that pigs and humans diverged approximately 100 million years ago, this observation is particularly notable, as it implies that certain

regulatory elements can be maintained over long evolutionary timescales despite divergence in surrounding sequences.

The apparent loss or modification of such elements in other species raises important questions

Species	Sequence similarity (%)					Original species
	R1	R2	R3	R4	Rm	
Bonobo	97.8	98.4	97.4	99.2	-	Human
Chimp	97.8	98.2	97.39	99.2	-	
Gorilla	99.0	98.8	98.4	-	-	
Orangutan	96.8	97.4	95.4	97.36	-	
Green monkey	-	-	-	-	-	
Crab eating macaque	92.6	95.99	92.34	-	-	
Squirrel monkey	90.89	90.65	91.89	91.1	-	
Rat	91.24	91.53	94.42	89.87	87.55	Mouse
Rabbit	-	88.37	-	-	-	
Hawaiian monk seal	-	85.5	-	77.5	-	Human
Dog	-	83.56	80.8	-	-	
Cat	91.89	83.5	88.14	-	-	
Minke whale	83.33	86.75	-	82.28	-	
Sheep	-	-	-	96.67	-	
Cow	-	-	-	-	-	
Pig	-	-	-	91.49	-	
Horse	-	82.67	-	90.0	-	
Hedgehog	-	-	-	-	-	
Elephant	-	-	-	-	-	
Tasmanian devil	-	-	-	-	-	
Opossum	-	-	-	-	-	
Turkey	-	96.88	93.36	-	-	Chicken
Zebra finch	-	83.57	78.76	-	-	

Table 5.1: Functionally and sequence-conserved elements within the alpha-globin locus. Candidate elements were identified by intersecting predicted enriched regions with sequences conserved at least by 70% across human, mouse, and chicken, and their similarity was assessed across 23 species.

about the evolutionary dynamics of regulatory landscapes and the selective pressures influencing their retention or disappearance.

These results highlight HYPER's ability to capture both sequence and functional conservation across species, while also pointing to previously uncharacterised regulatory elements. Notably, the model is able to identify key regulatory elements even when applied to genomes that differ from those used during training, indicating that predictive performance is not solely dependent on strong sequence identity. This suggests that HYPER can learn statistical sequence patterns associated with regulatory activity that remain informative despite evolutionary sequence divergence.

These findings open the door to using sequence-based models as tools for regulatory element discovery in species with limited or no epigenomic data, and for revisiting existing annotations across genomes. Future work could focus on experimental validation of these candidates, as well as extending this framework genome-wide to systematically map conserved and lineage-specific regulatory elements. Such efforts could provide deeper insight into how regulatory landscapes evolve, how gene expression is fine-tuned across species, and how sequence variation translates into functional divergence.

5.7 Discussion and conclusions

This chapter investigated HYPER's ability to learn and transfer regulatory grammar across species, using the alpha-globin locus as a model system. Building on previous validation of sequence-based predictions, the analysis focused on whether regulatory patterns learned in one species can be generalised to others across varying evolutionary distances.

First, species-specific models were trained or fine-tuned for mouse and chicken erythroid cells, for which experimental data were available, extending the human-trained framework to both a closely related mammal ($\sim$ 100 million years of divergence) and a more distant vertebrate ($\sim$ 300 million years of divergence). Because HYPER was already pretrained on both human and mouse data, evaluating its performance in these species alone did not fully assess its ability to generalise to out-of-distribution species. An additional test in chicken therefore provided a more stringent assessment of HYPER's cross-species generalisation beyond the species represented in its pretraining data. These models accurately reconstructed epigenetic

profiles, capturing key regulatory elements such as enhancers and promoters, while also highlighting limitations in modelling transcription-associated signals. Differences in performance were likely driven by data quality, with noisier datasets leading to increased background signal rather than fundamental limitations in the model.

Next, cross-species prediction was systematically evaluated by applying models to genomes they were not trained on, including predicting human regulatory landscapes using mouse and chicken models, and vice versa. Across all settings, core regulatory features, particularly distal enhancers and CTCF sites, were consistently recovered, demonstrating that HYPER captures a conserved regulatory grammar that generalises beyond species-specific sequence patterns. While species-specific models achieved higher fidelity within their own domain, cross-species models retained strong biological interpretability and, in some cases, revealed sensitivity to regulatory elements not captured by other models.

To interpret these predictions, nucleotide importance analysis (DeepLIFTShap) was performed across MCS-R elements, revealing conserved motif architectures despite sequence divergence. This was complemented by motif enrichment analysis (FIMO/TomTom), which consistently identified canonical erythroid regulators, particularly GATA1::TAL1, as the dominant regulatory drivers across species. The primary enhancer R2 showed the strongest conservation in both structure and motif composition, while facilitator elements (R3, R4, Rm) exhibited greater variability and lineage-specific adaptations. Cross-species motif analyses further demonstrated that conserved motifs are preferentially recovered even in out-of-distribution predictions, while additional motifs may reflect either biological divergence or methodological ambiguity.

Finally, the framework was extended to genomes lacking experimental data, including ancient hominins and additional mammals and birds. Predictions in Neanderthal, Denisovan, and early Homo Sapiens genomes revealed minimal regulatory differences relative to modern humans, consistent with their recent divergence and supporting strong conservation of the alpha-globin regulatory landscape. In contrast, broader comparisons across mammals and birds uncovered both conserved regulatory structures and lineage-specific elements. Notably, the identification of putative conserved elements, such as an Rm-like region in rat and an R4-like region in pig, demonstrates HYPER's ability to capture both sequence and functional conservation, while also suggesting previously unannotated regulatory regions.

Insights into model learning

The cross-species analyses presented in this chapter suggest that HYPER does not simply memorise species-specific sequence patterns, but instead learns multiple partially transferable layers of regulatory representation. At the most local level, the models consistently recover canonical erythroid TF motifs, particularly GATA1::TAL1 and related E-box-associated patterns, indicating robust learning of sequence features associated with core erythroid regulatory activity.

Beyond individual motifs, the nucleotide importance and motif enrichment analyses further suggest that the models capture aspects of combinatorial regulatory syntax, including motif clustering, spacing, and cooperative relationships between TFBSs. This is particularly evident within conserved enhancer-like elements such as R2, where groups of motifs remain consistently enriched across species despite underlying sequence divergence.

At a broader scale, the models also appear to learn sequence features associated with chromatin context and higher-order regulatory organisation. The ability to reconstruct both active and repressive chromatin domains, including enhancer regions, LADs, and pericentromeric regions, suggests that the learned representations incorporate sequence patterns correlated with distinct epigenomic environments, even when these relationships are only indirectly encoded in the underlying DNA sequence.

Finally, the successful transfer of predictions across large evolutionary distances implies that the models capture certain phylogenetically conserved regularities within regulatory systems. However, these representations should not be interpreted as direct mechanistic models of regulatory biology. Instead, they likely reflect statistical abstractions of partially conserved regulatory principles shaped by motif conservation, combinatorial redundancy, chromatin-associated sequence composition, and shared TF binding preferences across vertebrate genomes.

Together, these observations suggest that sequence-to-epigenome models learn a hierarchy of regulatory representations spanning local motif recognition, combinatorial enhancer organisation, chromatin-associated sequence context, and partially conserved evolutionary patterns.

Limitations of sequence-only modelling

Although sequence-to-epigenome models capture important regulatory features directly from DNA sequence, regulatory activity *in vivo* is influenced by multiple biological processes that are only partially encoded within the underlying nucleotide sequence. Consequently, sequence-only models necessarily provide an incomplete representation of gene regulation.

One important limitation involves three-dimensional chromatin organisation. Long-range enhancer–promoter interactions, chromatin looping, and higher-order nuclear architecture can strongly influence regulatory activity, yet many of these interactions depend on cell-state-specific structural configurations that are not fully predictable from local sequence alone. Similarly, nuclear compartmentalisation and chromosomal positioning within the nucleus contribute to gene regulation through mechanisms such as lamina association, TF organisation, and spatial chromatin accessibility.

In addition, regulatory activity is highly dynamic and depends on temporal variation in TF concentration, signalling pathways, developmental stage, and cellular environment. Sequence-based models generally operate on static representations of DNA and therefore cannot explicitly model dynamic transitions between regulatory states, such as differentiation, stress responses, or transient chromatin remodelling events.

Chromatin accessibility itself is also influenced by ATP-dependent chromatin remodelling complexes, histone modifications, nucleosome positioning, and cooperative interactions between regulatory proteins. Although some sequence features associated with these processes may be indirectly learned by the models, many aspects of chromatin state emerge from complex biochemical interactions that extend beyond DNA sequence alone.

Consequently, the predictions generated by HYPER and related architectures should be interpreted as probabilistic representations of regulatory potential derived from sequence-associated statistical regularities, rather than complete mechanistic reconstructions of cellular regulatory systems. Despite these limitations, the strong predictive performance observed throughout this thesis demonstrates that a substantial proportion of regulatory information is sufficiently encoded within genomic sequence to support biologically meaningful inference across species and regulatory contexts.

These results can be summarised in the following key conclusions:

- HYPER learns a conserved regulatory grammar that generalises across species, enabling accurate prediction in genomes not represented during training.
- Key regulatory elements can still be identified despite substantial evolutionary sequence divergence between species.
- Core elements, particularly R2 and GATA1::TAL1 motifs, are robustly preserved despite sequence divergence.
- Model performance is driven by data quality, with noise affecting the signal but not the underlying regulatory structure.
- Cross-species predictions remain biologically interpretable, even when applied to genomes outside the training distribution.
- Regulatory conservation operates at the level of motif logic and function, rather than strict sequence identity.
- The framework enables regulatory annotation and discovery in species lacking experimental epigenomic data.

These findings position HYPER as a powerful framework not only for prediction but also for studying the evolution of gene regulation. The ability to recover regulatory structure, motif composition, and functional elements across species suggests that sequence-based models can serve as tools for exploring genomes where experimental data are scarce or unavailable.

6 | Discussion

6.1 Overview and objectives

Gene regulation arises from complex interactions between multiple elements, many of which are located far from the genes they control, suggesting that regulatory instructions are encoded within the DNA sequence itself. Decoding these instructions to explain diverse and context-dependent gene expression programs remains a central challenge in molecular biology. These regulatory instructions can be conceptualised as a structured system, often referred to as regulatory grammar, which defines how combinations of sequence features—such as TFBSs, their arrangement, and local context—govern gene expression across cell types. This encompasses not only the presence of specific TFs, but also other features such as chromatin organisation and local sequence context. However, capturing this regulatory grammar in a systematic and generalisable way remains an ongoing challenge in the field.

From a computational perspective, DL approaches provide a powerful means to capture complex, nonlinear relationships directly from DNA sequence. However, the rapid pace at which new and improved methods are developed has outstripped their systematic evaluation in molecular biology, and there is still no clear consensus on which architectures best capture regulatory features across different biological contexts. Consequently, an additional challenge lies in identifying the most appropriate ML framework to model these processes.

Building on these challenges, translating insights from model organisms to human biology is a fundamental objective in molecular biology, with important implications for advancing healthcare and medicine. In the context of gene regulation, this requires identifying conserved regulatory principles and improving the interpretation of genetic variation. Accordingly, even if the regulatory grammar of a given cell type or tissue can be learned, a key question is the extent to which this knowledge generalises across species within the same cellular context. In such cases, the model would effectively map DNA sequence features to regulatory function, yet it remains unclear how comparable these mappings are between organisms. Although many regulatory mechanisms are conserved, differences in sequence composition and context-specific effects introduce additional complexity. Therefore, determining whether regulatory grammar learned in one species can be transferred to another remains an important open question.

In this context, this work addresses these challenges by applying and evaluating DL models to learn regulatory patterns from DNA sequence, assess their generalisability, and characterise the functional impact of genetic variation.

Under this framework, each chapter addresses a specific challenge, with subsequent chapters building on the findings. Chapter 3 establishes the computational foundation by benchmarking DL architectures for regulatory prediction, comparing CNN-based and transformer-based models. It evaluates their ability to predict CTCF binding, assesses genome-wide generalisation, and defines a robust validation framework for epigenetic signal prediction.

