Developing computational and machine learning techniques to robustly detect the genome's cell type-specific, protein coding and non-coding effects in Alzheimer's disease

Alan E. Murphy

A thesis submitted to Imperial College London for the degree of Doctor of Philosophy (PhD) 2024.

Department of Brain Sciences,

Faculty of Medicine,

Imperial College London.

Thesis abstract

Furthering our understanding into the cell's transcriptional and regulatory mechanisms and how changes in these relate to Alzheimer's disease (AD), promises to unravel the aetiology of this complex disease, potentially yielding improved therapeutics. However, doing so requires both AD-specific and broader advancements in experimental assays and computational approaches in the field. Here, we attempt to address such shortcomings in computational analyses, covering both the protein coding and non-coding genome. These approaches include the standardisation of processing and quality control of genetic information, the standardisation of analysis for cell type-specific transcriptional changes—both broadly and in the study of AD—prioritising functional and disease-relevant genomic loci in silico using deep learning to link epigenetics to transcription, and predicting cell type-specific effects of genetic variants while accounting for distal regulation with genomic deep learning models. Although the focus of this work is AD, the developed, open-source, computational and deep learning techniques are all broadly applicable to our comprehension of the cis-regulatory code.

The work presented in this thesis was funded by the UK Dementia Research Institute which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

Approximate word count: 65,000

Declarations

Chapter 2

Dr. Nathan Skene and I jointly conceived of the study. I undertook all analyses, and drafted the manuscript which was reviewed by Dr. Nathan Skene.

Chapter 3

Dr. Nathan Skene and I jointly conceived of the study whilst I executed it. Dr. Nurun Fancy conducted the quality control and processing of the data. I wrote the manuscript which was reviewed by both Dr. Nurun Fancy and Dr. Nathan Skene.

Chapter 4

Dr. Nathan Skene and I jointly conceived of the study whilst I undertook the analyses. Dr. Brian Schilder added functionality for integration with IEU GWAS's web server. I drafted the manuscript which was reviewed by Dr. Brian Schilder and Dr. Nathan Skene.

Chapter 5

Dr. Nathan Skene and I jointly conceived of the study whilst I undertook all analyses. Mr. William Beardall wrote an initial script to copy over Enformer's weights. Dr. Marek Rei and Dr. Mike Phuycharoen provided guidance on project direction. I undertook all other analyses and drafted the manuscript which was reviewed by Dr. Mike Phuycharoen and Dr. Nathan Skene.

Chapter 6

Dr. Sarah Marzi and I jointly conceived of the study whilst I undertook the analyses and drafted the manuscript. Ms. Aydan Askarova wrote the original scripts for running LDSC which I adapted for this work. Dr. Boris Lenhard, Dr. Nathan Skene and Dr. Sarah Marzi all reviewed the manuscript.

Acknowledgements

Completing a PhD thesis is an immense and, at times, seemingly insurmountable task that would not have been possible without the countless hours of help and support from so many people. I firmly believe that its completion is more a testament to the strength of the network and community surrounding the student, rather than a reflection of any individual's merit. The guidance, encouragement, and expertise I received have been invaluable throughout this journey, and I am deeply grateful to everyone who has contributed - sometimes unknowingly - to the development of this work. While I can only mention a few here, I recognise that this is but the tip of the iceberg.

I would firstly like to thank my PhD supervisor Nathan Skene, whose unwavering support, guidance, and mentorship have been instrumental in shaping this thesis. Your expertise, patience, and encouragement have not only enhanced my PhD thesis but also fostered my growth as a researcher. You have always viewed me as a peer rather than a student and your belief in my abilities has been a constant source of inspiration. I feel privileged to have worked under your supervision. Thank you for being an exceptional mentor and for making this journey a rewarding experience.

I would also like to extend my heartfelt thanks to my (unofficial) supervisor, Sarah, whose encouragement to pursue my interests ignited a genuine passion for addressing fundamental biological questions. Your guidance, input, and unwavering optimism were instrumental in the development of Chapter 6, which I dedicate to you. This chapter would not have been possible without your support. Thank you also to the brilliant scientists Alexi, Eugene and Mike for their support and advice but more importantly, for listening to my frequent rants

and complaints about my research – your counsel and constant encouragement helped ease my momentary pessimism.

Thank you to all my friends in the Skene, Nott and Marzi labs, especially Maria and Roxy who have been a massive support. To Brian, who I have learned so much from, thank you for all your knowledge and insight. To Kitty, you have been my closest friend, I don't think I could have survived in London, let alone in the group, without you. You have been a pillar of support to my time here. Moreover, to everyone on the 7th floor of the Sir Michael Uren Hub at Imperial, thank you for creating such a special, open work environment. I am grateful to have been a part of it. A special thank you goes to Jennifer for fostering this community and for her kind support and attentiveness.

To my family, thank you for your constant support, you always encouraged me to follow my passions even though you didn't really know what it is I was working on or even what I really do for that matter. Especially my mam, everything I have ever achieved, I owe to you. You have celebrated all of my academic achievements. A special thank you also to John and Dorothy, who watched me go from a 'proper' job to pursue my research passion, without your constant support, I'm not sure this would have been possible. To Zoe, who knew far too well what I was working on, thank you for your constant support, love and encouragement - I'm sorry you had to read every sentence I've ever written. This thesis is for you because without you, I would not even be in research. Thank you for making me the scientist I am today and for continuing to chase me across the world to pursue my research.

Finally, I would like to thank all co-authors I have not already mentioned; Nurun, William, Aydan, Marek, Mike Phuycharoen and Boris. Your contributions and feedback vastly improved this and the publications from this work.

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Statement of originality

I, Alan Murphy, declare that this thesis has been composed solely by myself and that it has not been submitted, either in its entirety or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgment, the work presented herein is entirely my own.

Publications arising from this thesis

Chapter 2

Murphy, A.E. and Skene, N.G. (2022) 'A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis', Nature Communications, 13(1), p. 7851. Available at: https://doi.org/10.1038/s41467-022-35519-4.

Chapter 3

Murphy, A.E., Fancy, N. and Skene, N. (2023) 'Avoiding false discoveries in single-cell RNA-seq by revisiting the first Alzheimer's disease dataset', eLife. Edited by J.-Y. An and M. Choi, 12, p. RP90214. Available at: https://doi.org/10.7554/eLife.90214.

Chapter 4

Murphy, A.E., Schilder, B.M. and Skene, N.G. (2021) 'MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics', Bioinformatics, 37(23), pp. 4593–4596. Available at: https://doi.org/10.1093/bioinformatics/btab665.

Chapter 5

Murphy, A.E., Beardall, W., Rei, M. et al. (2024) 'Predicting cell type-specific epigenomic profiles accounting for distal genetic effects'. Nat Commun 15, 9951. Available at: https://doi.org/10.1038/s41467-024-54441-5.

Chapter 6

Murphy, A.E., Askarova, A., et al. (2024) 'Predicting gene expression from histone marks using chromatin deep learning models depends on histone mark function, regulatory distance and cellular states'. Nucleic Acids Research, 2024;, gkae1212. Available at: https://doi.org/10.1093/nar/gkae1212.

List of abbreviations

Abbreviation	Meaning
AD	Alzheimer's disease
ALS	Amyotrophic Lateral Sclerosis
AP	Average-Precision
ARC	Abstraction and Reasoning Corpus
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
AUC	Area Under the receiver operating characteristics Curve
Aβ	β-amyloid peptide
caQTL	chromatin accessibility Quantitative Trait Loci study
cDNA	complementary DNA
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
cDNA	complementary DeoxyriboNucleic Acid
cM	centiMorgan
CNVs	Copy Number Variations
CRISPRi	CRISPR interference
DE	Differential Expression
DEG	Differentially Expressed gene
DNA	DeoxyriboNucleic Acid
DNNs	Deep Neural Networks
EMA	European Medicines Agency
eQTL	expression Quantitative Trait Loci study
ESC	Embryonic Stem Cell
EWCE	Expression Weighted Celltype Enrichment
FDA	Food and Drug Administration
FDR	False Discovery Rate
GLM	Generalised Mixed Model
gLMs	genomic Language Models
GPU	Graphical Processing Units
GTEx	The Genotype-Tissue Expression
GWAS	Genome-Wide Association Studies
GWAS-VCF	GWAS Variant Call Format
HDAC	Histone DeACetylase
HGP	The Human Genome Project
hQTL	histone mark Quantitative Trait Loci study
hQTL	histone mark Quantitative Trait Loci
IGVF	Genomic Variation on Function
Indel	insertion/deletions
LD	Linkage Disequilibrium
LDSC	Linkage Disequilibrium Score Regression

LFC	Log2 Fold Change
LLM	Large Language Model
MAD	Median Absolute Deviation
MAF	Minor Allele Frequency
MAGMA	Multi-marker Analysis of GenoMic Annotation
MAVEs	Multiplex Assays of Variant Effects
MCC	Matthews Correlation Coefficient
MHC	Major Histocompatibility Complex
MHC	Major Histocompatibility Complex
MMR	MisMatch Repair pathway
MPRA	Massively Parallel Reporter Assay
mQTL	methylation Quantitative Trait Loci study
mRNA	messenger RNA
MSE	Mean Squared Error
NAFLD	Non-Alcoholic Fatty Liver Disease
NPC	Neuronal Progenitor Cell
NPL	Natural Language Processing
OPC	Oligodendrocyte Precursor Cell
PIP	Posterior Inclusion Probability
QC	Quality Control
QTL	Quantitative Trait Loci
ReLU	Rectified Linear Unit
RNA	RiboNucleic Acid
RNA-Seq	RNA Sequencing
ROC	Receiver Operating Characteristics
RPKM	Reads Per Kilobase of transcript, per Million mapped reads
RVCD	Rare Variant associations with Common Diseases
s-LDSC	stratified LD Score regression
SCREEN	Search Candidate cis-Regulatory Elements by ENCODE
scRNA-Seq	single-cell RNA-Sequencing
SHH	*Sonic Hedgehog*
SLDP	Signed Linkage Disequilibrium Profile regression
SNP	Single Nucleotide Polymorphism
snRNA-Seq	Single-Nucleus RNA-Sequencing
SNV	Single Nucleotide Variant
T2T	Telomere-to-Telomere
tfQTL	transcription factor Quantitative Trait Loci study
TPM	Transcripts Per Million
TSS	Transcriptional Start Site
UMAP	Uniform-Manifold Approximation and Projection
VCF	Variant Call Format

xQTLs	differing Quantitative Trait Loci studies
ZRS	Zone of polarizing activity Regulatory Sequence

Table of contents

1. INTRODUCTION 24

1.1. Neurodegenerative diseases and Alzheimer's disease 24

1.2. The protein-coding genome 29

1.2.1. Gene expression 29

1.2.2. Single-nucleus transcriptomic analyses in Alzheimer's disease 32

1.3. The non-coding genome 34

1.3.1. Epigenetics 34

1.3.2. Epigenetics analyses and its role in Alzheimer's disease 37

1.4. Genetic variation 39

1.4.1. The study of various types of variation 39

1.4.2. The evolution of genetic research post-human genome project 40

1.4.3. Distribution of genetic variant information 41

1.4.4. GWAS – early results and limitations 43

1.4.5. GWAS results and downstream analyses in Alzheimer's disease 44

1.5. Machine learning and Deep learning for genomics 46

1.5.1. Motivating learning the genome's code 46

1.5.2. Deep learning's broad applicability in genomics 48

1.5.3. The necessity of deep learning in understanding genetic variants 49

1.5.4. How to utilise deep learning for genetic variant effect prediction 51

1.6. Linking epigenetics to expression – understanding the regulation of the coding genome 54

1.7. The study of the protein-coding and non-coding genome with respect to genetic variants as a tool for precision medicine in Alzheimer's disease 57

1.8. Aims of this thesis 60

2. A BALANCED MEASURE SHOWS SUPERIOR PERFORMANCE OF PSEUDOBULK METHODS IN SINGLE-CELL RNA-SEQUENCING ANALYSIS 62

2.1. Introduction 62
2.2. Methods 65
2.2.1. Differential expression techniques 65
2.2.2. Data simulation 67
2.2.3. Benchmarking 68
2.2.4. Evaluation 68
2.3. Results 69
2.4. Discussion 78
2.5. Data availability 80
2.6. Code availability 80

3. AVOIDING FALSE DISCOVERIES IN SINGLE CELL RNA-SEQ BY REVISITING THE FIRST ALZHEIMER'S DISEASE DATASET 81

3.1. Introduction 81
3.2. Methods 82
3.2.1. Processing of the snRNA-Seq dataset 82
3.2.2. Differential Expression analysis 86
3.3. Results and Discussion 90
3.4. Data availability 102
3.5. Code availability 102

4. MUNGESUMSTATS: A BIOCONDUCTOR PACKAGE FOR THE STANDARDISATION AND QUALITY CONTROL OF MANY GWAS SUMMARY STATISTICS 103

4.1. Introduction 103
4.2. Heterogeneity in GWAS Formats 104
4.3. Implementation 107
4.4. Usage 111
4.5. Discussion 112
4.6. Data availability 113
4.7. Code availability 113

5. PREDICTING CELL TYPE-SPECIFIC EPIGENOMIC PROFILES ACCOUNTING FOR DISTAL GENETIC EFFECTS 114

5.1. INTRODUCTION 114

5.2. METHODS 117

5.2.1. Data collection and processing 117

5.2.2. Calculating genetic variant distance from epigenetic effect peak in QTL associations 120

5.2.3. Training data 120

5.2.4. Pre-trained Enformer model 122

5.2.5. Enformer Celltyping architecture 123

5.2.6. Enformer Celltyping Training 126

5.2.7. Testing receptive field 127

5.2.8. Visualising cell type embedding 128

5.2.9. Benchmarking 128

5.2.10. Highly active enhancer (super-enhancer) prediction in cancer cell lines 130

5.2.11. Off-centre correlation analysis 131

5.2.12. Histone mark QTL SLDP 132

5.2.13. Global chromatin accessibility signal – cell type-specific motif enrichment 134

5.2.14. Cell type-specific genetic enrichment for complex traits 135

5.3. RESULTS 137

5.4. DISCUSSION 170

5.5. CONCLUSION 176

5.6. DATA AVAILABILITY 177

5.7. CODE AVAILABILITY 177

5.8. ACKNOWLEDGEMENTS 177

5.9. FUNDING 178

6. PREDICTING GENE EXPRESSION FROM HISTONE MARKS USING CHROMATIN DEEP LEARNING MODELS DEPENDS ON HISTONE MARK FUNCTION, REGULATORY DISTANCE AND CELLULAR STATES 179

6.1. INTRODUCTION 179

6.2. METHODS 184
6.2.1. Data collection and processing 184
6.2.2. Promoter model 185
6.2.3. Distal model 186
6.2.4. Model training 186
6.2.5. Histone mark levels 187
6.2.6. Gene expression state 187
6.2.7. Correlation analysis 188
6.2.8. In silico perturbation of histone mark activity levels 188
6.2.9. In silico perturbation enrichment in quantitative trait loci studies 189
6.2.10. In silico perturbation disease enrichment 191
6.3. RESULTS 193
6.3.1. Active histone marks prove most informative at promoter regions 193
6.3.2. The predictive performance of active and repressive marks differ based on cell state at promoter regions 196
6.3.3. Higher histone mark levels result in better predictive performance at promoter regions 198
6.3.4. Active marks are most predictive in the distal model. 202
6.3.5. Receptive field expansion and multi-histone mark predictions yield diminishing returns. 205
6.3.6. In silico histone mark perturbation prioritises functionally relevant genomic loci and disease relevant cell types. 209
6.4. DISCUSSION 218
6.5. DATA AND CODE AVAILABILITY 225
6.6. FUNDING 226
7. GENERAL DISCUSSION 227
7.1. KEY FINDINGS 228
7.1.1. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis 228

7.1.2. Avoiding false discoveries in single-cell RNA-Seq by revisiting the first Alzheimer's disease dataset 228

7.1.3. A tool for the standardisation and quality control of many GWAS summary statistics 229

7.1.4. Predicting cell type-specific epigenomic profiles accounting for distal genetic effects.....230

7.1.5. Predicting gene expression from histone marks using chromatin deep learning models depends on histone mark function, regulatory distance and cellular states.....231

7.2. IMPLICATIONS OF MY FINDINGS.....232

7.2.1. The standardisation of analysis for cell type-specific transcriptional changes 232

7.2.2. The standardisation of processing and quality control of genetic information.....233

7.2.3. Predicting the cell type-specific effects of genetic variants while accounting for distal regulation 234

7.2.4. Prioritising functional and disease relevant genomic loci in silico with deep learning by linking epigenetics to transcription.....234

7.3. FUTURE DIRECTIONS FOR THE FIELD 236

7.3.1. Uncovering the cis-regulatory code of AD.....236

7.3.2. Improving the utility of genomic deep learning models for personalised genomics 238

7.3.3. Benchmarking genomic deep learning models for personalised genomics.....244

7.3.4. An alternative approach to prioritising disease-related genomic loci.....248

7.4. FINAL WORDS 249

APPENDICES.....280

APPENDIX A.....281

APPENDIX B.....302

Table of figures

Figure 1-1: The substantial heritability of neurodegenerative diseases 26

Figure 1-2: The exponential increase of (brain) single-cell and single-nucleus RNA-Seq studies. 31

Figure 1-3: Overview of the different facets of Alzheimer's disease's regulatory code covered in this PhD thesis. 59

Figure 2-1: Type 1 error highlights the benefits of pseudobulk approaches. 64

Figure 2-2: Balanced performance of analysed models shows superiority of pseudobulk. 70

Figure 2-3: Pseudobulk mean is best performing at a constant type 1 error rate. 74

Figure 2-4*:* Pseudobulk mean is the best performing on imbalanced datasets 77

Figure 3-1: Differing processing results in disparate proportions of cell types across analyses. 85

Figure 3-2: Pseudobulk differential expression results in far less dubious disease-related genes. 91

Figure 3-3: The nuclei that were removed from our quality control approach as their proportion of mitochondrial reads were ≥10%, but kept in the authors'. 95

Figure 4-1: Most common summary statistic formats. 105

Figure 5-1 Enformer Celltyping predicts cell type-specific histone marks accounting for distal DNA regulation. 138

Figure 5-2 Enformer Celltyping predicts cell type-specific histone marks signals. 141

Figure 5-3 Enformer Celltyping makes best-in-class cell type-specific histone mark predictions. 145

Figure 5-4 Enformer Celltyping has superior performance at functionally relevant regions. 149

Figure 5-5 Enformer Celltyping can predict in cell types generated outside of ENCODE. . 153

Figure 5-6 Genomic deep learning models' genetic effect predictions predominately do not replicate QTL results. 159

Figure 5-7 Enformer Celltyping identifies motifs for cell type-specific transcription factors which are enriched based on the global chromatin accessibility signal. 162

Figure 5-8 Enformer Celltyping's predictions capture cell type-specific genetic enrichment for complex traits. 166

Figure 6-1: Understanding the complex relationship between histone mark activity and gene expression in silico. 182

Figure 6-2 The performance of the promoter model shows the best gene expression predictions based on active histone marks H3K4me3, H3K27ac and H3K9ac. 195

Figure 6-3 The predictive performance of the promoter model across samples clusters by histone mark function. 197

Figure 6-4 Higher histone mark levels are associated with better predictive performance in specific gene expression states. 200

Figure 6-5 Active histone modifications performed best at predicting gene expression in the distal model. 203

Figure 6-6 Pairwise combinatorial performance of distal model in predicting gene expression based on two histone marks. 208

Figure 6-7 Effect of in silico histone mark perturbation on expression. 212

Figure 6-8 In silico histone mark perturbation experiment highlights functional and disease enrichment. 215

Figure 6-9 In silico histone mark perturbation visualised for DSTN in lung carcinoma cell line. 217

Figure 7-1: Differing genetic variant data available for incorporation to genomic deep learning model's training regime 242

Table of tables

Table 1-1: The primary genomic location, transcriptional relationship, and proposed function of a variety of known histone marks. 56

Table 2-1: The different methods benchmarked in the analysis with their implementation approach and methodological types 66

Table 2-2: The area under the receiver operating characteristics curve (AUC) of all benchmarked models. 72

Table 3-1: Quality control summary. 84

Table 3-2: Differentially expressed genes from our reanalysis. 87

Table 3-3: Differentially expressed genes on reprocessed data. 89

Table 3-4: Strong concordance across different DE's significant genes' fold change values: 98

Table 4-1: Implemented quality control and standardisation checks. 110

1. Introduction

This chapter serves to introduce cells' transcriptional and regulatory mechanisms (known as the cis-regulatory code) and how changes in these relate to Alzheimer's disease (AD). It provides an overview of the protein-coding and the regulatory genome as well as genetic variants and their effect and presents the main aims of this PhD thesis to elucidate the genome's cell type-specific effects in AD.

1.1. Neurodegenerative diseases and Alzheimer's disease

AD is a neurodegenerative disease with a rapidly increasing global incidence that is expected to triple to approximately 150 million cases over the next 30 years (Nichols *et al.*, 2022). AD is defined by neuronal loss and gliosis and is thought to affect the hippocampus and entorhinal cortex at the early stages of the disease, regions with a major role in memory and learning (Rao *et al.*, 2022; Erkkinen, Kim and Geschwind, 2018; Kampmann, 2024). Importantly, there is selective vulnerability of specific neuronal subtypes in AD rather than equal loss. For example, cholinergic neurons in the nucleus basalis of Meynert which forms part of the basal forebrain (Whitehouse *et al.*, 1982) and pigmented cells, a form of noradrenergic neurons, in the locus coeruleus of the brainstem (Bondareff, Mountjoy and Roth, 1981; Tomlinson, Irving and Blessed, 1981) are both selectively lost (Kampmann, 2024). It is also hypothesised that synaptic loss is a crucial stage within AD pathogenesis that occurs prior to the loss of neurons which has been demonstrated in animal models and cell cultures (Meftah and Gan, 2023).

Despite the growing incidence of AD, there are limited treatment options available for those impacted. Moreover, the treatments available do not prevent the onset or progression of the disease rather they target symptoms (Breijyeh and Karaman, 2020). Two such treatments that

have been approved by the Food and Drug Administration (FDA) regulatory body are lecanemab and donanemab (Cummings et al., 2024). Although, the European Medicines Agency (EMA) has more recently rejected a licence for lecanemab due to its moderate benefits and known side effects. Both of these monoclonal antibodies have shown moderate slowing of cognitive decline, by clearing build-ups of extracellular plaque deposits of the β-amyloid peptide (Aβ) in the brain – a hallmark of AD (Nilsberth et al., 2001). However, despite the accumulation of extracellular Aβ plaques and intracellular tau-containing neurofibrillary tangles (Arnold et al., 1991) both being prominent features of AD, the question remains whether they have a causal role in the onset and progression of the disease? Evidence refuting this claim is that high levels of Aβ plaques and neurofibrillary tangles have also been found in older individuals who exhibit no cognitive impairment (Bennett et al., 2006). Thus, the mechanisms that underlie AD and lead to the initial synaptic loss followed by neuronal loss and gliosis and therefore resulting in cognitive impairment, remains elusive (Knopman *et al.*, 2021).

To study the mechanisms of AD, one can simplify and group the causal effects on a phenotype (P) into those resulting from genetics (G) or the environment (E) and their interaction (Wild, 2005):

$$P = G \times E \text{ or equivalently, } P = f(G, E)$$

An understanding of how much each component contributes, through methods such as twin studies (Craig *et al.*, 2020), is paramount for our understanding of the phenotype in question. Studies which aim to derive the contribution of genetics to this equation have found that most neurodegenerative diseases, including AD, are characterized by substantial heritability **Figure 1-1**.

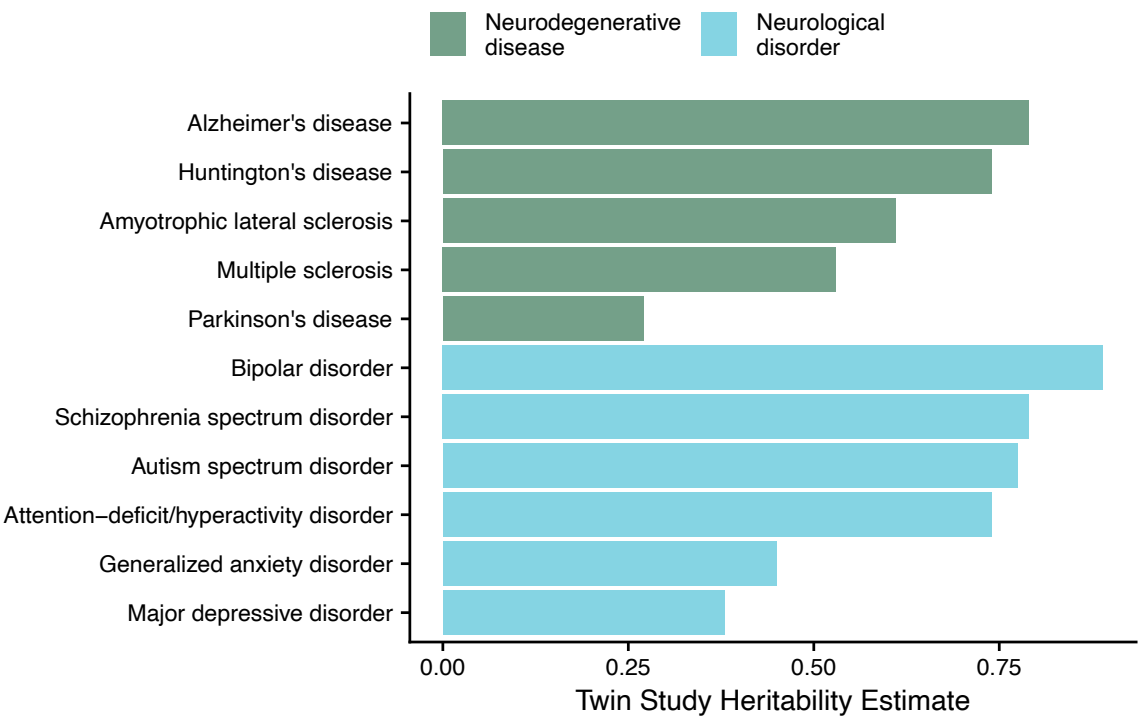


Figure 1-1: The substantial heritability of neurodegenerative diseases

The heritability (y-axis) of neurodegenerative diseases (green) and neurological disorders estimated by twin studies. Heritability estimated derived for Alzheimer's disease (Gatz *et al.*, 2006), Huntington's disease (Li *et al.*, 2006), Amyotrophic lateral sclerosis (Al-Chalabi *et al.*, 2010), multiple sclerosis (Ristori *et al.*, 2006), Parkinson's disease (Goldman *et al.*, 2019), bipolar disorder (McGuffin *et al.*, 2003), schizophrenia (Hilker *et al.*, 2018), autism spectrum disorder (Tick *et al.*, 2016), attention-deficit/hyperactivity disorder (Faraone and Larsson, 2019), generalised anxiety disorder (Davies *et al.*, 2015) and major depressive disorder (Kendler *et al.*, 2006).

Most neurodegenerative diseases have significant genetic components, such as frontotemporal dementia (Chow *et al.*, 1999), spinal muscular atrophy (Keinath, Prior and Prior, 2021) and spinocerebellar ataxia (Paulson, 2009), but are less prevalent and thus could not be estimated by classic twin studies. However, it is important to note that heritability estimates from twin studies can be inflated upwards due to assumption of a shared environment across monozygotic and dizygotic twins (i.e. the equal environments assumption) which may not be true for all cases (Boomsma, Busjahn and Peltonen, 2002; Craig *et al.*, 2020; Hagenbeek *et al.*, 2023). One major exception to the significant genetic component of neurodegenerative diseases, in **Figure 1-1**, is Parkinson's disease which has a relatively low heritability, estimated at 0.27 from twin studies (Goldman *et al.*, 2019). This highlights the strong contribution of environmental exposures which have been investigated, including pesticide exposures, smoking tobacco and brain injuries (Tsalenchuk, Gentleman and Marzi, 2023).

The high heritability of AD warrants substantial investment in the study of the genetic variants instigated in the disease. While recent large-scale genetic efforts – genome-wide association studies (GWAS), have associated genetic variants with a myriad of neurodegenerative diseases, these often lie in non-coding, regulatory regions and cannot be definitively linked with a specific gene. In fact, 98% of all single nucleotide polymorphisms (SNPs) identified in a recent genome-wide, meta-analysis of AD were found in non-coding, regulatory regions (I. E. Jansen *et al.*, 2019). Some of these variants have already been experimentally validated to occur at distal regulatory elements. For example, recently, Hi-C chromatin interaction assays have been conducted in the prefrontal cortex of AD patients and matched controls, finding distal chromatin looping interactions between known AD GWAS SNPs and disease associated genes (Meng *et al.*, 2023). Such interactions can be distal, for example, extremely distal

regulators have been extensively studied for other diseases/traits including the canonical example of the limb enhancer region which regulates expression of sonic hedgehog (*SHH*) gene. This enhancer, the Zone of Polarizing Activity Regulatory Sequence (ZRS), is approximately 800,000 to 1,000,000 base pairs away (in mouse and human) and found within the non-coding, intron of a neighbouring gene but when eliminated, causes truncation of limbs in mice (Sagai *et al.*, 2005; Visel, Rubin and Pennacchio, 2009). Moreover, this regulation can be altered by single base pair changes in humans - multiple SNPs in this same regulatory region have been identified in humans with preaxial polydactyly, a congenital malformation disorder (Lettice *et al.*, 2003). Thus, distal regulation should be considered to capture the effect of all AD GWAS SNPs.

Understanding the role of these regulatory variants is further burdened due to gene regulatory mechanisms' highly cell type-specific nature (Schreiber, Bilmes and Noble, 2020) – the selective vulnerability of AD emphasises this. Recent studies have repeatedly found that the genetic risk for AD acts primarily via microglia - the resident macrophages of the central nervous system (Skene and Grant, 2016; Nott *et al.*, 2019) but the function of regulatory variants in microglia will differ to that in other glia and neuronal cell types. Moreover, the genes and pathways by which these genetic variants act, leading to the onset and progression of the disease remain elusive. Thus, greater insight into the cell type-specific effects of non-coding genetic variants on cells' epigenome and transcriptome would further our understanding of the aetiology and treatment of AD.

1.2. The protein-coding genome

The protein-coding portion of the genome, also known as the exome, accounts for just 1.5% of the genome's entirety but contains all the sequence transcribed into RNA and subsequently translated into proteins (Marian, 2014). Thus, understanding the exome is a fundamental part of unravelling the complexities of neurodegenerative diseases. In the context of AD, and notably aberrant in gene expression profiling, alterations in specific protein-coding genes have already been implicated in the pathogenesis of the disease. These include *APOE*, *APP*, *PSEN1* and *PSEN2*. Susceptible variants in *APOE*, which encodes the ApoE protein and is produced predominantly by astrocytes and activated microglia in the brain, are the greatest genetic risk factor for AD (Kim, Basak and Holtzman, 2009). Other known risk factors include mutations in *APP* (encoding the amyloid precursor protein), *PSEN1* (encoding presenilin 1) and *PSEN2* (encoding presenilin 2) and account for the majority of cases of dominantly inherited AD (Haass *et al.*, 2012; Knopman *et al.*, 2021). The question remains whether further investigation into the exome, with larger studies and better techniques, will uncover the role of other AD relevant genes and the cell types in which they function.

1.2.1. Gene expression

Expression profiling using deoxyribonucleic acid (DNA) arrays has been utilised in the research of cell biology and disease since the early 1990s (Jordan, 2002). Its popularity grew with DNA microarrays which enabled high-throughput experiments, simultaneously measuring the relative concentrations of messenger ribonucleic acid (mRNA) relating to thousands of genes in a single experiment (Schena *et al.*, 1995). RNA has an integral role in the central dogma of biology which entails how DNA is transcribed to create RNA which is translated into proteins (CRICK, 1958). However, assaying expression levels also sheds light on non-coding RNAs,

which are not translated into proteins but possess regulatory functions (Delihas, 2015). The development of high-throughput, next-generation sequencing technologies and the RNA sequencing (RNA-Seq) method, which profiles RNA from the sequencing of complementary DNA (cDNA), further propelled the field (Wang, Gerstein and Snyder, 2009). RNA-Seq offered a full view of the whole transcriptome, i.e. the whole repertoire of transcripts for the particular tissue of cells, including allele-specific expression and alternative splicing. This was a notable improvement on microarrays which relied on pre-defined transcripts or genes, and thus, enabled discovery of new, rare transcripts (Rao *et al.*, 2019).

One drawback was that RNA-Seq could only detect the average gene expression of all the cells analysed; it cannot identify the heterogeneity between cells, masking the transcriptional signatures of minority cell types. Single-cell RNA-Seq (scRNA-Seq) addressed this issue by isolating cells using methods such as microfluidics, microwells, droplet-based or *in situ* barcoding to co-encapsulate it with a unique DNA barcoded bead (Jovic *et al.*, 2022). This meant mRNA could be linked back to the cell of origin, enabling the quantification of RNA abundance at a cellular level and opening up a new avenue of research. This meant that new insight into cellular transcriptomes could be measured avoiding confounding by cell type-specific signals and changes in proportional abundances found in bulk tissue RNA-Seq (Srinivasan *et al.*, 2016). This led to an explosion of research in the area which still exhibits exponential growth today (**Figure 1-2**).