Building on this, Chapter 4 applies the selected model to a more complex, cell-type-specific context by focusing on erythroid biology. The model is trained on TF binding and chromatin accessibility data to assess its ability to reconstruct regulatory landscapes and predict the impact of genetic variation on key erythroid regulators.

Finally, Chapter 5 extends this framework to an evolutionary perspective, investigating the extent to which regulatory grammar is conserved across species. By training models on mouse and chicken data, it explores patterns of conservation and divergence, with particular emphasis on the alpha-globin locus and recent regulatory changes in the human lineage.

6.2 Approach and analysis

This work establishes a robust benchmarking framework for evaluating DL models in regulatory genomics, enabling a systematic comparison of three models: DeepSEA, BPNet, and HYPER for predicting CTCF binding from high-resolution erythrocyte data. By introducing a multi-level validation strategy—spanning signal reconstruction, motif discovery, and variant effect prediction—it provides a comprehensive assessment of model performance, particularly in capturing peak shape, resolution, and variant-induced changes. This framework not only identifies HYPER as the most reliable model but also contributes a generalisable pipeline for evaluating DL approaches in epigenetic signal prediction.

Building on this, this study advances the application of HYPER to more complex, cell-type-specific regulatory contexts. It introduces an extended validation strategy for modelling TF binding and chromatin accessibility in erythroid cells, including a novel approach based on

enriched region segmentation to assess reconstructed signals. In addition, it demonstrates the model's ability to capture both active and repressive chromatin states and to generalise to genetic variation. Through the integration of nucleotide-level attribution and motif analysis, this work further contributes to the interpretability of DL predictions by linking model outputs to underlying biological mechanisms.

Finally, this work provides new insights into the conservation and transferability of regulatory grammar across species. By applying models across human, mouse, and chicken genomes, it demonstrates that core regulatory features can be learned and generalised beyond species-specific contexts. The combination of cross-species prediction, nucleotide importance, and motif analysis enables the identification of both conserved regulatory architectures and lineage-specific adaptations. Extending this framework to species lacking experimental data further highlights its potential to uncover previously unannotated regulatory elements, offering a scalable approach for studying gene regulation across evolutionary distances.

Overall, this work advances both the methodological and biological understanding of sequence-driven gene regulation, offering a scalable framework to study regulatory mechanisms and their variation across cell types and species.

6.3 Key findings and implications

The benchmarking results indicate that transformer-based architectures provide a clear advantage over CNN-based models for modelling gene regulation from DNA sequence. In particular, HYPER's ability to capture peak shape, signal resolution, and variant-induced effects suggests that accurate modelling of regulatory activity benefits from incorporating long-range interactions, which are not fully captured by traditional convolutional approaches. However, HYPER's improved performance may also be partly attributable to its pretraining on a large epigenetic dataset, which could provide a richer prior over regulatory sequence patterns and enhance its ability to learn regulatory relationships beyond those captured through task-specific training alone.

Extending this framework to erythroid-specific, more complex regulatory features further demonstrates that regulatory predictions are highly dependent on data quality and signal definition. The model's strong performance in chromatin accessibility, alongside its ability to distinguish

active and repressive chromatin states, highlights that DNA sequence encodes both activating and silencing regulatory information. Moreover, the consistent identification of GATA1–TAL1 as a central driver, together with the sensitivity of accessibility to single-nucleotide changes, reinforces the importance of combinatorial and context-dependent TF binding. The integration of attribution methods also shows that these models can move beyond prediction, providing mechanistic insight at base-pair resolution.

Finally, the cross-species analysis reveals that regulatory grammar is largely conserved at the level of motif interactions rather than exact sequence identity. The ability of HYPER to generalise across species and recover core regulatory elements supports the existence of transferable regulatory principles, while also highlighting the presence of lineage-specific adaptations. This demonstrates that sequence-based models can be used not only to predict regulatory activity, but also to study its evolution and to extend regulatory annotation to species lacking experimental data.

6.4 Future work

While this work provides a comprehensive framework for modelling and interpreting gene regulation from DNA sequence, several avenues remain for further development. This section outlines key directions for future work aimed at improving the biological validity, generalisability, and applicability of these approaches.

6.4.1 Experimental validation

A direct next step would be to generate experimental data to validate the findings from the cross-species analysis. This would involve producing epigenomic datasets in additional vertebrate species analysed in this study, particularly those lacking existing data, such as chicken or other non-model organisms. In particular, generating high-resolution TFBS data would enable a precise assessment of whether HYPER accurately predicts both the location and intensity of regulatory signals. Such validation would be essential to confirm the biological relevance of the inferred cross-species regulatory grammar and to determine the extent to which model predictions reflect true regulatory activity.

The most immediate species to extend this analysis is chicken, as a model has already been trained on erythroid chicken data, although currently limited to CTCF and lacking additional TFs. Expanding this dataset to include other TFs would provide a valuable validation framework. Additionally, pig represents a strong candidate for extending the analysis to a non-rodent mammalian system. This is particularly relevant given the identification of an R4-like element in pig in Section 5.6.2, and the overall similarity of the pig alpha-globin locus to that of humans, making it a compelling model for further investigation.

6.4.2 Data quality and resolution

The availability of high-quality epigenetic datasets—particularly for TF binding—remains limited. Two main strategies can be considered to address this constraint. The first, and most rigorous, is the generation of new experimental data, although this requires substantial time, cost, and technical resources. The second, more immediate approach involves leveraging computational methods to expand existing datasets, including data augmentation techniques such as *in silico* sequence perturbations, reverse-complement augmentation, and synthetic data generation, which have been successfully applied in previous studies (e.g. BPNet, DeepSEA).

Incorporating systematically mutated sequences, as performed in Section 4.4, together with their corresponding experimental signals, could help expand the range of causal relationships between sequence and function captured by the model. Similarly, integrating reverse-complement sequences would effectively double the dataset size without altering the underlying biology, thereby improving representation and generalisation, although this strategy has already been implemented in Basenji. Additionally, generative models such as generative adversarial networks (GANs) have been used to create synthetic DNA sequences and improve class balance in genomic ML tasks [238], and could provide a further avenue for data augmentation.

While these approaches can enhance model performance and robustness, their outputs would still require careful validation to ensure biological relevance.

6.4.3 Extending the biological scope

This study has focused on erythroid cells, particularly erythrocytes, which represent a well-characterised system, well-defined differentiation pathways, and extensively studied regulatory

elements. The alpha-globin locus was used as a primary validation system for similar reasons, given its well-understood regulatory architecture. While this provides a strong foundation for model development and validation, it also raises the question of how well these approaches generalise to other genomic loci and biological contexts. In particular, it would be important to evaluate HYPER's performance on loci with less conserved sequences or different regulatory architectures.

Loci such as the HOX gene clusters exhibit strong conservation at the gene level, while their regulatory elements remain flexible and dynamic, varying across developmental stages and cell types, particularly during embryogenesis and differentiation of stem and progenitor cells [239]. In contrast, loci such as MYC contain cancer-associated regulatory regions that are highly context-dependent, with enhancer activity varying widely across cell types, including proliferative tissues, stem cells, and a range of tumour environments [240]. Similarly, FOXP2 is highly conserved at the protein level but displays regulatory differences across species, particularly in neural cell types. It plays a key role in brain development, especially in cortical neurons and basal ganglia circuits, and its disruption is associated with severe speech and language disorders [241].

From a cellular perspective, a natural extension of this work is to investigate additional cell types. One such example is the regulation of erythropoietin (EPO) production under hypoxic conditions. In humans, identifying the cellular source of EPO has been challenging due to its transient expression and rapid degradation following reoxygenation, limiting experimental accessibility.

Recent studies in mouse models have provided new insights through the development of a hypoxia-inducible lineage tracing system. In this approach, cells expressing EPO under hypoxic conditions activate a tamoxifen-inducible Cre recombinase (CreERT2), leading to permanent labelling with a fluorescent reporter (tdTomato). This enables the tracking of EPO-producing cells and their descendants over time. Using single-cell multi-omic profiling, these studies identified a distinct population of kidney fibroblast-like cells responsible for EPO production, termed Norn cells [242], which appear to be evolutionarily conserved.

Building on this, an important question is whether models such as HYPER can capture the regulatory grammar underlying these pathways. In particular, evaluating whether such models can identify analogous regulatory signatures in human data and predict the existence of Norn-like

cell populations would represent a significant step toward extending sequence-based modelling to less characterised systems. This approach could enable the investigation of regulatory mechanisms in cell populations that are experimentally inaccessible or not yet identified in humans, while reducing the reliance on direct human experimentation.

6.4.4 Attribution methods for transformers

Recent work has raised questions about whether gradient-based methods are the most appropriate approach for feature attribution in transformer-based models. These concerns arise because transformers rely on attention mechanisms and complex, context-dependent interactions between sequence elements. Gradient-based methods capture local sensitivity of the output to small input perturbations, but may not fully reflect the importance of features that emerge through nonlinear and long-range interactions mediated by attention layers. As a result, these methods can underestimate the contribution of sequence features involved in cooperative or context-dependent regulatory effects [243].

Although there is currently no strong evidence directly comparing gradient-based and transformer-specific attribution methods, or demonstrating clear superiority of one over the other, a few transformer-tailored approaches have been proposed in recent years [243, 244]. However, these methods remain relatively underexplored and lack thorough validation in genomic contexts. In addition, many of these approaches have not yet been integrated into widely used software packages, making their practical application more challenging.

Nevertheless, systematically comparing these attribution methods would be a valuable direction for future work, as it could clarify whether transformer-specific approaches provide more accurate or biologically meaningful interpretations of model predictions.

6.4.5 Integration of genomic languages

An additional direction for future work is the integration of large language model (LLM) approaches for genomic sequence modelling. These models, trained on large-scale sequence data, have the potential to capture long-range dependencies and learn generalisable regulatory representations across cell types and species. In particular, their token-based architectures offer a

natural framework for modelling regulatory grammar, aligning closely with the objectives of this study. Incorporating such models into the existing benchmarking framework would enable systematic comparisons between task-specific and pretrained architectures, particularly in low-data settings, while also allowing the evaluation of more recent models trained on larger and more diverse datasets.

However, their high computational cost and limited interpretability remain important challenges. In many academic settings, the resources required to train such models are not readily available, although their application may become more feasible as computational methods and infrastructure continue to advance.

6.4.6 Deeper cross-species analysis

The results of this study suggest that HYPER is capable of learning regulatory grammar, defined as the set of rules governing gene expression encoded within DNA sequences in a cell-type-specific context. As shown previously, the model can also generalise across sequence variants and accurately predict their functional effects, opening the door to a wide range of potential applications.

Given that regulatory networks are composed of interacting elements whose instructions are largely encoded in DNA sequence, a model that effectively captures this regulatory grammar could be used to predict the consequences of regulatory rewiring—that is, changes in the interactions between enhancers, TFs, promoters, and chromatin structure that control gene expression. Since HYPER has demonstrated the ability to simulate the effects of genetic variants, it should also be capable of predicting the impact of more complex sequence perturbations, such as altering the spacing or orientation of motifs, as well as the addition or removal of regulatory elements.

In this context, the model could be used to systematically evaluate how such changes reshape regulatory interactions and influence downstream chromatin accessibility or TF binding. Therefore, this approach provides a framework for linking sequence variation to changes in regulatory architecture, offering a means to computationally explore and interpret regulatory rewiring events.

Building on this, incorporating a temporal dimension would enable the exploration of how reg-

ulatory networks for a given gene evolve over time, both within the human lineage and across species. This would allow the investigation of how regulatory elements are gained, lost, or modified throughout evolution, providing insights into the mechanisms underlying regulatory conservation and divergence. Such an approach could help identify key events of regulatory rewiring, link sequence changes to functional outcomes, and improve our understanding of how gene regulation adapts over evolutionary timescales.

Implementing this framework would require extensive high-quality datasets, detailed characterisation of regulatory elements, and a comprehensive understanding of regulatory network properties across multiple species and time points, which currently remains a significant challenge.

6.4.7 Motif-based representation of DNA sequences

From a computational perspective, a model capable of learning regulatory grammar would enable the representation of DNA sequences in terms of biologically meaningful features, such as cell-type-specific motif importance associated with chromatin accessibility or TF binding.