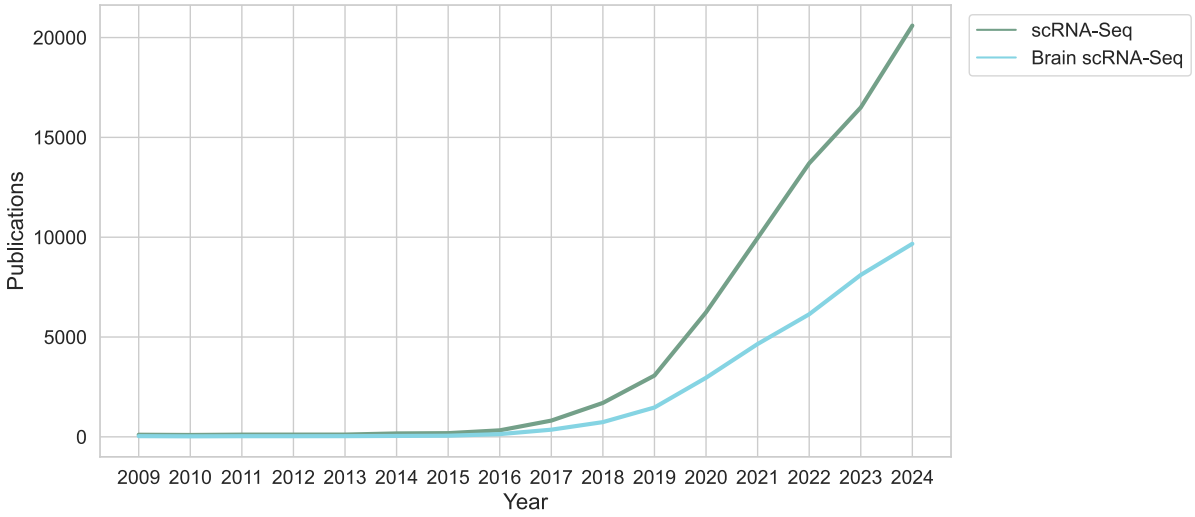


Figure 1-2: The exponential increase of (brain) single-cell and single-nucleus RNA-Seq studies.

The number of publications (Y-axis) from Google Scholar mentioning single-cell or single-nucleus RNA-Seq (scRNA-Seq) by year (X-axis) up to 2024. Run on January 1st, 2025.

Despite the growth of research in scRNA-Seq, there was little consensus for standards of the bioinformatic techniques used to analyse the data in this new field. Although standardisation and reproducibility has improved for data pre-processing (Khozoie *et al.*, 2021; Peltzer *et al.*, 2024), there continued to be little agreement on methods to identifying cell type-specific differentially expressed genes (DEGs). There are three main approaches which have become popular in the field – pseudoreplication, mixed models and pseudobulk approaches. Pseudoreplication approaches, such as ROTS (Suomi *et al.*, 2017), take cells from the same individual as independent replicates. Mixed models, such as generalised mixed models (GLMs), aim to account for both subject and cell-level information often using a random effect for samples to account for the group's subject-level heterogeneity. Whilst, pseudobulk approaches aggregate a cell type's reads to an individual often by sum or mean to help avoid issues with cell dropout and low sequencing depths. The field has predominately accepted that pseudoreplication approaches lead to over estimation of confidence and thus substantial numbers of falsely called DEGs (Lazic, 2010; Zimmerman, Espeland and Langefeld, 2021). Whereas, debate has continued between the use of pseudobulk approaches and mixed models.

1.2.2. Single-nucleus transcriptomic analyses in Alzheimer's disease

A multitude of studies have conducted single-nucleus RNA-Seq (snRNA-Seq) in AD (Mathys *et al.*, 2019, 2023; Zhou *et al.*, 2020; Gerrits *et al.*, 2021; Smith *et al.*, 2022) however the techniques used to pre-process, integrate and identify the cell types and the DEGs have varied dramatically. Unsurprisingly, this has resulted in the specific dysregulated genes and instigated cell types differing dramatically across studies. Differences in differential expression techniques can particularly affect the identified DEGs (Squair *et al.*, 2021).

Consensus is needed around standardisation for analysing snRNA-Seq data so that results from these studies can be compared and possibly, combined to give better insight into AD dysregulation. Sc/snRNA-Seq offers previously unprecedented resolution of cellular, transcriptional changes in AD but for further discoveries for AD, consensus on analytical approaches and increases in sample size and coverage are paramount.

1.3. The non-coding genome

Since the vast majority of AD SNPs are outside of the exome, forming part of the non-coding genome, they are expected to affect cellular regulation (I. E. Jansen *et al.*, 2019). Thus, further insight into what was once considered predominately 'junk DNA' should elucidate the aetiology of the disease (Dunham *et al.*, 2012). The methods by which the non-coding genome regulates gene expression, and by which these AD related SNPs can alter this regulation, is covered in this section.

1.3.1. Epigenetics

The term epigenetics was first introduced by Conrad Waddington over 80 years ago as the study of "the causal interactions between genes and their products which bring the phenotype into being" (Waddington, 1942, 1968). Epigenetics' meaning eventually became more precise, from encompassing all molecular pathways modulating the expression of a genotype and its related phenotype, to the study of "changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence" (Wu Ct and Morris, 2001). This refined definition encompassed transcriptional changes that can be passed down through somatic (mitotically) or germ (meiotically) cell division, without altering the underlying DNA sequence. Thus, the study of epigenetics is that of dynamic cellular regulation, inclusive of gene regulation and other mechanisms, such as alternative splicing.

Importantly, these epigenetic mechanisms are erased and reestablished after fertilisation which is paramount for embryonic development and cell differentiation. However, conversely, more recent studies have indicated that certain epigenetic marks can be inherited, perhaps due to partial escape of complete reprogramming, resulting in cross-

generational changes in developmental and cellular features (Bohacek and Mansuy, 2013; Fukushima, Takeda and Nakamura, 2023).

This dynamic transcriptional regulation is paramount for cellular function and development. Epigenetics' role in cell differentiation, where cell regulation changes to promote cell specialisation, was first illustrated as a landscape some 67 years ago to highlight the process of cell fate determination (Waddington, 1957). Epigenetic regulation is also instrumental for all other cellular processes, like cell cycle entry and exit (Coller, 2019) and immune cell activation (Buckner and Ziegler, 2004). Moreover, when epigenetic mechanisms are dysregulated, this can result in disease. For example, tumorigenesis has recently been found to arise even in the absence of any genetic mutations following transient loss of epigenetic gene silencing (Parreno *et al.*, 2024).

Some fundamental methods of epigenetic regulation are DNA methylation, DNA-binding proteins, non-coding RNAs and histone modifications. DNA methylation occurs where methyl groups are added to cytosine bases and is commonly associated with gene silencing (Jin, Li and Robertson, 2011). DNA-binding proteins, such as transcription factors, regulate gene expression by binding to specific regulatory DNA sequences (Reményi, Schöler and Wilmanns, 2004). Non-coding RNAs, including microRNAs, small interfering RNAs and long non-coding RNAs, modulate gene expression through various mechanisms, such as mRNA splicing or degradation (Kaikkonen, Lam and Glass, 2011). Finally, histone modifications are covalent post-translational modifications of histone proteins, such as acetylation and methylation, that alter chromatin structure, affecting gene accessibility and thus transcription (Bannister and Kouzarides, 2011). Despite substantial research into each of these mechanisms, the order in which they occur is still debated. However, recent examples of pioneer transcription factors binding to compacted chromatin, inducing nucleosome structural changes and recruiting

histone marks, have been studied (Zaret, 2020; Sinha *et al.*, 2023). This proposes an ordered approach to gene activation of chromatin decompaction by DNA binding proteins, followed by activating histone marks and the removal of methyl marks, leading to greater gene expression.

Although epigenetic mechanisms are vast, the empirical chapters in this thesis will concentrate on gene regulation, specifically histone modifications and DNA-binding proteins; transcription factors. Techniques for experimentally measuring such mechanisms of cellular regulation have developed rapidly. With the advent of next-generation sequencing technologies, sequencing of tens to hundreds of millions of short DNA fragments in a single run became a possibility. This lead to assays for whole genome sequencing of histone modifications and DNA-binding proteins, as well as the related measure of chromatin accessibility, becoming widely adopted. Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) became the prevalent technique for protein binding experiments including transcription factors and histone marks, enabling high-resolution, genome-wide profiling (Johnson *et al.*, 2007; Park, 2009). This method involves cross-linking proteins to DNA, cleaving the chromatin, immunoprecipitating with protein-specific antibodies and finally, amplifying and sequencing the associated DNA fragments. For chromatin accessibility, the assay for transposase-accessible chromatin using sequencing (ATAC-Seq) has also became a prevalent technique (Buenrostro *et al.*, 2013). Open chromatin regions are tagged with sequencing adaptors and cleaved by hyperactive Tn5 transposases. The tagged DNA fragments are next purified, amplified and sequenced. These techniques catalysed the creation of large, standardised data repositories by consortia like ROADMAP and ENCODE, resulting in comprehensive assaying of transcription factor binding, histone mark mapping and chromatin accessibility assays across a variety of cell types and tissues (ENCODE Project

Consortium, 2012; Kundaje *et al.*, 2015). These single sources of uniformly processed, rich epigenetic assays have greatly advanced research into gene regulation, disease mechanisms, and developmental biology.

Despite their utility, these efforts concentrated on obtaining the average protein binding signal across a group of possibly diverse cells. Similar to the mass adoption of single-cell transcriptomic research (**Figure 1-2**), single-cell epigenetic assays have now been popularised such as single-cell ATAC-Seq (Lareau *et al.*, 2019) for chromatin accessibility and single-cell ChIP-Seq (Grosselin *et al.*, 2019) and single-cell CUT&TAG (Bartosovic, Kabbe and Castelo-Branco, 2021) for DNA protein binding. These approaches have further refined our understanding by enabling base-pair resolution mapping at the single-cell level, offering insight into epigenetic heterogeneity within cell populations and the dynamic regulatory landscapes in development and disease. However, the single-cell epigenetic field currently lacks consensus on a standardised approach, with a multitude of proposed assays, as well as a relatively high cost which has hindered the collation of large data repositories prevalent with bulk techniques (Mehrmohamadi *et al.*, 2021). For this reason, this work will focus on data from bulk ATAC-Seq and ChIP-Seq assays for tissues, cell lines and isolated cell types.

1.3.2. Epigenetics analyses and its role in Alzheimer's disease

Epigenetic mechanisms have long been linked with AD, relating the dysregulation of genes involved in processes like synaptic plasticity and memory storage and retrieval (Mehler and Mattick, 2007; Vecsey *et al.*, 2007; Guan *et al.*, 2009; Michán *et al.*, 2010; Gao *et al.*, 2022). Moreover, total histone mark levels, including acetylation, have been noted as significantly altered in the disease (Narayan *et al.*, 2015). Histone modifications associated with noncoding regulatory regions such as enhancers have also been assayed and found to harbour AD

variants in cell type-specific analyses (Nott *et al.*, 2019). Furthermore, a multitude of transcription factor binding profiles have been shown to be dysregulated or disrupted in AD such as that of Sp1, AP-2 and PPARγ (Citron *et al.*, 2008; Chen *et al.*, 2013). Sp1 has been shown to regulate *BACE1* which processes APP and is essential for Aβ production (Zhang and Song, 2013; Das and Yan, 2017). As previously mentioned, susceptible variants in *APOE* are the strongest genetic risk factors for late-onset AD (Kim, Basak and Holtzman, 2009). Furthermore, changes in *APOE* expression level have long been associated with AD (Laws *et al.*, 2003). AP-2 and PPARγ and other developmental and hormonal factors have been noted to regulate *APOE* by binding to its proximal promoter region (García *et al.*, 1996; Yue *et al.*, 2004). Thus, these works instigate cellular regulation as having a key role in the pathogenesis of the disease.

1.4. Genetic variation

As previously outlined, AD is highly heritable (Gatz *et al.*, 2006). Thus, an understanding of the genetic component of the disease promises to elucidate the noted dysregulation of the epigenome and transcriptome. The field of human genetics is primarily focused on the understanding of inherited genetic variation. Although the study of genetic variation plays a role in most diseases, the vast majority of the genome is identical across individuals - It was thought that on average, a person's set of genomic variants will account for just 0.4% of their whole genome (Lander *et al.*, 2001), when compared to a reference genome. More recent work put this figure closer to 1%, with those genomes arising from African and other ancestral populations exhibiting greater genetic variation (Trost, Loureiro and Scherer, 2021).

1.4.1. The study of various types of variation

These aforementioned genetic variations are broadly categorised into two types - variation between individuals (germline variants) or within individuals (somatic mutations) (Yu *et al.*, 2024). Somatic mutations are not heritable, unless they occur in germ cells, and are instead often caused by DNA damage or errors in DNA replication and repair such as depurination and deamination. These sporadic mutations are commonly studied in cancer research but are less studied in neurodegenerative disease due to the lack of cell division in neuronal populations. Although somatic mutations have recently been linked to neurodegenerative disease where these mutations occur in nervous tissue during development (Verheijen, Vermulst and van Leeuwen, 2018) and in aging due to neurogenesis; the *de novo* formation of neurons (Kumar *et al.*, 2019). Moreover, their effect on DNA repair has been well-studied in Huntington's disease where the mismatch repair (MMR) pathway has been instigated in somatic expansions of CAG repeats, which were found to progressively increase in length, particularly in the brain (Telenius *et al.*, 1994; Maiuri *et al.*, 2019).

Germline variants, which are present in germ cells, are heritable and are found in every cell of the offspring. Such mutations are subject to generations of evolutionary pressures such as drift, selection and recombination, shaping their frequency in whole populations (Lynch *et al.*, 2016). Therefore, studying germline mutations across populations promises to give insight into their relationship with the prevalence of phenotypes such as disease and may even uncover the instigated biological processes.

These genetic variants, whether they are germline or somatic, encompass multiple differing types of DNA mutations. SNPs, also known as single nucleotide variants (SNVs), are single base-pair changes in the DNA. SNPs are both the most common form of genetic variant and the most commonly studied (Eichler, 2019). A SNP can have multiple alternative alleles or bases observed in the population (multi-allelic) or just a single allele (bi-allelic). Other forms of genetic variants which have been observed are copy number variations (CNVs) where sequences are repeated, small insertion/deletions (Indels) where multiple base-pairs are inserted instead of a single base-pair or vice-versa and translocations and inversions where major chromosomal rearrangements result in the relocation of genes or regulatory DNA on other chromosomes (translocations) or on the same chromosome in reverse (inversion). The work in this thesis will focus on germline SNPs due to their prevalence and extensive research in AD.

1.4.2. The evolution of genetic research post-human genome project

The study of such genetic variants across populations, and the field of genetics as a whole, was fundamentally changed in 2001 by the Human Genome Project (HGP) (Lander *et al.*, 2001). By providing the first, albeit incomplete, map of the human genome, the HGP revolutionised large-scale genetic studies. The most prominent example of this is the

facilitation of GWAS for the first time, thanks to the availability of the HGP's reference genome. Further work such as the 1000 genomes project, which increased the number of samples to more accurately represent the range of human diversity (Auton *et al.*, 2015), and the Telomere-to-Telomere (T2T) Consortium, which completed sequencing of the remaining 8% of the genome in 2022, made up of heterochromatic regions (Nurk *et al.*, 2022), have iteratively improved the reference genome. These efforts have not only enhanced the accuracy and representation of diversity in genetic analyses of the entire genome, but they have also enabled the development and application of advanced sequencing technologies, facilitating further genetic research that benefits the study of all diseases, including AD.

1.4.3. Distribution of genetic variant information

To effectively leverage the advancements brought by the HGP and large-scale genetic studies, it is crucial to consider how genetic variant information is distributed and utilised within the research community. This distribution is predominately done in the form of summary statistic files. For example, these summary statistic files are used to distribute outputs of GWAS and quantitative trait loci studies (xQTLs). However, herein the focus is on GWAS summary statistics due to their prevalence in research. Summary statistics from past studies do not require the transfer of individual-level, personally identifiable information from participants and tend to become more valuable over time as it becomes possible to meta-analyse and integrate them with new annotation information through approaches such as Linkage Disequilibrium Score Regression (LDSC) (Bulik-Sullivan *et al.*, 2015; Garg *et al.*, 2021), Generalized Gene-Set Analysis of GWAS Data with Multi-marker Analysis of GenoMic Annotation (MAGMA) (Leeuw *et al.*, 2015), and multi-phenotype investigations (Tanigawa *et al.*, 2019; Aguirre *et al.*, 2021). Summary statistics are also commonly integrated for use in the

meta-analysis of GWAS. However, these tools and this integration require a standardised data format which was historically lacking from the field. The diversity of data formats in summary statistics has been a result of the phenotypes in question, for example disease-control or quantitative trait, the software used to perform the analysis, such as PLINK (Purcell *et al.*, 2007) and GCTA (Yang *et al.*, 2011) or just the preference of the consortium in question.

There have been movements to standardise the summary statistic file format such as the NHGRI-EBI GWAS Catalogue standardised format (Buniello *et al.*, 2019), the MRC IEU OpenGWAS infrastructure and catalogue (Elsworth *et al.*, 2020) and the SMR Tool binary format (Zhu *et al.*, 2016). More recently, the variant call format (VCF) to store GWAS summary statistics (GWAS-VCF) (Lyon *et al.*, 2021) has been developed which has manually converted over 10,000 GWAS to this format. While GWAS-VCF offers a standardised format that future GWAS consortium may adopt, there are still a multitude of past, publicly available GWAS which have not been standardised (Lin *et al.*, 2018; McCormack *et al.*, 2018; P. R. Jansen *et al.*, 2019; Luciano *et al.*, 2021). For instance, although their summary statistics are publicly available, the GWAS for Cerebral small vessel disease (Sargurupremraj *et al.*, 2020) is not yet available in VCF format via IEU OpenGWAS (Elsworth *et al.*, 2020). Furthermore, as VCF is not yet the standard for sharing files between geneticists, unpublished GWAS shared internally within genetics consortia or provided by personal genetics companies are still found in a variety of summary statistic formats. As such, there is a need for tools to move between the various formats in which summary statistics are stored.

The standardisation of GWAS summary statistics also requires quality control to ensure cohesive integration. For example, checking if the non-effect allele from the summary statistics matches the reference sequence from a reference genome to ensure consistent directionality of allelic effects across GWAS. In addition, downstream analysis tools often

require a degree of quality control which, in the case of meta-analysis, must be applied across all GWAS. One such example is the removal of all non-biallelic SNPs which is a common requirement of downstream analyses (Lyon *et al.*, 2021). Therefore, there is a pressing need, in the field and more specifically for the work undertaken in this thesis, for a tool to uniformly process and apply quality controls to summary statistics.

1.4.4. GWAS – early results and limitations

Despite the revolutionary effect of GWAS in the field of genetics, there remains limitations to its findings. As previously mentioned, large-scale GWAS have primarily focused on SNPs and as such, SNPs make up the majority of associations with disease, leaving other types of genetic variants such as CNVs and other structural variations potentially understudied (Zhang *et al.*, 2009; Stankiewicz and Lupski, 2010). Moreover, what has become clear from GWAS, is that the vast majority of the disease-related SNPs are found in non-coding regions of the genome. For example, one study recently estimated the proportion of non-coding SNPs to be as high as 95% across complex diseases (Watanabe *et al.*, 2019). This makes it challenging to accurately associate the majority of SNPs identified in a disease with the affected gene(s).

A further complication when considering the instigated SNPs from GWAS is confounding by linkage disequilibrium (LD). LD occurs when alleles are co-inherited based on their physical proximity, making it difficult to identify the causal SNPs (Slatkin, 2008; Reshef *et al.*, 2018). Thus, it is common to have genomic loci associated with a disease from GWAS but it may be unclear which of multiple SNPs are causal for the disease and which are just in LD (i.e. tag-SNPs). Computational approaches to identify causal SNPs have been developed such as FINEMAP (Benner *et al.*, 2016). However, recent reviews have highlighted that despite considerable advancements, there is still work to be done to improve current fine-mapping

approaches (Schaid, Chen and Larson, 2018; Broekema, Bakker and Jonkers, 2020; Burgess, 2022).

Furthermore, even assuming a causal SNP is identified, genetic variants which are associated with the function or activity of regulatory elements exert their effects in a cell type-specific manner. Thus, understanding the cell type-specific effect of genetic variants is paramount.

1.4.5. GWAS results and downstream analyses in Alzheimer's disease

A multitude of genetic studies have been undertaken in AD (J. Zhang *et al.*, 2023). These have included large GWAS which continue to uncover associated variants with increased sample sizes, now exceeding half a million cases and controls (I. E. Jansen *et al.*, 2019; Wightman *et al.*, 2021; Bellenguez *et al.*, 2022a). However, as previously stated, 98% of these AD related genetic variants lie in non-coding, regulatory regions and cannot be definitively linked with a specific gene (I. E. Jansen et al., 2019). Thus, they primarily influence gene expression by altering genomic regulatory elements, i.e. the epigenome, rather than directly affecting protein function. Understanding the role of these regulatory variants is further burdened because of gene regulatory mechanisms' highly cell type-specific nature (Schreiber, Bilmes and Noble, 2020). Thus more work is required to fully utilise and comprehend these rich genetic resources for AD. A standardised approach to processing and quality control would be a necessary first step to enable greater utility of these rich genetic resources, removing any barriers to access and analysis. Some downstream analyses have already been conducted on these genetic variants using transcriptomics and epigenetics, instigating the resident macrophages of the central nervous system microglia in the genetic risk of the disease (Skene and Grant, 2016; Nott *et al.*, 2019). This has supported a recent shift in the focus of AD research to glia cells rather than neuronal cell types which are subsequently depleted in the

disease. However, the genes and pathways by which these genetic variants act, leading to the onset and progression of the disease remain elusive. Thus, greater insight into the cell type-specific genes and cell type-specific effects of non-coding genetic variants on cells' epigenome would help in our understanding of the disease aetiology and treatment. One method by which new insight can be gained into the cell type-specific effects of genetic variants and understanding the cis-regulatory code more generally is machine learning.

1.5. Machine learning and Deep learning for genomics

Over the past decade, deep learning, a subdiscipline of machine learning, has experienced a surge of application within the field of genomics with an aim to learn the regulatory information of the genome, i.e. the cis-regulatory code.

1.5.1. Motivating learning the genome's code

Deep learning's application is not the first attempt to understand these cis-regulatory mechanisms, research has been ongoing in the field for more than the past half-century. In the 1960s, it was discovered that bacterial genes are regulated by nearby DNA sequences; i.e. motifs (Jacob and Monod, 1961; Lewis, 2005). The progression of functional genomic assays some 20 years later enabled the discovery that enhancers could function, at least to a certain extent, independently of their native genomic context, orientation or precise distance from the gene (Banerji, Rusconi and Schaffner, 1981). It would later be proven that enhancer activity depends on the arrangement of the specific bases and distance to other motifs; i.e. motif syntax (Farley *et al.*, 2015). This ability to cut and insert regulatory motifs forms the basis of multiplex assays of variant effects (MAVEs): Including massively parallel reporter assays (MPRAs) which test the activities of thousands of candidate genomic regulatory elements simultaneously via the next-generation sequencing of a barcoded reporter transcript (Mulvey, Lagunas and Dougherty, 2021; Zheng and VanDusen, 2023). In the same decade, it was discovered that these regulatory regions contain short sequence motifs with a median length of 9 bases, calculated from the human JASPAR 2024 database (Rauluseviciute *et al.*, 2024), that are recognized and preferentially bound by transcription factors (Mitchell and Tjian, 1989). Today, there are whole databases of manually curated, high-quality and non-redundant DNA-binding profiles for transcription factors across differing species, like JASPAR

(Rauluseviciute *et al.*, 2024). Transcription factors provide a means by which cells can control the activation of regulatory regions like enhancers and their associated genes through processes like extracellular signals or transcriptional regulation in a spatial and temporal fashion, for example during embryonic development (Levine and Davidson, 2005; Zeitlinger, 2020).

Despite extensive experimental efforts combined with advancements in functional genomic assays, there has been no breakthrough in understanding of the complete cis-regulatory code (Slattery *et al.*, 2014; Zeitlinger, 2020). This may be due to the sheer complexity of the regulation where a multitude of regulatory motifs interact to, for example, control expression of a gene. Or perhaps, how this regulation occurs in a cell type or cell state-specific manner (Levine and Tjian, 2003). However, given the vast experimental data and the understanding that these genomic motifs are paramount to cellular regulation, whether that be regulatory motifs like enhancers, the instigated motif syntax, transcription factor binding motifs or more likely combinations of all of these, it is conceivable to consider that a model could learn the relationships between combinations of these motifs to understand the cis-regulatory code (Zeitlinger, 2020). These models can generalise to new genomic sequences by learning the sequence grammar - the relationships and interactions between proteins, like transcription factors, and DNA, as well as higher-order complexes, and how these regulate cell type-specific functional genomic outputs (Sasse, Chikina and Mostafavi, 2024). The methods by which deep learning models have been exploited to do just this in genomic research are discussed in the following section.

1.5.2. Deep learning's broad applicability in genomics

The deployment of deep neural networks (DNNs) in genomics research has been significantly facilitated by the growing availability of graphical processing units (GPUs) (Shi *et al.*, 2016; Eraslan *et al.*, 2019) and open-source, user-friendly frameworks for their implementation like PyTorch (Paszke *et al.*, 2017) and TensorFlow (Abadi *et al.*, 2016). Similar to these open-source frameworks, recent efforts have focused on standardising the application of DNNs in genomics specifically. Notable contributions include Kipoi, tangermeme, gReLU and EUGENe (Avsec *et al.*, 2018; Klie *et al.*, 2023; Lal *et al.*, 2024; Schreiber, 2024). These frameworks have enhanced the efficiency of data processing such as loading and manipulating genomic sequences (Kipoi, Tangermeme, EUGENe), model training (EUGENe, gReLU), the reuse of trained models (Kipoi, gReLU), and downstream evaluation and interpretation tasks such as variant effect prediction or sequence design (EUGENe, Tangermeme, gReLU). Such efforts are crucial for ensuring efficient, standardised and defect-free analysis in the field.

Given this work, it is unsurprising that the applications of DNNs in genomics have been very broad – the Kipoi repository now houses more than 2,000 models (Avsec *et al.*, 2018). These approaches have included epigenetic imputation (Ernst and Kellis, 2015; Schreiber, Durham, *et al.*, 2020), synthetic sequence design e.g. promoters and enhancers (Gosai *et al.*, 2023; P. Zhang *et al.*, 2023; Taskiran *et al.*, 2024), functional genomic predictions for epigenetics and transcriptomics (Minnoye *et al.*, 2020; Avsec, Agarwal, *et al.*, 2021; Avsec, Weilert, *et al.*, 2021; Linder *et al.*, 2023) and even self-supervised learning models pretrained on DNA sequence, also known as genomic language models (gLMs) (Dalla-Torre *et al.*, 2023; Karollus *et al.*, 2024). For imputation and synthetic sequence design models, the outputs are useful in their own right – the predicted genomic assays can be used in place of experimental data for the cell type-specific epigenetic signal/regulatory DNA motifs. Moreover, the synthetic

biology field has received growing attention of late with more emphasis on cell viability under augmented gene expression modules (Ceroni *et al.*, 2015, 2018). However, functional genomic models which predict in held out chromosomes e.g. Enformer (Avsec, Agarwal, *et al.*, 2021) and gLMs which predict held out sequences of the input DNA e.g. Nucleotide Transformer (Dalla-Torre *et al.*, 2023), do not produce novel outputs – their outputs have already been experimentally measured. Instead, these models are of use for tasks like transfer learning, genetic variant imputation predictions or model interpretation. There has been a recent shift in the field, placing an increased emphasis on interpretability of these models for biological insight (Atak *et al.*, 2021; Avsec, Weilert, *et al.*, 2021; Nair *et al.*, 2023; Taskiran *et al.*, 2024) as opposed to primarily performance benchmarks which is typical in the machine learning research field (Li *et al.*, 2023; Wu *et al.*, 2024). Interpretability approaches for these models are paramount for their use to unravel the cis-regulatory code, giving insight into the genomic drivers from genetic variants to whole motifs or genomic loci (Shrikumar, Greenside and Kundaje, 2019; Shrikumar *et al.*, 2020).

1.5.3. The necessity of deep learning in understanding genetic variants

One aim of my work focuses on applying DNNs to understand the effect of genetic variants' genomic regulation in a cell type-specific manner which necessitates creating an imputation model which makes functional genomic predictions using DNA sequence. And as previously outlined, genetic sequence variants which are associated with the function or activity of regulatory elements, quantitative trait loci (QTLs), are predominately found in non-coding regions and will tend to exert their effects in a cell type-specific manner. Thus, mapping out molecular and regulatory QTLs comprehensively in disease-relevant cell types would enable

the interpretation of functional outcomes of genetic variants on gene expression and regulation. Techniques to study QTLs include CRISPR interference (CRISPRi) (Larson *et al.*, 2013) and MPRAs (Mulvey, Lagunas and Dougherty, 2021). However, these approaches lack the *in vivo* capabilities or scalability to large sample sizes thus limiting their efficacy. Currently, the most suitable approach is to conduct population studies measuring individuals' genomic variance and its correlation to regulatory elements, i.e. xQTL mapping studies. However, epigenetic QTL studies have proven less popular than that of gene expression (eQTL mapping studies), which include full catalogues of studies with sample sizes in the high hundreds (Kerimov *et al.*, 2021), and some reaching as many as 30,000 (Võsa *et al.*, 2021). To note, one exception in epigenetic studies is methylation QTL studies (mQTL) which, similar to eQTLs, have exceeded 30,000 participants (Min *et al.*, 2021). This preference for eQTL over epigenetic QTLs is surprising given these QTL studies on intermediate levels of regulation of gene expression have been found to have a greater overlap with GWAS hits than eQTLs (Arthur *et al.*, 2024). The largest epigenetic QTL study was conducted in blood immune cell types (Kundu *et al.*, 2022), with a sample size of just 196, far less than the average expression QTL study. Moreover, the study is one of just a small number of histone mark (hQTL) or transcription factor (tfQTL) studies conducted. Robustly identifying regulatory QTLs from primary data in epigenetic QTL studies requires very large sample numbers due to the high-dimensional number of pairwise tests for associations between all SNPs and all regulatory elements. Thus, undertaking such studies in large sample sizes and on every possible cell type of interest is impractical.

1.5.4. How to utilise deep learning for genetic variant effect prediction

An alternative to experimentally measuring effects is to predict the effect of genetic variants using deep learning approaches, known as *in silico* mutagenesis. Numerous DNNs have been developed to predict the effect of genetic variants across different tasks such as transcription factor binding (Keilwagen, Posch and Grau, 2019; Li and Guan, 2021), histone marks (Kelley *et al.*, 2018) and gene expression (Kelley, Snoek and J. Rinn, 2016) profiles. The main advancement of more recent approaches in the field has focused on the amount of DNA sequence which models can take into account simultaneously to make the prediction. This has generally led to greater performance on held-out genomic positions and better genetic variant effect predictions. Intuitively, these larger window sizes mean that when predicting the profile at a given location, distal effects of DNA sequence can be modelled. The maximum distance from the predicted profile to DNA sequence in either direction of such models has increased from 500 base-pairs for DeepSea (Zhou and Troyanskaya, 2015) to 20,000 base-pairs in Basenji2 (Kelley *et al.*, 2018) and in the current best-in-class approach, Enformer (Avsec, Agarwal, *et al.*, 2021), to 100,000 base-pairs. Enformer swapped out convolutional layers from Basenji2's original architecture with multi-headed attention layers and three relative positional encoding functions (taken from past work (Dai *et al.*, 2019)) to infer genomic context and dramatically increase the receptive field.

However, notably, these models were all trained on wild type profiles and the effect of genetic variants were predicted after model training, by comparing the wild type prediction against that based on the DNA sequence with the imputed SNP or using model interpretation techniques like DeepLIFT (Shrikumar, Greenside and Kundaje, 2019), to generate sequence importance scores and TF-MoDISco (Shrikumar *et al.*, 2020), to identify the related motifs. Some approaches have chosen to predict the effect of genetic variants directly, one such

notable approach by Wang *et al.* (Q. S. Wang *et al.*, 2021). The authors first identified the causal SNP from those that were in LD through a statistical fine-mapping approach; FINEMAP (Benner *et al.*, 2016), to avoid the model learning spurious effects of genetic variants. Although this showed promise, recent reviews have highlighted that despite considerable advancements, there is still work to be done to improve current fine-mapping approaches (Schaid, Chen and Larson, 2018; Broekema, Bakker and Jonkers, 2020; Burgess, 2022). Apart from the issue of defining causal SNPs from those in LD, this approach of training directly on genetic variants suffers from a lack of diversity in the samples and, more importantly, a lack of *in vivo* experimental of functional genomic assays across samples. For example, ENCODE's data repositories where, despite hosting greater than 10,000 assays, only has up to three biological replicates per assay (ENCODE Project Consortium, 2012).

Importantly, all of the aforementioned models were trained solely on DNA sequence and thus cannot predict cell type profiles that were not seen in the training phase. It would be beneficial to predict epigenetic profiles and the effect of genetic variants on these in previously unseen cell types where collecting cell type epigenetic profiles is invasive, time-consuming and costly. The necessity of such cell type epigenetic imputation is apparent even in large consortia like ENCODE's (ENCODE Project Consortium, 2012) data repositories where, despite hosting greater than 10,000 assays, cell types have not been experimentally assayed for every epigenetic mark of interest. Moreover, the cell types present, mostly whole tissues or cell lines, do not capture the breadth of cell types studied in disease research, for example in the brain. Some past models have attempted to address this by generalising predictions of epigenetic profiles to other cell types using chromatin accessibility information from the cell type of interest, often DNase-Seq or ATAC-Seq, such as Leopard (Li and Guan, 2021), Catchitt (Keilwagen, Posch and Grau, 2019) (both predict transcription factor binding profiles) and

Epitome (Morrow *et al.*, 2021) (histone mark profiles). Notably, these approaches do not consider distal genetic information, instead using local genetic windows, far shorter than that of Enformer, and other local features to the infer the cell type variability. However, this local information may not be enough for a model to fully capture cell type representations.

1.6. Linking epigenetics to expression – understanding the regulation of the coding genome

As previously outlined in **Section 1.1**, understanding the cis-regulatory code and effect of genetic variants in AD requires insight into cells' epigenetic regulation and transcription. However, these two omic layers are not independent of one another – non-coding genetic variants can dysregulate epigenetic mechanisms which in turn cascades to differential gene expression. Thus, linking epigenetics to expression would enable a wholistic view of this effect. Moreover, this approach offers an alternative to interpreting functionally relevant loci, by perturbing epigenetic signals, compared to genomic deep learning models trained on DNA sequence where substituting single base pairs such as genetic variants are notoriously difficult to interpret (Toneyan, Tang and Koo, 2022; Huang *et al.*, 2023; Linder *et al.*, 2023; Sasse, Ng, A. E. Spiro, *et al.*, 2023).

Post-translational modifications on the N-terminal tails of histone proteins, i.e. histone marks, are a key epigenetic mechanism by which eukaryotic cells regulate transcriptional activity, via altering chromatin structure and interacting with other transcriptional regulators (Bannister and Kouzarides, 2011; Miller and Grant, 2013). These epigenetic modifications enable cell plasticity without changes to the underlying DNA sequence. Histone mark dynamics in a given cell are mediated by both internal and extracellular cues (McBrian *et al.*, 2013; Niu *et al.*, 2015). Alterations in histone modifications have been found to strongly associate with cellular differentiation, cell cycle stages and the development of different diseases (Jenuwein and Allis, 2001; Dunham *et al.*, 2012; Marzi *et al.*, 2018; Zhao and Shilatifard, 2019). For example, as cells mature and differentiate, chromatin accessibility and histone acetylation become progressively restricted throughout their lineage (Atlasi and Stunnenberg, 2017).

Individual regulatory effects of histone marks on transcription have been widely studied: While H3K9ac is associated with active promoter regions (Karmodiya *et al.*, 2012), H3K4me1 is found at distal enhancers (Bae and Lesch, 2020) (summarised in **Table 1-1**). However, less emphasis has been placed on the extent to which these histone marks directly regulate gene expression levels. To investigate if transcriptional levels in different cellular contexts can be determined solely from histone modification states, one could correlate levels of histone modifications in regulatory elements with mRNA expression individually. However, this ignores two additional levels of complexity: Firstly, it is well-known that histone modifications interact with other epigenetic factors such as pioneer transcription factors, which in turn have been linked to enhancer activation (Atlasi and Stunnenberg, 2017). Secondly, histone marks themselves act in concert and the interaction between different regulatory elements is not necessarily additive.

To circumvent these challenges it is possible to conduct *in silico* experiments, predicting transcriptional levels from histone mark signals where the observed chromatin state is assumed to capture the other contributing epigenetic regulators without directly experimentally observing them. Such *in silico* experimentation could enable the knock-out and thus prioritisation of cell type-specific regulatory genomic loci, possibly leading to new insight into disease, including AD.

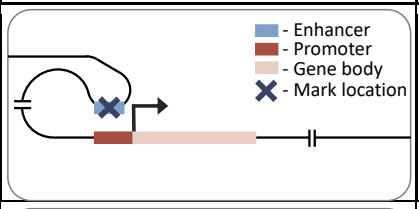
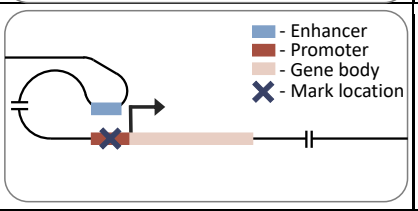
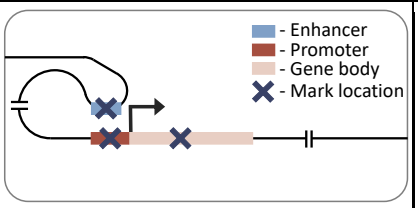
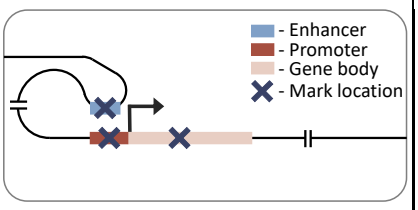
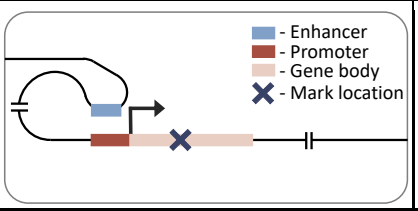
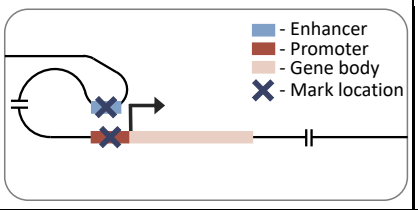
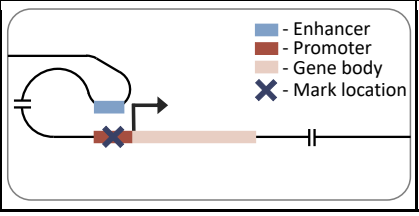
Histone mark	Genomic location	Transcript-ional relationship	Proposed function	References
H3K4me1		Activating	• Enriched at active and poised enhancers. • Suggested to fine-tune, rather than tightly control, enhancer activity and function by recruiting key transcription factors.	(Rada-Iglesias, 2018; Bae and Lesch, 2020)
H3K4me3		Activating	• Found at promoter regions • Has a direct preferential association with the PHD finger of nucleosome remodelling factor (NURF) complex which remodels chromatin, making the DNA accessible for gene transcription.	(Schneider *et al.*, 2004; Wysocka *et al.*, 2006)
H3K9me3		Repressive	• Involved in the formation of heterochromatin, • Found at transposable elements, satellite repeats and genes, where it ensures transcriptional silencing. • These heterochromatin has also been found to relate to cell lineage-dependent, transcriptional silencing.	(Foret *et al.*, 2014; Wang *et al.*, 2018; Padeken, Methot and Gasser, 2022)
H3K27me3		Repressive	• Act as a gene silencer, widely distributed but particularly abundant near promoters and regulates gene expression via proximity or looping. • Function has been linked to Polycomb repressive complexes (PCR1,PCR2) which can be recruited and contribute to chromatin compaction.	(Cao *et al.*, 2002; Ku *et al.*, 2008; Sanz *et al.*, 2008; Bosselut, 2016; Cai *et al.*, 2021)
H3K36me3		Repressive	• Enriched in gene bodies. • A binding partner for histone deacetylases (HDACs) which prevent run-away RNA polymerase II (Pol II) transcription.	(Carrozza *et al.*, 2005; Joshi and Struhl, 2005; Li, Notani and Rosenfeld, 2016)
H3K27ac		Activating	• Enriched at active enhancer and promoter regions (differing from H3K4me1 which also indicates poised enhancers). • Recruits transcription factors to increase transcription. For example, bromodomain-containing protein 4 (BRD4) which enhances Pol II recruitment and increases transcription.	(Herrera-Uribe *et al.*, 2020; Ibragimov, Bylino and Shidlovskii, 2020)
H3K9ac		Activating	• Enriched at promoter regions. • Mediates super elongation complex (SEC) and pol II chromatin occupancy on the proximal promoter region thus aiding in the switch from transcription initiation to elongation.	(Karmodiya *et al.*, 2012; Gates *et al.*, 2017)

Table 1-1: The primary genomic location, transcriptional relationship, and proposed function of a variety of known histone marks.

This table is taken from our previous publication (Alan E Murphy et al., 2024).

1.7. The study of the protein-coding and non-coding genome with respect to genetic variants as a tool for precision medicine in Alzheimer's disease

AD is characterised by selective vulnerability of specific neuronal subtypes and despite the recent availability of treatments to clear build-ups of Aβ, these only postpone the onset of symptoms rather than treat the underlying disease (Bennett *et al.*, 2006; Breijyeh and Karaman, 2020). Hence, there is a pressing need to further investigate the protein-coding and non-coding genome, especially in relation to the known genetic variants, to better understand this selective vulnerability, the affected cell types and the molecular mechanisms driving AD. These disease variants are predominately located in non-coding regions of the genome (I. E. Jansen *et al.*, 2019) and it has been experimentally shown that they often regulate the genome through long-range interactions (Meng *et al.*, 2023). However, this regulation is cell type-specific, highlighting the necessity of functional data to elucidate their impact on the transcriptome or the epigenome. Understanding these interactions at a cell-specific level is crucial for identifying the precise molecular drivers of AD.

Enabling such investigation requires approaches for the standardisation of the multitude of AD genetic information (I. E. Jansen *et al.*, 2019; Wightman *et al.*, 2021; Bellenguez *et al.*, 2022a) as the first step towards increasing its utility and insight from these rich genetic resources. Additionally, improved approaches for linking this genetic data to the functionally relevant genomic loci in a cell type-specific manner are critical. Such approaches may provide novel insight into AD progression and regulation to support the translation of this information into treating the underlying disease. For instance, *in silico* experiments using deep learning models, as explored in this thesis, represent a promising direction for these analyses.

Furthermore, a standardised approach to analysing the wealth of snRNA-Seq data from AD transcriptomic studies could significantly enhance our understanding of cell type-specific

gene drivers. These computational advancements, which are both AD-specific and applicable to broader genomic studies, are necessary for gaining new insight into the molecular drivers of AD. An overview of the work presented in this thesis and our efforts to bridge these gaps is illustrated in **Figure 1-3**.

By integrating these multifaceted approaches, we can move closer to precision medicine in AD, leveraging detailed genetic and functional insights to develop more effective and targeted treatments enhanced by greater insight into disease aetiology.

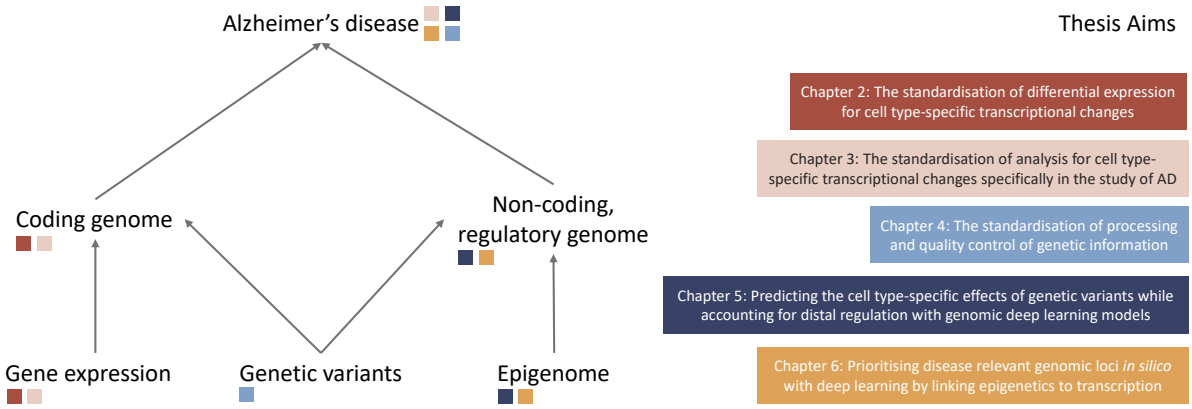


Figure 1-3: Overview of the different facets of Alzheimer's disease's regulatory code covered in this PhD thesis.

The different thesis chapters, their aims and their relationships are colour-coded. Certain aims are general bioinformatic analyses, applicable in all disease research, whereas some are specific to Alzheimer's disease.

1.8. Aims of this thesis

The expanding body of genetic research, including large-scale studies (GWAS) (P. R. Jansen *et al.*, 2019; Bellenguez *et al.*, 2022a), into the molecular drivers of AD point to genetic variations primarily located in regulatory, non-coding region of the genome. Some early analyses have highlighted microglia as holding the genetic burden of the disease (Skene and Grant, 2016; Nott *et al.*, 2019) however, the instigated regulatory pathways, transcription factors and genes remain elusive. Fundamental advancements in computational approaches are necessary to cohesively analyse this genetic data and shed light on the AD's cis-regulatory code. This PhD thesis aims to address such shortcomings in computational analyses, covering both the protein coding and non-coding genome, for both AD-specific and broader advancements (depicted in **Figure 1-3**). Although the focus of this work is AD, the developed, open-source, computational and deep learning techniques are all applicable to our comprehension of the cis-regulatory code. This work is highlighted under the following aims:

Aim 1: The standardisation of differential expression for cell type-specific transcriptional changes. **Chapter 2** conducts a benchmark to highlight best-practice approaches for differential expression analysis of scRNA-Seq to inform a standardised approach for disease research. This chapter has been published in Nature Communications (Murphy and Skene, 2022).

Aim 2: The standardisation of analysis for cell type-specific transcriptional changes specifically in the study of AD. **Chapter 3** is a case-study of the first scRNA-Seq analysis in AD to highlight the effect of differing approaches to quality control and differential expression analysis. This

reanalysis gives new insight into the dysregulated genes based on the original experimental data. This chapter has been published in eLife (Murphy, Fancy and Skene, 2023).

Aim 3: The standardisation of processing and quality control of genetic information. All downstream analyses of genetic data requires standardisation and quality control, something which is lacking from the broader computational biology field. **Chapter 4** details our approach and publicly available tool for this standardisation and processing. This chapter has been published in Bioinformatics (Murphy, Schilder and Skene, 2021).

Aim 4: Predicting the cell type-specific effects of genetic variants while accounting for distal regulation with genomic deep learning models. **Chapter 5** details a new deep learning model to predict epigenetic signals in previously unseen cell types. The work highlights methods by which to use the model to uncover disease relevant genomic loci. This chapter has been published in Nature Communications (Alan E. Murphy *et al.*, 2024).

Aim 5: Prioritising functional and disease relevant genomic loci *in silico* with deep learning by linking epigenetics to transcription. **Chapter 6** presents an in-depth analysis of deep learning approaches linking epigenetics to transcription. The work highlights alternative approaches to identifying disease and functionally relevant genomic loci. This chapter has been published in Nucleic Acids Research (Alan E Murphy et al., 2024).

2. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis

2.1. Introduction

The work in this chapter is adapted from that set out in our previous publication (Murphy and Skene, 2022). Recently, Zimmerman *et al*., highlighted the importance of accounting for the dependence between cells from the same individual when conducting differential expression analysis on single-cell RNA-sequencing data (Zimmerman, Espeland and Langefeld, 2021). Their work proved the inadequacy of pseudoreplication approaches for such analysis – This was an important step forward that was conclusively proven by them. However, the authors reported that mixed models were superior in their work, in contrast to other recent benchmarking studies which highlighted pseudobulk approaches (Crowell *et al.*, 2020; Squair *et al.*, 2021). This questions whether there is a single, consistently best performing approach to differential expression analysis of pseudoreplication, mixed models and pseudobulk approaches or whether this varies based on the study in question?

In their study, Zimmerman *et al.*, performed a systematic analysis of differential expression methods' type 1 error rates; the proportion of non-differentially expressed genes indicated as differentially expressed by a model (Zimmerman, Espeland and Langefeld, 2021). Their analysis was conducted on simulated, single-cell expression data across 20,000 iterations. The authors tested iterations of 5 to 40 individuals and 50 to 500 cells using an unadjusted p-value cut-off of 0.05 for significance. Plotting their results showed pseudobulk approaches had the lowest type 1 error at every iteration (**Figure 2-1**). However, evaluating such models on their type 1 or type 2 error rate in isolation is insufficient to determine their true performance. For

example, a method with low type 1 error may have a high type 2 error rate. Therefore, we need to consider both type 1 and type 2 error rate to accurately benchmark the models. Moreover, because no seed was set for the pseudo-random number generator used in their hierarchical single-cell expression simulation approach (hierarchicell), the different methods evaluated by Zimmerman *et al.* were compared on different simulated datasets. In summary, there appear to be limitations in both their benchmarking and simulation approaches. Here, we corrected these issues, reran the author's analysis to determine the best performing methodological approach for differential expression analysis.

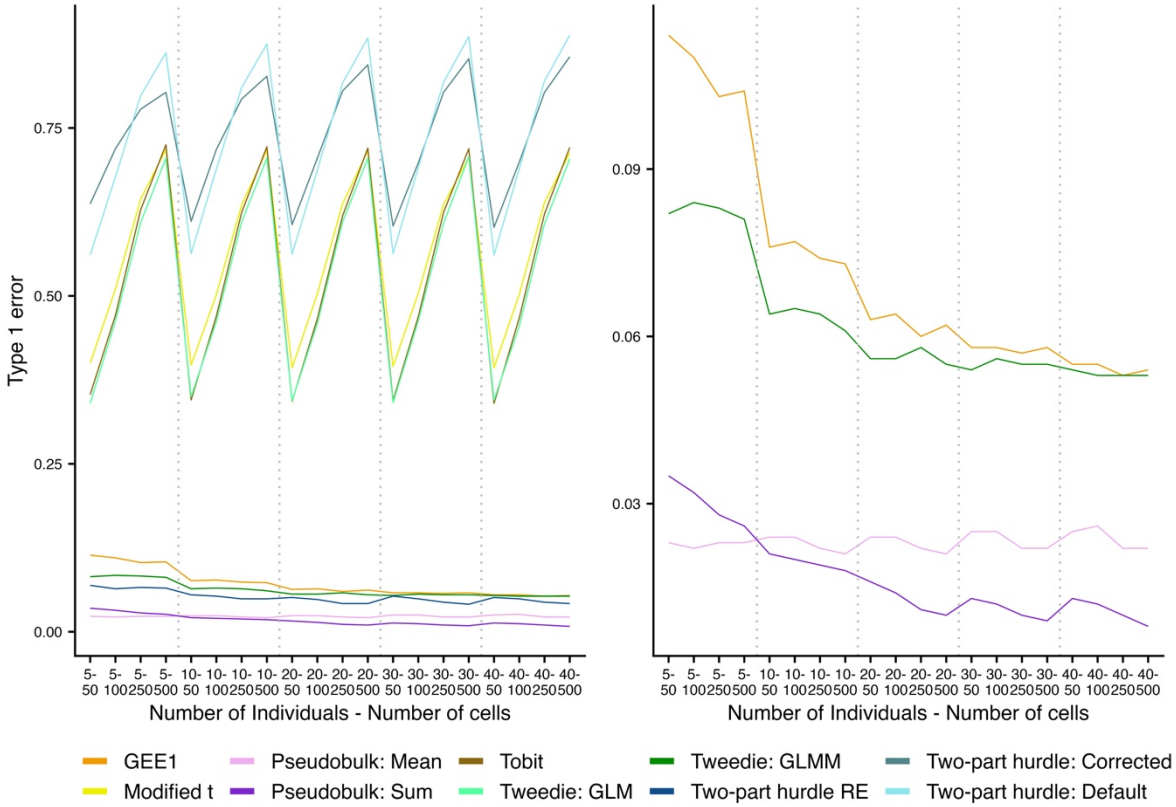


Figure 2-1: Type 1 error highlights the benefits of pseudobulk approaches.

The average type 1 error from the 20,000 iterations; 50 runs for each of the 5 to 40 individuals and 50 to 500 cells at a p-value cut-off of 0.05 on 5,000 genes reported by Zimmerman *et al.,* (Zimmerman, Espeland and Langefeld, 2021). Left shows all benchmarked models whereas right focuses on the top four approaches. The different models are pseudoreplication approaches; 'Modified t' (ROTS), 'Tobit' (Monocole), 'Two-part hurdle: Default' (MAST), 'Two-part hurdle: Corrected' (MAST & ComBat), 'GEE1' (Geepack), 'Tweedie: GLM' (glmmTMB), pseudobulk approaches; 'Pseudobulk: Mean' (DESeq2), 'Pseudobulk: Sum' (DESeq2) and mixed model approaches; 'Tweedie: GLMM' (glmmTMB) and 'Two-part hurdle: RE' (MAST).

2.2. Methods

2.2.1. Differential expression techniques

Commonly used pseudoreplication, mixed models and pseudobulk approaches in the literature were used for this benchmark study, matching those used by the original authors (Zimmerman, Espeland and Langefeld, 2021). These were implemented in R (v3.5.1) following the default implementations and parameters unless otherwise specifically stated. All techniques used are listed in **Table 2-1**.

Method	Description	Method Type	Implementation
GEE1	Generalised linear models - generalized estimating equation	Pseudoreplication	Geepack (Højsgaard, Halekoh and Yan, 2006)
Modified t	Reproducibility-Optimized Statistical Testings (ROTS)	Pseudoreplication	ROTS (Suomi *et al.*, 2017)
Pseudobulk: Mean	Pseudobulk with mean aggregation	Pseudobulk	DESeq2 (Love, Huber and Anders, 2014)
Pseudobulk: Sum	Pseudobulk with sum aggregation	Pseudobulk	DESeq2 (Love, Huber and Anders, 2014)
Tobit	Monocole	Pseudoreplication	Monocole (Trapnell *et al.*, 2014)
Tweedie: GLM	Generalised linear model with Tweedie distribution.	Pseudoreplication	glmmTMB (Brooks *et al.*, 2017)
Tweedie: GLMM	Generalised linear mixed-effects model with Tweedie distribution.	Mixed Model	glmmTMB (Brooks *et al.*, 2017)
Two-part hurdle RE	Model-based analysis of single-cell transcriptomics (MAST) with a random effect for individuals.	Mixed Model	MAST (Finak *et al.*, 2015)
Two-part hurdle: Corrected	MAST batch-corrected for individuals.	Pseudoreplication	MAST & ComBat (Johnson, Li and Rabinovic, 2007; Finak *et al.*, 2015)
Two-part hurdle: Default	MAST without random effects.	Pseudoreplication	MAST (Finak *et al.*, 2015)

Table 2-1: The different methods benchmarked in the analysis with their implementation approach and methodological types

2.2.2. Data simulation

The data simulation technique hierarchicell is described in Zimmerman *et al.*'s publications (Zimmerman and Langefeld, 2021; Zimmerman, Espeland and Langefeld, 2021) but in short - The approach used a single-cell dataset, a Smart-Seq2 assay of pancreatic alpha cells (Segerstolpe *et al.*, 2016), as a reference to estimate default simulation parameters. After normalisation and filtering of low quality cells, genes were filtered leaving an uncorrelated subset that captures the most representative patterns of detectable transcripts per million (TPM) values, without the resulting parameter estimates being primarily driven by dropout. The non-zero TPM values were used to calculate the within-sample and across sample variance and mean to enable the calculation of synthetic data following a negative binomial distribution. By using a reference dataset, the approach ensures that true intra- and inter-individual correlation patterns are maintained in the synthetic data. Thereafter, the simulated data is adjusted based on the user's choice of gene dropout, number of individuals, number of cells per individual, the distribution for the variation in the number of cells and the true fold change of the genes.

We modified Zimmerman *et al*.'s hierarchicell approach in two ways. First, we enabled the simulation of both differentially expressed and non-differentially expressed genes. The DEGs are simulated with uniform sampling of fold change between a max and minimum value. We also enabled the variation in proportion of DEGs and non-DEGs. Secondly, we modified hierachicell to correct the seeding of the pseudo-random number generator to enable fair comparisons across models, ensuring the simulated data was consistent for each model.

2.2.3. Benchmarking

We utilised our updated version of hierarchicell to simulate 20,000 scRNA-Seq datasets for testing; 50 iterations of each combination of 5–40 individuals and 50–500 cells with 5,000 DEGs and 5,000 non-DEGs. The differentially expressed genes were randomly simulated with a fold change between 1.1 and 10 ($\log_2$ fold change 0.14-3.32).

Moreover, to test varying the proportions of DEGs, we created 50 iterations of datasets for 5%, 10%, 20% and 30% DEGs with 20 individuals and 100 cells. And to test an imbalance in the number of cells, scRNA-Seq datasets were simulated 50 times for 20 individuals and 5,000 genes. The number of cells were randomly chosen using a gamma distribution with shape 4 and scale 45 separately for cases and controls to produce the imbalanced dataset (giving a mean 150-200 cells).

2.2.4. Evaluation

To perform a fair comparison between the different pseudoreplication, mixed models and pseudobulk approaches, we used the Matthews Correlation Coefficient (MCC) giving a balanced measure of performance. MCC is a well-known and frequently adopted metric in the machine learning field, which offers a more informative and reliable score on binary classification problems (Chicco and Jurman, 2020). MCC produces scores in [-1,1] and will only assign a high score if a model performs well on both non-differentially and differentially expressed genes. Moreover, MCC scores are proportional to both the size of the differentially and non-differentially expressed genes, so it is robust to imbalanced datasets. We also benchmarked the models using receiver operating characteristics (ROC) curves for different proportions of differentially expressed genes to study the model performances at differing p-value cut-offs i.e. for statistical test comparisons.

2.3. Results

Our analysis based on MCC demonstrates that pseudobulk approaches achieve highest performance across individuals and cells variations (**Figure 2-2**). There is one exception for sum pseudobulk, which performs worse than Tobit at 5 individuals and 10 cells. **Figure 2-2** also highlights a trend whereby pseudoreplication models; 'Modified t', 'Tobit', 'Two-part hurdle: Default' and 'Two-part hurdle: Corrected' (which take cells as independent replicates) showed increasingly poor performance as the number of cells increases. This trend is likely due to the over-estimation of power driven by the dependence between cells from the same individual (Lazic, 2010) and agrees with Zimmerman *et al.*'s findings (Zimmerman, Espeland and Langefeld, 2021). On the other hand, both pseudobulk approaches; 'Pseudobulk: Mean' and 'Pseudobulk: Sum', showed improved performance as the number of cells increases. This trend was also noted in two of the other models; 'GEE1' and 'Tweedie: GLMM'.

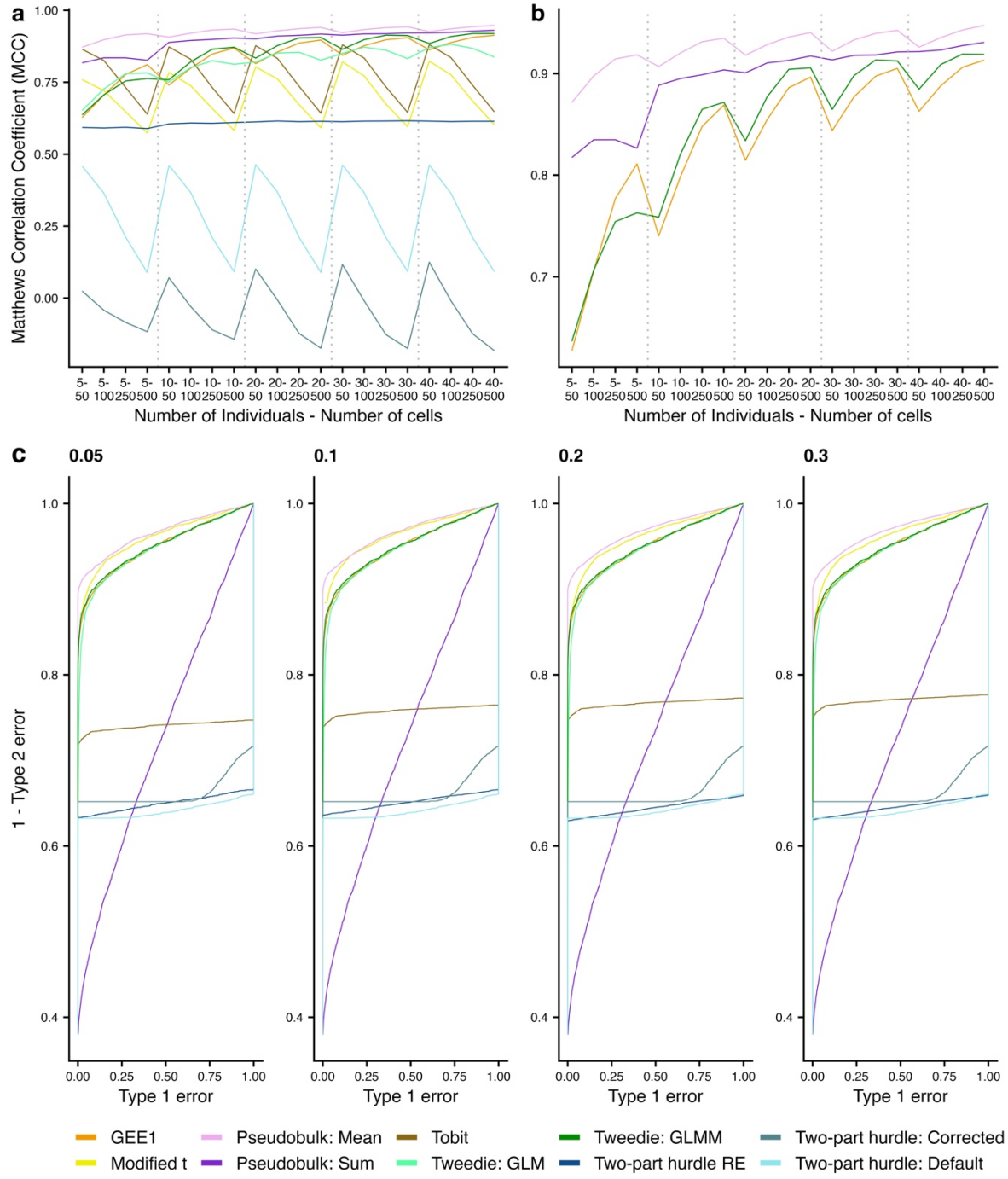


Figure 2-2: Balanced performance of analysed models shows superiority of pseudobulk.

(a, b) give the average Matthews correlation coefficient from the 20,000 iterations; 50 runs for each of the 5–40 individuals and 50–500 cells at a p-value cut-off of 0.05 on 10,000 genes. (a) shows all benchmarked models whereas (b) focuses on the top four approaches. (c) gives the receiver operating characteristics (ROC) curve across 50 runs each for different

proportions of simulated differentially expressed genes — 0.05, 0.1, 0.2, 0.3. Twenty individuals were simulated for case and controls, each with 100 cells. The different models are pseudoreplication approaches; 'Modified t' (ROTS), 'Tobit' (Monocole), 'Two-part hurdle: Default' (MAST), 'Two-part hurdle: Corrected' (MAST & ComBat), 'GEE1' (Geepack), 'Tweedie: GLM' (glmmTMB), pseudobulk approaches; 'Pseudobulk: Mean' (DESeq2), 'Pseudobulk: Sum' (DESeq2) and mixed model approaches; 'Tweedie: GLMM' (glmmTMB) and 'Two-part hurdle: RE' (MAST). More detail on these models is given in **Table 2-1**. The area under the receiver operating characteristics curve (AUC) performance split by each iteration, corresponding to (c) above is given in **Table 2-2**.

Method	0.05	0.1	0.2	0.3
GEE1	0.950138	0.950138	0.950138	0.950138
Modified t	0.956821	0.960299	0.957801	0.957978
Pseudobulk: Mean	**0.965886**	**0.967704**	**0.968249**	**0.968664**
Pseudobulk: Sum	0.731316	0.731252	0.730117	0.729725
Tobit	0.740201	0.758228	0.76654	0.770334
Tweedie: GLM	0.947437	0.947437	0.947437	0.947437
Tweedie: GLMM	0.951053	0.951053	0.951053	0.951053
Two-part hurdle RE	0.649905	0.651323	0.644665	0.645651
Two-part hurdle: Corrected	0.662429	0.662429	0.662429	0.662429
Two-part hurdle: Default	0.641707	0.641707	0.641707	0.641707

Table 2-2: The area under the receiver operating characteristics curve (AUC) of all benchmarked models.

Results shown across different proportions of simulated differentially expressed genes (DEGs) - 0.05, 0.1, 0.2, 0.3. 20 individuals were simulated for case and controls, each with 100 cells. The top performing method at each proportion is in bold. The proportion of DEGs are of the 5,000 non-DEGs simulated. The values are rounded to six decimal places. The different models are pseudoreplication approaches; 'Modified t' (ROTS), 'Tobit' (Monocole), 'Two-part hurdle: Default' (MAST), 'Two-part hurdle: Corrected' (MAST & ComBat), 'GEE1' (Geepack), 'Tweedie: GLM' (glmmTMB), pseudobulk approaches; 'Pseudobulk: Mean' (DESeq2), 'Pseudobulk: Sum' (DESeq2) and mixed model approaches; 'Tweedie: GLMM' (glmmTMB) and 'Two-part hurdle: RE' (MAST).