As a preliminary exploration of this idea, I extracted k-mer importance profiles from different regulatory elements, including enhancers, promoters, and CTCF sites identified using LanceOtron and annotated with REgulamentary [245], in the human genome. These were then compared with the representations of the alpha-globin regulatory elements studied throughout this thesis. To further expand the analysis, equivalent elements from mouse and chicken were also included. The resulting feature representations were reduced in dimensionality using UMAP to assess their similarity.

Interestingly, all alpha-globin regulatory elements clustered together, regardless of whether they originated from human, mouse, or chicken (Figure 6.1). This suggests that the model captures a shared, functionally relevant representation of these elements that is conserved across species. In other words, despite sequence divergence, these regulatory regions retain a common underlying regulatory grammar that can be identified through their learned feature representations. This observation supports the idea that such representations could be used to define biologically meaningful encodings of DNA sequences.

Building on HYPER's ability to capture both sequence and functional conservation, as demon-

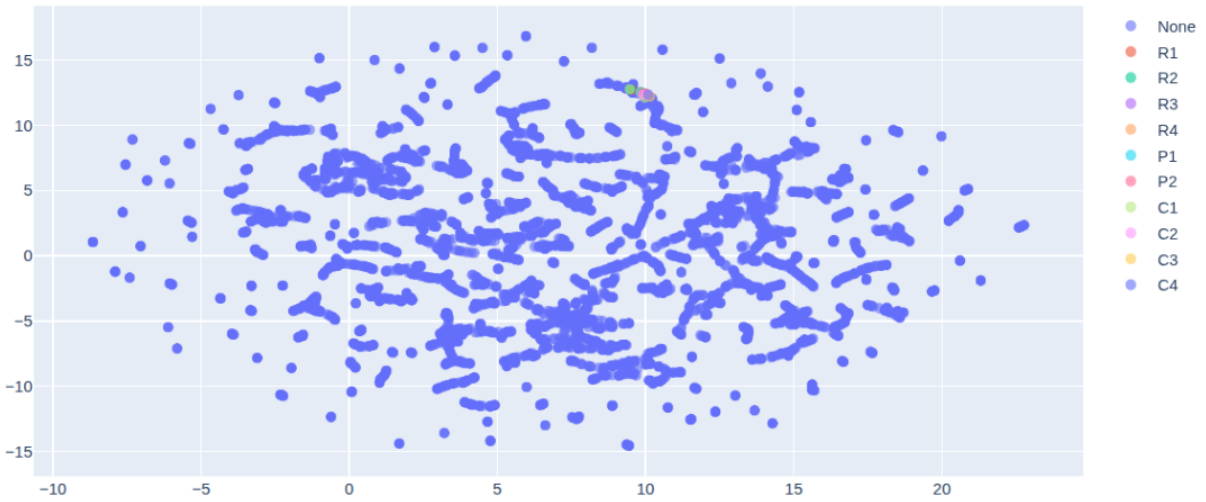


Figure 6.1: UMAP projection of k-mer importance–based representations for regulatory element sequences. Purple points correspond to regulatory elements genome-wide in human, while coloured points represent alpha-globin elements, including enhancer-like elements (R1, R2, R3, R4, Rm), NPRL3 and alpha-globin promoters (P1 and P2), and CTCF sites within the alpha-globin locus.

strated in Section 5.6, this representation could be further leveraged to identify equivalent regulatory elements across species using k-mer or motif importance profiles. This extends the previous proposal on cross-species analysis, providing a framework to detect functionally conserved elements even in the presence of sequence divergence.

From a computational perspective, this approach also offers opportunities to improve existing methods for genomic modelling. As previously discussed, LLMs rely on tokenisers to convert DNA sequences into numerical representations. However, conventional tokenisation methods are typically based on statistical properties of the sequence, such as k-mers or frequency patterns, and do not incorporate biological context. In contrast, a tokenisation strategy based on HYPER-derived importance representations would encode sequences according to their functional and regulatory properties, capturing cell-type-specific and context-dependent information.

Such a biologically informed tokenisation approach has not yet been explored and could provide more meaningful inputs for genomic language models, enabling them to learn from functionally relevant biological units rather than purely statistical representations. This, in turn, has the potential to improve the modelling of regulatory grammar and enhance generalisation across different cell types and species.

6.4.8 Clinical translation

From a clinical perspective, these approaches have the potential to fundamentally transform how regulatory biology is studied by shifting from experimental-driven to prediction-driven frameworks. By enabling large-scale *in silico* simulation of genetic variation and regulatory perturbations, models such as HYPER can prioritise functionally relevant variants and regulatory elements at high resolution before any experimental intervention. This is particularly powerful for investigating rare, transient, or otherwise inaccessible human cell types, where direct experimentation is limited or infeasible.

Importantly, the ability to train models on non-human species and apply them to human genomes provides a pathway to translate regulatory insights without the need for extensive human experimentation. By leveraging cross-species generalisation, it becomes possible to infer human regulatory mechanisms from model organisms, significantly reducing experimental burden. While experimental validation remains essential, such approaches can substantially lower the cost, ethical constraints, and scale of human studies, accelerating the discovery of disease mechanisms and the application of genomic insights in clinical settings.

A | Appendix

Chapter 3

Supplementary Figures

1. Signal quantification in human CTCF training and test sets

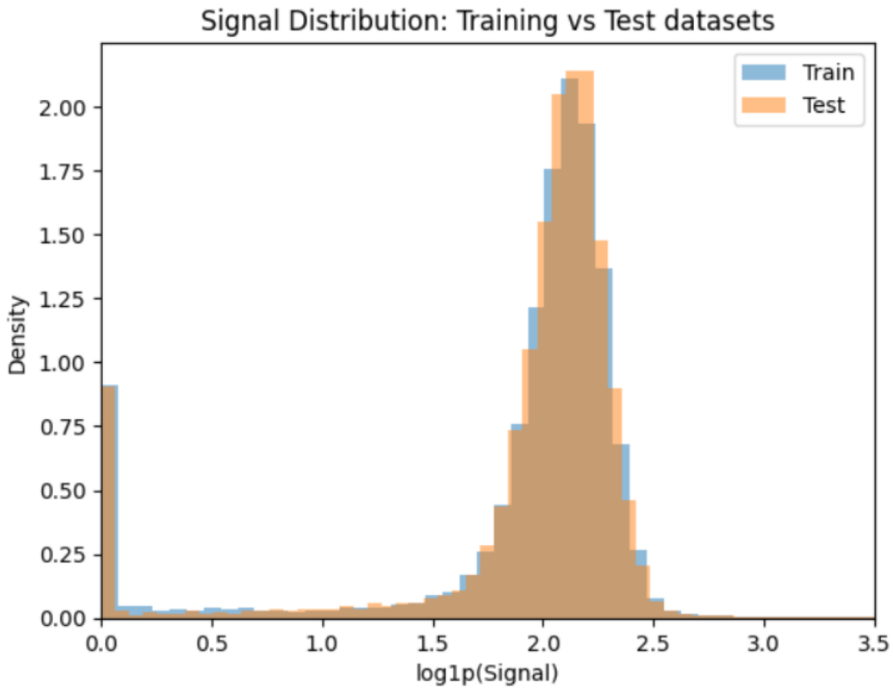


Figure S1: Distribution of log-transformed CTCF binding signal values in the training and test datasets used for HYPER model training. The strong overlap between the two distributions indicates consistency between training and evaluation data and suggests minimal distributional bias between dataset partitions.

Chapter 4

Supplementary Figures

1. Signal quantification in human training and test sets

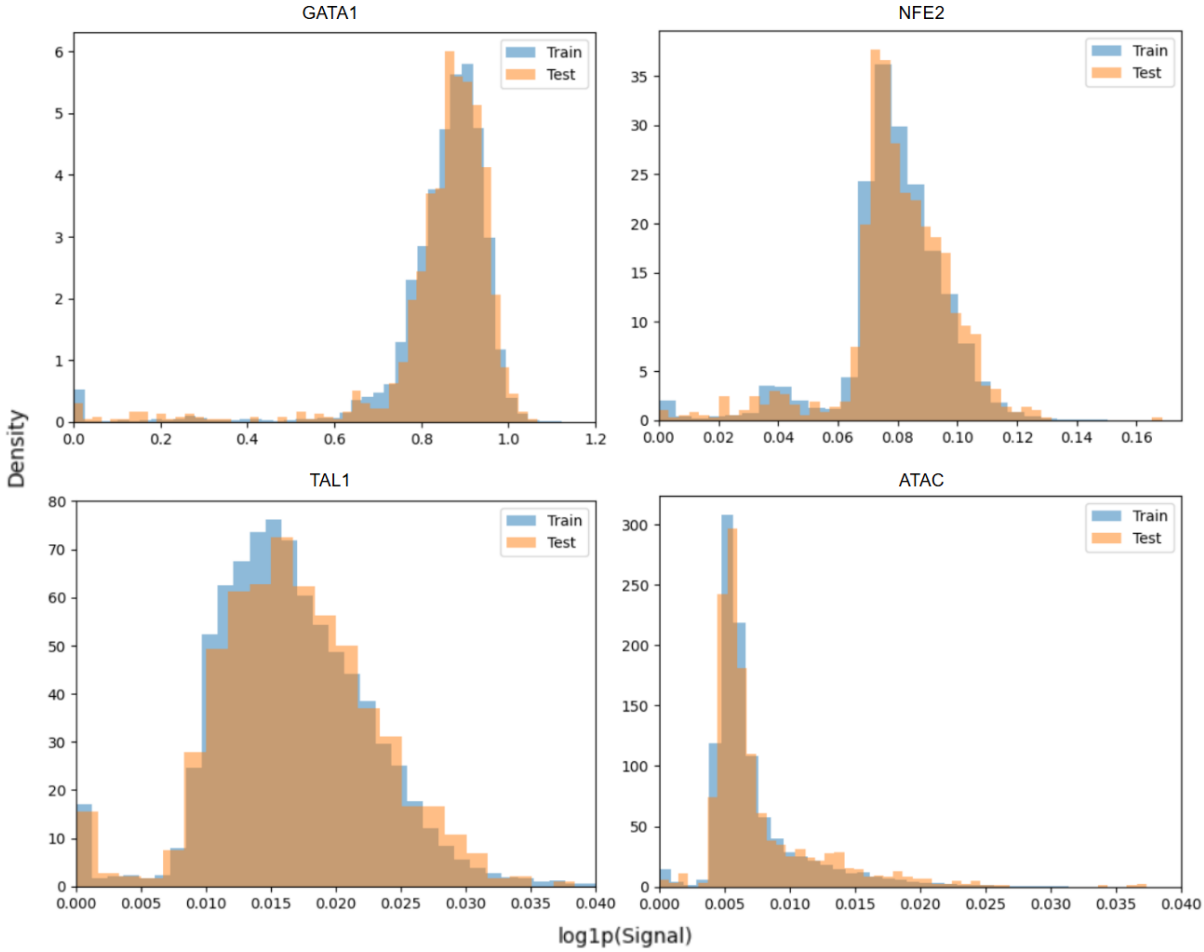


Figure S2: Distributions for the training set (blue) and test set (orange) are shown for GATA1 (top left), NF-E2 (top right), TAL1 (bottom left), and ATAC-seq (bottom right).

Chapter 5

Supplementary Figures

1. Signal quantification in mouse training and test sets

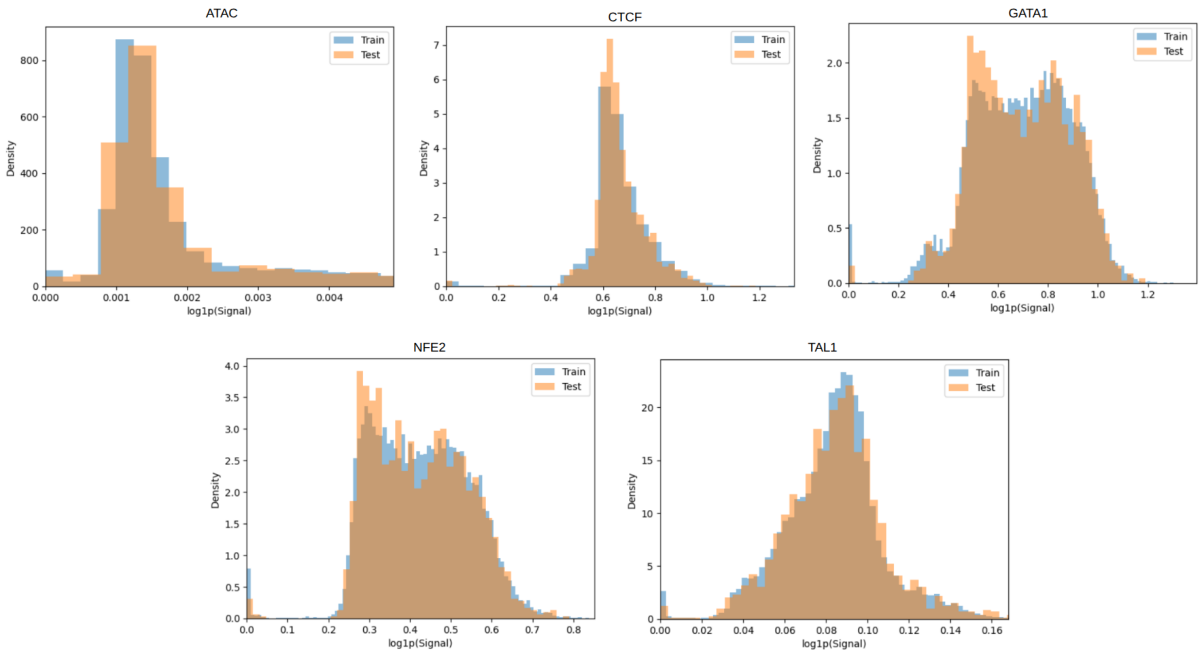


Figure S3: Density estimates of signal values for the mouse epigenetic tracks in training and test sets. The close agreement between distributions confirms minimal dataset shift. ATAC-seq, CTCF, and TAL1 exhibit background-enriched, skewed distributions, while GATA1 and NF-E2 display wider dynamic ranges, consistent with increased signal heterogeneity.

2. Signal quantification in chicken training and test sets

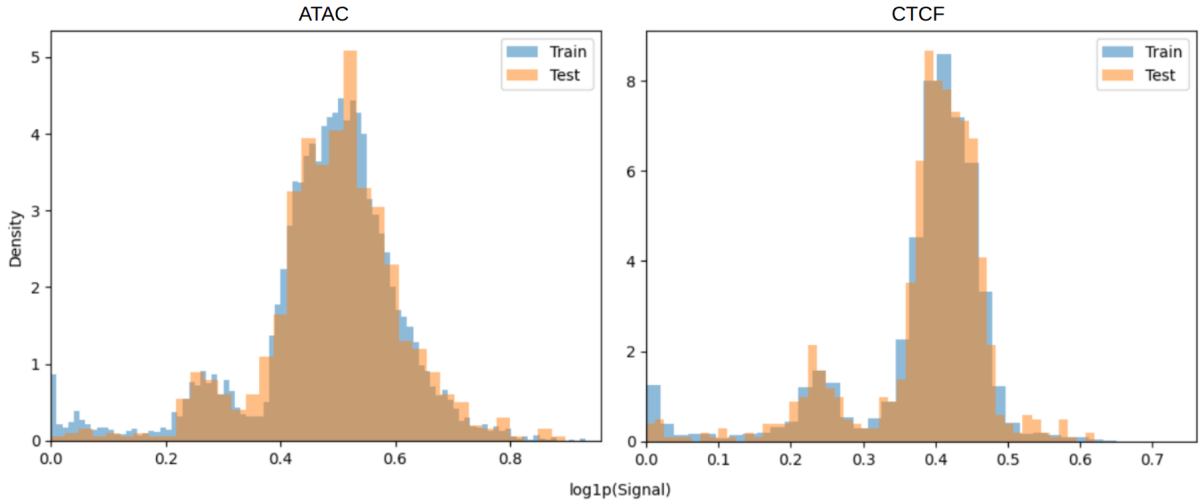


Figure S4: Distribution of signal values for chicken tracks in training and test datasets. The close overlap indicates consistent data splitting. ATAC-seq exhibits a broader distribution with greater variability, while CTCF shows a narrower, more tightly centred distribution, reflecting more constrained binding patterns.

Supplementary Tables

1. Hominin genomes

Specimen	Group	Sample tissue	Sex	Age	Estimated sample age (years)	Region	Sequencing coverage	Accession number
Altai	Neanderthal	Unknown bone fragment	Male	Adult	110,000-130,000	Denisova Cave, Siberia	37×	PRJEB1265 [227]
Denisova 11 (Denny)	Denisova-Neanderthal	Unknown bone fragment	Female	∼12 y.o.	90,000	Denisova Cave, Siberia	2.6×	PRJEB24663 [228]
Denisova 3	Denisova	Distal phalanx (finger) bone	Female	Juvenile	60,000-70,000	Denisova Cave, Siberia	30×	PRJEB3092 [246]
Mezmaiskaya 1	Neanderthal	Infant skeleton	Female	1-2 y.o.	60,000-70,000	Mezmaiskaya Cave, Russia	0.5×	PRJEB2065 [247]
Chagyrskaya 7	Neanderthal	Fragmentary skeletal remains	Female	Adult	50,000-60,000	Chagyrskaya Cave, Siberia	1-2×	PRJEB55327 [248]
Vindija 33.19	Neanderthal	Unknown bone fragment	Female	Adult	50,000-60,000	Vindija Cave, Croatia	30×	PRJEB21157 [249]
Ust Ishim 1	Early modern human	Left femur bone	Male	Adult	45,000	Irtysh River, Siberia	42×	PRJEB6622 [250]
Loschbour	Early modern human	Nearly complete skeleton	Male	Adult	8,000	Loschbour rock shelter, Luxembourg	20-22×	GSE96833 [251]
LBK	Early modern human	Whole skeleton	Female	20-30 y.o.	7,000-7,500	Viesenhäuser Hof, Germany	19×	PRJEB105335 [252]

Table S1: Summary of the hominin individuals for the cross-species analysis. For each specimen (or individual name as reported in the original study), the taxonomic group, sex, estimated age at death, estimated age of the sample, geographic origin, sequencing coverage, and accession number are provided. Three specimens are highlighted: the Altai Neanderthal, which has the highest sequencing coverage among Neanderthals; Ust-Ishim, the early modern human with the highest coverage; and Denisova 11 (“Denny”), a first-generation Neanderthal-Denisovan hybrid.

2. Cross-species genomes

Species	Assembly	Year	Evolutionary distance
Bonobo	panPan3	2021	6-7 million years
Cat	felCat9	2020	80-100 million years
Chimpanzee	panTro6	2018	6-7 million years
Cow	bosTau8	2009	80-100 million years
Crab eating macaque	macFas5	2013	23-25 million years
Dog	canFam3	2011	80-100 million years
Elephant	loxAfr3	2009	80-100 million years
Gorilla	gorGor6	2019	8-10 million years
Green monkey	chlSab2	2014	23-25 million years
Hawaiian monk seal	neoSch1	2017	60-100 million years
Hedgehog	eriEur2	2012	60-100 million years
Horse	equCab2	2007	60-100 million years
Minke whale	balAcu1	2013	60-100 million years
Opossum	monDom5	2007	180 million years
Orangutan	ponAbe3	2018	14-18 million years
Pig	susSCR3	2011	80–90 million years
Rabbit	oryCun2	2009	80–90 million years
Rat	rn6	2014	80–90 million years
Sheep	oviAri4	2015	80–90 million years
Squirrel monkey	saiBol1	2011	35–40 million years
Tasmanian devil	sarHar1	2011	140–160 million years
Turkey	melGal5	2019	300–320 million years
Zebra finch	taeGut2	2020	300–320 million years

Table S2: List of species genomes used in the cross-species analysis. The 23 species names are followed by their assembly names, year of release, and evolutionary distance from humans (in millions of years).

3. Cross-species predictions

Species	UCSC Genome Browser
Bonobo	https://genome-euro.ucsc.edu/s/arodrig/Bonobo_nprl3
Cat	https://genome-euro.ucsc.edu/s/arodrig/Cat_nprl3
Chimpanzee	https://genome-euro.ucsc.edu/s/arodrig/Chimp_nprl3
Cow	https://genome-euro.ucsc.edu/s/arodrig/Bull_nprl3
Crab eating macaque	https://genome-euro.ucsc.edu/s/arodrig/Macaque_nprl3
Dog	https://genome-euro.ucsc.edu/s/arodrig/Dog_nprl3
Elephant	https://genome-euro.ucsc.edu/s/arodrig/Elephant_nprl3
Gorilla	https://genome-euro.ucsc.edu/s/arodrig/Gorilla_nprl3
Green monkey	https://genome-euro.ucsc.edu/s/arodrig/Green_Monkey_nprl3
Hawaiian monk seal	https://genome-euro.ucsc.edu/s/arodrig/Seal_nprl3
Hedgehog	https://genome-euro.ucsc.edu/s/arodrig/Hedgehog_nprl3
Horse	https://genome-euro.ucsc.edu/s/arodrig/Horse_nprl3
Minke whale	https://genome-euro.ucsc.edu/s/arodrig/Minke_whale_nprl3
Opossum	https://genome-euro.ucsc.edu/s/arodrig/Opossum_nprl3
Orangutan	https://genome-euro.ucsc.edu/s/arodrig/Orangutan_nprl3
Pig	https://genome-euro.ucsc.edu/s/arodrig/Pig_nprl3
Rabbit	https://genome-euro.ucsc.edu/s/arodrig/Rabbit_nprl3
Rat	https://genome-euro.ucsc.edu/s/arodrig/Rat_nprl3
Sheep	https://genome-euro.ucsc.edu/s/arodrig/Sheep_nprl3
Squirrel monkey	https://genome-euro.ucsc.edu/s/arodrig/Squirrel_Monkey_nprl3
Tasmanian devil	https://genome-euro.ucsc.edu/s/arodrig/Tasmanian_Devil_nprl3
Turkey	https://genome-euro.ucsc.edu/s/arodrig/Turkey_nprl3
Zebra finch	https://genome-euro.ucsc.edu/s/arodrig/Zebra_Finch_nprl3

Table S3: UCSC genome browsers with predictions of the alpha-globin locus across the 23 species using the human, mouse, and chicken models.

Bibliography

[1] B. Alberts, A. Johnson, J. Lewis *et al.*, "Molecular biology of the cell, 4th edition," 2002.

[2] J. Zingg, "Genetic and epigenetic aspects of DNA methylation on genome expression, evolution, mutation and carcinogenesis," *Carcinogenesis*, vol. 18, no. 5, p. 869–882, May 1997.

[3] M. Missinato, S. Murphy, M. Lynott *et al.*, "Conserved transcription factors promote cell fate stability and restrict reprogramming potential in differentiated cells," *Nature Communications*, vol. 14, no. 1, Mar. 2023.

[4] T. W. Harris, "WormBase: a multi-species resource for nematode biology and genomics," *Nucleic Acids Research*, vol. 32, no. 90001, p. 411D – 417, Jan. 2004.

[5] G. Orphanides and D. Reinberg, "A unified theory of gene expression," *Cell*, vol. 108, no. 4, p. 439–451, Feb. 2002.

[6] P. Kolovos, T. Knoch, F. Grosveld *et al.*, "Enhancers and silencers: an integrated and simple model for their function," *Epigenetics Chromatin*, vol. 5, no. 1, Jan. 2012.

[7] E. Friman, I. Flyamer, D. Marenduzzo *et al.*, "Ultra-long-range interactions between active regulatory elements," *Genome Research*, vol. 33, no. 8, p. 1269–1283, 2023.

[8] N. Matharu and N. Ahituv, "Minor loops in major folds: Enhancer–promoter looping, chromatin restructuring, and their association with transcriptional regulation and disease," *PLOS Genetics*, vol. 11, no. 12, p. e1005640, Dec. 2015.

[9] T. Pollex, A. Rabinowitz, M. Gambetta *et al.*, "Enhancer–promoter interactions become more instructive in the transition from cell-fate specification to tissue differentiation," *Nature Genetics*, vol. 56, no. 4, p. 686–696, Mar. 2024.