Moreover, for statistical test comparisons, another approach is to compare the power of tests at the same test size. That is, to compare the models' sensitivity (1 – type 2 error) at a consistent type 1 error rate. Therefore, we generated ROC curves for the different approaches, enabling such comparisons. For example, in **Figure 2-3**, we highlight the different sensitivity scores (1 – type 2 error) of the models obtained at a consistent type 1 error rate of 0.05. We find that pseudobulk mean performs best at this type 1 error rate (with a sensitivity >0.9 whereas all other methods had < 0.9) and at all other type 1 error rates (**Figure 2-2**). Interestingly, we show that the two mixed model approaches ('Two-part hurdle: RE' and 'GLMM Tweedie') perform relatively poorly even compared to some pseudoreplication approaches. This analysis demonstrates how pseudobulk mean obtains low type-2 error rates, even at the lowest type-1 error rates of the methods benchmarked, supporting our MCC results.

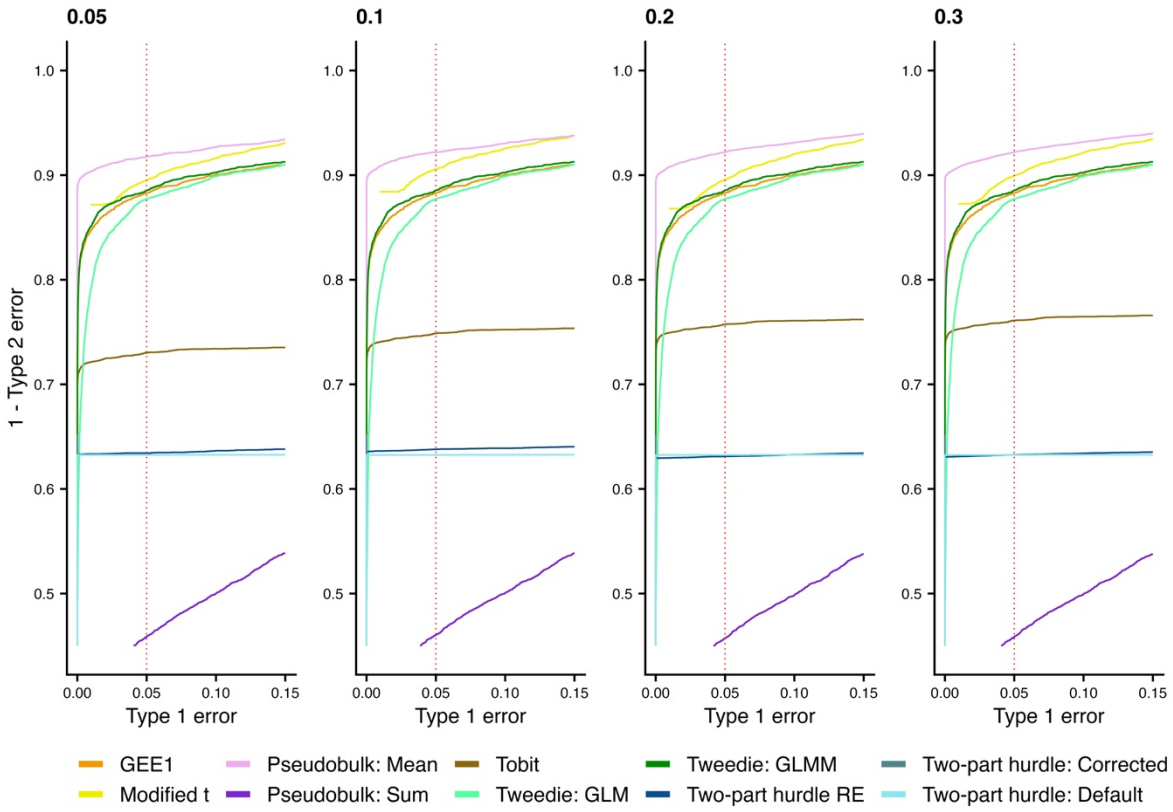


Figure 2-3: Pseudobulk mean is best performing at a constant type 1 error rate.

The four images give the receiver operating characteristics (ROC) curve across 50 runs each for different proportions of simulated differentially expressed genes (DEGs) - 0.05, 0.1,0.2,0.3. 20 individuals were simulated for case and controls, each with 100 cells. The proportion of DEGs are of the 5,000 non-DEGs simulated. The sensitivity (1 – type 2 error) of the different approaches at a 0.05 type 1 error value are highlighted by the red dashed line. The different models are pseudoreplication approaches; 'Modified t' (ROTS), 'Tobit' (Monocole), 'Two-part hurdle: Default' (MAST), 'Two-part hurdle: Corrected' (MAST & ComBat), 'GEE1' (Geepack), 'Tweedie: GLM' (glmmTMB), pseudobulk approaches; 'Pseudobulk: Mean' (DESeq2), 'Pseudobulk: Sum' (DESeq2) and mixed model approaches; 'Tweedie: GLMM' (glmmTMB) and 'Two-part hurdle: RE' (MAST).

Zimmerman *et al.* argued that pseudobulk methods are “overly conservative" relative to mixed models in their work. Specifically, they refer to pseudobulk approaches’ lower than nominal levels of type 1 error rates, demonstrated in their results where, based on a consistent p-value cut-off of 0.05, they benchmark the performance of different methods at identifying non-differentially expressed genes (**Figure 2-1**). Their analysis showed pseudobulk approaches’ type 1 error rates were below the expected 0.05 of false positives at each number of individuals and number of cells combination. In this analysis, it is true that pseudobulk approaches have mis-calibrated confidence intervals, obtaining fewer false positives than expected at a 0.05 p-value cut-off. Given this conservative 95% confidence intervals of pseudobulk methods, they could, as a result, have a higher type 2 error than other methods. However, our ROC analysis disproves this. It shows how, at equal type 1 error rates, pseudobulk mean has the lowest type 2 error rate of all tested methods (**Figure 2-2**, **Figure 2-3**).

All analysis to this point have been on simulations with an equal number of cells in each sample. However, in real datasets this would never be the case (’t Hoen *et al.*, 2013). To mirror this, we simulated data with an imbalanced number of cells between case and controls. We find that pseudobulk mean outperforms all other approaches on this analysis (**Figure 2-4**). The pseudobulk approach which aggregated by averaging rather than taking the sum appears to be the top performing overall. However, it is worth noting that hierarchicell does not normalise the simulated datasets before passing to the pseudobulk approaches. This is a standard step in such analysis to account for differences in sequencing depth and library sizes (Love, Huber and Anders, 2014). This approach was taken by Zimmerman *et al.* as their data is simulated one independent gene at a time without considering differences in library size. The effect of this step is more apparent on the imbalanced number of cells where pseudobulk

sum's performance degrades dramatically. Pseudobulk mean appears invariant to this missing normalisation step because of the averaging's own normalisation effect. Importantly, this was a flaw in the simulation software strategy and does not show an improved performance of pseudobulk mean over sum. We believe this approach also affected the performance of pseudobulk sum on the different proportions of differentially expressed genes (**Figure 2-2**).

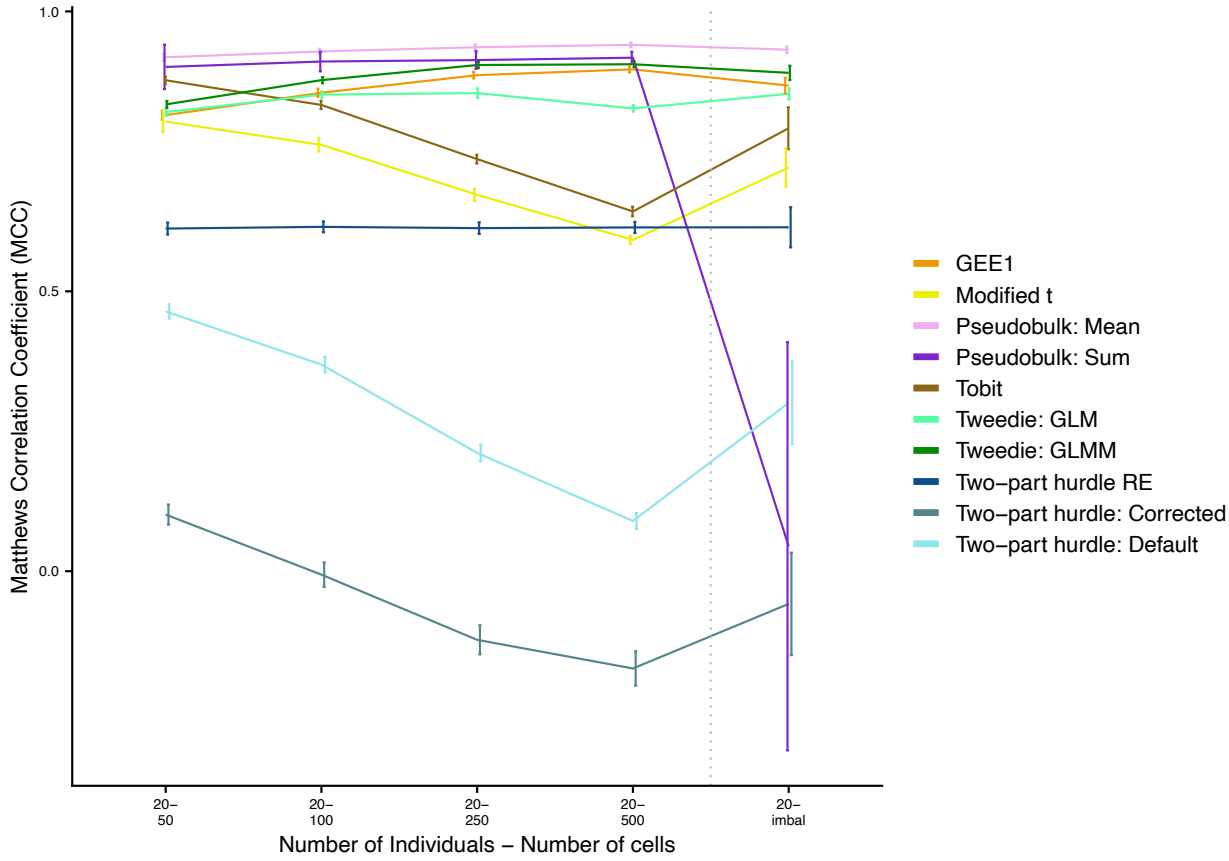


Figure 2-4*:* Pseudobulk mean is the best performing on imbalanced datasets

The average Matthews correlation coefficient of all benchmarked models across all balanced number of cells and the imbalanced number of cells for 20 individuals; 50 runs for each at a p-value cut-off of 0.05 on 5,000 genes. The number of cells were randomly chosen using a gamma distribution with shape 4 and scale 45 separately for cases and controls to produce the imbalanced dataset (giving a mean 150-200 cells). The error bars give 1 standard deviation around the mean. The different models are pseudoreplication approaches; 'Modified t' (ROTS), 'Tobit' (Monocole), 'Two-part hurdle: Default' (MAST), 'Two-part hurdle: Corrected' (MAST & ComBat), 'GEE1' (Geepack), 'Tweedie: GLM' (glmmTMB), pseudobulk approaches; 'Pseudobulk: Mean' (DESeq2), 'Pseudobulk: Sum' (DESeq2) and mixed model approaches; 'Tweedie: GLMM' (glmmTMB) and 'Two-part hurdle: RE' (MAST).

2.4. Discussion

Our analysis found that pseudobulk methods outperformed both pseudoreplication techniques and mixed models. Based on these findings, we recommend the use of pseudobulk approaches for differential expression in single-cell RNA-sequencing analyses. Given mixed models' well known computational inefficiency, our work promotes a quick, computationally less intensive and more accurate approach for researchers. Pseudobulk approaches were also found to be top performing in a recent review by Squair *et al.*, (Squair *et al.*, 2021). Notably, the pseudobulk method used here, DESeq2 (Love, Huber and Anders, 2014), performed worse than other pseudobulk models in Squair *et al.*'s analysis and so their adoption may further increase the performance of pseudobulk approaches on our dataset. Conversely, Squair *et al.*, did not consider all models included in our analysis or the different forms of pseudobulk aggregation. Therefore, our results on sum and mean pseudobulk extend their findings and indicate that mean aggregation may be the best performing. However, the reader should be cognisant that the lack of a normalisation step based on the flaw in the simulation software strategy likely causes the increased performance of mean over sum aggregation. Further, the use of simulated datasets in our analysis may not accurately reflect the differences between individuals that are present in biological datasets. Thus, despite both our results and those reported by Squair *et al.*, there is still room for further analysis, benchmarking more models, including different combinations of pseudobulk aggregation methods and models, on more representative simulated datasets and biological datasets to identify the optimal approach. Specifically, we would expect pseudobulk sum with a normalisation step to outperform pseudobulk mean since it can account for the intra-individual variance which is otherwise lost with pseudobulk mean, but this should be tested, including on imbalanced datasets and at consistent type 1 error rates.

In conclusion, our results demonstrate that pseudobulk approaches lead to the best performance for the analysis of single-cell expression data based on power at equivalent type 1 error rates and on MCC for both balanced and imbalanced number of cells, from this simulated dataset.

2.5. Data availability

All data underlying figures and table produced in this chapter are available at https://github.com/Al-Murphy/reanalysis_scRNA_seq_benchmark, DOI: (Murphy, 2022).

2.6. Code availability

The modified version of hierarchicell which returns the Matthews correlation coefficient performance metric as well as the type 1 error rates, uses the same simulated data across approaches and has checkpointing capabilities (so runs can continue from where they left off if aborted or crashed) is available at: https://github.com/neurogenomics/hierarchicell, DOI: (Zimmerman and Murphy, 2022). The benchmarking script along with the results is available at:

https://github.com/Al-Murphy/reanalysis_scRNA_seq_benchmark, DOI: (Murphy, 2022).

3. Avoiding false discoveries in single cell RNA-Seq by revisiting the first Alzheimer's disease dataset

3.1. Introduction

The work in this chapter is adapted from that set out in our previous publication (Murphy, Fancy and Skene, 2023). Having helped define a best-practice approach to differential expression analysis, our attention turned to applying it to neurodegenerative disease, using the first snRNA-Seq study of AD (Mathys *et al.*, 2019) as a case study. Mathys *et al*., undertook the first snRNA-Seq of AD (Mathys *et al.*, 2019). The authors profiled the transcriptomes of approximately 80,000 cells from the prefrontal cortex, collected from 48 individuals – 24 of which presented with varying degrees of AD pathology. Mathys *et al*.'s data processing and quality control strategy for their single-nucleus RNA-Seq data was state-of-the-art at this time. Furthermore, the authors took extra measures in an attempt to ensure the reliability of their results. Here, we re-analysis these data, not a criticism of the study, but as an endeavour to raise awareness and provide recommendations for rigorous analysis of sc/snRNA-Seq data for future studies. Most importantly, we aim to ensure that the AD research field does not focus on spuriously identified genes.

3.2. Methods

3.2.1. Processing of the snRNA-Seq dataset

The data re-processing was conducted with scFlow (Khozoie *et al.*, 2021), the steps of which are discussed in the following two sections:

3.2.1.1 Quality control of snRNA-Seq data

The raw single-nucleus RNA-Sequencing data (10.7303/syn18485175) and the ROSMAP metadata (10.7303/syn3157322) were downloaded from https://www.synapse.org/ upon acquiring appropriate approval. Downstream primary analyses of gene-cell matrices were performed using our scFlow pipeline (Khozoie *et al.*, 2021). To determine ambient RNA profile and distinguish true nuclei from empty droplets, emptyDrops was used with a lower parameter of <100 counts, an alpha cutoff of ≤0.001, and with 10,000 Monte-Carlo iterations (Lun *et al.*, 2019). This approach has been recommended as best practice in the literature (Amezquita *et al.*, 2020). Nuclei were then filtered for ≥200 total counts and ≥200 total expressed genes, which was defined as a minimum of 2 counts in at least 3 cells. We excluded any nuclei with total counts or total expressed genes with more than 4 median absolute deviation (MAD) defined by an adaptive thresholding method. Nuclei were excluded if the proportion of counts mapping to mitochondrial genes was more than 10%, as set out in best-practice guidelines (Amezquita *et al.*, 2020). Doublets were identified using the DoubletFinder algorithm, with a doublets-per-thousand-cells increment of 8 cells (recommended by 10X Genomics), a pK value of 0.005 (McGinnis, Murrow and Gartner, 2019). DoubletFinder was shown to be the best overall performing method in a recent benchmark (Xi and Li, 2021). The aggregated number of cells and proportions dropped at each step is given in **Table 3-1** while a comparison to the proportion of cells in each cell type after reprocessing compared to the author's processed data is in **Figure 3-1**. All files from the

scFlow run, including quality control statistics are available in the Github repository in the scFlow_files folder: https://github.com/neurogenomics/reanalysis_Mathys_2019. This includes sample level genes and cells quality control numbers.

3.2.1.2 Integration and clustering

The linked inference of genomic experimental relationships (LIGER) package was used to calculate integrative factors across samples (Welch *et al.*, 2019). LIGER was recently found to be one of the top performing methods for batch-effect correction (Tran *et al.*, 2020). LIGER parameters used included: k: 30, lambda: 5.0, thresh: 0.0001, max_iters: 100, knn_k: 20, min_cells: 2, quantiles: 50, nstart: 10, resolution: 1, num_genes: 3000, center: false. Two-dimensional embeddings of the LIGER integrated factors were calculated using the uniform-manifold approximation and projection (UMAP) algorithm with the following parameters: pca_dims: 50, n_neighbours: 35, init: spectral, metric: euclidean, n_epochs: 200, learning_rate: 1, min_dist: 0.4, spread: 0.85, set_op_mix_ratio: 1, local connectivity: 1, repulsion_strength: 1, negative_sample_rate: 5, fast_sgd: false (McInnes, Healy and Melville, 2020). The Leiden community detection algorithm was used to detect clusters of cells from the 2D UMAP (LIGER) embeddings; a resolution parameter of 0.001 and a k value of 50 was used (Traag, Waltman and van Eck, 2019). This approach has been noted as best practice by a recent review (Heumos *et al.*, 2023). Automated cell-typing of the detected clusters was performed as previously described using the Expression Weighted Celltype Enrichment (EWCE) algorithm in scFlow against a previously generated cell-type data reference from the Allan Human Brain Atlas (Skene and Grant, 2016; Hodge *et al.*, 2019). The top five marker genes for each automatically annotated cell-type were determined using Monocle 3 and validated against canonical cell-type markers (Trapnell *et al.*, 2014).

Quality Control Steps	Total Cells	Percentage
Pre-Quality Control	35,389,440	
Total failed	35,337,874	99.85%
- Minimum Library Size (n<200)	35,307,281	99.77%
- Maximum Library Size	4,742	0.01%
- Minimum Expressed Genes (n<200)	35,312,434	99.78%
- Maximum Library Size/Expressed Genes (MAD>4)	2,149	0.01%
- Proportion Mitochondrial Genes (>=0.1)	1,097,738	3.10%
- Multiplets (pK = 0.0054)	581	0.00%
Total passed	51,566	0.15%

Table 3-1: Quality control summary.

Overview of the aggregated number of cells across samples removed at each step of the quality control (QC) as part of scFlow. Note that cells can fail QC for more than one check so only the total failed and Total passed rows will sum to 100%.

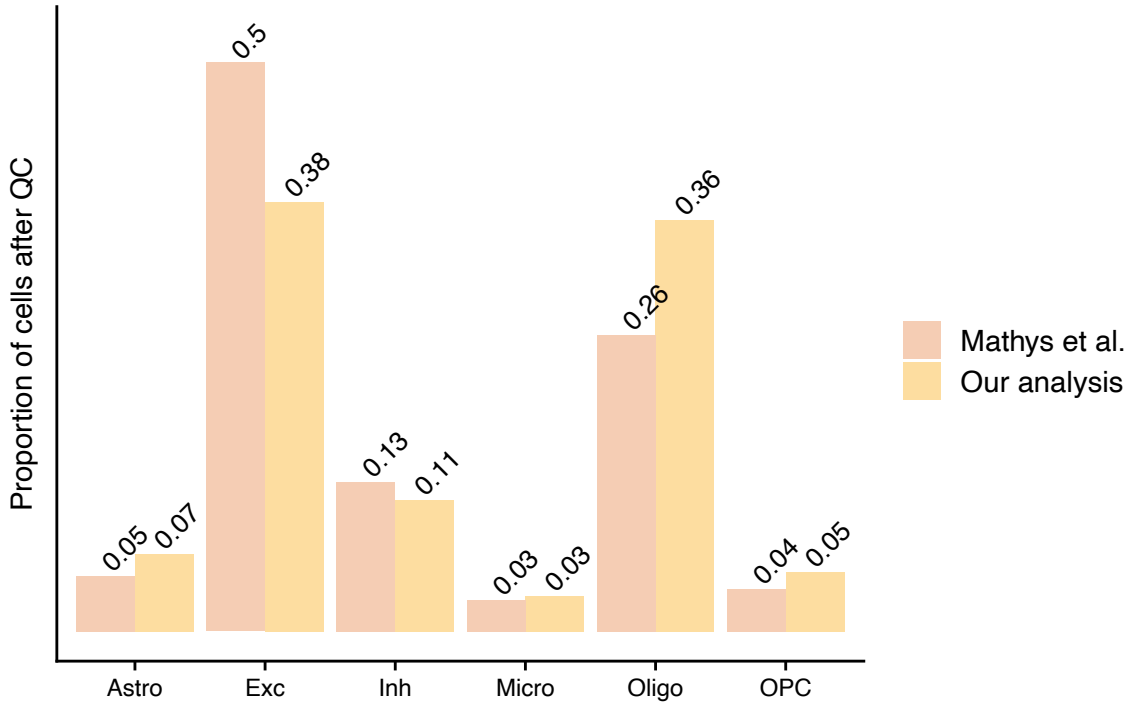


Figure 3-1: Differing processing results in disparate proportions of cell types across analyses. The proportion of cells (Y-axis) left after quality control (QC) from the author's processing approach (Mathys et al.) and our standardised pipeline approach - scFlow (Our analysis). The original author's analysis resulted in far greater a proportion of excitatory neurons (Exc). Here, Astro are astrocytes, Exc are excitatory neurons, Inh are inhibitory neurons, Micro are microglia, Oligo are oligodendrocytes and OPCs are oligodendrocyte precursor cells.

3.2.2. Differential Expression analysis

All differential expression (DE) analysis was conducted using pseudobulk DE approach with sum aggregation and edgeR LRT (Chen, Lun and Smyth, 2016). Pseudobulk aggregates nuclei within a biological replicate (an individual) for each cell type, reducing the dropout issue in single cell data and avoiding the false inflation of confidence from non-independent samples of pseudoreplication approaches (Squair *et al.*, 2021; Murphy and Skene, 2022). The DE analysis pipeline is available at: https://github.com/neurogenomics/reanalysis_Mathys_2019. This is a general use pipeline which can be run for any single-nucleus or single-cell transcriptomic dataset. Note that we report differentially expressed genes across AD and controls using the same processed data the authors used (**Table 3-2**) and using our reprocessed data (**Table 3-3**).

Cell	logFC	logCPM	LR	PValue	adj_pval	HGNC
Mic	2.70178913	6.99794619	26.1418415	3.17E-07	0.00061349	ACRBP
Mic	1.48930071	8.06240877	28.6361217	8.73E-08	0.00019303	APOC1
Mic	1.09327669	8.64199769	21.5323014	3.48E-06	0.00336416	CD81
Mic	-1.4157681	7.93884875	23.9955467	9.66E-07	0.00135806	CD83
Mic	3.3782727	6.86183548	32.0804401	1.48E-08	4.58E-05	CLEC1B
Mic	2.84072452	6.74370542	21.7745509	3.07E-06	0.00316269	EGF
Mic	2.55769658	6.78345087	18.0468872	2.16E-05	0.01699007	ELOVL7
Mic	-1.2056098	8.33197499	22.6644045	1.93E-06	0.00229576	IFI44L
Mic	-1.6616069	7.15366639	16.4801274	4.92E-05	0.03306938	IFI6
Mic	-1.9809425	7.00396289	17.9180823	2.31E-05	0.01699007	IFIT3
Mic	2.76502672	6.72978805	20.6543637	5.50E-06	0.00472825	ITGA2B
Mic	1.90963403	7.01552233	16.3200189	5.35E-05	0.03448474	MAP1A
Mic	-1.8194508	8.26208887	45.2221008	1.76E-11	1.36E-07	NAMPT
Mic	2.0945044	7.11048456	20.8068524	5.08E-06	0.00462318	NEXN
Mic	-2.3789762	6.93896985	22.3912441	2.22E-06	0.00245752	NR4A2
Mic	-2.8553462	6.73713862	22.8029868	1.79E-06	0.00229576	NR4A3
Mic	3.32873829	6.84942721	30.955327	2.64E-08	6.81E-05	PF4
Mic	3.4213986	6.87326383	33.2621657	8.05E-09	3.11E-05	PKHD1L1
Mic	3.64525677	6.93422174	38.661272	5.04E-10	2.60E-06	PPBP
Mic	2.30482679	8.10570443	60.7932697	6.34E-15	9.81E-11	PTPRG
Mic	-1.0382468	8.11450266	15.5968273	7.84E-05	0.04850839	RORA
Mic	2.54636649	6.69202981	17.2532606	3.27E-05	0.02300507	SDPR
Mic	-0.9629617	8.8434334	17.9319131	2.29E-05	0.01699007	SYTL3
Mic	-1.4215374	7.99629806	25.4736272	4.48E-07	0.00077092	TMEM2
Mic	2.98901596	6.77276641	24.2100819	8.64E-07	0.00133637	TUBB1
OPC	-2.8274718	5.03371292	22.1334581	2.54E-06	0.04176231	EGR1

Table 3-2: Differentially expressed genes from our reanalysis.

Table of our differentially expressed genes using the same processed data the authors used and pseudobulk differential expression approach. Where OPCs are oligodendrocyte precursor cells and Micro are microglia.

Cell	logFC	logCPM	LR	PValue	adj_pval	ensembl_id	HGNC
OPC	-4.1544663	4.92100803	21.6911445	3.20E-06	0.04985906	ENSG00000166573	GALR1
Astro	-4.5845276	4.7965143	22.2367847	2.41E-06	0.037634	ENSG00000137959	IFI44L
Micro	-3.7616619	7.32875316	26.8149688	2.24E-07	0.00077905	ENSG00000077238	IL4R
Micro	-2.0681446	7.88736441	17.5929095	2.74E-05	0.0346187	ENSG00000105835	NAMPT
Micro	-1.6757556	7.58472506	19.1736829	1.19E-05	0.02076348	ENSG00000118257	NRP2
Micro	-3.1556403	6.85232653	19.2064627	1.17E-05	0.02076348	ENSG00000135363	LMO2
Micro	-3.4339265	6.9290472	19.5975589	9.56E-06	0.02076348	ENSG00000138135	CH25H
Micro	-2.8183109	6.77500676	16.907959	3.92E-05	0.04550806	ENSG00000142408	CACNG8
Micro	2.90076647	8.34560617	45.5144266	1.52E-11	2.11E-07	ENSG00000144724	PTPRG
Micro	3.25867589	6.91671013	16.5519147	4.73E-05	0.0490155	ENSG00000163106	HPGDS
Micro	-2.0290905	7.12321166	16.4746746	4.93E-05	0.0490155	ENSG00000171612	SLC25A33
Micro	-3.4657301	6.93307221	19.7883301	8.65E-06	0.02076348	ENSG00000172243	CLEC7A
Micro	-4.172807	7.16813583	34.3515807	4.60E-09	3.20E-05	ENSG00000174600	CMKLR1
Micro	-3.1984588	6.87310555	18.5335889	1.67E-05	0.0232342	ENSG00000227531	RP11-202G18.1
Micro	3.40562887	6.9381703	18.5526502	1.65E-05	0.0232342	ENSG00000228058	RP11-552D4.1
Micro	4.46073301	7.66559163	29.7716679	4.86E-08	0.00022549	ENSG00000253496	RP11-13N12.1

Table 3-3: Differentially expressed genes on reprocessed data.

The differentially expressed genes from our reanalysis, using the reprocessed data and pseudobulk differential expression approach. Where OPCs are oligodendrocyte precursor cells, Astro are astrocytes and Micro are microglia.

3.3. Results and Discussion

Our questions of Mathys *et al.*, focus around their data processing and their differential expression analysis (**Figure 3-2**) (Mathys *et al.*, 2019). Firstly, in relation to their processing approach, the authors discussed the high percentages of mitochondrial reads and low number of reads per cell present in their data. This is indicative of low cell quality (Ilicic *et al.*, 2016) however, we believe the authors' quality control (QC) approach may not capture all of these low quality cells. Moreover, the authors did not integrate the cells from different individuals to account for batch effects. As the field has matured since the author's work was published, dataset integration has become a common step in single-cell RNA-Seq protocols and is recommended by some to remove confounding sources of variation (Heumos *et al.*, 2023; Amezquita *et al.*, 2020; Tran *et al.*, 2020). To gain advantage of these recent approaches, we used scFlow (Khozoie *et al.*, 2021) to reprocess the author's data. This pipeline included the removal of empty droplets, nuclei with low read counts and doublets followed by embedding and integration of cells from separate samples and cell typing. ScFlow combines best practice approaches for processing sc/snRNA-Seq datasets, see **3.2 Methods** for a detailed explanation of these steps. Re-processing resulted in 50,831 cells passing QC, approximately 20,000 less the authors' post-processing set with differing cell type proportions (**Figure 3-1**).

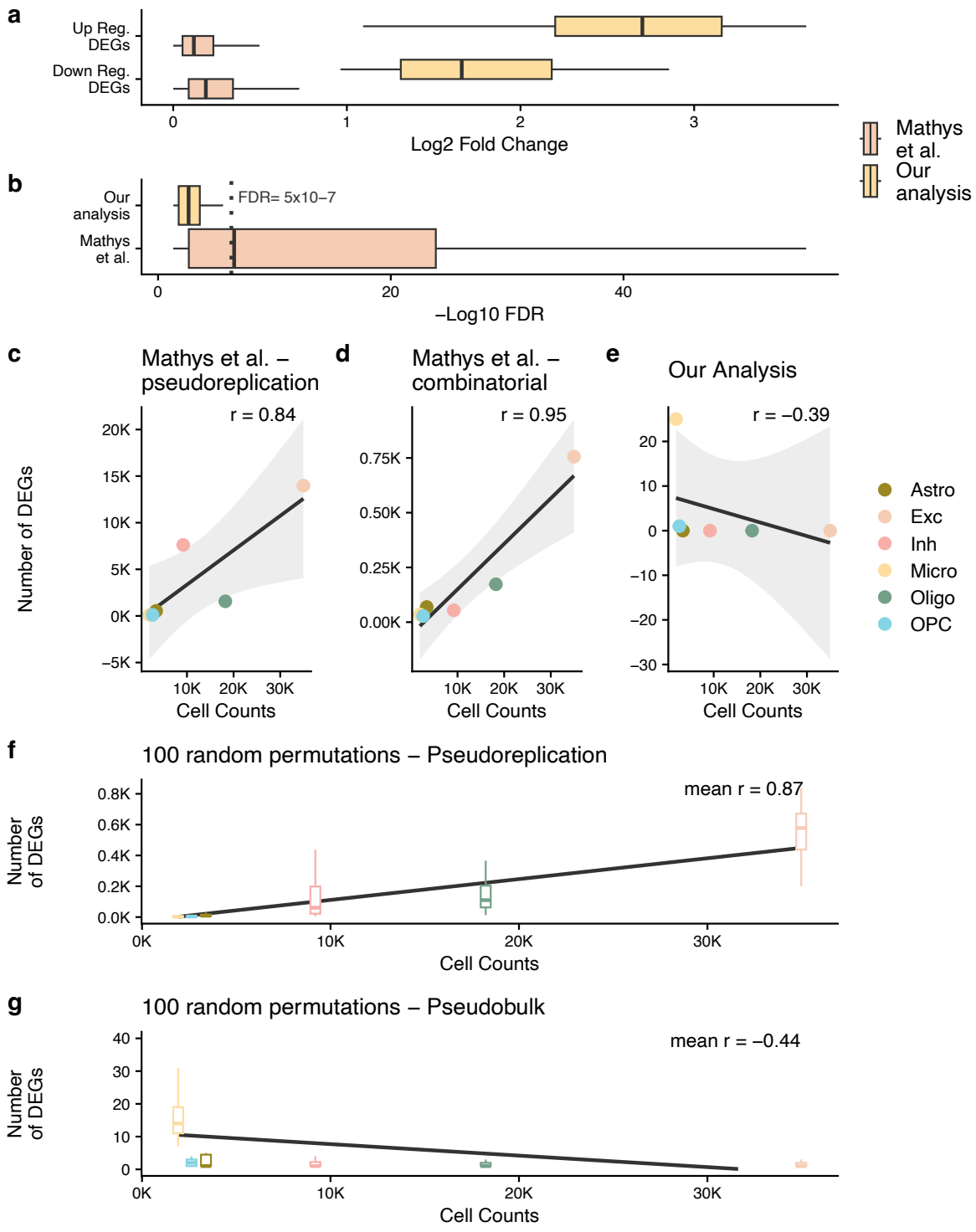


Figure 3-2: Pseudobulk differential expression results in far less dubious disease-related genes.

(a, b) highlight the $\log_2$ fold change and -$\log_{10}$ false discovery rate (FDR) of the differentially expressed genes from the author's original work (Mathys et al.) and our reanalysis (Our

analysis). In **b**, we have marked an FDR of $5x10^{-7}$, dashed grey line, to highlight how small the p-values from Mathys et al.'s analysis are. (c, d, e, f, g) show the Pearson correlation between the cell counts after QC and the number of DEGs identified. For (f, g) analysis, the samples have been randomly mixed between case and control patients. The different cell types are astrocytes (Astro), excitatory neurons (Exc), inhibitory neurons (Inh), microglia (Micro), oligodendrocytes (Oligo) and oligodendrocyte precursor cells (OPC).