[10] J. Wang, J. Zhuang, S. Iyer *et al.*, "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors," *Genome Research*, vol. 22, no. 9, p. 1798–1812, 2012.

[11] I. Georgakopoulos-Soares, C. Deng, V. Agarwal *et al.*, "Transcription factor binding site orientation and order are major drivers of gene regulatory activity," *Nature Communications*, vol. 14, no. 1, Apr. 2023.

[12] Y. Yin, E. Morgunova, A. Jolma *et al.*, "Impact of cytosine methylation on DNA binding specificities of human transcription factors," *Science*, vol. 356, no. 6337, May 2017.

[13] N. Zabet, R. Foy, and B. Adryan, “The influence of transcription factor competition on the relationship between occupancy and affinity,” *PLoS ONE*, vol. 8, no. 9, p. e73714, 2013.

[14] S. Weingarten-Gabbay and E. Segal, “The grammar of transcriptional regulation,” *Human Genetics*, vol. 133, no. 6, p. 701–711, Jan. 2014.

[15] T. Graf and T. Enver, “Forcing cells to change lineages,” *Nature*, vol. 462, no. 7273, p. 587–594, Dec. 2009.

[16] J. Xu, Y. Du, and H. Deng, “Direct lineage reprogramming: strategies, mechanisms, and applications,” *Cell Stem Cell*, vol. 16, no. 2, p. 119–134, Feb. 2015.

[17] D. Li, Y. Yang, Y. Li *et al.*, “Epigenetic regulation of gene expression in response to environmental exposures: From bench to model,” *Science of The Total Environment*, vol. 776, p. 145998, 2021.

[18] B. Strober, R. Elorbany, K. Rhodes *et al.*, “Dynamic genetic regulation of gene expression during cellular differentiation,” *Science*, vol. 364, no. 6447, p. 1287–1290, 2019.

[19] A. F. Cunha, A. F. Brugnerotto, A. Duarte *et al.*, “Global gene expression revealed a set of genes involved in the modification of cells during erythropoiesis.” *Blood*, vol. 110, no. 11, p. 4074–4074, Nov. 2007.

[20] A. Cardona, M. Peixoto, T. Borjigin *et al.*, “Bridging spatial and temporal scales of developmental gene regulation,” *Current Opinion in Genetics &; Development*, vol. 92, p. 102328, 2025.

[21] N. M. A. Aboud, C. Tupper, and I. Jialal, “Genetics, epigenetic mechanism,” 2026.

[22] J. Pfluger and D. Wagner, “Histone modifications and dynamic regulation of genome accessibility in plants,” *Current Opinion in Plant Biology*, vol. 10, no. 6, p. 645–652, Dec. 2007.

[23] A. D’Urso and J. Brickner, “Mechanisms of epigenetic memory,” *Trends in Genetics*, vol. 30, no. 6, p. 230–236, 2014.

[24] R. Krishnakumar and R. Blelloch, “Epigenetics of cellular reprogramming,” *Current Opinion in Genetics & Development*, vol. 23, no. 5, p. 548–555, Oct. 2013.

[25] T. Zhang and M. Meaney, “Epigenetics and the environmental regulation of the genome and its function,” *Annual Review of Psychology*, vol. 61, no. 1, p. 439–466, Jan. 2010.

[26] D. Liang, A. Elwell, N. Aygun *et al.*, “Cell-type-specific effects of genetic variation on chromatin accessibility during human neuronal differentiation,” *Nature Neuroscience*, vol. 24, no. 7, p. 941–953, May 2021.

[27] Y. Chen, R. Liang, Y. Li *et al.*, "Chromatin accessibility: biological functions, molecular mechanisms and therapeutic application," *Signal Transduction and Targeted Therapy*, vol. 9, no. 1, Dec. 2024.

[28] J. Mattick and P. Amaral, "RNA, the epicenter of genetic information: A new understanding of molecular biology," 2022.

[29] K. H. Brettingham-Moore, P. C. Taberlay, and A. Holloway, "Interplay between transcription factors and the epigenome: insight from the role of RUNX1 in leukemia," *Frontiers in Immunology*, vol. 6, 2015.

[30] R. Galupa and E. Heard, "Topologically associating domains in chromosome architecture and gene regulatory landscapes during development, disease, and evolution," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 82, p. 267–278, 2017.

[31] E. R. Gibney and C. M. Nolan, "Epigenetics and gene expression," *Heredity*, vol. 105, no. 1, p. 4–13, May 2010.

[32] L. Dauban, M. Eder, M. de Haas *et al.*, "Interactions between the genome and the nuclear lamina are multivalent and cooperative," *Nature Structural & Molecular Biology*, vol. 32, no. 11, p. 2335–2348, 2025.

[33] J. W. Blayney, H. Francis, A. Rampasekova *et al.*, "Super-enhancers include classical enhancers and facilitators to fully activate gene expression," *Cell*, vol. 186, no. 26, pp. 5826–5839.e18, Dec. 2023.

[34] M. Kassouf, H. Francis, M. Gosden *et al.*, "The α-globin super-enhancer acts in an orientation-dependent manner," *Nature Communications*, vol. 16, no. 1, Jan. 2025.

[35] M. Benton, S. Talipineni, D. Kostka *et al.*, "Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function," *BMC Genomics*, vol. 20, no. 1, 2019.

[36] Q. Li, K. P. X., Fang *et al.*, "Locus control regions," *Blood*, vol. 100, no. 9, p. 3077–3086, Nov. 2002.

[37] W. Whyte, D. Orlando, D. Hnisz *et al.*, "Master transcription factors and mediator establish super-enhancers at key cell identity genes," *Cell*, vol. 153, no. 2, p. 307–319, Apr. 2013.

[38] M. Kassouf, S. Ford, J. Blayney, and D. Higgs, "Understanding fundamental principles of enhancer biology at a model locus: Analysing the structure and function of an enhancer cluster at the α-globin locus," *BioEssays*, vol. 45, no. 10, 2023.

[39] H. Szutorisz, N. Dillon, and L. Tora, "The role of enhancers as centres for general transcription factor recruitment," *Trends in Biochemical Sciences*, vol. 30, no. 11, p. 593–599, Nov. 2005.

[40] M. Tsompana and M. Buck, "Chromatin accessibility: a window into the genome," *Epigenetics &; Chromatin*, vol. 7, no. 1, Nov. 2014.

[41] S. Goodwin, J. McPherson, and W. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, p. 333–351, May 2016.

[42] R. Stark, M. Grzelak, and J. Hadfield, "RNA sequencing: the teenage years," *Nature Reviews Genetics*, vol. 20, no. 11, p. 631–656, 2019.

[43] R. Thurman, E. Rynes, R. Humbert *et al.*, "The accessible chromatin landscape of the human genome," *Nature*, vol. 489, no. 7414, p. 75–82, 2012.

[44] J. Buenrostro, P. Giresi, L. Zaba *et al.*, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nature Methods*, vol. 10, no. 12, p. 1213–1218, Oct. 2013.

[45] Z. Li, M. Schulz, T. Look *et al.*, "Identification of transcription factor binding sites using ATAC-seq," *Genome Biology*, vol. 20, no. 1, Feb. 2019.

[46] H. Zhang, T. Lu, S. Liu *et al.*, "Comprehensive understanding of Tn5 insertion preference improves transcription regulatory element identification," *NAR Genomics and Bioinformatics*, vol. 3, no. 4, Oct. 2021.

[47] S. Subramaniyam, M. DeJesus, A. Zaveri *et al.*, "Statistical analysis of variability in TnSeq data across conditions using zero-inflated negative binomial regression," *BMC Bioinformatics*, vol. 20, no. 1, Nov. 2019.

[48] M. Bentsen, P. Goymann, H. Schultheis *et al.*, "ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation," *Nature Communications*, vol. 11, no. 1, Aug. 2020.

[49] H. Hojo and S. Ohba, "Runt-related transcription factors and gene regulatory mechanisms in skeletal development and diseases," *Current Osteoporosis Reports*, vol. 21, no. 5, p. 485–492, 2023.

[50] L. Song and G. Crawford, "DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells," *Cold Spring Harbor Protocols*, vol. 2010, no. 2, p. pdb.prot5384, Feb. 2010.

[51] M. Cooper, A. Ray, A. Bhattacharya *et al.*, "ATAC-seq optimization for cancer epigenetics research," *Journal of Visualized Experiments*, no. 184, 2022.

[52] A. K. Calviello, A. Hirsekorn, R. Wurmus *et al.*, "Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modelling," *Genome Biology*, vol. 20, no. 1, Feb. 2019.

[53] J. Buenrostro, B. Wu, H. Chang *et al.*, "ATAC-seq: A method for assaying chromatin accessibility genome-wide," *Current Protocols in Molecular Biology*, vol. 109, no. 1, Jan. 2015.

[54] A. Boyle, S. Davis, H. Shulha *et al.*, "High-resolution mapping and characterization of open chromatin across the genome," *Cell*, vol. 132, no. 2, p. 311–322, Jan. 2008.

[55] Y. Kashima, Y. Sakamoto, K. Kaneko *et al.*, "Single-cell sequencing techniques from individual to multiomics analyses," *Experimental &; Molecular Medicine*, vol. 52, no. 9, p. 1419–1427, 2020.

[56] S. Baek and I. I. Lee, "Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation," *Computational and Structural Biotechnology Journal*, vol. 18, p. 1429–1439, 2020.

[57] T. Stuart, A. Butler, P. Hoffman *et al.*, "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, pp. 1888–1902.e21, 2019.

[58] B. Ren, F. Robert, J. Wyrick *et al.*, "Genome-wide location and function of DNA binding proteins," *Science*, vol. 290, no. 5500, p. 2306–2309, Dec. 2000.

[59] N. C. for Biotechnology Information, "SRA database growth," 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

[60] S. Salzberg, X. Chen, J. Henderson *et al.*, "Finding genes in DNA using decision trees and dynamic programming," *ISMB 1996*, 1996.

[61] E. Lander, L. Linton, B. Birren *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, p. 860–921, Feb. 2001.

[62] M. Brown, W. Grundy, D. Lin *et al.*, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences*, vol. 97, no. 1, p. 262–267, Jan. 2000.

[63] Z. Yao and W. Ruzzo, "A regression-based K nearest neighbour algorithm for gene function prediction from heterogeneous data," *BMC Bioinformatics*, vol. 7, no. S1, Mar. 2006.

[64] M. Ozarar, V. Atalay, and R. Atalay, “Prediction of protein subcellular localisation based on primary sequence data,” in *Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, 2004.* IEEE, 2004, p. 118–120.

[65] A. Guerra, J. Páez, and N. Campillo, “Artificial neural networks in ADMET modelling: Prediction of blood–brain barrier permeation,” *QSAR & Combinatorial Science*, vol. 27, no. 5, p. 586–594, May 2008.

[66] B. Alipanahi, A. Delong, M. Weirauch *et al.*, “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,” *Nature Biotechnology*, vol. 33, no. 8, p. 831–838, 2015.

[67] J. Zhou and O. Troyanskaya, “Predicting effects of noncoding variants with deep learning–based sequence model,” *Nature Methods*, vol. 12, no. 10, p. 931–934, Aug. 2015.

[68] Z. Avsec, M. Weilert, A. Shrikumar *et al.*, “Base-resolution models of transcription-factor binding reveal soft motif syntax,” *Nature Genetics*, vol. 53, no. 3, p. 354–366, Feb. 2021.

[69] C. Chen, J. Hou, X. Shi *et al.*, “DeepGRN: Prediction of transcription factor binding site across cell-types using attention-based deep neural networks,” *BMC Bioinformatics*, vol. 22, no. 1, Feb. 2021.

[70] D. Quang and X. Xie, “DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences,” *Nucleic Acids Research*, vol. 44, no. 11, p. e107–e107, Apr. 2016.

[71] A. Vaswani, N. Shazeer, N. Parmar *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[72] S. Choi and M. Lee, “Transformer architecture and attention mechanisms in genome data analysis: A comprehensive review,” *Biology*, vol. 12, no. 7, p. 1033, 2023.

[73] Z. Avsec, V. Agarwal, D. Visentin *et al.*, “Effective gene expression prediction from sequence by integrating long-range interactions,” *Nature Methods*, vol. 18, no. 10, p. 1196–1203, Oct. 2021.

[74] Y. Ji, Z. Zhou, H. Liu *et al.*, “DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome,” *Bioinformatics*, vol. 37, no. 15, p. 2112–2120, Feb. 2021.

[75] H. Zeng, M. Edwards, G. Liu *et al.*, “Convolutional neural network architectures for predicting DNA–protein binding,” *Bioinformatics*, vol. 32, no. 12, p. i121–i127, 2016.

[76] A. Barredo-Arrieta, N. Díaz-Rodríguez, J. D. Ser *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, p. 82–115, 2020.

[77] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," 2017.

[78] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2017.

[79] H. Cha, "Erythropoiesis: Insights from a genomic perspective," *Experimental & Molecular Medicine*, vol. 56, no. 10, p. 2099–2104, Oct. 2024.

[80] I. McNiece, K. Langley, and K. Zsebo, "Recombinant human stem cell factor synergises with GM-CSF, G-CSF, IL-3 and epo to stimulate human progenitor cells of the myeloid and erythroid lineages," *Experimental Hematology*, vol. 19, no. 3, pp. 226–231, mar 1991.

[81] M. Weiss, C. Yu, and S. Orkin, "Erythroid-cell-specific properties of transcription factor GATA-1 revealed by phenotypic rescue of a gene-targeted cell line," *Molecular and Cellular Biology*, vol. 17, no. 3, p. 1642–1651, Mar. 1997.

[82] C. Porcher, W. Swat, K. Rockwell *et al.*, "The T cell leukemia oncoprotein SCL/TAL-1 is essential for development of all hematopoietic lineages," *Cell*, vol. 86, no. 1, p. 47–57, 1996.

[83] J. Armstrong and B. Emerson, "NF-E2 disrupts chromatin structure at human β-globin locus control region hypersensitive site 2 in vitro," *Molecular and Cellular Biology*, vol. 16, no. 10, p. 5634–5644, Oct. 1996.

[84] G. Liu, Y. Hou, X. Jin *et al.*, "PI3K/HSCB axis facilitates FOG1 nuclear translocation to promote erythropoiesis and megakaryopoiesis," *eLife*, vol. 13, May 2024.

[85] B. Nuez, D. Michalovich, A. Bygrave *et al.*, "Defective haematopoiesis in fetal liver resulting from inactivation of the EKLF gene," *Nature*, vol. 375, no. 6529, p. 316–318, May 1995.

[86] D. Higgs, D. Vernimmen, and B. Wood, "Long-range regulation of α-globin gene expression," p. 143–173, 2008.

[87] C. Kiefer, C. Hou, J. Little *et al.*, "Epigenetics of β-globin gene regulation," *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 647, no. 1–2, p. 68–76, Dec. 2008.

[88] D. Higgs, M. Vickers, A. Wilkie *et al.*, "A review of the molecular genetics of the human alpha-globin gene cluster," *Blood*, vol. 73, no. 5, pp. 1081–1104, apr 1989.

[89] S. Liebhaber and J. Russell, "Expression and developmental control of the human α-globin gene cluster," *Annals of the New York Academy of Sciences*, vol. 850, no. 1, p. 54–63, 1998.

[90] M. Albitar, F. Cash, C. Peschle *et al.*, "Developmental switch in the relative expression of the alpha 1- and alpha 2-globin genes in humans and in transgenic mice," *Blood*, vol. 79, no. 9, pp. 2471–2474, may 1992.

[91] M. Mahajan, S. Karmakar, P. Newburger *et al.*, "Dynamics of α-globin locus chromatin structure and gene expression during erythroid differentiation of human CD34+ cells in culture," *Experimental Hematology*, vol. 37, no. 10, pp. 1143–1156.e3, Oct. 2009.

[92] S. Mettananda, R. Gibbons, D. Higgs *et al.*, "Understanding α-globin gene regulation and implications for the treatment of β-thalassemia," *Annals of the New York Academy of Sciences*, vol. 1368, no. 1, p. 16–24, Dec. 2015.

[93] S. Capasso, G. Cardiero, G. Musollino *et al.*, "Functional analysis of three new alpha-thalassemia deletions involving MCS-R2 reveals the presence of an additional enhancer element in the 5' boundary region," *PLOS Genetics*, vol. 19, no. 5, p. e1010727, May 2023.

[94] V. Viprakasit, A. Kidd, H. Ayyub *et al.*, "De novo deletion within the telomeric region flanking the human α globin locus as a cause of α-thalassaemia," *British Journal of Haematology*, vol. 120, no. 5, p. 867–875, Mar. 2003.

[95] A. Jarman, W. Wood, J. Sharpe *et al.*, "Characterization of the major regulatory element upstream of the human α-globin gene cluster," *Molecular and Cellular Biology*, vol. 11, no. 9, p. 4679–4689, 1991.

[96] H. Voon and J. Vadolas, "Controlling α-globin: a review of α-globin expression and its impact on β-thalassemia," *Haematologica*, vol. 93, no. 12, p. 1868–1876, 2008.

[97] M. Hall, D. Curtis, D. Metcalf *et al.*, "The critical regulator of embryonic hematopoiesis, SCL, is vital in the adult for megakaryopoiesis, erythropoiesis, and lineage choice in CFU-S12," *Proceedings of the National Academy of Sciences*, vol. 100, no. 3, p. 992–997, Jan. 2003.

[98] D. Hesslein and D. Schatz, "Factors and forces controlling V(D)J recombination," p. 169–232, 2001.

[99] P. Fraser, S. Pruzina, M. Antoniou *et al.*, "Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes." *Genes & Development*, vol. 7, no. 1, p. 106–113, Jan. 1993.

[100] C. Jia, S. Huang, J. Yan *et al.*, "Effects of human locus control region elements HS2 and HS3 on human β-globin gene expression in transgenic mouse," *Blood Cells, Molecules, and Diseases*, vol. 31, no. 3, p. 360–369, Nov. 2003.

[101] K. Johnson, J. Grass, M. Boyer *et al.*, "Cooperative activities of hematopoietic regulators recruit RNA polymerase II to a tissue-specific chromatin domain," *Proceedings of the National Academy of Sciences*, vol. 99, no. 18, p. 11760–11765, Aug. 2002.

[102] Z. Xu, S. Huang, L. Chang *et al.*, "Identification of a TAL1 target gene reveals a positive role for the LIM domain-binding protein Ldb1 in erythroid gene expression and differentiation," *Molecular and Cellular Biology*, vol. 23, no. 21, p. 7585–7599, Nov. 2003.

[103] G. Stamatoyannopoulos, "Control of globin gene expression during development and erythroid differentiation," *Experimental Hematology*, vol. 33, no. 3, p. 259–271, Mar. 2005.

[104] G. Frati and A. Miccio, "Genome editing for β-hemoglobinopathies: Advances and challenges," *Journal of Clinical Medicine*, vol. 10, no. 3, p. 482, Jan. 2021.

[105] N. Tesio and D. Bauer, "Molecular basis and genetic modifiers of thalassemia," *Hematology/Oncology Clinics of North America*, vol. 37, no. 2, p. 273–299, Apr. 2023.

[106] C. Craddock, P. Vyas, J. Sharpe *et al.*, "Contrasting effects of alpha and beta globin regulatory elements on chromatin structure may be related to their different chromosomal environments." *The EMBO Journal*, vol. 14, no. 8, p. 1718–1726, Apr. 1995.

[107] E. Epner, A. Reik, D. Cimbora *et al.*, "The β-globin LCR is not necessary for an open chromatin structure or developmentally regulated transcription of the native mouse β-globin locus," *Molecular Cell*, vol. 2, no. 4, p. 447–455, Oct. 1998.

[108] K. Mattioli, W. Oliveros, C. Gerhardinger *et al.*, "Cis and trans effects differentially contribute to the evolution of promoters and enhancers," *Genome Biology*, vol. 21, no. 1, Aug. 2020.

[109] F. He, K. Steige, V. Kovacova *et al.*, "Cis-regulatory evolution spotlights species differences in the adaptive potential of gene expression plasticity," *Nature Communications*, vol. 12, no. 1, 2021.

[110] A. Barrière, K. Gordon, and I. Ruvinsky, "Distinct functional constraints partition sequence conservation in a cis-regulatory element," *PLoS Genetics*, vol. 7, no. 6, p. e1002095, 2011.

[111] J. Hughes, J. Cheng, N. Ventress *et al.*, "Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences," *Proceedings of the National Academy of Sciences*, vol. 102, no. 28, p. 9830–9835, 2005.

[112] L. Taher, D. McGaughey, S. Maragh *et al.*, "Genome-wide identification of conserved regulatory function in diverged sequences," *Genome Research*, vol. 21, no. 7, p. 1139–1149, May 2011.

[113] S. Domené, V. Bumaschny, F. de Souza *et al.*, "Enhancer turnover and conserved regulatory function in vertebrate evolution," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1632, p. 20130027, Dec. 2013.

[114] C. Attanasio, A. Reymond, R. Humbert *et al.*, "Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells," *Genome Biology*, vol. 9, no. 12, Dec. 2008.

[115] A. Karollus, J. Hingerl, D. Gankin *et al.*, "Species-aware DNA language models capture regulatory elements and their evolution," *Genome Biology*, vol. 25, no. 1, Apr. 2024.

[116] E. Dutrow, D. Emera, K. Yim *et al.*, "Modelling uniquely human gene regulatory function via targeted humanization of the mouse genome," *Nature Communications*, vol. 13, no. 1, Jan. 2022.

[117] M. Phan, T. Zehnder, F. Puntieri *et al.*, "Conservation of regulatory elements with highly diverged sequences across large evolutionary distances," *Nature Genetics*, vol. 57, no. 6, p. 1524–1534, May 2025.

[118] R. Ordoñez, W. Zhang, G. Ellis *et al.*, "Genomic context sensitizes regulatory elements to genetic disruption," *Molecular Cell*, vol. 84, no. 10, pp. 1842–1854.e7, May 2024.

[119] S. Li, E. Kvon, A. Visel *et al.*, "Stable enhancers are active in development, and fragile enhancers are associated with evolutionary adaptation," *Genome Biology*, vol. 20, no. 1, 2019.

[120] S. Li and I. Ovcharenko, "Enhancer jungles establish robust tissue-specific regulatory control in the human genome," *Genomics*, vol. 112, no. 3, p. 2261–2270, May 2020.

[121] M. Miyata, N. Gillemans, D. Hockman *et al.*, "An evolutionarily ancient mechanism for regulation of haemoglobin expression in vertebrate red cells," *Blood*, vol. 136, no. 3, p. 269–278, 2020.

[122] A. Moon and T. Ley, "Conservation of the primary structure, organization, and function of the human and mouse beta-globin locus-activating regions." *Proceedings of the National Academy of Sciences*, vol. 87, no. 19, p. 7693–7697, Oct. 1990.

[123] A. Oudelaar, R. Beagrie, M. Kassouf *et al.*, "The mouse alpha-globin cluster: A paradigm for studying genome regulation and organization," *Current Opinion in Genetics &; Development*, vol. 67, p. 18–24, Apr. 2021.

[124] D. Nebor, K. Gillinder, R. Robledo *et al.*, "Identifying novel modifiers of embryonic globin expression by combining ChIP-seq, RNA-seq and eQTL mapping in the adult Nan mouse model," *Blood*, vol. 128, no. 22, p. 398–398, Dec. 2016.

[125] C. Lechauve, J. Keith, M. Dudley *et al.*, "Ancestral β-globin gene haplotypes modify β-thalassemia severity in a mouse model," *Blood Advances*, vol. 8, no. 23, p. 5988–5992, Nov. 2024.

[126] M. Bender, A. Reik, J. Close *et al.*, "Description and targeted deletion of 5' hypersensitive site 5 and 6 of the mouse beta-globin locus control region," *Blood*, vol. 92, no. 11, pp. 4394–4403, dec 1998.

[127] P. Chan, A. Wai, S. Philipsen *et al.*, "5'HS5 of the human β-globin locus control region is dispensable for the formation of the β-globin active chromatin hub," *PLoS ONE*, vol. 3, no. 5, p. e2134, May 2008.