In regards to data quality, it is worth noting that over 99% of nuclei had less than 200 genes expressed (**Table 3-1**). While this QC step was not unique to our reprocessing, the authors made the same exclusion in their analysis (Mathys *et al.*, 2019), it highlights the relatively low quality of the data which may be attributable to the early stage of snRNA-Seq technology of the time. For example, Brase *et al.*'s recent study of snRNA-Seq of autosomal dominant AD (Brase *et al.*, 2023) used a more stringent cut-off for the minimum number of expressed genes and still kept 27% (122 times more) of the assayed cells after all QC steps. Moreover, the authors discussed the high percentages of mitochondrial reads in their data. The differences in approaches to filtering based on the proportion mitochondrial reads accounts for the notable discrepancy in the number of nuclei after QC between our approach and the authors. Our approach used a 10% cut-off for the proportion of mitochondrial reads in a nuclei, as set out in Amezquita *et al.*'s best-practice guidelines (Amezquita *et al.*, 2020), which is less stringent than Seurat's 5% guideline (Hao *et al.*, 2021) or that from Heumos *et al.* - 8% from a median absolute deviations (MAD) based cut-off selection (Heumos *et al.*, 2023). Conversely, the authors filtered out high mitochondrial read nuclei based on clusters from their t-SNE projection of the data (Mathys *et al.*, 2019). Even at our lenient cut-off, over 16,000 nuclei that were removed in our QC pipeline were kept by the author's analysis (**Figure 3-3**), explaining the discrepancy in the number of nuclei after QC. Based on **Figure 3-3**, it is clear the author's approach was ineffective at removing nuclei with high proportions of mitochondrial reads which is indicative of cell death (Ilicic *et al.*, 2016; Heumos *et al.*, 2023) – both excitatory and inhibitory nuclei with higher than 75% of reads from the mitochondria were kept in the final, processed dataset by the authors. We have made the data from our alternative processing approach publicly available (through Synapse:

https://doi.org/10.7303/syn51758062.1) so researchers can utilise this resource free of low quality nuclei.

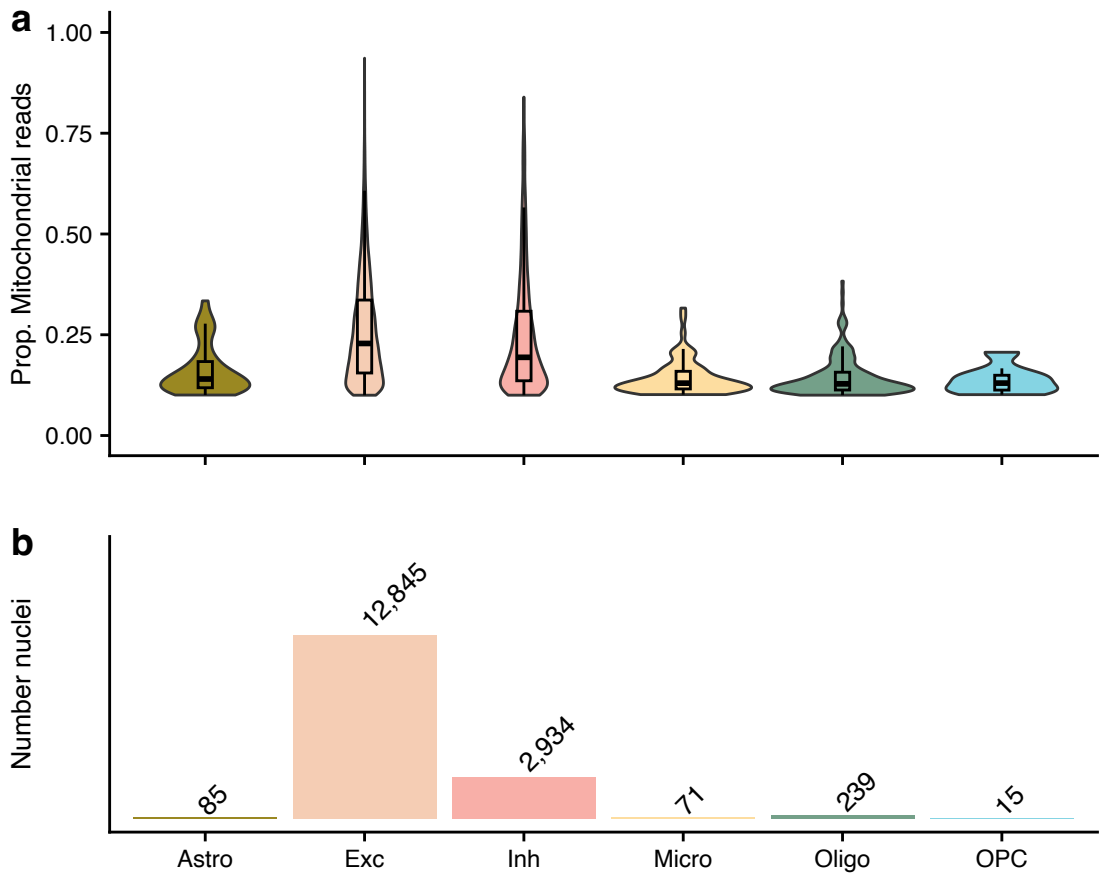


Figure 3-3: The nuclei that were removed from our quality control approach as their proportion of mitochondrial reads were ≥10%, but kept in the authors'.

The nuclei that were removed from our quality control approach, as their proportion of mitochondrial reads were >= 10%, but kept in the authors. (a) shows the proportion of mitochondrial reads across the different cell types. (b) gives the number of removed nuclei which were kept by the authors. The different cell types are astrocytes (Ast), excitatory neurons (Ex), inhibitory neurons (In), microglia (Mic), oligodendrocytes (Oli) and oligodendrocyte precursor cells (Opc).

Our second question of Mathys *et al*., is based on their differential expression approach. The authors conducted a differential expression analysis between the controls and the patients with AD pathology, concentrating on six neuronal and glial cell types; excitatory neurons, inhibitory neurons, astrocytes, microglia, oligodendrocytes and oligodendrocyte precursor cells, derived from the Allen Brain Atlas (Tasic *et al.*, 2018). They performed downstream analysis on their identified DEGs and investigated some of the most compelling genes in more detail. Therefore, all findings put forward by their paper were based upon the validity of their differential expression approach. However, for this approach, the authors conducted a two-part, cell and patient level analysis. The cell-level analysis took each cell as an independent replicate and the results of which were compared for consistency in directionality and rank of their DEGs against their patient level analysis, a Poisson mixed model. The authors identified 1,031 DEGs using this combinatorial approach – DEGs requiring an FDR <0.01 in the cell-level and an FDR<0.05 in the patient level analysis. It is important to note that this cell-level differential expression approach, also known as pseudoreplication, over-estimates the confidence in DEGs due to the statistical dependence between cells from the same patient not being considered (Murphy and Skene, 2022; Squair *et al.*, 2021; Zimmerman, Espeland and Langefeld, 2021; Lazic, 2010). When we inspect all DEGs identified at an FDR of 0.05 from the authors' cell-level analysis, this number increases to 14,274. Pseudobulk DE analysis has recently been proven to give optimal performance compared to both mixed models and pseudoreplication approaches (Murphy and Skene, 2022; Squair *et al.*, 2021; Crowell *et al.*, 2020; Soneson and Robinson, 2018). It aggregates counts to individuals thus accounting for the dependence between an individual's cells.

Here, to compare the effect of the different DE approaches in isolation, we apply a pseudobulk DE approach (Chen, Lun and Smyth, 2016) to the authors original processed data.

We found 26 unique DEGs when considering the six cell types used by the authors (**Table 3-2**). This was 549 times fewer DEGs than that reported originally at an FDR of 0.05. When we compare these DEGs, we can see that the absolute log2 fold change (LFC) of our DEGs is 15 times larger than the authors'; median LFC of 2.34 and 0.16, despite the authors' DEGs having an FDR score 8,000 times smaller; median FDR of 2.89×10^{-7} and 0.002 (**Figure 3-2a-b**). Although we examined a high correlation in the genes' fold change values across our pseudobulk analysis and the author's pseudoreplication analysis (Pearson R of 0.87 for an adjusted p-value of 0.05, **Table 3-4**), the p-values and resulting DEGs vary considerably. The correspondence in fold change values is expected given the approaches are applied to the same dataset whereas the probabilities, which pertain to the likelihood that a gene's expressional changes is related to the case/control differences in AD, importantly do not align. We can show that this stark contrast is just an artefact of the authors taking cells as independent replicates and thus artificially inflating confidence by considering the Pearson correlation between the number of DEGs found and the cell counts (**Figure 3-2c-e**). There is a near perfect, positive correlation between DEG and cell counts for the authors' pseudoreplication analysis (**Figure 3-2c**) and for the 1,031 genes from the authors' combinatorial approach (**Figure 3-2d**) which is not present in our pseudobulk re-analysis (**Figure 3-2e**).

Pseudoreplication adjusted p-value cut-off	Number of genes compared	Pearson correlation
0.01	20152	0.8646269
0.05	23903	0.8708275
0.1	26382	0.8721126
0.25	32117	0.8764692
0.5	42022	0.8751554
1	84467	0.826248

Table 3-4: Strong concordance across different DE's significant genes' fold change values:

Pearson correlation on fold change values between our pseudobulk differential expression analysis and the author's pseudoreplication analysis on all genes found to be significant at different adjusted p-value cut-offs from the author's pseudoreplication analysis.

A further point which questions the authors' DE approach is that they identified the vast majority of DEGs in the more abundant, neuronal cell types (Mathys *et al.*, 2019). However, an increase in the number of cells is not the same as an increase in sample size since these cells are not independent from one another - they come from the same sample. Therefore, an increase in the number of cells should not necessarily result in an increase in the number of DEGs whereas an increase in the number of samples would. This point is the major issue with pseudoreplication approaches which over-estimate confidence when performing differential expression due to the statistical dependence between cells from the same patient not being considered (Squair *et al.*, 2021; Lazic, 2010). In our opinion, it makes more sense to identify the majority of large effect size DEGs in microglia which recent work has established is the primary cell type by which the genetic risk for Alzheimer's disease acts (Skene and Grant, 2016; McQuade and Blurton-Jones, 2019). This is exactly what we found with our pseudobulk DE approach - 96% of all DEGs were in microglia (**Table 3-2**) whereas only 3% of the authors' DEGs were in microglia.

Although it has been proven that pseudoreplication approaches result in false positives by artificially inflating the confidence from non-independent samples, we wanted to investigate the effect of the approach on the authors' dataset. We ran the same cell-level analysis approach – a Wilcoxon rank-sum test and FDR multiple-testing correction, 100 times whilst randomly permuting the patient identifiers (**Figure 3-2f**). We would expect to find minimal DEGs with this approach given the random mixing of case and control patients. However, this pseudoreplication approach consistently found high numbers of DEGs and we observe the same correlation between the number of cells and number of DEGs as with the authors results. We did not observe the same pattern when running the same analysis with pseudobulk differential expression (**Figure 3-2g**). As a result, we conclude that integrating this

pseudoreplication approach with a mixed model like the authors proposed just artificially inflates the test confidence for a random sample of the genes resulting in more false discoveries in cell types with bigger counts.

Up to this point, to compare the effect of the DE approaches in isolation, we analysed the same processed data from the authors as opposed to our reprocessed data. We also performed pseudobulk DE on our reprocessed data and found 16 unique DEGs (**Table 3-3**). It is worth noting that the fold change correlation between our two DE analyses (reprocessed data vs authors processed data) on the identified DEGs is only moderate (Pearson R of 0.57) and is lower than that of the correlation between pseudoreplication and pseudobulk on the same dataset (**Table 3-4**). This highlights the effect that the low quality, high mitochondrial read cells have on DE analysis.

In conclusion, the authors' analysis has been highly influential in the field with numerous studies undertaken based on their results, something we show has uncertain foundations. However, we would like to highlight that the use of pseudoreplication in neuroscience research is not isolated to the author's work, others have used this approach (Fernandes *et al.*, 2020; Wakhloo *et al.*, 2020; Lui *et al.*, 2021) and their results should be similarly scrutinised. Here, we provide our processed count matrix with metadata and also, the DEGs identified using an independently validated, DE approach so that other researchers can use this rich dataset free from spurious nuclei or DEGs. While the number of DEGs found here are significantly lower, much greater confidence can be had that these are AD relevant genes. The low number of DEGs found may also cause concern given the sample size and cost of collection and sequencing of such datasets. However, the increasing number of snRNA-Seq studies being conducted for AD, creates the opportunity to conduct differential meta-analyses to increase power. Further work is required in the field to develop methods to

conduct such analysis, integrating studies and accounting for their the heterogeneity, similar to that which has been done for bulk RNA-Seq (Rau, Marot and Jaffrézic, 2014). Some such approaches have already been made in COVID-19 research which could be leveraged for neurodegenerative disease .

3.4. Data availability

Mathys et al.'s raw snRNA-Seq data (10.7303/syn18485175) and the ROSMAP metadata (10.7303/syn3157322) were downloaded from https://www.synapse.org/ upon acquiring appropriate approval. Our reprocessed version of Mathys et al.'s snRNA-Seq data has also been made publicly available through Synapse: https://doi.org/10.7303/syn51758062.1.

3.5. Code availability

All files from the scFlow run, including quality control statistics are available in the Github repository in the scFlow_files folder: https://github.com/neurogenomics/reanalysis_Mathys_2019. The DE analysis pipeline is available at: https://github.com/neurogenomics/reanalysis_Mathys_2019.

4. MungeSumstats: A Bioconductor package for the standardisation and quality control of many GWAS summary statistics

4.1. Introduction

The work in this chapter is adapted from that set out in our previous publication (Garg *et al.*, 2021; Murphy, Schilder and Skene, 2021). GWAS summary statistics have popularised and accelerated genetic research. However, a lack of standardisation of the file formats used and tools for formatting and quality control have proven problematic when running secondary analysis tools or performing meta-analysis studies. To address these issues, we introduce MungeSumstats a Bioconductor R package for the rapid standardisation and quality control of many GWAS summary statistics. MungeSumstats can handle the most common summary statistic formats as well as GWAS-VCFs to enable the integrative meta-analysis of diverse GWAS. MungeSumstats also offers a comprehensive and tuneable quality control protocol with defaults for common, best-practice approaches. MungeSumstats capitalises on R's familiar interface, is readily accessible through Bioconductor and utilises an intuitive approach, running with a single line of input code. MungeSumstats produces a reformatted, standardised, tabular summary statistic file, VCF or R native data object.

4.2. Heterogeneity in GWAS Formats

To demonstrate the diversity in summary statistics across GWAS, we analysed a public repository of over 200 publicly available GWAS (Gloudemans, 2021). From this, the most common summary statistics were derived, see

Figure 4-1 for the twelve most common file header formats.

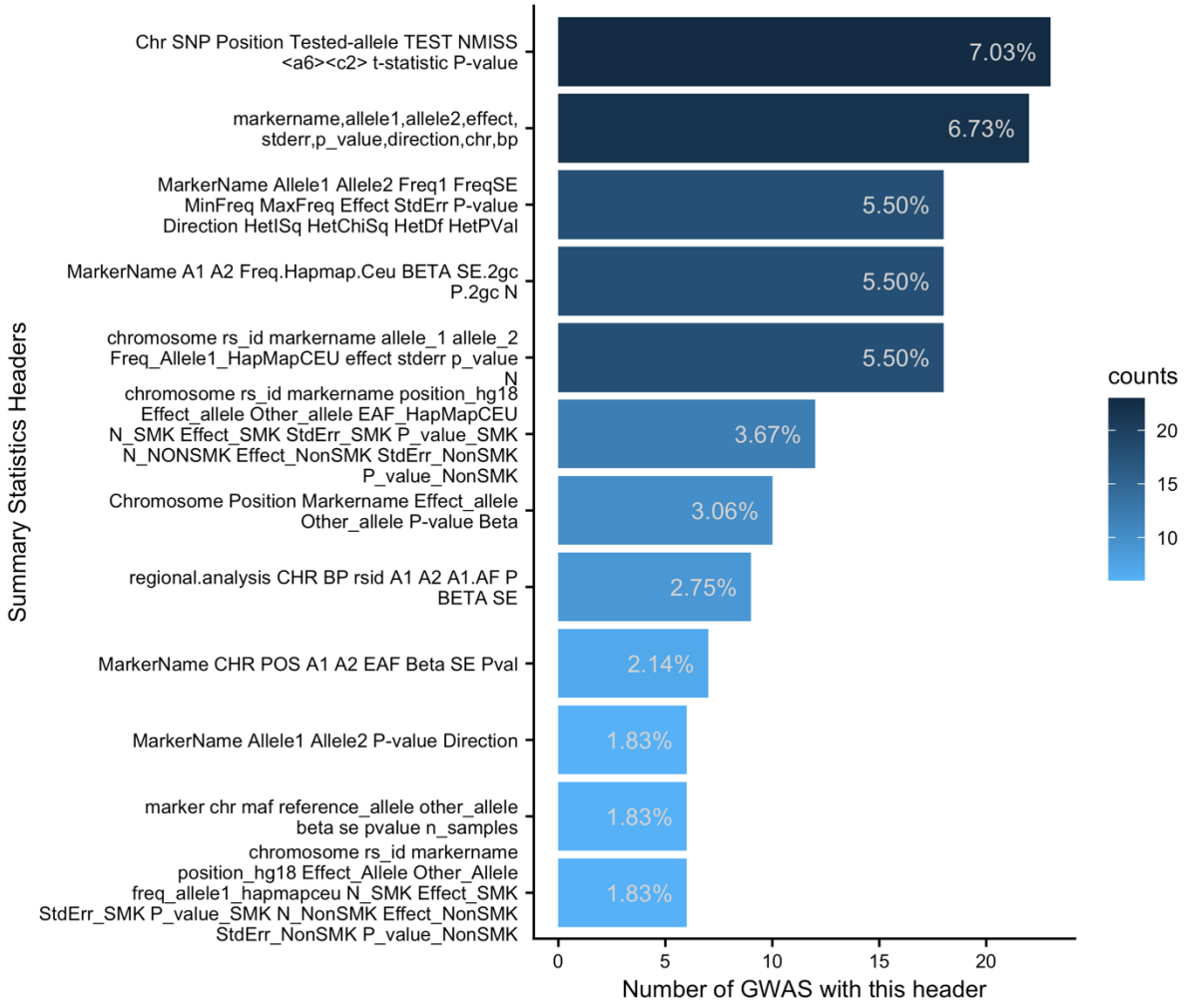


Figure 4-1: Most common summary statistic formats.

Highlights the twelve most common summary statistic formats from a repository of over 200 publicly available GWAS (Gloudemans, 2021). The summary statistics' headers are displayed on the Y-axis and their proportional counts on the X-axis. Note that a GWAS can have more than 1 summary statistics file and `<a6><c2>` is the symbol `¶¬` read into R.

A total of 327 summary statistic files were derived from the analysis which corresponded to 127 unique formats. Thus on average, every 2.5 summary statistic files had a unique format, showing the clear disparity across GWAS. The 12 most common formats, shown in **Figure 4-1**, accounted for approximately 47% all summary statistics. MungeSumstats has been tested on these 12 most common formats and is able to standardise their summary statistics.

4.3. Implementation

MungeSumstats was implemented using the R programming language (v 4.0) and Bioconductor S4 data infrastructure (v 3.13) enabling the full analysis of summary statistics within the R environment. The package removes the need for external software to perform the standardisation and quality control steps.

MungeSumstats' implementation ensures both memory and speed efficiency through the use of R data.table (v1.14.0) (Dowle and Srinivasan, 2021), which can take advantage of multi-core parallelization. Moreover, MungeSumstats benefits from Bioconductor's infrastructure for efficient representation of full genomes and their SNPs, using BSgenome (v1.59.2) SNP reference genomes (Pagès, 2021). Either Ensembl's GRCh37 or GRCh38 are queried along with dbSNP reference SNPs dataset versions 144 or 155 dependent on the build for the particular GWAS and the user's preference. Numerous of MungeSumstats' quality control steps for summary statistics require the use of a reference genome. For example, an allele flipping test is run, see

Table 4-1, to ensure consistent directionality of allelic effect and frequency variables. The effect or alternative allele is always implemented as the second allele (A2), in line with the approach for GWAS-VCF (Lyon *et al.*, 2021). Moreover, MungeSumstats can impute any missing, essential information like SNP ID, base-pair position and effect/non-effect allele.

Using these two infrastructures, MungeSumstats conducts more than 40 checks on the inputted summary statistics file, see

Table 4-1 for a description of their use. Note that some of these checks were added subsequent to MungeSumstats and the related publication's release (Murphy, Schilder and Skene, 2021) and as such, are not reflected in early documentation. MungeSumstats is also written to ensure the ease of addition of further checks so if users have summary statistics

which can't currently be handled in MungeSumstats, these can be incorporated easily in future releases. Finally, MungeSumstats returns a reformatted, tabular summary statistics file, a VCF or an R native data object (data.table, VRanges or GRanges) with standardised columns for the information necessary for downstream analysis.

Step	MungeSumstats Check	Description
1	Check VCF format	If the input file is in variant call format (VCF), if so import
2	Check tab, space or comma delimited	If input is space or comma delimited convert to tab delimited. Can handle .tsv, .txt, .csv, .tsv.gz, .txt.gz, .csv.gz, .tsv.bgz, .txt.bgz, .csv.bgz, .vcf, .vcf.gz, .vcf.bgz files.
3	Check for header name synonyms	If any alternative names are found for SNP, BP, CHR, A1, A2, P, Z, OR, BETA, LOG_ODDS, SIGNED_SUMSTAT, N, N_CAS, N_CON, NSTUDY, INFO or FRQ convert them to a standard name. Robust conversion approach with 176 unique mappings
4	Check for multiple models or traits in GWAS	If multiple, user must specify one to analyse
5	Check for uniformity in SNP ID	Ensure no mix of RS ID, missing 'rs' prefix and/or CHR:BP
6	Check for CHR:BP:A2:A1 all in one column (or variations)	Split into separate columns if found
7	Check for CHR:BP in one column	Split into separate columns if found
8	Check for A1/A2 in one column	Split into separate columns if found
9	Check if CHR and/or BP is missing	If so, infer from the chosen reference genome
10	Check if SNP ID is missing	If so, infer from the chosen reference genome
11	Check if A1 and/or A2 are missing	If so, infer from the chosen reference genome
12	Check that vital columns are present	Check for the necessary columns; SNP, CHR, BP, P, A1, A2. Missing columns will be attempted to be imputed from reference.
13	Check for one signed/effect column	Effect columns Z, OR, BETA, LOG_ODDS, SIGNED_SUMSTAT
14	Check for missing data	If data is missing from any entry, remove the SNP
15	Check for duplicated columns	If there are any remove one
16	Check for p-values lower > 5e-324	These are not recognised in R and cause issues with downstream analysis software like LDSC/MAGMA. User can convert to 0.
17	Check N column	Ensure it is an integer and check if the sample size for a SNP isn't greater than mean multiplied by five times the standard deviation. Removes SNPs that have substantial more samples than the rest.
18	Check SNPs are RS ID's	Checks validity of SNP IDs as RS IDs, other IDs can still be used
19	Check for uniformity of SNP ID format	Any different formats are standardised
21	Check for duplicated rows, based on SNP ID	Duplicates are removed
22	Check for duplicated rows, based on base-pair position	Duplicates are removed
23	Check for SNPs on reference genome	Correct any missing from reference genome using BP and CHR
24	Check INFO score	Remove SNPs with imputation score less than 0.9 or user-specificed
25	Check FRQ value	Ensure all SNPs have frq scores above a user set threshold. Check frq refers to the minor allele frequency (MAF) not major
26	Check SNPs' standard error (SE)	Ensure SNP's SE values are positive, remove SNP if not
27	Check SNPs' beta	Check that SNPs' effect columns (like BETA) aren't equal to 0 and option to and impute beta if missing from other effect columns
28	Infer direction of effect columns	Since allele columns (A1/A2 and synonyms) are often ambiguously named but MungeSumstats assumes A2 is the minor/effect allele. Infer if the chosen A1/A2 columns correspond to this interpretation and relabel if not. This infer is based on three checks (which run in this order): 1. Check if ambiguous naming conventions are used (i.e. allele 1 and 2 or equivalent). If not exit, otherwise continue to next checks. 2. Look for effect column/frequency column where the A1/A2 explicitly mentioned, if found then the direction is known and should update A1/A2 naming so A2 is the effect column.

		3. If not found in 2, a final check should be against the reference genome, whichever of A1 and A2 has more matches with the reference genome should be taken as ***not*** the effect allele.
29	Check for strand-ambiguous SNPs	Remove strand-ambiguous SNPs if found
30	Check for non-biallelic SNPs	Infer from chosen reference genome and remove any if found
31	Check for allele flipping	The effect/alternative/minor allele is assumed to be A2. The allele flipping function checks A1 against a reference genome. For a given SNP, if A1 doesn't match the reference genome sequence (i.e. it is the alternative allele, not the reference allele for example), A1 and A2 along with the effect and frequency columns are flipped, creating consistent directionality of allelic effects across GWAS.
32	Check for SNPs on chromosome X, Y, and mitochondrial SNPs (MT)	If any are found these are removed.
33	Check for z-score (Z)	Ensure present and option impute if missing from other effect columns
34	Check for sample size (N)	Check for presence of N and impute if missing
35	Set CHR style	Convert style of CHR values to one of NCBI, UCSC, dbSNP or Ensembl
36	Check Indels	Ensure handling of Indels and exclusion from certain checks and optionally remove these.
37	Check on dbSNP version	Choose dbSNP reference dataset version to use for other checks and optionally remove SNPs missing from dbSNP reference set
38	Check output format is LDSC ready	Standardised file can be passed to LDSC without pre-processing
39	Check output format is IEU OpenGWAS ready	Standardised file will match IEU OpenGWAS format
40	Check effect column values	Ensure effect columns (like BETA) aren’t equal to 0
41	Check Standard Error	Ensure standard error (SE) is positive
42	Check dropped and imputed values	Return indicators of the imputed values for a SNP and return the SNPs and the reason for exclusion because of quality control.
43	Check build and liftover	Check build and perform liftover to desired reference genome if necessary

Table 4-1: Implemented quality control and standardisation checks.

Lists the quality control and standardisation checks conducted. Most checks are optional and can be set by the user. Here CHR is chromosome, BP is Base-pair position, A1 is the non-effect allele, A2 is the effect allele, N is the sample size, INFO is imputation information score, FRQ is the minor allele frequency (MAF) of the SNP, SNP ID is the single nucleotide polymorphism reference ID, N is the sample size, P is the unadjusted p-value, Z is z-score, OR is odds ratio, LOG_ODDS is the log odds ratio, BETA is the effect size estimate relative to the alternative allele and SIGNED_SUMSTAT is the directional effect size estimate for the summary statistics.

4.4. Usage

Once MungeSumstats is installed, usage involves a single line of code or one function call (*format_sumstats*) with the path to the summary statistics file of interest. Then, the path to the reformatted, standardised summary statistic file is returned. MungeSumstats also offers adjustable parameters to manage the quality control steps. These include options to adjust the imputation information score (INFO) cut-off threshold, the number of samples (N) outliers cut-off threshold and whether to remove mitochondrial SNPs or SNPs on the X or Y chromosome (and many others, see

Table 4-1). Quality control steps which use a reference genome can also be adjusted such as whether to filter SNPs based on their RS ID's presence on the reference genome, whether to check for allele flipping and whether to remove multi-allelic or strand-ambiguous SNPs. These parameters ensure MungeSumstats can be adjusted to the user's analysis pipelines.

4.5. Discussion

Here, we presented MungeSumstats, a Bioconductor package for the standardisation and quality control of GWAS and other genetic studies summary statistics. This package enables integration of summary statistics of vastly different formats, simplifying meta-analysis and summary statistics use in other secondary research applications. The package also provides an efficient, user-friendly R-native approach, returning a standardised, tabular format file, VCF or R native data object. This ensures that the summary statistics are accessible to the average user. Moreover, MungeSumstats is written to permit future development of additional standardisation steps if users encounter issues with their specific GWAS.

4.6. Data availability

The data to derive the summary statistic formats is open source and collated at:

https://github.com/mikegloudemans/gwas-download.

4.7. Code availability

The analysis deriving the most common summary statistic formats is available at:

https://al-murphy.github.io/SumstatFormats. MungeSumstats is available on Bioconductor (>= v3.13) and can also be found on Github at: https://neurogenomics.github.io/MungeSumstats.

5. Predicting cell type-specific epigenomic profiles accounting for distal genetic effects

5.1. Introduction

The work in this chapter is adapted from that set out in our publication (Alan E. Murphy *et al.*, 2024). As previously detailed in **Section 1.5** but summarised here, understanding genetic variants' effects in a cell type-specific manner is crucial for interpreting GWAS results. However, profiling these effects across the non-coding genome remains challenging due to the scalability limits of experimental methods, assaying a cell's epigenome. Deep learning offers an alternative to experimental methods for predicting the effects of genetic variants, a process referred to as in silico mutagenesis. Several DNNs have been applied to the prediction of variant effects on transcription factor binding, histone modifications, and gene expression (Kelley, Snoek and J. L. Rinn, 2016; Kelley *et al.*, 2018; Keilwagen, Posch and Grau, 2019; Li and Guan, 2021). A key advancement in recent models has been the ability to process larger DNA sequences, allowing for the consideration of distal genomic elements, which improves prediction accuracy. For example, Enformer, a state-of-the-art model, has extended the receptive field to 196,608 base-pairs, using multi-headed attention layers to capture long-range genomic context (Avsec, Agarwal, *et al.*, 2021).

However, these models cannot predict cell type profiles that were not seen in the training phase. It would be beneficial to predict epigenetic profiles and the effect of genetic variants on these in previously unseen cell types where collecting cell type epigenetic profiles is invasive, time-consuming and costly. The necessity of such cell type epigenetic imputation is apparent even with ENCODE's expansive data repository where the cell types present do not capture the breadth of cell types studied in disease research, for example in the brain (ENCODE Project Consortium, 2012). Some past models have attempted to address this by

generalising predictions of epigenetic profiles to other cell types such as Leopard (Li and Guan, 2021), Catchitt (Keilwagen, Posch and Grau, 2019) and Epitome (Morrow *et al.*, 2021). However, these approaches do not consider distal genetic information, instead using local genetic windows, far shorter than that of Enformer, and other local features to the infer the cell type variability. However, this local information may not be enough for a model to fully capture cell type representations.

Here, we propose Enformer Celltyping, a self-attention based neural network model to predict histone mark activity in previously unseen cell types. Our model predicts six histone mark profiles from DNA sequence, using chromatin accessibility information from the cell type of interest and is trained on 104 ENCODE samples, hereafter referred to as cell types despite containing tissue samples, isolated cell types and cell lines, sourced from EpiMap (Boix *et al.*, 2021). We use a custom transfer learning approach on a pre-trained Enformer (Avsec, Agarwal, *et al.*, 2021) model to account for the distal effect of genetic code on these regulatory signals. Enformer Celltyping also embeds, using a method similar to that of Avocado's (Schreiber, Durham, *et al.*, 2020) and taking inspiration from the field of Natural Language Processing (NLP) (Ghannay *et al.*, 2016), a genome-wide and local representation of cell types from chromatin accessibility information. This work makes the model applicable to any cell type of interest where chromatin accessibility data is available. We validate Enformer Celltyping's performance against best-in-class approaches and in a number of differing biological settings and inspect the cell type-specific information the model learns.

Moreover, we apply a new framework to validate the predictive performance of epigenetic deep learning models at genetic variant effect predictions using QTL studies with known associations and accounting for LD (Slatkin, 2008), creating a benchmark for use by the broader field. Researchers hoping to study the epigenome and the effect of genetic variants

on cell or sub-cell types' not captured in repositories like ENCODE, can not use current models or have to use the closest available cell type from current model's outputs as a proxy. Here, we take a step toward addressing this problem by creating Enformer Celltyping, a model which can predict epigenetic signals in previously unseen cell types, using a large receptive field to capture genomic regulatory information. We have made a pre-trained version of Enformer Celltyping available along with a fine-tuneable version of the Enformer model, something which was previously not available, so others may make use of these in their research (code: https://github.com/neurogenomics/EnformerCelltyping and pre-trained model: https://figshare.com/account/home#/projects/159143).

5.2. Methods

5.2.1. Data collection and processing

The training data was derived from GRCh37 DNA sequence and p-value continuous tracks after peak calling in EpiMap (Boix *et al.*, 2021), encompassing data from the ENCODE project (ENCODE Project Consortium, 2012). The p-value was used over fold change down to its greater signal-to-noise ratio. The DNA sequence was one-hot encoded following the same approach as previous models (Zhou and Troyanskaya, 2015; Kelley *et al.*, 2018; Avsec, Agarwal, *et al.*, 2021; Li and Guan, 2021). The aim for the histone mark tracks data collection from EpiMap was to maximise both the number of histone marks the model could predict and also the number of cell types which due to the sparse nature of ENCODE data, results in a trade-off between the two (**Fig. A-12**). Any cell types with assays marked as low quality by EpiMap were removed from the training set. The final training set included ChIP-Seq histone mark data for H3K27ac, H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3 from 104 cell types (**Table A-1**). Given the current preference of ATAC-Seq over DNase-Seq to measure chromatin accessibility (ATAC-Seq has similar sensitivity and specificity to DNase-Seq with three to five times fewer cells (Buenrostro *et al.*, 2013)), we opted to train our model using ATAC-Seq data so it would be more applicable to future users. To avoid losing cell types, imputed chromatin accessibility ATAC-Seq data sourced from EpiMap (Boix *et al.*, 2021), was used for any where there was not an observed track which was the case for the majority (98 out of 104) of cell types. All tracks were converted from 25 base-pair average signals to 128 base-pair averaging to match the resolution of the Enformer (Avsec, Agarwal, *et al.*, 2021). All training data tracks measure the $-\log_{10}$ adjusted p-value from MACS2 peak calling (Zhang *et al.*, 2008), indicating the statistical significance of a genomic position i.e. is the site likely to be a real binding site when compared against a background. The -log10 p-value derived after

MACS 2.0 peak calling was chosen over the fold change measure as it typically has a higher signal-to-noise ratio (Ernst and Kellis, 2015). This is the same reason that ENCODE chose -log10 p-value as the primary signal tracks for analyses (Kundaje *et al.*, 2015). Moreover, -log10 p-values are more interpretable which is useful for downstream analysis of predictions, for example thresholding for peaks for functional regions done with ChromHMM (Ernst and Kellis, 2017). Such thresholding of output can not be easily done with fold change values.

Despite EpiMap's uniform reprocessing, to avoid possible issues like sequencing depth biases, we followed previous approaches of training on the arcsinh-transformed signal (Hoffman *et al.*, 2012; Durham *et al.*, 2018; Schreiber, Durham, *et al.*, 2020):

$$\sinh^{-1} x = \ln(x + \sqrt{1 + x^2})$$

We achieved clear separation across epigenetic marks on our training cell types after this transformation which was not present otherwise (**Fig. A-13**). The average chromatin accessibility and histone mark signals, used for model training, were derived by averaging these 104 cell types' tracks (see **Methods Section 5.2.5**).

Two separate sets of cells were used as test cases for the model; 3 immune cell types sourced from EpiMap (Boix *et al.*, 2021); CD14+ monocytes, CD16+ neutrophils and naive CD4+T cells and Nott *et al.*'s isolated cell types from resected cortical brain tissue (Nott *et al.*, 2019); PU.1+ microglia, NeuN+ neuronal, OLIG2+ oligodendrocyte and NeuN- LHX2+ astrocytes. The Nott *et al*., data was processed to produce consensus peak p-value tracks across samples (different individuals were taken as separate replicates) using the ENCODE ChIP-Seq and ATAC-Seq pipelines (https://github.com/ENCODE-DCC/chip-seq-pipeline2 and https://github.com/ENCODE-DCC/atac-seq-pipeline) with the supplied conda environment and default parameters. Note, this was the same processing that was conducted on the

training data. One exception to the default parameters was subsampling the signals to 30 million reads for the ChIP-Seq and 50 million for the ATAC-Seq to match the processing of the EpiMap data and avoiding performance loss due to sequencing depth issues (Whalen *et al.*, 2022). Identically to the training cell types, the test sets were averaged to 128 base-pair resolution and arcsinh-transformed. However, none of these test cell types or any similar cell types, such as other immune cells, from EpiMap were used in the training set. Moreover, these were not used to derive the average histone mark or chromatin accessibility tracks to avoid data leakage between training and test sets (Whalen *et al.*, 2022).

The histone mark QTL (hQTL) Blueprint phase 2 data (Kundu *et al.*, 2022) was downloaded from the European Genome-Phenome Archive (ID: EGAD00001005199). The six hQTL sets (CD14+ monocytes, CD16+ neutrophils and naive CD4+T cells for histone marks H3K4me1 and H3K27ac) were uniformly processed using MungeSumstats (v1.6.0) (Murphy, Schilder and Skene, 2021), using the format_sumstats function and default parameters with the exception of the check_dups parameter which was set to false to enable quality control and formatting of QTL datasets. To transform the hQTL datasets to be compared with Enformer Celltyping's SNP predictions, all SNP to histone mark peak interactions further than Enformer Celltyping's receptive field, or any trans interactions, were removed along with all SNPs not included in HapMap3 (Reshef *et al.*, 2018). SNPs effects were then aggregated to a single value per SNP using the following transformation, as done in previous work (Reshef *et al.*, 2018; Kelley, 2020; Avsec, Agarwal, *et al.*, 2021):

$$\widehat{\alpha_m} = \frac{1}{\sqrt{|G_m|}} \sum_{k \epsilon G_m} \hat{\alpha}_m^{(k)}$$

Where G_m is the set of all histone mark peaks for which a cis-hQTL test was performed for variant m and $\hat{\alpha}_m^{(k)}$ is the marginal correlation of SNP m and histone mark peak k's signal (i.e. the Beta effect value from the study) (Reshef *et al.*, 2018).

5.2.2. Calculating genetic variant distance from epigenetic effect peak in QTL associations

The distance from the SNP to the associated histone mark peak for the six histone mark hQTL BLUEPRINT Phase 2 study (Kundu *et al.*, 2022) were aggregated to one genetic variant per LD block to avoid double counting. Aggregation was done to keep SNPs with the lowest p-value for each LD block and if multiple with the same, the shortest distance to the effect region (to avoid biases towards distal interactions). The Pickrell LD blocks (Pickrell, 2014) (available from https://bitbucket.org/nygcresearch/ldetect-data) were used for this analysis (**Fig. A-2**). This highlights the need for models which can account for these distal interactions when predicting epigenetic profiles, not only transcriptional profiles.

5.2.3. Training data

Enformer Celltyping was trained on 67,007 GRCh37 genomic regions across chromosomes 1-22 and across the 104 training cell types. The regions were picked by first binning the genome based on the size of the model's predictive window (114,688 base-pairs). These bins were then filtered at a DNA level and also a cell type level. The DNA filters were to exclude bins in ENCODE blacklist regions (ENCODE Project Consortium, 2012) and ensuring subsequent genomic positions up and downstream of the region so the model can utilise the full input window size (196,608 base-pairs). The cell type level filters were to ensure a coverage of histone mark signals > 12.5% in the region (based on a -$\log_{10}$ p-value cut-off of 2 to mark a

peak (Boix *et al.*, 2021) for a mark in a 128 base-pair bin). This was done for each histone mark independently and then all histone mark regions were down-sampled to match that of the histone mark with the lowest number of peaks, resulting in 11,903 positions for each histone mark. This approach ensured equal representation and avoided the model prioritising training on a subset of over-represented marks (Whalen *et al.*, 2022). The filtering resulted in 67,007 genomic regions across all training cell types and corresponds to 14,188 unique DNA positions which were split into a training and validation set by chromosome, a similar number of unique genomic regions to past approaches (Kelley *et al.*, 2018; Avsec, Agarwal, *et al.*, 2021). The forward and reverse complement sequence and small, random sequence shifts up and downstream for each were also generated for each genomic region to be included in the training set. The validation set was created by splitting 20% of the training cell types and the start position for these was randomly permuted around the chosen regions to avoid the model overfitting to the training bins. A list of the training positions and cell types is available for download through figshare (https://figshare.com/account/home#/projects/159143).

For each position in each histone mark, the average and distribution of arcsinh-transformed, $-\log_{10}$ p-values across the training cell types were calculated for the pre-training step (see **Methods Section 5.2.5**). This distribution was created by binning signals for each histone mark and genomic region, based on their $-\log_{10}$ p-value signal into 10 groups at 0.5 intervals from 0 to 4.5. The distributions for each histone mark were calculated before model training, the code for which (and all other data processing and model training/testing) is available in our Github repository (https://github.com/neurogenomics/EnformerCelltyping).

5.2.4. Pre-trained Enformer model

Enformer Celltyping uses a pre-trained version of Enformer (Avsec, Agarwal, *et al.*, 2021), which has had its pre-trained layers removed after the multi-headed attention layers and its weights frozen (**Fig. A-1**). A custom approach was developed to be able to apply transfer learning to Enformer. The Enformer model available from Tensorflow Hub has not been formatted to allow modifying of its layers or unfreezing weights (it doesn't expose a *__call__* method, or any of its internal variables, and lacks a model signature). To circumvent this issue, we first extracted the weight matrices from the available, trained version of Enformer and rebuilt an untrained version of Enformer following the architecture provided in the original study (Avsec, Agarwal, *et al.*, 2021). We next manually mapped layers across the two models and once completed, updated the weights in each layer of the untrained Enformer model to the values of the trained version of Enformer. Once complete, we tested the predictions against the trained version of Enformer to ensure the weights were mapped correctly. From here, we removed the final layers after the multi-headed attention layers and froze the weights of the other layers and incorporated this pre-trained, chopped version of Enformer into our Enformer Celltyping model. This recreated version of Enformer is now fully customisable allowing flexibility in the transfer learning approach such as fine-tuning through Tensorflow API. We have made an automated script to create this version of Enformer available in our Github repository (https://github.com/neurogenomics/EnformerCelltyping) so others can similarly use Enformer in their own problem domains.

5.2.5. Enformer Celltyping architecture

Enformer Celltyping was implemented in Tensorflow v2.4 with Sonnet v2 functionality in the pre-trained Enformer model. Enformer Celltyping was trained with a two-step approach – a pre-training step with a separate training of DNA sequence and chromatin accessibility submodules, and a subsequent cell type-specific training step, combining the output of the two model branches with subsequent layers. The pre-training step improved model performance overall (**Fig. A-4**).

For the pre-training, Enformer Celltyping's architecture is split into two separate submodules (**Fig. A-1a**), both with their own optimiser and loss function. The first "DNA" submodule uses the DNA sequence to predict the average and distribution of each histone mark. The second "celltyping" submodule predicts the difference between the average histone mark signal and the cell type-specific signal. These are explained in more detail below; how the data is constructed is discussed in Methods section Training data. The output blocks of the first DNA submodule is shown in **Fig A-1a** as "Output avg" and "Output distribution" while the second celltyping submodule's pretraining output is shown as "Output delta". It is the first "DNA" submodule which uses the pre-trained, chopped, frozen version of Enformer (Avsec, Agarwal, *et al.*, 2021), through a transfer learning approach (discussed above in **Methods Section 5.2.4**).

The first is a "DNA" submodule takes one-hot encoded DNA sequence of length 196,608 base-pairs (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]) as its input, and processes it through four main parts: (1) seven convolutional blocks with pooling, (2) eleven transformer blocks, and (3) a cropping layer and (4) a further dense layer and convolutional block and output layer. During both training steps, the layers corresponding to the first three parts of

the architecture (the pre-trained Enformer layers) were frozen and only the final part's weights were updated through backpropagation. The DNA submodule outputs an average cell type prediction and distribution across the training cell types for each histone mark at the genomic region. The average cell type -log10 p-value score for each histone mark, used to train the model, was derived from the 104 EpiMap, training cell types (see **Methods Section 5.2.1**). The distribution data corresponds to the proportion of training cells whose histone mark, arcsinh-transformed $-\log_{10}$ p-value fall into each of 10 bins between 0 and >=4.5 at 0.5 intervals for a given genomic location. For example, if for H3K27ac signal at a genomic location, all cell types have no signal i.e. $-\log_{10}$ p-value = 0, then the distribution to predict would be [**1**,0,0,0,0,0,0,0,0,0]. Whereas, if two cell types of the 104 had $-\log_{10}$ p-value = 5.5, then the distribution would be [**.98**,0,0,0,0,0,0,0,0,**0.019**]. Including the distribution of histone mark signals as well as the average in the DNA module prediction for the pre-training step helps capture variability of regions across cell types and improved the overall performance of Enformer Celltyping (**Fig. A-5**).

The second "celltyping" submodule takes two inputs; a local and global representation of ATAC-Seq for the cell type. The local chromatin accessibility is the ATAC-Seq from the 199,936 base-pairs averaged at 128 base-pair resolution of the same genomic location as the DNA (1,562 positions). This local chromatin accessibility signal is first pre-processed by calculating the difference between the cell type-specific signal and the average chromatin accessibility signal of the training before passing to the model. Whereas, the global chromatin accessibility corresponds to the chromatin accessibility for 3,000 base-pairs around the transcriptional start site of 1,216 marker genes, averaged at 250 base-pair resolution (3.648 million base-pairs, input size of 14,592 positions). These marker genes were taken from a database of all known human marker genes derived from over 1,000 single-cell RNA-Seq experiments for any

cell type, collated by PangloaDB (Franzén, Gan and Björkegren, 2019). Human marker genes were excluded if they fell outside of chromosomes 1-22 or in ENCODE blacklist regions (ENCODE Project Consortium, 2012), leaving 1,216 genes. The list of PangloaDB genes and code to derive the global chromatin accessibility is available at: https://github.com/neurogenomics/EnformerCelltyping. We tested the performance of using each of the global and local chromatin accessibility signals separately and combined for a subset of 300 training steps, highlighting the benefit of incorporating both (**Fig. A-3**). Note that the global chromatin accessibility signal will be the same for a cell type regardless of the genomic region it is predicting in. Similar to Avocado's approach to capture differing information about the epigenetic landscape (Schreiber, Durham, *et al.*, 2020) and more broadly, other approaches in the field of NLP (Odena *et al.*, 2021), the global and local chromatin accessibility information were embedded at different resolutions. These embeddings are then flattened, concatenated and passed through two dense layers before being passed to a separate output head per histone mark. Note that this flattening step after the embedding layers does not affect the positional information of the chromatin accessibility information. The celltyping submodule outputs the difference between the average histone mark prediction and the cell type of interest for the genomic region.

For the full training step, Enformer Celltyping combines the three outputs; "Output avg", "Output distribution" and "Output delta", of the "DNA" and "celltyping" submodules with subsequent convolutional and dense layers (**Fig. A-1b**) to predict the cell type-specific histone mark signal for six histone mark (H3K27ac, H3K4me1, H3K4me3, H3K9me3, H3K27me3 and H3K36me3), in a given genomic region. Enformer Celltyping's output has six channels, one for each histone mark and 896 positions, corresponding to the centre 114,688 base-pairs of the input window, aggregated into 128 base-pair resolution bins. Only the centre 114,688 base-

pairs are predicted on to avoid predicting on the edge positions which do not have as many neighbouring positions upon which to make a prediction (Avsec, Agarwal, *et al.*, 2021). Enformer Celltyping is trained to predict arcsinh-transformed -$\log_{10}$ p-value, however, this can be converted back to a raw -$\log_{10}$ p-value which Enformer Celltyping outputs by default.

5.2.6. Enformer Celltyping Training

Enformer Celltyping was trained using Tensorflow API for 1,000 steps of a batch size of 128 cell type-specific genomic positions (128,000 positions) as a pre-training stage taking approximately 4 days (see **Methods Section 5.2.5**). The combined architecture, full training stage was run for 4 full epochs (6,940 steps with a batch size of 128 cell type-specific genomic positions equating to 888,320 positions) stopping when the model started to overfit on the validation dataset, taking approximately 1.5 days on a Nvidia A100, 80GB RAM GPU. A learning rate of 0.0002 was used for pre-training and 0.005 for the full training stage and the ADAM optimiser (Kingma and Ba, 2017) was used for both. The low initial learning rate was chosen to allow for learning rate warmup and avoid early overfitting (Gotmare *et al.*, 2018) after which we matched the learning rate to similar models (Kelley, Snoek and J. Rinn, 2016; Avsec, Agarwal, *et al.*, 2021). In the pre-training stage, the loss function differed across the submodules; a Poisson negative log-likelihood loss function was used for the average signal prediction (the same as past approaches (Kelley *et al.*, 2018; Avsec, Agarwal, *et al.*, 2021)) and cross-entropy loss was used for the distribution in the DNA submodule whereas a mean squared error (MSE) loss function was used for the celltyping submodule following other epigenetic embedding approaches (Schreiber, Bilmes and Noble, 2020; Schreiber, Durham, *et al.*, 2020), given the possibility of negative values. The Poisson negative log-likelihood loss

function was used for the full training stage. Since Enformer Celltyping freezes the weights relating to the chopped, pre-trained Enformer model (see **Methods Section 5.2.4**), the output from these layers for each genomic position was computed just once for each genomic location and then load as needed for each epoch. As a result of this, training was substantially quicker than for Enformer in the original study (Avsec, Agarwal, *et al.*, 2021) despite our model seeing more positions for multiple epochs compared to Enformer's single epoch approach, totalling 132 GPU hours for pre and full training stages versus 5,376 GPU hours.

5.2.7. Testing receptive field

To test the receptive field of both Enformer and Enformer Celltyping, we evaluated the effect of random genetic variants on histone mark predictions across increasing distances (from 0 to 98,304 base-pairs up and downstream). Specifically, we measured the average change in prediction across reference and alternative sequence on the centre four positions (512 base-pairs) of the output signal. DNA sequences were simulated by randomly sampling base-pairs. Next, a single random genetic variant per simulated DNA sequence was inserted and the effect on prediction measured. The location of the genetic variants were spread across 1,000 evenly spaced locations across the 196,608 base-pair input. This random sequence and random variant simulation was repeated 100 times for each genetic variant location (i.e. 100,000 simulations overall) to capture the average effect of any variant. All six output channels of Enformer Celltyping were averaged whereas only the output channels from Enformer corresponding to histone mark signals were considered. The same random iteration seed was used for both models so that the same DNA sequences and genetic variants would

be simulated. The average change across reference and alternative was shown at the position of the genetic variant (**Figure 5-1b**).

Similarly, we tested the effect of changes in the local chromatin accessibility on Enformer Celltyping by measuring the average change in prediction across reference and alternative chromatin accessibility signal on the centre four positions (512 base-pairs) of the output signal. We replaced the cell type-specific local chromatin accessibility signal of 640 base-pairs with the average signal at increasing distances up to the full receptive field of the model (**Figure 5-1c**). The code for testing the DNA and chromatin accessibility receptive fields are available at: https://github.com/neurogenomics/EnformerCelltyping.

5.2.8. Visualising cell type embedding

A side-effect of embedding the global chromatin accessibility information in Enformer Celltyping is that these latent representations of cell types is continuously improved by the model during backpropagation. We can then visualise these latent representations with a 2D projection. For all 104 training cell types in our training set we plotted the global chromatin accessibility embeddings using UMAP (McInnes, Healy and Melville, 2020) with default parameters (**Figure 5-2a**).

5.2.9. Benchmarking

Enformer Celltyping was benchmarked against Epitome (Morrow *et al.*, 2021) across the six histone marks for three immune cell types from EpiMap (Boix *et al.*, 2021) (CD14+ monocytes, CD16+ neutrophils and naive CD4+T cells) (**Figure 5-3a**). Epitome predicts transcription factor binding and histone marks for a cell type using chromatin accessibility (DNase or ATAC-Seq)

data and was shown to have best-in-class predictive performance at cell type-specific histone mark predictions (Morrow *et al.*, 2021).

The weights for a pre-trained version of Epitome, trained on Panc1, PC-9, OCI-LY7, MCF-7, Karpas-422, IMR-90, HepG2, HeLa-S3, HCT116, H9, H1, GM23338, GM23248, GM12878, A549 cells from ENCODE(ENCODE Project Consortium, 2012), were supplied by the authors and which we have made available through figshare (https://figshare.com/account/home#/projects/159143). Both Epitome and Enformer Celltyping were used to make genome-wide predictions of the same six histone marks for the three immune cell types. Epitome predicts in 200 base-pair bins in a classification setting – whether there is a peak in a 200 base-pair interval or not whereas Enformer Celltyping has been developed to predict signal values directly ($-\log_{10}$ p-values) at 128 base-pair resolution. To compare the methods, Enformer Celltyping's predictions were converted into classification scores using a $-\log_{10}$ p-value cut-off of 2, similar to previous work (Boix *et al.*, 2021). The predictions for the two models were averaged to the lowest common multiple between the two (3,200 base-pairs). Both of these differences led to a blurring of signal and suboptimal predictions for both approaches but were necessary for comparison. As used in previous work (Schreiber, Durham, *et al.*, 2020), the performance was evaluated using a balanced measure of performance, accounting for the disparity in the number of regions with and without peaks, the average-precision (AP) for each cell type – histone mark combination. AP is a single score that represents the precision-recall curve based on the rank of predictions using the mean of precisions for each threshold, with the increase in recall from the previous threshold considered as the weight.

We did not benchmark Enformer Celltyping against Enformer as they perform different, non-comparable tasks - Enformer predicts in held-out chromosomes and so does not have to extrapolate predictions to previously unseen cell types whereas Enformer Celltyping does on whole, held-out cell types. However, we did test both models for genetic variant effect predictions (see **Methods Section 5.2.12**).

5.2.10. Highly active enhancer (super-enhancer) prediction in cancer cell lines

To determine how informative Enformer Celltyping's embeddings are given the strong performance at enhancer regions (**Figure 5-4c**), we trained a random forest classifier to distinguish highly active enhancers, sometimes referred to as super-enhancers, from normal enhancers in cancer cell lines, using four-fold cross-validation, holding out all data from one cell line in each fold. We used the scikit-learn 1.0.2 implementation (Pedregosa *et al.*, 2011), with default parameters, expect setting the maximum features considered per decision tree split to $\log_2$ of the total number of features due to the large input size. These highly active enhancers were defined as enhancer regions with a multitude of transcription factor binding sites, often identified by occupation of high levels of master transcription factors and coactivators (Hnisz *et al.*, 2013; Lovén *et al.*, 2013) and were obtained from the dbSuper database (Khan and Zhang, 2016), for the cell lines MCF-7 (https://asntech.org/dbsuper/data/bed/hg19/MCF-7.bed), K562 (https://asntech.org/dbsuper/data/bed/hg19/K562.bed), HEPG2 (https://asntech.org/dbsuper/data/bed/hg19/HepG2.bed) and GM12878 (https://asntech.org/dbsuper/data/bed/hg19/GM12878.bed). While the identification of all enhancer regions was from their ENCODE H3K27ac bed files MCF-7

(https://www.encodeproject.org/files/ENCFF024TEH/@@download/ENCFF024TEH.bed.gz), K562 (https://www.encodeproject.org/files/ENCFF044JNJ/@@download/ENCFF044JNJ.bed.gz), HEPG2 (https://www.encodeproject.org/files/ENCFF392KDI/@@download/ENCFF392KDI.bed.gz) and GM12878 (https://www.encodeproject.org/files/ENCFF367KIF/@@download/ENCFF367KIF.bed.gz). We down-sampled the number of enhancer regions (negative samples) to match the number of highly active enhancers (positive samples) in the cell line and visualised performance on the held-out test cell line with ROC curves.

5.2.11. Off-centre correlation analysis

Enformer Celltyping makes histone mark predictions for 128 base-pair bins in a 114,688 base-pairs wide predictive window. We tested whether sliding the genomic location, i.e. the DNA and local chromatin accessibility information, would result in a change in the histone mark prediction at a given genomic location. We tested this by randomly sampling 1,000 genomic positions from our three immune cell types (CD14+ monocytes, CD16+ neutrophils and naive CD4+T cells) which have SNP information in the Blueprint phase 2 hQTL studies data (Kundu *et al.*, 2022). For each position, we measured the correlation in matching genomic regions between the centred prediction and two off-centre predictions, one up and one down-stream. The positions up and down-stream where chosen to capture the full input window as output (**Fig. A-10**), to capture as much overlap as possible. Finally we reported the range of these correlations for both genomic data with and without a SNP to ensure genetic variants

did not cause a bias in the results (**Figure 5-6c**). The code for testing the off-centre correlation analysis is available at: https://github.com/neurogenomics/EnformerCelltyping.

5.2.12. Histone mark QTL SLDP

Although multiple approaches have now been developed for *in silico* mutagenesis, there is little consensus on how best to benchmark performance. Experimental datasets including CRISPR, MPRAs and xQTL mapping studies have all been proposed (Avsec, Agarwal, *et al.*, 2021; Karollus, Mauermeier and Gagneur, 2023; Sasse, Ng, A. E. Spiro, *et al.*, 2023). However, currently only xQTL mapping studies can be carried out *in vivo* on large sample sizes. Thousands of genetic variants are tested in one study based on the natural variation in the population. However, associations detected in xQTL mapping studies are confounded by LD, where alleles are co-inherited based on their physical proximity (Slatkin, 2008), making it difficult to identify the causal SNPs (Reshef *et al.*, 2018). Therefore, LD needs to be taken into account when comparing these datasets against model predictions.

For our benchmarking framework, we predicted the effect of each genetic variant from the six hQTL datasets (Kundu *et al.*, 2022) with Enformer Celltyping by predicting the histone mark signal for both the reference and the alternative sequence, calculating their difference and summing this effect across the full genomic window to get a single, signed score for each SNP. Despite our model predicting the measured change in histone mark binding across the full receptive field, aggregating the effect to a single value is necessary to match the format of the hQTL data and for SLDP. We predicted in six cell types, three matching immune cells and three unrelated cell type from EpiMap: Pancreas, Leg Muscle and Stomach tissue, to act as negative controls. To obtain a prediction across the full input window size, and thus utilising

the full receptive field of Enformer Celltyping, we predicted with the SNP first centred in the input window and then performed two further predictions with the SNP slid off-centre (**Fig. A-10**) and appended these three output windows before calculating the score. Finally, we averaged scores computed using the forward and reverse complement sequence and with small, random sequence shifts up and downstream. This meant for each SNP, we predicted (2x reference and alt, 3x full receptive field, 4 x reverse complement and random shift) 24 predictions. We predicted for 867,568 SNPs across the six hQTL datasets for the six cell types, resulting in approximately 125 million predictions.

We utilised signed linkage disequilibrium profile regression (SLDP) (Reshef *et al.*, 2018) to measure the statistical concordance between the signed variant effects (our model's predictions) and the genome-wide association study's marginal correlations (the aggregated hQTL SNP values, see Methods section Data collection and processing). SLDP uses generalized least-squares regression to measure the agreement between these, iteratively inverting the direction of the signed variant effect measures along with their neighbouring entries in blocks to derive a null distribution. The method accounts for population LD so does not require any fine-mapping strategies. The measured agreement defines how important the variants are to the phenotype's heritability (Reshef *et al.*, 2018). We ran SLDP for all combinations of our cell type predictions and histone mark with each hQTL set (6x cell types, 6x histone mark predictions and 6x hQTL sets, resulting in 216 tests) and report the results in **Figure 5-6a**. All code to run the genetic effect predictions and SLDP analysis (with a conda environment) is available at: https://github.com/neurogenomics/EnformerCelltyping.

5.2.13. Global chromatin accessibility signal – cell type-specific motif enrichment

We inspected the cell type-specific motifs based on the global chromatin accessibility signal to infer what was driving the cell type-specific signal the model captures. We first predicted genome-wide tracks for histone marks acting outside of the transcriptional start site (TSS) or gene body (H3K27ac, H3K4me1 and H3K9me3), as these would be more likely to harbour cell type-specific transcription factors, using Enformer Celltyping and identified peaks in the data with -log10 p-value score >2, at 1024 base pair resolution. For each peak, the influence of the global signal was approximated by calculating the partial derivatives of the model with respect to the input, i.e. the gradient on the input (Hechtlinger, 2016). The results for each peak was ordered by absolute value and the top 10% of peaks reliant on the global signal were identified. The DNA at these positions were run through Homer's (Heinz *et al.*, 2010) known motif analysis to search for motif enrichment using default settings. The Homer motif analysis is available at: https://github.com/neurogenomics/EnformerCelltyping.

The transcription factors' genes relating to the resulting motifs, using an FDR cut-off of 0.05 for the Homer known motif results, were tested for cell type specificity using EWCE (Skene and Grant, 2016) bootstrapping (repeated 10,000 times with an FDR adjusted p value threshold of 0.05). All known motifs from Homer were used as the background set (approximately 400 transcription factor genes) after mapping non-human genes to human based on one-to-one orthologs and any motifs overlapping all 7 cell types were removed to get the cell type-specific motifs (4 in total). The scRNA-Seq reference dataset used to determine cell type enrichment was the Descartes human whole body, fetal dataset containing approximately 377,000 cells and 77 distinct cell types (Cao *et al.*, 2020). This approach was repeated for all cell types of interest and FDR multiple test correction was implemented to account for the repeated tests (**Figure 5-7**).

5.2.14. Cell type-specific genetic enrichment for complex traits

To test for genetic enrichment of complex traits, Enformer Celltyping's H3K27ac predictions for Nott *et al.*'s (Nott *et al.*, 2019) microglia, neurons and oligodendrocytes (continuous signals) were converted to bedGraph and finally to bed files, using a peak cut-off of arcsinh(1) for inclusion in the bed file; this led to better enrichment than more stringent peak cut-offs. To enable a fair comparison, the ATAC-Seq and experimental H3K27ac signals for the same cell types were converted using the same approach.

Summary statistics for GWAS for glial diseases - Alzheimer's (I. E. Jansen *et al.*, 2019; Bellenguez *et al.*, 2022a), neuronal diseases - autism spectrum disorder (Grove *et al.*, 2019), major depressive disorder (Wray *et al.*, 2018) and schizophrenia (Trubetskoy *et al.*, 2022), neuronal traits – cognitive function (Davies *et al.*, 2018) and intelligence (Savage *et al.*, 2018) and immune diseases/traits – neutrophil count (Vuckovic *et al.*, 2020), eczema (Loh *et al.*, 2018) and peptic ulcer disease (Wu *et al.*, 2021) were downloaded and uniformly processed with MungeSumstats v1.11.3 (Murphy, Schilder and Skene, 2021); default settings, converting build to hg19 where necessary and saving in 'LDSC' format.

We applied stratified LD score regression (s-LDSC) (Finucane *et al.*, 2015) v1.0.1 (https://github.com/bulik/ldsc) to test for disease enrichment. Specifically, annotation files for each of the signals for each cell type of genomic loci were first created with Phase 3 of the 1000 genomes reference. Followed by the generation of LD scores with a window size of 1 centiMorgan (cM) i.e. approximately 1 million base pairs, filtering to HapMap3 SNPs to match the baseline model. Finally, the enrichment analysis was run for the GWAS summary statistics across the different cell types and signals as well as those in the baseline model whilst excluding the major histocompatibility complex (MHC) (due to the known difficulties

predicting LD in this region) (Finucane *et al.*, 2015). All code to process the data and run the s-LDSC analysis (with conda environments) is available at: https://github.com/neurogenomics/EnformerCelltyping.

5.3. Results

Enformer Celltyping uses Enformer (Avsec, Agarwal, *et al.*, 2021) through a transfer learning approach to expand on its capabilities, predicting in previously unseen cell types (that were not part of the training set). Enformer Celltyping benefits from Enformer's attention layers which produce its large receptive field (approximately 100k base-pairs), accounting for distal DNA regulatory elements in its predictions (**Figure 5-1a** and **Fig. A-1**). We measure the sensitivity of this receptive field experimentally with simulated DNA sequences (**Figure 5-1b**). The need for such large receptive field of distal regulatory elements can be inferred by observing the distance between SNPs and the peaks they regulate in histone-QTL studies (**Fig. A-2**): while these studies are confounded by LD, the prevalence of long-range interactions does not appear to be consistent with a model whereby most histone QTLs act via local motif disruption. Enformer Celltyping can predict in new cell types, imputing their epigenetic signal, by embedding global and local chromatin accessibility (ATAC-Seq) signals for the cell type of interest. We experimentally show how the model is also sensitive to permutations of the local chromatin accessibility with simulated signals, following a similar trend to the DNA sequence (**Figure 5-1c**). Using this combined approach, Enformer Celltyping shows best-in-class performance at predicting six histone mark signals for any genomic region of interest (**Figure 5-1d**).

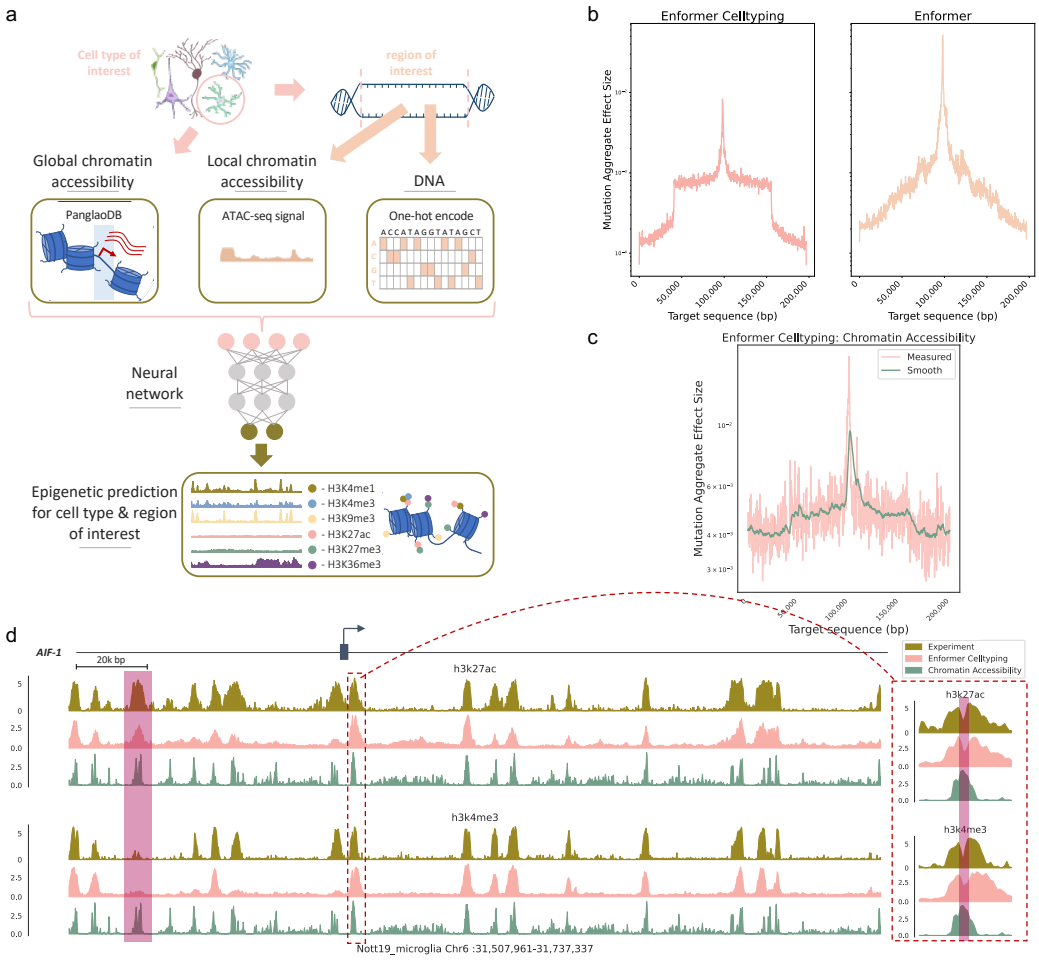


Figure 5-1 Enformer Celltyping predicts cell type-specific histone marks accounting for distal DNA regulation.