[128] B. Hug, A. Moon, and T. Ley, "Structure and function of the murine β-globin locus control region 5' HS-3," *Nucleic Acids Research*, vol. 20, no. 21, p. 5771–5778, 1992.

[129] M. Bender, J. Roach, J. Halow *et al.*, "Targeted deletion of 5'HS1 and 5'HS4 of the β-globin locus control region reveals additive activity of the DNaseI hypersensitive sites," *Blood*, vol. 98, no. 7, p. 2022–2027, Oct. 2001.

[130] T. Beacon and J. Davie, "Chicken erythrocyte: Epigenomic regulation of gene activity," *International Journal of Molecular Sciences*, vol. 24, no. 9, p. 8287, May 2023.

[131] O. Svoboda and P. Bartunek, "Origins of the vertebrate erythro/megakaryocytic system," *BioMed Research International*, vol. 2015, p. 1–10, 2015.

[132] J. Knezetic and G. Felsenfeld, "Identification and characterization of a chicken alpha-globin enhancer," *Molecular and Cellular Biology*, vol. 9, no. 3, p. 893–901, Mar. 1989.

[133] M. Prioleau, M. Gendron, and O. Hyrien, "Replication of the chicken β-globin locus: Early-firing origins at the 5' HS4 insulator and the ρ- and β^A-globin genes show opposite epigenetic modifications," *Molecular and Cellular Biology*, vol. 23, no. 10, p. 3536–3549, May 2003.

[134] S. Razin, S. Ulianov, E. Ioudinkova *et al.*, "Domains of α- and β-globin genes in the context of the structural-functional organization of the eukaryotic genome," *Biochemistry (Moscow)*, vol. 77, no. 13, p. 1409–1423, Dec. 2012.

[135] J. Flint, C. Tufarelli, J. Peden *et al.*, "Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster," *Human Molecular Genetics*, vol. 10, no. 4, p. 371–382, Feb. 2001.

[136] M. Reitman, E. Lee, and H. Westphal, "Function of the upstream hypersensitive sites of the chicken β-globin gene cluster in mice," *Nucleic Acids Research*, vol. 23, no. 10, p. 1790–1794, May 1995.

[137] N. Sjakste and T. Sjakste, "Structure of globin gene domains in mammals and birds," *Russian Journal of Genetics*, 2002.

[138] M. Gomes, A. Kaushik, A. Chirayil *et al.*, "Application of machine learning in cancer epigenetics: A deeper look into the epigenome," p. 155–180, 2025.

[139] D. Alsayaydeh, V. Narayanamurthy, A. Futane *et al.*, "Advances of artificial intelligence-enabled epigenetics," *Health and Technology*, vol. 15, no. 4, p. 663–676, 2025.

[140] S. Riva, E. Georgiades, E. Gur *et al.*, "CATCH-UP: A high-throughput upstream-pipeline for bulk ATAC-seq and ChIP-seq data," *Journal of Visualized Experiments*, no. 199, 2023.

[141] S. Andrews, "FastQC: A quality control tool for high throughput sequence data," https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2010.

[142] A. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, no. 15, p. 2114–2120, Apr. 2014.

[143] B. Langmead and S. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, p. 357–359, Mar. 2012.

[144] H. Li, B. Handsaker, A. Wysoker *et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, p. 2078–2079, 2009.

[145] F. Ramírez, F. Dündar, S. Diehl *et al.*, "deepTools: a flexible platform for exploring deep-sequencing data," *Nucleic Acids Research*, vol. 42, no. W1, p. W187–W191, May 2014.

[146] H. Amemiya, A. Kundaje, and A. Boyle, "The ENCODE blacklist: Identification of problematic regions of the genome," *Scientific Reports*, vol. 9, no. 1, 2019.

[147] L. Hentges, M. Sergeant, C. Cole *et al.*, "LanceOtron: A deep learning peak caller for genome sequencing experiments," *Bioinformatics*, vol. 38, no. 18, p. 4255–4263, 2022.

[148] Kundaje Lab. (2019) BPNet preprocessing. GitHub. [Online]. Available: https://github.com/kundajelab/bpnet/wiki/Preprocessing

[149] Calico Labs. (2026) Basenji preprocessing. GitHub. [Online]. Available: https://github.com/calico/basenji/tree/master

[150] A. F. Agarap, "Deep learning using Rectified Linear Units (ReLU)," 2018.

[151] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," 2016.

[152] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015.

[153] J. Cordonnier, A. Loukas, and M. Jaggi, "Multi-head attention: Collaborate instead of concatenate," 2020.

[154] R. Schwessinger, M. Gosden, D. Downes *et al.*, "DeepC: Predicting 3D genome folding using megabase-scale transfer learning," *Nature Methods*, vol. 17, no. 11, p. 1118–1124, Oct. 2020.

[155] D. Kelley, "Cross-species regulatory sequence activity prediction," *PLOS Computational Biology*, vol. 16, no. 7, p. e1008050, 2020.

[156] M. Kagda, B. Lam, C. Litton *et al.*, "Data navigation on the encode portal," *Nature Communications*, vol. 16, no. 1, Oct. 2025.

[157] A. Kundaje, W. Meuleman, J. Ernst *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, 2015.

[158] F. Aguet, S. Anand, K. Ardlie *et al.*, "The GTEx consortium atlas of genetic regulatory effects across human tissues," *Science*, vol. 369, no. 6509, pp. 1318–1330, 2020.

[159] S. Riva, E. Sanders, T. Wilson *et al.*, "REnformer, a single-cell ATAC-seq predicting model to investigate open chromatin sites," *openRxiv*, 2025.

[160] J. M. Gaspar, "Improved peak-calling with MACS2," Dec. 2018.

[161] N. Kokhlikyan, V. Miglani, M. Martin *et al.*, "Captum: A unified and generic model interpretability library for pytorch," 2020.

[162] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," 2017.

[163] Kundaje Lab. (2025) BPNet-refactor package. GitHub. [Online]. Available: https://github.com/kundajelab/bpnet-refactor

[164] ——. (2026) TF-MoDISco: Transcription factor motif discovery from importance scores. GitHub. [Online]. Available: https://github.com/kundajelab/tfmodisco

[165] T. Bailey, L. Johnson, J.Grant *et al.*, "The MEME suite," *Nucleic Acids Research*, vol. 43, no. W1, pp. W39–W49, 2015.

[166] B. Zimmermann, I. Bilusic, C. Lorenz *et al.*, "Genomic SELEX: A discovery tool for genomic aptamers," *Methods*, vol. 52, no. 2, p. 125–132, Oct. 2010.

[167] R. Pantier, K. Chhatbar, G. Alston *et al.*, "High-throughput sequencing SELEX for the determination of DNA-binding protein specificities in vitro," *STAR Protocols*, vol. 3, no. 3, p. 101490, 2022.

[168] I. Rauluseviciute, R. Riudavets-Puig, R. Blanc-Mathieu *et al.*, "JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles," *Nucleic Acids Research*, vol. 52, no. D1, p. D174–D182, Nov. 2023.

[169] I. Vorontsov, I. Eliseeva, A. Zinkevich *et al.*, "HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors," *Nucleic Acids Research*, vol. 52, no. D1, p. D154–D163, Nov. 2023.

[170] V. Matys, "TRANSFAC(R): transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, no. 1, p. 374–378, Jan. 2003.

[171] M. Gosden, "An exploration of non-coding sequence variation across human erythroid differentiation," *University of Oxford D.Phil. thesis*, 2024.

[172] B. van de Geijn, G. McVicker, Y. Gilad *et al.*, "WASP: allele-specific software for robust molecular quantitative trait locus discovery," *Nature Methods*, vol. 12, no. 11, p. 1061–1063, 2015.

[173] C. Boeckel, X. Pastor, M. Heinig *et al.*, "Differential analysis of protein-DNA binding using ChIP-seq data," p. 63–89, 2024.

[174] D. Taliun, D. Harris, M. Kessler *et al.*, "Sequencing of 53, 831 diverse genomes from the NHLBI TOPMed program," *Nature*, vol. 590, no. 7845, p. 290–299, Feb. 2021.

[175] D. Prawitt, T. Enklaar, B. Gärtner-Rupprecht *et al.*, "Microdeletion of target sites for insulator protein CTCF in a chromosome 11p15 imprinting center in Beckwith–Wiedemann syndrome and Wilms' tumor," *Proceedings of the National Academy of Sciences*, vol. 102, no. 11, p. 4085–4090, Mar. 2005.

[176] E. Konrad, N. Nardini, A. Caliebe *et al.*, "CTCF variants in 39 individuals with a variable neurodevelopmental disorder broaden the mutational and clinical spectrum," *Genetics in Medicine*, vol. 21, no. 12, p. 2723–2733, Dec. 2019.

[177] S. Anisimov, "A prevalence of imprinted genes within the total transcriptomes of human tissues and cells," *Molecular Biology International*, vol. 2012, p. 1–29, 2012.

[178] J. Boyes, J. Omichinski *et al.*, "Perturbation of nucleosome structure by the erythroid transcription factor GATA-1," *Journal of Molecular Biology*, vol. 279, no. 3, p. 529–544, 1998.

[179] Y. Kang, Y. Kim, J. Kang *et al.*, "Erythroid specific activator GATA-1-dependent interactions between CTCF sites around the β-globin locus," *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, vol. 1860, no. 4, p. 416–426, Apr. 2017.

[180] R. Pop, A. Pisante, D. Nagy *et al.*, "Identification of mammalian transcription factors that bind to inaccessible chromatin," *Nucleic Acids Research*, vol. 51, no. 16, p. 8480–8495, 2023.

[181] G. Nagy, E. Czipa, L. Steiner *et al.*, "Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of ctcf and cohesin proteins on DNA," *BMC Genomics*, vol. 17, no. 1, Aug. 2016.

[182] G. Singh, S. Mullany, S. Moorthy *et al.*, "A flexible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells," *Genome Research*, vol. 31, no. 4, p. 564–575, Mar. 2021.

[183] E. Nora, A. Goloborodko, A. Valton *et al.*, "Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization," *Cell*, vol. 169, no. 5, p. 930–944.e22, May 2017.

[184] A. Thibodeau, S. Khetan, A. Eroglu *et al.*, "CoRE-ATAC: A deep learning model for the functional classification of regulatory elements from single cell and bulk ATAC-seq data," *PLOS Computational Biology*, vol. 17, no. 12, p. e1009670, Dec. 2021.

[185] C. Merrill, A. Montgomery, M. Pabon *et al.*, "Harnessing changes in open chromatin determined by ATAC-seq to generate insulin-responsive reporter constructs," *BMC Genomics*, vol. 23, no. 1, May 2022.

[186] Q. Liu, K. Hua, X. Zhang *et al.*, "DeepCAGE: Incorporating transcription factors in genome-wide prediction of chromatin accessibility," *Genomics, Proteomics & Bioinformatics*, vol. 20, p. 496–507, Mar. 2022.

[187] H. Hassanzadeh and M. Wang, "DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins," p. 178–183, Dec. 2016.

[188] A. Maslova, R. Ramirez, K. Ma *et al.*, "Deep learning of immune cell differentiation," *Proceedings of the National Academy of Sciences*, vol. 117, no. 41, p. 25655–25666, 2020.

[189] A. Pampari, A. Shcherbina, E. Kvon *et al.*, "ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants," Dec. 2024.

[190] Y. Xu, C. Wang, K. Xu *et al.*, "TRAFICA: an open chromatin language model to improve transcription factor binding affinity prediction," *Bioinformatics*, vol. 41, no. 11, Aug. 2025.

[191] Z. Gao, Q. Liu, W. Zeng *et al.*, "EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics," *Genome Biology*, vol. 25, no. 1, Dec. 2024.

[192] E. Nguyen, M. Poli, M. Faizi *et al.*, "HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution," 2023.

[193] A. Olijnik, N. Roy, C. Scott *et al.*, "Genetic and functional insights into CDA-I prevalence and pathogenesis," *Journal of Medical Genetics*, vol. 58, no. 3, p. 185–195, 2020.

[194] C. Palii, C. Perez-Iratxeta, Z. Yao *et al.*, "Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages: Cell environment influences TAL1 genomic targeting," *The EMBO Journal*, vol. 30, no. 3, p. 494–509, Dec. 2010.