(a) Enformer Celltyping uses transformer modules on DNA sequence and embeds both local and genome-wide chromatin accessibility signal to predict histone mark signals in cell types outside of the training set. Image created with the help of SciDraw (https://scidraw.io/) (Chilton, 2020a, 2020b, 2020c; Galliano, 2020) (b) The experimentally derived receptive field for both Enformer and Enformer Celltyping. Random permutations are made at increasing distances from the centre point (100,000 on the x-axis) and the effective change in the prediction on that central point is measured (y-axis). (c) Experimentally derived local chromatin accessibility receptive field for Enformer Celltyping following a similar permutation

approach to (b). (d) An example of the observed (Experiment), predicted (Enformer Celltyping) and the chromatin accessibility signal at the transcriptional start site (blue arrow) of the Allograft inflammatory factor 1 (*AIF-1)* marker gene in microglia. The y-axis denotes the $-\log_{10}$ p-value, indicating the statistical significance of a protein binding at a genomic position. Highlighted in red is a H3K27ac-specific peak not found in H3K4me3 highlighting that the model uses more than chromatin accessibility signal to identify peaks. Also highlighted in the red dashed box is a region at the transcriptional start site of AIF-1 showing how the histone peaks (actual and predicted) occur around the ATAC peaks even at the 128 base-pair resolution of the model.

Enformer Celltyping, when predicting the histone mark signals, embeds a global as well as the local representation of the inputted chromatin accessibility signal for the cell type of interest. This global cell type embedding shows co-localisation of similar cell types (**Figure 5-2a**). Using this combination of global and local chromatin accessibility signal improved the model's predictive performance (**Fig. A-3**), though we note that the local chromatin signal had a larger effect on model performance than the global. Enformer Celltyping is capable of making accurate predictions in these previously unseen cell types using chromatin accessibility data and DNA, such as that for microglia at the transcriptional start site of the Allograft inflammatory factor 1 (*AIF-1)* marker gene where a microglia-specific H3K27ac peak was distinct from the predictions for the other brain cell types and replicated in the experimental data (**Figure 5-2b**). Enformer Celltyping's cell type-specific predictive capabilities was notably aided by a pre-training step (**Fig. A-4**). This step included training the DNA and chromatin accessibility inputs as separate submodules; the DNA module predicting the average histone mark signal and distribution of signal across the training set by incorporating Enformer's pre-trained layers and the chromatin accessibility module predicting the difference of the average and cell type of interest. The pre-training step sensibly initialised the weights of the chromatin accessibility layers before combining with the DNA layers which contained the Enformer architecture for the full training step (see the Methods section Enformer Celltyping architecture).

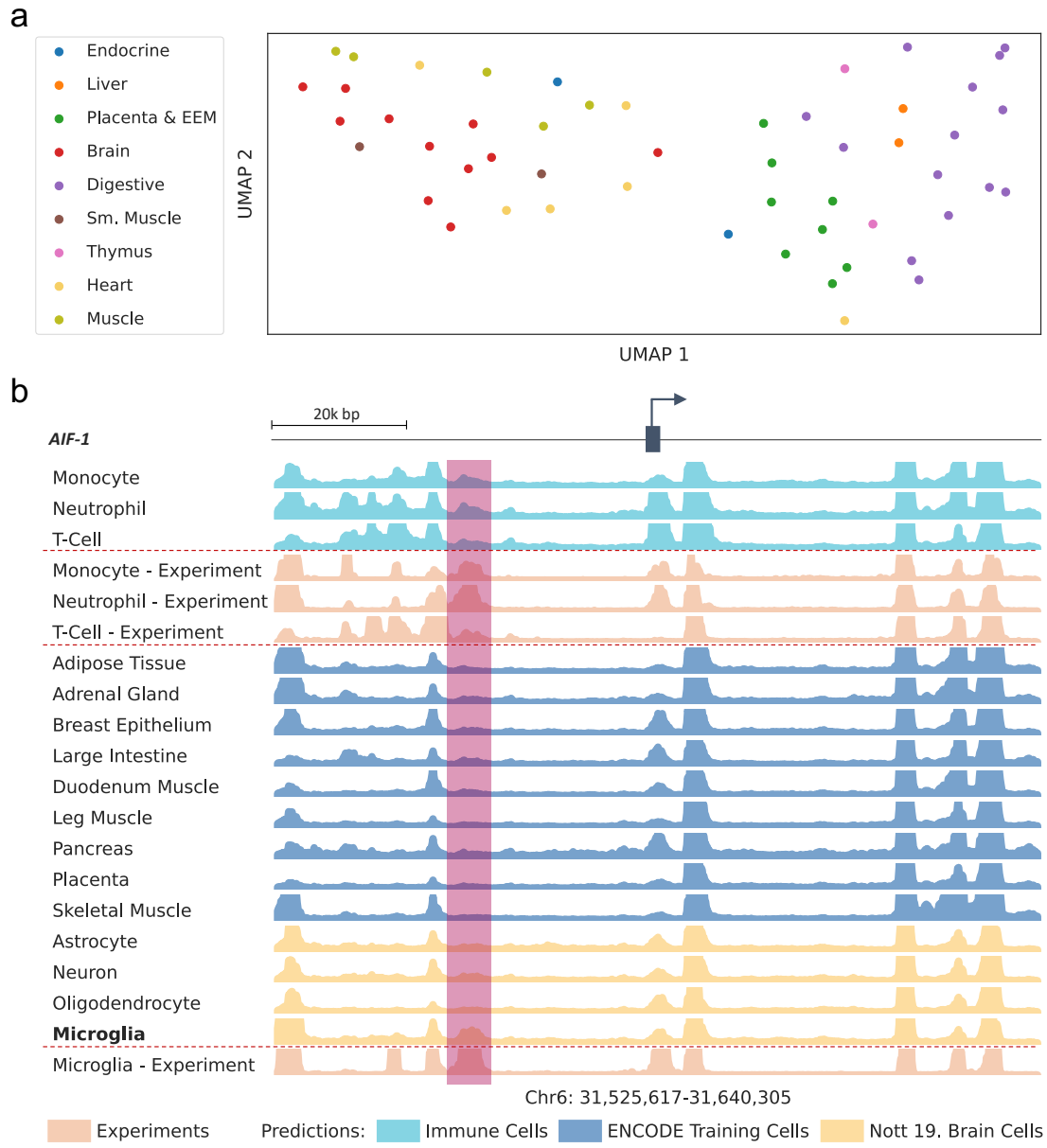


Figure 5-2 Enformer Celltyping predicts cell type-specific histone marks signals.

(a) A UMAP projection of Enformer Celltyping's global chromatin accessibility embeddings. Cell types from EpiMap are coloured by their broad tissue type. (b) Enformer Celltyping's H3K27ac predictions (turquoise, navy and gold) and the experimental assay signals (apricot) at the transcriptional start site of *AIF-1* gene across varying cell types. *AIF-1* is the canonical

marker of microglia. Highlighted in red is a microglia-specific peak not found in the other brain cell types but present in immune cells and which is validated in the experimental results.

We benchmarked Enformer Celltyping against Epitome (Morrow *et al.*, 2021), the current best-in-class approach for predicting histone mark signals in previously unseen cell types, in a genome-wide prediction task on three immune cell types (**Figure 5-3a**). Enformer Celltyping (and Epitome) generalise predictions of epigenetic profiles in cell types which were not seen during training by using their chromatin accessibility information as input. To avoid the performance of our predictions being inflated for this and subsequent benchmarks, we tested three immune cell types (Monocytes, Neutrophils and T-cells) for which data was generated by ENCODE, but which we held out of the training dataset for Enformer Celltyping. For a fair comparison, we binarised the prediction of our model to match that of Epitome (see **Methods Section 5.2.9**). We report the ROC curves and the AP score for both models. We evaluated performance based on the AP for each cell type and histone mark combination to get a balanced measure of the models' performances. Enformer Celltyping out-performed Epitome across all cell types and histone mark predictions.

Due to differences in predictive resolution between the two models, large genomic intervals (3,200 base-pairs) had to be used for the classification task so performance could be compared. This would result in suboptimal performance for both models, blurring their signals. To get a clearer picture of how well Enformer Celltyping predicts, following the approach of previous work (Schreiber, Bilmes and Noble, 2020; Schreiber, Singh, *et al.*, 2020), we benchmarked Enformer Celltyping's performance against the average signal (Avg) for the same histone mark from the training dataset used to train Enformer Celltyping (**Figure 5-3b**). Here we used the same three immune cell types but benchmarked performance based on the MSE between the prediction and the actual $-\log_{10}$ p-value score, genome-wide at 128 base-pair resolution, utilising Enformer Celltyping's regression output. We split performance by each histone mark, avoiding reporting a single performance score, to investigate if the model

had any specific biases. Overall, Enformer Celltyping performed similarly to the average signal. The predictions for H3K27ac and H3K4me3 signals across the three cell types does consistently out-perform the average, with lower MSE, whereas predictive performance on H3K27me3 consistently under-performs.

One explanation for the inconsistent out-performance relative to the average, is that a model can generally perform well genome-wide by predicting the average signal. To explore this further, we note that during pre-training one of the Enformer Celltyping submodules is explicitly trained to predict the average: it's performance is shown in **Figure 5-3b** as "EC Avg". This pre-training step was necessary as Enformer Celltyping uses Enformer as a pre-trained model for the DNA input module. The Enformer-derived layers had sensible weight initialisations whereas the model's chromatin accessibility module weights were randomly initialised. Thus the short pre-training step avoids issues where all predictions were made based on the DNA information and acts as a warm-up step for the chromatin accessibility layers. During pre-training, the DNA input module predicted the average and distribution of histone mark signals across the training cell types and the chromatin accessibility input module predicted the difference between the average signal and the cell type specific signal (see Methods section Enformer Celltyping architecture and **Fig. A-1a**). Subsequent layers combining the two modules were next added for the second, full training step (**Fig. A-1b**). This pre-training step improved the overall performance of Enformer Celltyping (**Fig. A-4**), as did including the distribution of histone mark signals as well as the average in the DNA module predictions (**Fig. A-5**).

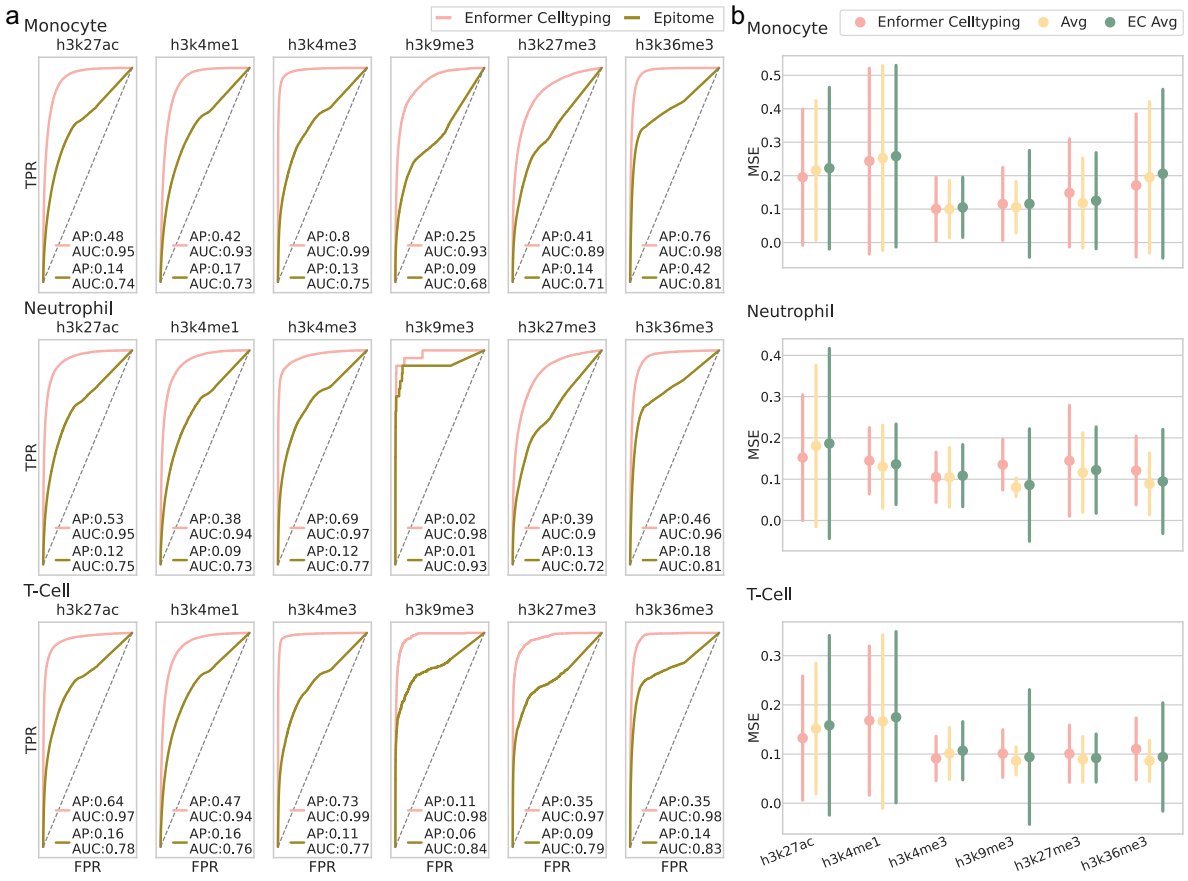


Figure 5-3 Enformer Celltyping makes best-in-class cell type-specific histone mark predictions.

(a) Enformer Celltyping's outperforms Epitome at genome-wide epigenetic predictions for all three immune cell types (held out from model training). Enformer Celltyping's predictions were binarized and averaged to 3,200 base-pairs to enable a comparison with Epitome. The receiver operating characteristics (ROC) curves with the area under ROC score and the average precision (AP) score are reported. (b) Enformer Celltyping frequently, but not consistently, predicts cell type-specific values (for three immune cell types) better than the average signal of the training cell types. The three cell types were not in the training data used to calculate the average. Moreover, Enformer Celltyping's superior performance against the average predictions from pre-training (EC Avg) highlight the model's ability to pick up on cell type-specific signals. Performance is measured by the mean squared error (MSE) for genome-

wide predictions with error bars for the standard deviation of the error of the predictions genome-wide.

While these results show Enformer Celltyping performs similar to the average signal, this is genome-wide performance, something that the average signal would be expected to perform strongly at given the overlap between cell's histone mark signals. Moreover, in **Figure 5-4a**, we highlight that these MSE values, for both the average signal and our model, are lower in regions of no genomic activity ('No peak' regions) which contribute to the genome-wide scores. Thus, to investigate Enformer Celltyping's performance in functionally relevant regions, we measured the model's correlation, which avoids issues of data scale and focuses on the shape i.e. is there a peak present, with the true signal at promoters, local and distal regulatory regions (**Figure 5-4b-c**). Here, we focused on the histone marks with known functional roles in these regions; H3K27ac at active enhancers and promoters, H3K4me3 at active promoters and H3K4me1 at active enhancers (Alan E Murphy et al., 2024), and derived the biological regions from Search Candidate cis-Regulatory Elements by ENCODE (SCREEN) (Moore *et al.*, 2020), available from: https://screen.encodeproject.org/. This work highlights Enformer Celltyping's superior performance in functionally relevant regions, where the average signal fails to capture the cell type-specific signals. This is particularly notable for distal enhancer regions for the blood immune cell types where Enformer Celltyping outperformed the average in 96% of the cell type-chromosome comparisons (**Figure 5-4c**, panel 3). Distal regulatory regions are highly cell type-specific thus highlighting the model's utility for epigenetic imputation. This distinction is further emphasised by the genome plots of selected SCREEN distal enhancer regions for the immune cell types (**Fig. A-6**).

Moreover, from **Figure 5-3b**, we can also see Enformer Celltyping's cell type-specific predictions (Enformer Celltyping) outperform that of the average predictions from pre-training (EC Avg) in the vast majority of cell types and histone marks. Moreover, the variance in the error of the predictions genome-wide is far smaller. However, the difference is minimal

and based on genome-wide measure of MSE, which, in the same manner as for the training set average, EC Avg would be expected to perform strongly in. Thus, we also compared Enformer Celltyping's performance in functionally relevant regions against EC Avg (**Figure 5-4b**). Enformer Celltyping outperformed EC Avg in all categories, with higher correlation with experimental signal in 98% of the comparisons tested for distal enhancer regions (**Figure 5-4b**, panel 3), highlighting how the model picks up on cell type-specific signals during the full training step.

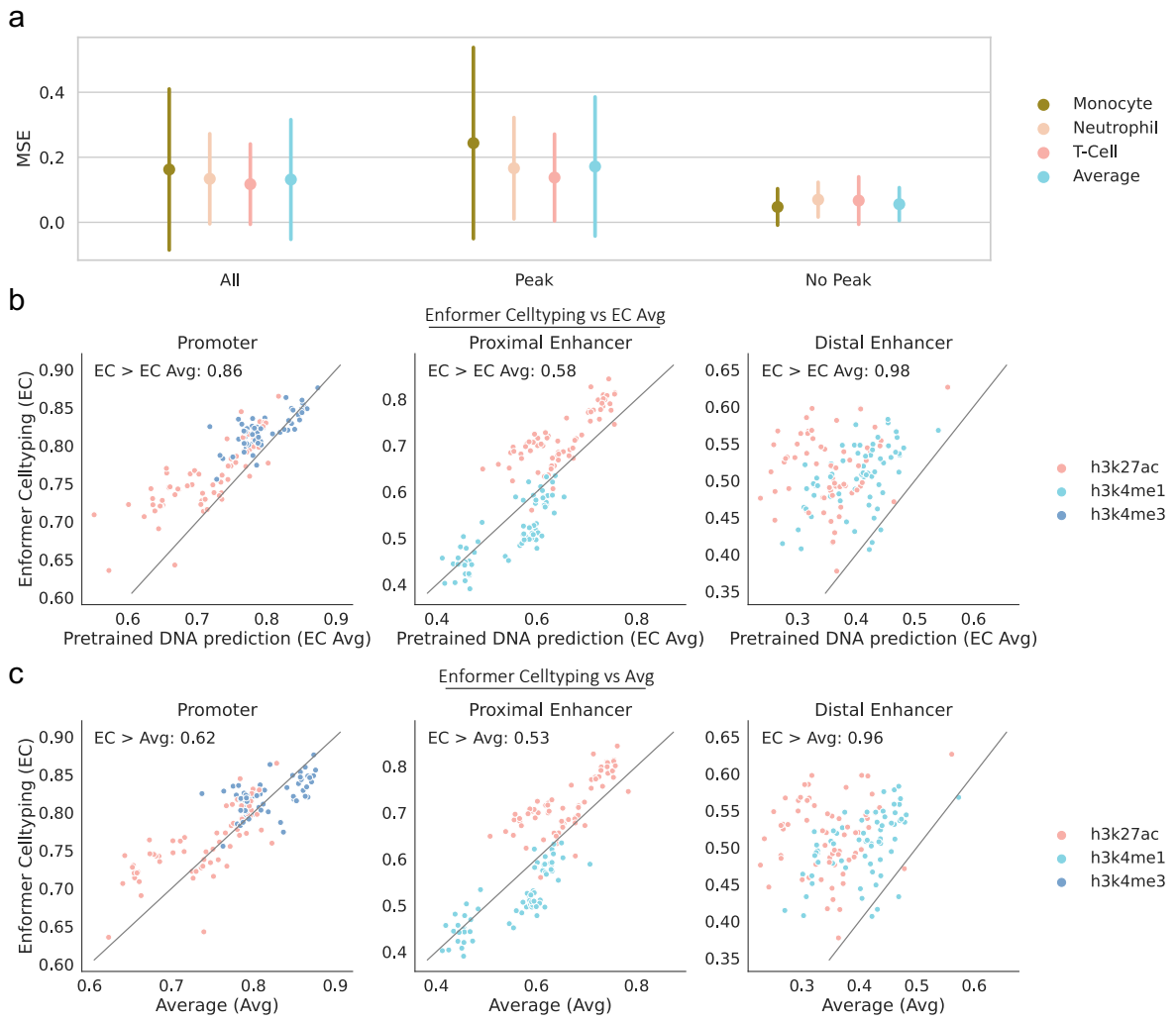


Figure 5-4 Enformer Celltyping has superior performance at functionally relevant regions.

(a) Enformer Celltyping's (EC) mean squared error (MSE) genome-wide ('All') and regions of histone mark activity ('Peak') or no activity ('No Peak'). Peak and No Peak regions were derived based on a -$\log_{10}$ p-value cut-off of 2(Boix *et al.*, 2021). Performance is shown against the average signal of the training cell types, for the three immune cell types excluded from the training dataset. The bars denote the standard deviation of the error of the predictions across all predicted regions that were labelled with the given activity. (b-c) Enformer Celltyping's Pearson R correlation with the true signal (y-axis) compared against the average predictions from pre-training's (EC Avg) (b) or the training set experimental values average's (avg) (c) Pearson R (y-axis) for the three blood cell types across different functional regions

sourced from Search Candidate cis-Regulatory Elements by ENCODE (SCREEN)(Moore *et al.*, 2020) and derived from a diverse range of ENCODE cell types. Performance was measured for histone marks with known functional roles in these regions for the same three immune cell types. Pearson R values were calculated per chromosome for each cell type (representing a dot in the figure). The 'EC > EC Avg' or 'EC > Avg' represents for how many cell type-chromosome values our model outperformed the alternative in.

Up to this point, Enformer Celltyping's performance has only been tested on cell types generated from ENCODE. We next measured its performance in a manner that replicates how researchers can use Enformer Celltyping to predict histone marks in their cell types of interest for which they have chromatin accessibility information (Schreiber, Singh, *et al.*, 2020). To this end, we utilised H3K27ac and H3K4me3 ChIP-Seq and ATAC-Seq data for isolated cell types (PU.1+ microglia, NeuN+ neuronal, OLIG2+ oligodendrocyte and NeuN- LHX2+ astrocytes) from resected cortical brain tissue (Nott *et al.*, 2019). Consensus peak p-value tracks (taking individuals as replicates) were derived and the data processed in the same manner as for the training cell types. Enformer Celltyping has not been trained on any isolated cell types from the brain and this data was not collected following the ENCODE protocols so its performance across the two measured histone marks should give a fair indication for any future applications of the model (**Figure 5-5a-b**). However, we did process the data and derive the -log10 p-value probability of binding using the same ENCODE pipeline to make it comparable. Moreover, it is worth noting that no microglia/brain macrophage track was used to train Enformer, see https://raw.githubusercontent.com/calico/basenji/master/manuscripts/cross2020/targets_human.txt. There are generic brain tissue tracks like 'brain, adult', 'midbrain male adult (78 years) and male adult (84 years)' or 'brain female embryo (96 days)', as well as tracks for astrocytes and neurons differentiated from mesenchymal cells (as opposed to isolated cell types as we have used) and one CAGE track for 'Oligodendrocyte – precursors'. Thus, it is possible that some data leakage may affect performance on these cell types.

To enable a fair comparison between ENCODE and non-ENCODE cell types, we show performance against the three ENCODE immune cell types which were held out of the Enformer Celltyping training dataset: grouped and they are referred to as "Blood immune

cells". Overall, Enfomer Celltyping's performance did not worsen drastically (relative to the ENCODE blood immune cells) despite the cells' dissimilarity to ENCODE samples. This becomes more impressive when we consider the performance of the average signal which showed a sharp increase in the MSE for H3K4me3 in the non-ENCODE cell types, relative to the ENCODE blood immune cells. Enformer Celltyping's ability to predict well in the presence of such a domain shift compared to the average signal of the ENCODE training cell types, highlights its ability to use the cell type-specific chromatin accessibility effectively when making predictions. Moreover, Enformer Celltyping outperforms the average signal at distal enhancer regulatory regions in three of the four cell types in at least 95% of the cell type-chromosome comparisons (**Figure 5-5b**), highlighting its superior performance and utility predicting in functionally relevant, cell type-specific regions. The model did underperform for astrocytes which we believe was, at least partly, contributed to by the lower ATAC-Seq sequencing depth, at approximately 36 million across the two replicates compared to the 50 million read depth Enformer Celltyping was trained on. To investigate how chromatin accessibility (ATAC-Seq) data sequencing depth, an input for Enformer Celltyping, impacts model performance more generally, we down-sampled the ATAC signal at varying levels 5 times with differing seeds on oligodendrocytes (**Figure 5-5c**). Even at a sequencing depth of 50% of what the model was trained to predict with, Enformer Celltyping outperforms the average signal in the cell type-specific, distal enhancer regions, highlighting the model's applicability under data quality variations.

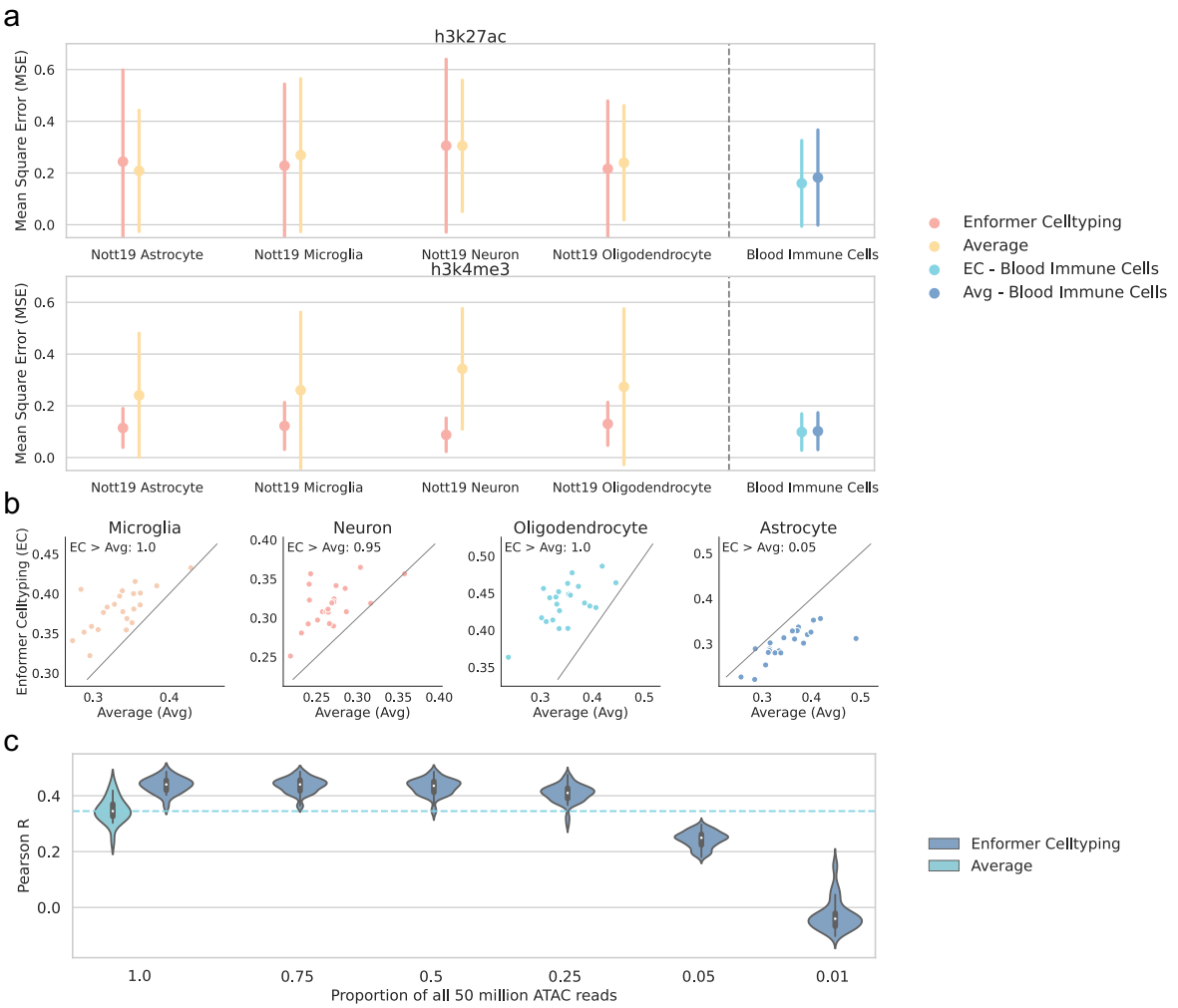


Figure 5-5 Enformer Celltyping can predict in cell types generated outside of ENCODE.

(a) Enformer Celltyping's mean squared error (MSE) across four isolated cortical brain cell types from Nott *et al.* (Nott *et al.*, 2019). The average performance for Enformer Celltyping and the average signal across the three ENCODE immune cell types (Blood Immune Cells) is also shown for comparison. The results for the "Blood immune cells" is the union of prediction results for T-cells, Neutrophils and Monocytes. "Average" here refers to the MSE calculated using the average values for the histone mark, across all cell types used in the training of Enformer Celltyping. Performance is measured for H3K27ac and H3K4me3 which were both assayed in the study. Performance is based on genome-wide predictions (over 16,000 per cell type) with error bars for the standard deviation. (b) Enformer Celltyping's Pearson R

correlation with the true signal (y-axis) compared against the Average's (avg) Pearson R (y-axis) in distal regulatory regions sourced from Search Candidate cis-Regulatory Elements by ENCODE (SCREEN) (Moore *et al.*, 2020) and derived from a diverse range of ENCODE cell types. Performance was measured for H3K27ac which has a known functional role in this region. Pearson R values were calculated per chromosome for each cell type (representing a dot in the figure). (c) Pearson R correlation (y-axis) on distal regulatory regions from SCREEN for Oligodendrocytes at differing proportions of the 50 million reads (x-axis). Subsampling was repeated 5 times with different seeds.

To further test Enformer Celltyping's robustness and broader applicability, we considered cancer research, validating its genome-wide predictive performance on ENCODE cancer cell lines which were not part of the training data (**Fig. A-7a-b**). The model showed no deterioration in its predictive quality compared to the immune cells lines (**Fig. A-7a**). Moreover, Enformer Celltyping outperformed the average signal at all regulatory regions, especially in enhancer regions, known to be cell type-specific; with better performance in 91% of proximal and 100% of distal enhancer cell type-chromosome comparisons (**Fig. A-7b**). These results highlight Enformer Celltyping's efficacy at predicting histone mark signals across varied cellular environments. Moreover, given its strong predictive performance at enhancer regions (**Figure 5-4a**, **Fig. A-7b**) and to investigate the usability of Enformer Celltyping's embeddings, these features were used with a random forest classifier to distinguish highly active enhancers, sometimes referred to as super-enhancers albeit this terminology is still debated (Pott and Lieb, 2015), from other enhancer regions in cancer cell lines. These super-enhancer regions were defined by a multitude of TF binding sites, often occupied by master transcription factors/coactivators like MED1 and BRD4 which are fundamental to cell type-identity (Khan and Zhang, 2016). We found these super-enhancers to be highly cell type-specific (**Fig. A-7c**), so a method to predict cell type-specific super-enhancers would be useful, highlighting the benefit of Enfomer Celltyping's ability to predict in previously unseen cell types. Enformer Celltyping's embeddings showed strong performance on these held-out cell lines (mean area under ROC of 0.85) (**Fig. A-7d**). Finally, Enformer Celltyping's performance across all 9 cell types tested is highlighted in **Fig. A-9**, demonstrating the model's consistent performance across these diverse cell types.

Having validated Enformer Celltyping's ability to predict the cell type-specific, genome-wide histone mark signals conditioned with chromatin accessibility information, we proceeded to

test its primary application of predicting the cell type-specific effect of non-coding, disease-relevant genetic variants on the epigenome. The requirements of this task are why Enformer Celltyping was designed in the manner that it was: able to predict in novel cell types (such that it can predict SNP effects in disease relevant cells); as well as incorporating the effect of distal genetic regulators (which we showed to be necessary to capture the effect of genetic variants in **Fig. A-2**).