[195] S. Goh, M. Josleyn, Y. Lee *et al.*, "The human reticulocyte transcriptome," *Physiological Genomics*, vol. 30, p. 172–178, 2007.

[196] I. Listerman, A. Bledau, I. Grishina *et al.*, "Extragenic accumulation of RNA Polymerase II enhances transcription by RNA Polymerase III," *PLoS Genetics*, vol. 3, no. 11, p. e212, Nov. 2007.

[197] K. Zaret and J. Carroll, "Pioneer transcription factors: establishing competence for gene expression," *Genes & Development*, vol. 25, no. 21, p. 2227–2241, Nov. 2011.

[198] Y. Chen, M. Jørgensen, R. Kolde *et al.*, "Prediction of RNA Polymerase II recruitment, elongation and stalling from histone modification data," *BMC Genomics*, vol. 12, no. 1, Nov. 2011.

[199] D. Cimini, B. Howell, P. Maddox *et al.*, "Merotelic kinetochore orientation is a major mechanism of aneuploidy in mitotic mammalian tissue cells," *The Journal of Cell Biology*, vol. 153, no. 3, p. 517–528, Apr. 2001.

[200] R. Ma and B. Zhu, “Pericentromeric sequences, where a conservation paradox occurs,” *Trends in Cell Biology*, vol. 35, no. 12, p. 1007–1015, Dec. 2025.

[201] A. Buchwalter, J. Kaneshiro, M. Hetzer *et al.*, “Coaching from the sidelines: the nuclear periphery in genome regulation,” *Nature Reviews Genetics*, vol. 20, no. 1, p. 39–50, Oct. 2018.

[202] B. van Steensel and A. Belmont, “Lamina-associated domains: Links with chromosome architecture, heterochromatin, and gene repression,” *Cell*, vol. 169, no. 5, p. 780–791, May 2017.

[203] J. Zhu, Q. Guo, M. Choi *et al.*, “Centromeric and pericentric transcription and transcripts: their intricate relationships, regulation, and functions,” *Chromosoma*, vol. 132, no. 3, p. 211–230, 2023.

[204] G. Chaudhry, J. Chen, L. Snipes *et al.*, “Genomic and epigenomic maps of mouse centromeres and pericentromeres,” *Nature Communications*, vol. 16, no. 1, Nov. 2025.

[205] H. Scherthan, K. Schöfisch, T. Dell *et al.*, “Contrasting behavior of heterochromatic and euchromatic chromosome portions and pericentric genome separation in pre-bouquet spermatocytes of hybrid mice,” *Chromosoma*, vol. 123, no. 6, p. 609–624, Aug. 2014.

[206] A. Rodriguez and P. Bjerling, “The links between chromatin spatial organization and biological function,” *Biochemical Society Transactions*, vol. 41, no. 6, p. 1634–1639, Nov. 2013.

[207] Y. Bozhilov, D. Downes, J. Telenius *et al.*, “A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker,” *Nature Communications*, vol. 12, no. 1, 2021.

[208] H. Loya, G. Kalantzis, and F. Cooper, “A scalable variational inference approach for increased mixed-model association power,” *Nature Genetics*, vol. 57, no. 2, p. 461–468, Jan. 2025.

[209] D. Vuckovic, E. Bao, P. Akbari *et al.*, “The polygenic and monogenic basis of blood traits and diseases,” *Cell*, vol. 182, no. 5, p. 1214–1231.e11, 2020.

[210] M. Pietzner, E. Wheeler, J. Carrasco-Zanini *et al.*, “Mapping the proteo-genomic convergence of human diseases,” *Science*, vol. 374, no. 6569, Nov. 2021.

[211] M. Hemberg and G. Kreiman, “Conservation of transcription factor binding events predicts gene expression across species,” *Nucleic Acids Research*, vol. 39, no. 16, p. 7092–7102, May 2011.

[212] H. Chiba, R. Yamashita, K. Kinoshita *et al.*, “Weak correlation between sequence conservation in promoter regions and in protein-coding regions of human-mouse orthologous gene pairs,” *BMC Genomics*, vol. 9, no. 1, Apr. 2008.

[213] T. Ryba, I. Hiratani, J. Lu *et al.*, "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types," *Genome Research*, vol. 20, no. 6, p. 761–770, Apr. 2010.

[214] N. Ahituv, S. Prabhakar, F. Poulin *et al.*, "Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny," *Human Molecular Genetics*, vol. 14, no. 20, p. 3057–3063, 2005.

[215] A. Tanay, A. Regev, and R. Shamir, "Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast," *Proceedings of the National Academy of Sciences*, vol. 102, no. 20, p. 7203–7208, May 2005.

[216] J. Li, J. Wang, P. Zhang *et al.*, "Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types," *Nature Genetics*, vol. 54, no. 11, p. 1711–1720, Oct. 2022.

[217] M. Wu, C. Kwoh, T. Przytycka *et al.*, "Integration of genomic and epigenomic features to predict meiotic recombination hotspots in human and mouse," p. 297–304, Oct. 2012.

[218] X. Xu, S. Hoang, M. Mayo *et al.*, "Application of machine learning methods to histone methylation ChIP-seq data reveals H4R3me2 globally represses gene expression," *BMC Bioinformatics*, vol. 11, no. 1, 2010.

[219] A. van der Velde, K. Fan, J. Tsuji *et al.*, "Annotation of chromatin states in 66 complete mouse epigenomes during development," *Communications Biology*, vol. 4, no. 1, Feb. 2021.

[220] M. Xiang, R. Grosso, A. Takeda *et al.*, "A single-cell transcriptional roadmap of the mouse and human lymph node lymphatic vasculature," *Frontiers in Cardiovascular Medicine*, vol. 7, Apr. 2020.

[221] X. Jiang, Y. Wang, Z. Xiao *et al.*, "A differentiation roadmap of murine placentation at single-cell resolution," *Cell Discovery*, vol. 9, no. 1, Mar. 2023.

[222] T. T. M. Consortium *et al.*, "Single-cell transcriptomics of 20 mouse organs creates a tabula muris," *Nature*, vol. 562, no. 7727, p. 367–372, Oct. 2018.

[223] L. Minnoye, I. Taskiran, D. Mauduit *et al.*, "Cross-species analysis of enhancer logic using deep learning," *Genome Research*, vol. 30, no. 12, p. 1815–1834, 2020.

[224] T. Lv, Q. Han, Y. Li *et al.*, "Cross-species prediction of histone modifications in plants via deep learning," *Genome Biology*, vol. 27, no. 1, Jan. 2026.

[225] L. Weinstock, J. Schambach, A. Fisher *et al.*, “Hybrid CNN and multi-head attention model for analyzing epigenetic mechanisms and gene expression across fungal phylogenetic distances,” Dec. 2024.

[226] Y. Cheng, Z. Ma, B. Kim *et al.*, “Principles of regulatory information conservation between mouse and human,” *Nature*, vol. 515, no. 7527, p. 371–375, Nov. 2014.

[227] K. Prüfer, F. Racimo, N. Patterson *et al.*, “The complete genome sequence of a Neanderthal from the Altai mountains,” *Nature*, vol. 505, no. 7481, p. 43–49, Dec. 2013.

[228] V. Slon, F. Mafessoni, B. Vernot *et al.*, “The genome of the offspring of a Neanderthal mother and a Denisovan father,” *Nature*, vol. 561, no. 7721, p. 113–116, Aug. 2018.

[229] A. King, D. Songdej, D. Downes *et al.*, “Reactivation of a developmentally silenced embryonic globin gene,” *Nature Communications*, vol. 12, no. 1, 2021.

[230] J. Hughes, N. Roberts, S. McGowan *et al.*, “Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment,” *Nature Genetics*, vol. 46, no. 2, p. 205–212, Jan. 2014.

[231] F. Yue, Y. Cheng, A. Breschi *et al.*, “A comparative encyclopedia of DNA elements in the mouse genome,” *Nature*, vol. 515, no. 7527, p. 355–364, Nov. 2014.

[232] E. Gushchanskaya, A. Artemov, S. Ulyanov *et al.*, “The clustering of CpG islands may constitute an important determinant of the 3D organization of interphase chromosomes,” *Epigenetics*, vol. 9, no. 7, p. 951–963, Apr. 2014.

[233] L. Li, T. Comi, R. Bierman *et al.*, “Recurrent gene flow between Neanderthals and modern humans over the past 200,000 years,” *Science*, vol. 385, no. 6705, 2024.

[234] D. Reich, N. Patterson, M. Kircher *et al.*, “Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania,” *The American Journal of Human Genetics*, vol. 89, no. 4, p. 516–528, Oct. 2011.

[235] K. Mack and M. Nachman, “Gene regulation and speciation,” *Trends in Genetics*, vol. 33, no. 1, p. 68–80, Jan. 2017.

[236] D. Dimcheff, S. Drovetski, and D. Mindell, “Phylogeny of tetraoninae and other galliform birds using mitochondrial 12S and ND2 genes,” *Molecular Phylogenetics and Evolution*, vol. 24, no. 2, p. 203–215, Aug. 2002.

[237] M. G. S. Consortium, R. Waterston *et al.*, *Nature*, vol. 420, no. 6915, p. 520–562, Dec. 2002. [Online]. Available: http://dx.doi.org/10.1038/nature01262

[238] T. Murad, S. Ali, and M. Patterson, "Exploring the potential of GANs in biological sequence analysis," *Biology*, vol. 12, no. 6, p. 854, 2023.

[239] T. Montavon and D. Duboule, "Chromatin organization and global regulation of Hox gene clusters," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1620, p. 20120367, 2013.

[240] R. Jha, F. Kouzine, and D. Levens, "MYC function and regulation in physiological perspective," *Frontiers in Cell and Developmental Biology*, vol. 11, Oct. 2023.

[241] A. Caporale, C. Gonda, and L. Franchini, "Transcriptional enhancers in the FOXP2 locus underwent accelerated evolution in the human lineage," *Molecular Biology and Evolution*, vol. 36, no. 11, p. 2432–2450, 2019.

[242] B. Kragesteen, A. Giladi, E. David *et al.*, "The transcriptional and regulatory identity of erythropoietin producing cells," *Nature Medicine*, vol. 29, no. 5, p. 1191–1200, Apr. 2023.

[243] B. Azarkhalili and M. Libbrecht, "Generalized attention flow: Feature attribution for transformer models via maximum flow," 2025.

[244] F. Mehri, M. Baghshah, M. Pilehvar *et al.*, "LibraGrad: Balancing gradient flow for universally better vision transformer attributions," 2024.

[245] S. Riva, E. Sanders, E. Georgiades *et al.*, "Deciphering cis-regulatory elements using REgulamentary," *Bioinformatics Advances*, vol. 6, no. 1, 2026.

[246] M. Meyer, M. Kircher, M. Gansauge *et al.*, "A high-coverage genome sequence from an archaic Denisovan individual," *Science*, vol. 338, no. 6104, p. 222–226, Oct. 2012.

[247] R. Green, J. Krause, A. Briggs *et al.*, "A draft sequence of the Neandertal genome," *Science*, vol. 328, no. 5979, p. 710–722, May 2010.

[248] L. Skov, S. Peyrégne, D. Popli *et al.*, "Genetic insights into the social organization of Neanderthals," *Nature*, vol. 610, no. 7932, p. 519–525, Oct. 2022.

[249] K. Prüfer, C. de Filippo, S. Grote *et al.*, "A high-coverage Neandertal genome from Vindija Cave in Croatia," *Science*, vol. 358, no. 6363, p. 655–658, Nov. 2017.

[250] Q. Fu, H. Li, P. Moorjani *et al.*, “Genome sequence of a 45,000-year-old modern human from western Siberia,” *Nature*, vol. 514, no. 7523, p. 445–449, Oct. 2014.

[251] D. Gokhman, M. Nissim-Rafinia, L. Agranat-Tamir *et al.*, “Differential DNA methylation of vocal and facial anatomy genes in modern humans,” *Nature Communications*, vol. 11, no. 1, Mar. 2020.

[252] I. Olalde, E. Altena, Q. Bourgeois *et al.*, “Lasting lower Rhine–Meuse forager ancestry shaped Bell Beaker expansion,” *Nature*, vol. 652, no. 8111, p. 938–946, Feb. 2026.