We developed an framework, building on past approaches utilising QTL datasets (Reshef *et al.*, 2018; Kelley, 2020; Avsec, Agarwal, *et al.*, 2021), systematically measuring the genome-wide correlation between our model's predictions and the measured effects from the QTL studies using SLDP (Reshef *et al.*, 2018) (see **Methods Section 5.2.12**). We used QTL datasets as they directly measure the effect of extant human genetic variation. Furthermore, we have not concentrated on predicting the effect of genetic variants in isolation due to the confounding effect of LD, obscuring the coinherited and causal SNPs (Reshef *et al.*, 2018). Approaches like fine-mapping could help identify the causal SNPs, however recent works have highlighted that despite considerable advancements there is still opportunity for improvement (Schaid, Chen and Larson, 2018; Broekema, Bakker and Jonkers, 2020; Burgess, 2022). Note that past work concentrated on using eQTL studies and predicting genetic variant effects in the same cell types that the models were trained on. On the other hand, we utilised hQTL datasets and our model's ability to predict in previously unseen cell types using their chromatin accessibility data, matching the cell types studied in the hQTL to gain insight into the model's predictive ability of SNP effects on the epigenome. Moreover, we implemented strict pre-processing criteria, filtering out any interactions from the hQTL set where the distance was greater than the model's receptive field and measured the effect across the full receptive field of the model.

We used hQTL data from the Blueprint project phase 2 (Kundu *et al.*, 2022), for two histone marks (H3K4me1 and H3K27ac), each of which was obtained for three cell types (T-cells, Monocytes and Neutrophils). These were the same cell types which were held out of the Enformer Celltyping training data, as discussed above. These hQTL datasets were generated using whole genome sequencing from neutrophils (n = 196), monocytes (n = 194) and T-cells (n = 169) individuals. We note that while these are the largest published hQTL datasets, they remain small by the standards of modern eQTL datasets, and so could be considered underpowered to detect QTL effects. We tested 867,568 genetic effect predictions for all of Enformer Celltyping's predicted histone marks for six cell types against the six hQTL datasets (Kundu *et al.*, 2022). Similarly to model training, we tested the effect on the forward, reverse strand and with small random shifts in genomic starting position to return the average effect. We expected to see a significant result for the prediction in the same cell type and histone mark for which the hQTL study was completed (**Figure 5-6a**).

The results of the SLDP genetic effect predictions were worse than what was expected with only one of six of the matching histone mark and cell types being nominally significant and none after FDR correction of 0.05 (**Figure 5-6a**). We noted that one possible causes of differences, relates not to the quality of SNP effect predictions, but to differences in predictions of peaks for the reference sequence: sometimes Enformer Celltyping doesn't predict peaks which are seen in the experimental data. If a peak is missed in the predictions, then the effect of a SNP which decreases the size of that (missing) peak, cannot be captured. To assess whether this affects the results, we tried filtering out SNP-to-histone QTLs associated with peaks that were missed by Enformer Celltyping, and re-running the SLDP analysis. However, this did not improve the significant matched cell type and mark cases (**Fig. A-9**).

We next tested whether the inability to replicate the hQTL SNP effect signals are specific to Enformer Celltyping or is more generally a trait of genomic deep learning models trained using variation across the genome (which has been noted by other recent work (Sasse, Ng, A. Spiro, *et al.*, 2023)). We reran our framework on the same hQTL sets using the matched cell type and histone mark output tracks from Enformer (**Figure 5-6b**). Interestingly, Enformer similarly fails to capture the hQTL SNP effect signals with only one expected significant result for H3K27ac in T-Cells. This finding highlights how the current training paradigm of genomic deep learning models is insufficient to capture the effect of genetic variants sufficiently with or without extrapolating predictions to previously unseen cell types.

In making these predictions, we altered our approach to *in silico* mutagenesis, relative to the standard in the field (**Fig. A-10**). One issue with the standard approach, which centres the SNP in the input window, is that it does not utilise the full receptive field of these models (**Fig. A-10a-b**). Our approach captures the full receptive field by making multiple predictions for each SNP, sliding the input window (**Fig. A-10c**). Given our approach results in multiple predictions of the same genomic locations, we sought to confirm that predictions were consistent for the same genomic region, regardless of where it appears in the input window (**Figure 5-6c**) (also see **Methods Section 5.2.11**). This resulted in extremely high correlations for all tested cell types, confirming that centring the model's predictions on the SNP and shifting the input window in the hQTL dataset would not affect the histone mark binding predictions.

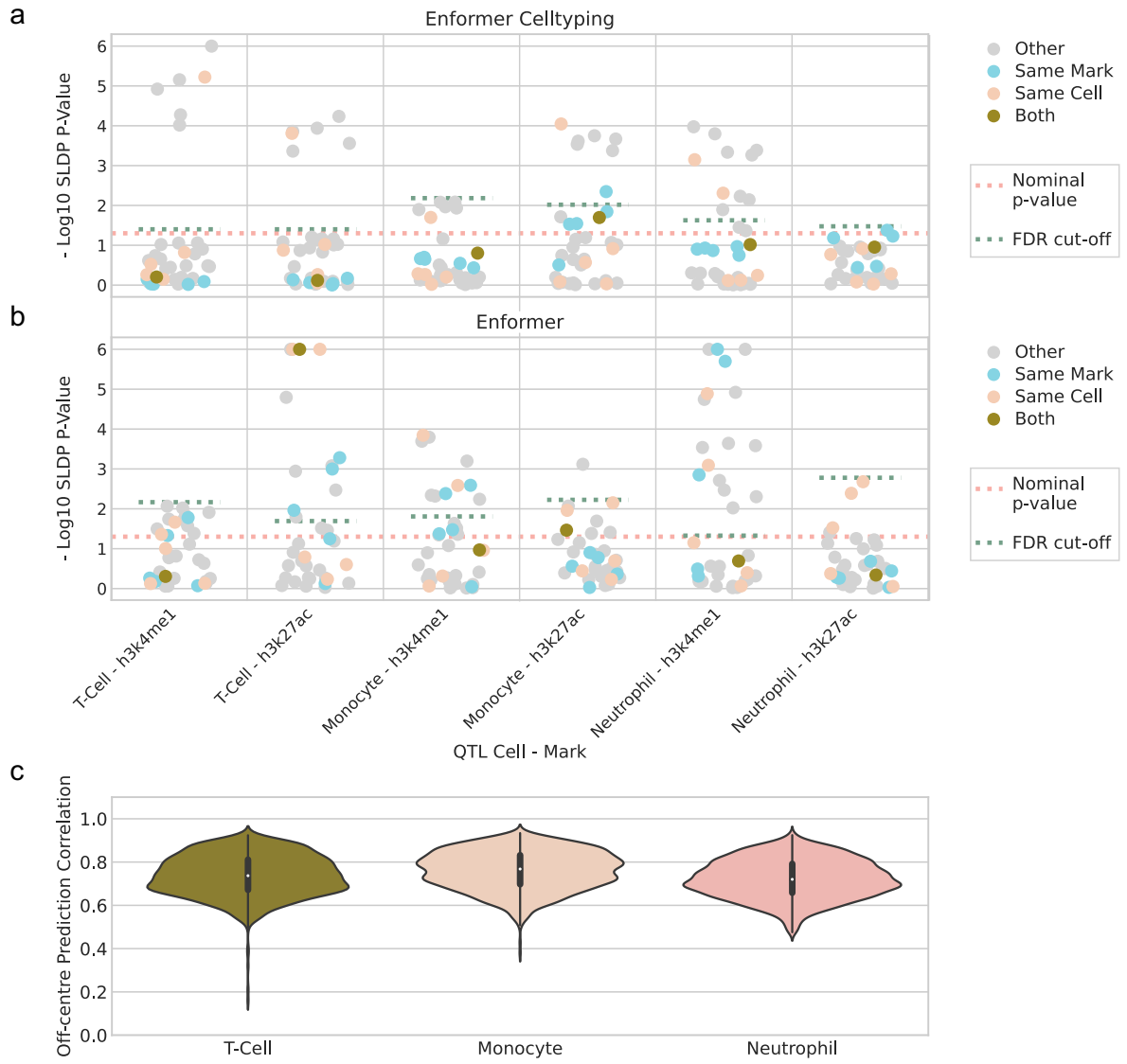


Figure 5-6 Genomic deep learning models' genetic effect predictions predominately do not replicate QTL results.

Statistical significance (y-axis) of SLDP (Reshef *et al.*, 2018) genome-wide concordance between Enformer Celltyping's (a) or Enformer's (b) genetic variant predictions and measured hQTL effect sizes. The x-axis depicts the six Blueprint phase 2 hQTL studies (Kundu *et al.*, 2022) and the cell-histone mark predictions are coloured by their relationship to the hQTL study: 'Both' – both the histone mark and the cell predicted in were the same as the hQTL study, 'Same Mark' – the mark matched, 'Same Cell' – the cell matched and 'Other' – neither the cell

nor mark matched. If there was perfect agreement between our model and the hQTL studies, all 'Both' entries would be significant. A nominal p-value (pink dashed line) and the cut-off after Benjamini and Hochberg correction based on the minimum, non-significant p-value, corrected at the level of each hQTL (Turquoise dashed line) are both added. (b) Enformer Celltyping makes consistent predictions for a genomic region, regardless of where it lies within the model's receptive field. The correlation (y-axis) in Enformer Celltyping's histone mark predictions at matching genomic regions when the model's input window is moved. This was tested for the three immune cell types (x-axis) for 1,000 genomic positions each with and without genetic variants added.

We have demonstrated that Enformer Celltyping accounts for changes in local chromatin accessibility for its predictions (**Figure 5-1c**) but also that it considers global, cell type-specific signals (**Figure 5-2**, **Figure 5-5**). To investigate what was driving the model's cell type-specificity, we identified the genomic regions with distal acting histone mark peaks that had the highest contribution from the global chromatin accessibility signal for a selection of cell types (**Figure 5-7a** and **Methods Section 5.2.13**). The corresponding DNA sequences for these peaks were assessed for enrichment of known motifs using Homer (Heinz *et al.*, 2010) (**Figure 5-7b-c**). Finally for each cell type, based on the set of significant known motifs, we looked for cell type specificity of the related transcription factors using EWCE (Skene and Grant, 2016) (**Figure 5-7d**).

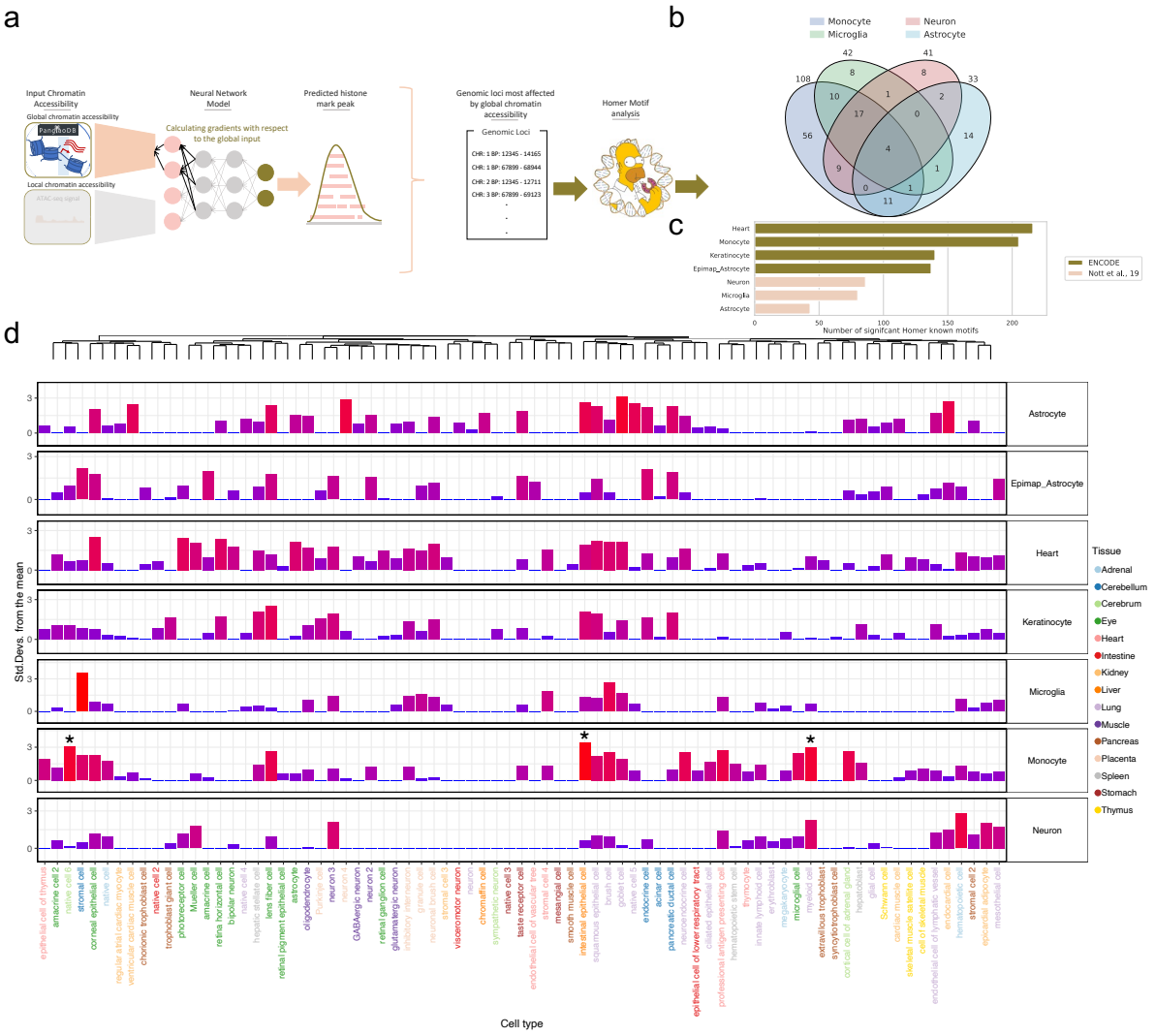


Figure 5-7 Enformer Celltyping identifies motifs for cell type-specific transcription factors which are enriched based on the global chromatin accessibility signal.

(a) For each peak, the influence of the global signal was approximated by calculating the partial derivatives of the model with respect to the input, i.e. the gradient on the input. From this, peaks and their genomic loci that rely the most on the global input signal were identified. These top loci were analysed with Homer (Heinz *et al.*, 2010) for each cell type. (b) The overlap between significant transcription factors across the immune and brain cell types tested for all histone marks. (c) The number of significant transcription factors per cell type. (d) Cell type-specificity of the transcription factors from Homer for each cell type, calculated using EWCE

(Skene and Grant, 2016). The rows are the differ cell types for which predictions were made – note the brain cell types are taken from Nott *et al.* (Nott *et al.*, 2019) with the exception of Epimap_Astrocyte which was from ENCODE (ENCODE Project Consortium, 2012) sample available through EpiMap (Boix *et al.*, 2021). The top row is a dendrogram displaying the clustering of cell types from the Descartes human whole body single-cell RNA–seq reference dataset used (Cao *et al.*, 2020). The columns, ordered by the dendrogram, are the cell types tested from the reference dataset with the colour representing the tissue of origin. Significant associations found by EWCE (false discovery rate adjusted p-value<0.05), indicating enrichment in a cell type are marked with an asterisk. The background geneset were the approximately 400 transcription factor genes from Homer.

Overall, **Figure 5-7** highlights how Enformer Celltyping identifies DNA motifs relating to distinct transcription factors for the different cell types tested and some of these transcription factors match the same cell types in their transcriptional specificity. For example, monocytes had significant enrichment in myeloid cells. Furthermore, the strongest enrichment for the heart, which contains epithelial cells in the epicardium, had the largest enrichment in corneal epithelial cells. Neurons showed strong enrichment in neuronal cell types (neuron 3). Keratinocytes had the largest enrichment in lens fibre cells which have been noted to resemble keratin-rich keratinocytes in the stratum corneum of the epidermis and share biological processes including organelle degradation during their terminal differentiation (Bassnett, 2009). The differentiated astrocyte cell from ENCODE (ENCODE Project Consortium, 2012) (Epimap_Astrocyte) and the isolated astrocyte from Nott *et al.* (Nott *et al.*, 2019), showed large cell type enrichments in brain cell types, in neurons 3 and 4 respectively but not in the reference astrocyte cell type. Moreover, we see overlap in the transcription factors identified with Homer in similar cell types such as across the immune cells; microglia and monocytes (**Figure 5-7b**). Interestingly, the cell types from Nott *et al.* (Nott *et al.*, 2019) (Astrocyte, Microglia and Neurons) showed found fewer transcription factors from the global chromatin dependent peak motifs (**Figure 5-7c**). This could indicate a bias in Enformer Celltyping whereby, although the domain shift of predicting in isolated, single cell types outside of ENCODE does not seem to affect the genome-wide predictive performance, it does affect the global cell type-identity the model learns. We can also see this effect when we consider the projection of the model's embedded global chromatin accessibility signal for the Nott *et al.* brain cell types versus the ENCODE brain cell types which seem to separate in a lower dimensional space (**Fig. A-11**).

Finally, Enformer Celltyping showed strong performance in distal regulatory regions (**Figure 5-4c**, **Figure 5-5b**, **Fig. A-7b**) so, despite having limited success at accurately capturing the effect of genetic variants (**Figure 5-6**), we queried whether the predicted enhancer regions were enriched for genetic variants associated with complex traits and diseases in cell type-specific regulatory regions (**Figure 5-8**).

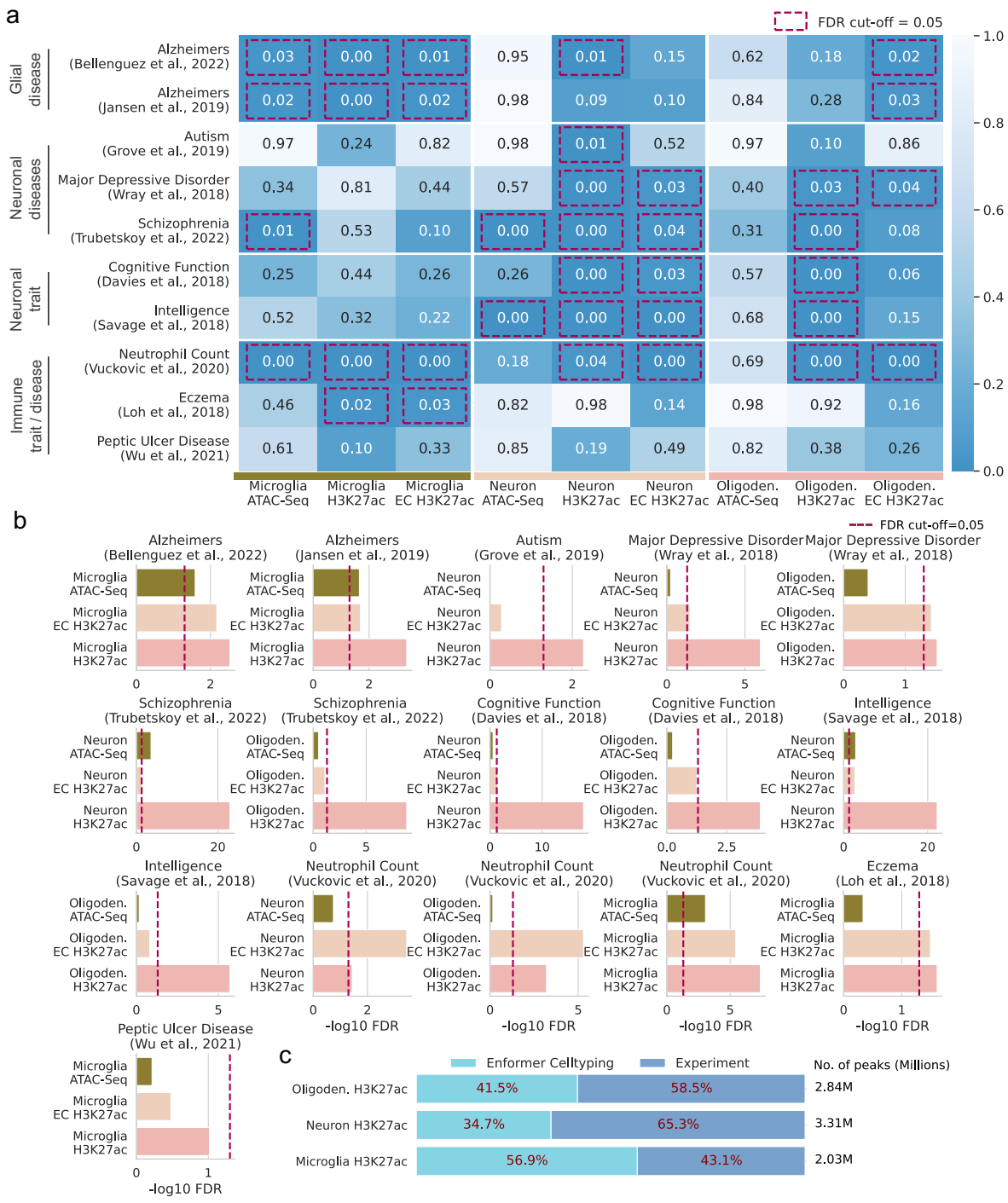


Figure 5-8 Enformer Celltyping's predictions capture cell type-specific genetic enrichment for complex traits.

(a) Heatmap of stratified LD score regression (s-LDSC)(Finucane *et al.*, 2015) analysis for genetic variants associated with brain and immune diseases/traits and behavioural traits (sourced from associated GWAS) displayed as false discovery rate (FDR) value for significance

of enrichment for ATAC-Seq chromatin accessibility signal, H3K27ac signal and Enformer Celltyping's (EC) predictions of H3K27ac for microglia, neurons and oligodendrocytes (oligoden.). (b) $-\log_{10}$(FDR) genetic enrichment for the complex traits from s-LDSC. (c) Proportion of peaks in derived peak files used for s-LDSC analysis.

We first aimed to confirm that we could recapitulate the genetic enrichment results of Nott *et al.* (Nott *et al.*, 2019); despite using the same dataset, to ensure comparability with our predictions, we had to omit samples and alter their analysis. Specifically, our predictions are continuous, genome-wide signals (i.e. in bigwig format) whereas we ran s-LDSC with binary annotations (i.e. peaks in bed format): To ensure the real and predicted signals were comparable, they were uniformly processed. Using the experimental H3K27ac signal, we recapitulated Nott *et al.*'s results, finding known associations in glial and neuronal diseases and traits like microglia with Alzheimer's disease, neurons with schizophrenia and neurons with intelligence (**Figure 5-8a**). Moreover, we included immune GWAS traits and diseases (eczema and neutrophil count) and found enrichments for microglia, the resident macrophage of the brain. Given this, we assumed the significant associations for the experimental H3K27ac signals were true. Thus, we could next test whether EC's predictions, which use ATAC-Seq chromatin accessibility data, improve the capture of genetic variants for these known trait-to-celltype relationships over the enrichment seen in ATAC-Seq.

Enformer Celltyping's predictions captured the vast majority of known enrichments with significant associations (**Figure 5-8a**). Furthermore, the predictions improved on the ATAC-Seq data in 14 of the 16 expected enrichments (**Figure 5-8b**). It is worth noting that for some cases the level of enrichment of the predicted H3K27ac signal did not match the experimental H3K27ac (**Figure 5-8b**). This was most notable for neurons in Schizophrenia and Major depressive disorder. We believe this was, at least in part, due to the relatively lower number of peaks in the derived bed files for the predictions compared to the experimental data (**Figure 5-8c**). Overall, these results show how Enformer Celltyping can be used along with the breadth of publicly available ATAC-Seq data to derive histone mark signals which harbour greater enrichment for genetic variants associated with complex traits and diseases than

chromatin accessibility alone, highlighting it's utility for the study of complex traits and disease.

5.4. Discussion

We report Enformer Celltyping, a deep learning model to predict epigenetic signals in previously unseen cell types, using DNA sequence and chromatin accessibility data from the cell type of interest to make its predictions, with the largest receptive field to date of 100,000 base-pairs. Enformer Celltyping achieves this receptive field through a transfer learning approach based on Enformer, a model which predicts gene expression and epigenetic signals limited to the same cell types it was trained on (Avsec, Agarwal, *et al.*, 2021). Here, we recreated Enformer; removing the output and convolutional layers after the transformer block and freezing the weights in pre-trained layers. This approach differs from past applications which mostly fit a linear model on top of Enformer's output (Karollus, Mauermeier and Gagneur, 2023; Sasse, Ng, A. Spiro, *et al.*, 2023), offering more flexibility and aligning with current approaches in the machine learning domain such as NLP (Devlin *et al.*, 2019). Our approach shows the power of transfer learning using large, pre-trained models in computational biology. Here, we trained Enformer Celltyping on a data set several times larger than Enformer's, using a fraction of computational resources (132 versus 5,376 GPU hours). We believe this approach, which we have made openly available, shows promise for what researchers can achieve fine-tuning large, pre-trained models in the face of limited resources.

Enformer Celltyping achieves best-in-class performance for histone mark predictions (**Figure 5-3a**), strong predictive performance in functional genomic regulatory regions (**Figure 5-4**) and maintains accuracy even with a domain shift to isolated cell types assayed outside of ENCODE (**Figure 5-5**). Developed as a quantitative model, Enformer Celltyping predicts a continuous $-\log_{10}$ p-value signal, which has been shown to yield better generalisation and interpretability than binary, classification models (Toneyan, Tang and Koo, 2022). Moreover,

the -$\log_{10}$ p-value derived after MACS 2.0 peak calling was chosen over the fold change measure as it typically has a higher signal-to-noise ratio (Ernst and Kellis, 2015; Schreiber, Durham, *et al.*, 2020). However, there are some limitations to both our approach to model training and Enformer Celltyping's performance. Firstly, due to a lack of availability, we used imputed chromatin accessibility ATAC-Seq data for the majority of cell types from EpiMap (Boix *et al.*, 2021). A possible improvement could be to use assayed chromatin accessibility data or use a better performing imputation technique (Schreiber, Durham, *et al.*, 2020; Schreiber *et al.*, 2023). Our reliance here on imputed data highlights the importance of generating new biological datasets to supplement those already available which have been paramount for the training genomic deep learning models like ENCODE. Secondly, Enformer Celltyping's predictions are made in 128 base-pair bins. This is lower resolution than the standard bin size after peak calling from ENCODE (25 base-pairs) (ENCODE Project Consortium, 2012). A major goal of developing Enformer Celltyping was to use it to predict the effect of genetic variants. Doing so at a lower resolution will dilute the effective change on the epigenetic signal which may have contributed to the suboptimal performance of Enformer Celltyping on the genetic variant analysis. Predicting in 25 base-pair bins was not possible as a result of the transfer learning approach using the frozen layers from Enformer. Possible solutions to enable this would be to either allow the weights to update in Enformer so it can adjust to the desired resolution which would incur a much higher a computational cost, equivalent to that which the original authors faced or to use a U-Net architecture on top of Enformer to increase the resolution, as proposed recently (Linder *et al.*, 2023). Furthermore, upon investigating the causes of the model's predictions based on global cell type information, we found that cell-type specific histone mark predictions are influenced by the presence of cell type-specific transcription factor motifs. However, there were only a

limited number of expected significant enrichments, which could be due to confounders such as the limited number of transcription factors with known motifs. Moreover, mapping these back to human based on orthologs and the fact that some motifs will span whole transcription factor families may affect the quality of the results.

A primary goal of machine learning models that predict gene expression or epigenetic signals from DNA, including Enformer Celltyping, is to predict the effect of genetic variants (Reshef *et al.*, 2018; Kelley, 2020; Avsec, Agarwal, *et al.*, 2021). Usually, these models can only make predictions in the same cell types they were trained in. Here, we have expanded this utility to predict in any cell type based on its chromatin accessibility data, enabling genetic variant predictions in any cell type of interest which is more aligned with researcher's needs for investigation into disease-relevant cell types. For use with our model and for the wider community, we created the first framework for genetic variant effect evaluation using hQTL data in combination with SLDP (Reshef *et al.*, 2018). Previous studies have evaluated performance using eQTL data, which is several layers further removed from the presumed direct biological effect of SNPs (eQTLs aggregate over multiple layers of gene regulation, while hQTLs will be more directly affected). Moreover, other studies have used fine-mapping to evaluate model performance, but such approaches are compromised by their reliance on fine-mapping approaches, which differ depending on the methodology used (Schaid, Chen and Larson, 2018; Broekema, Bakker and Jonkers, 2020; Burgess, 2022). Our approach utilises a model's full receptive field (**Fig. A-10**), along with interpretable p-value cut-offs to measure associations, as opposed to only considering z-scores as previously considered (Reshef *et al.*, 2018; Avsec, Agarwal, *et al.*, 2021). This framework can be used to evaluate any future genomic deep learning models as the field focuses on improving their performance in genetic variant effect predictions.

Overall, Enformer Celltyping's performance in predicting the effect of genetic variants predominately did not replicate the QTL results using the SLDP measure of concordance (Reshef *et al.*, 2018) (**Figure 5-6a**). We hypothesised that two issues contributed to this: Firstly, that the hQTL set contained interactions which caused a histone mark binding position to be removed by a genetic variant but these histone mark binding positions were not captured in Enformer Celltyping's prediction on the major allele. Secondly, that genomic deep learning models, including Enformer Celltyping, inherently struggle to accurately predict the effect of genetic variants based on their current training paradigm. We tested the effect of the first issue by removing such interactions before running SLDP but this had little effect on increasing the number of expected significant associations (**Fig. A-9**). The second issue has already received growing attention in the literature, specifically in relation to Enformer. Firstly, Karollus *et al.*, showed how the predicted effect of regulatory elements decreased with distance and did not match the measured causal effects of these elements, such as enhancers (Karollus, Mauermeier and Gagneur, 2023). This means that on average, Enformer struggles to identify distal regulatory genomic regions, weighting their importance less than local regulatory information. This may give insight into Enformer Celltyping's relatively poor performance in predicting the effect of genetic variants where the majority of genetic variants tested were distal (**Fig. A-2**). Secondly, Sasse *et al.*, comprehensively showed how Enformer underperforms at predicting the effect of genetic variants on transcription, even predicting the incorrect direction of effect in up to 40% of tested cases (Sasse, Ng, A. Spiro, *et al.*, 2023). The authors hypothesised that this was a result of Enformer being trained on points across the reference genome and then needing to make out-of-sample predictions for individuals' personal genetic variants at any locus. In our work, these issues could be compounded by the fact that models like Enformer Celltyping, which predict the effect of genetic variants in

previously unseen cell types, need alternative functional genomic information to infer the cell type it is predicting in. For Enformer Celltyping, this is chromatin accessibility information. However, this chromatin accessibility data used for the genetic variant effect predictions did not account for the genetic variant thus further disadvantaging the model's ability to predict the genetic effect accurately. In our work, we also tested the extent to which this affected our results by running the same genetic variant effect prediction test on Enformer (**Figure 5-6b**). However, Enformer performed similarly highlighting that it did not play a major role in the results. We believe the real limitation is the lack of genetic variation seen in the current approach to genomic deep learning model's training which we have noted for epigenetics and which was highlighted for transcription by Sasse *et al*. Therefore, we agree with Sasse *et al.*, that future work hoping to use a machine learning model to predict the effect of genetic variants should concentrate on training models on data which accounts for genetic variants (Sasse, Ng, A. Spiro, *et al.*, 2023) and secondly, that research should focus on predicting in the same cell types upon which the model was trained rather than extrapolating to previously unseen cell types using functional genomic data from the cell type of interest like chromatin accessibility information. This is, of course, limited by the current lack of suitable public data, specifically for epigenetics. Despite genomic deep learning model's inability to predict the effect of genetic variants, our results highlight Enformer Celltyping's utility for complex trait research; the predicted histone mark signals harbour greater enrichment for genetic variants associated with complex traits and diseases than chromatin accessibility signals (**Figure 5-8**).

In summary, current genomic deep learning models like Enformer are not suitable to study the epigenome and the effect of genetic variants on disease-relevant, cell or sub-cell types not captured in public repositories. Here, we introduce Enformer Celltyping, a model which can predict epigenetic signals in previously unseen cell types, using a large receptive field to

capture genomic regulatory information. Enformer Celltyping requires just chromatin accessibility data from the cell type of interest to enable predictions of regulatory effects on six histone mark signals. Moreover, we introduce a framework to benchmark methods for genetic variant effect predictions and show how the training paradigm for current genomic deep learning models is insufficient to accurately capture these genetic variant effects. Despite this, we show Enformer Celltyping's ability to capture genetic enrichment of complex traits.

5.5. Conclusion

Enformer Celltyping is a self-attention based, neural network model which predicts histone mark activity in previously unseen cell types. It makes these predictions based on DNA sequence and chromatin accessibility information from the cell type of interest. Our model uses a transfer learning approach on a pre-trained Enformer (Avsec, Agarwal, *et al.*, 2021) model to account for the distal genetic regulatory effects and embeds both a global and local representation of cell types from chromatin accessibility information. We show Enformer Celltyping's predictions out-perform current approaches and generalises to cell types assayed separately to the original training set. However, our developed framework using population QTL data highlights how Enformer Celltyping and current genomic deep learning models, can not yet accurately predict the effect of genetic variants. Despite this, Enformer Celltyping's predicted signals do capture cell type-specific genetic enrichment for complex traits, highlighting its utility for research.

5.6. Data availability

All training and test cell types from EpiMap, along with scripts to download and complete all pre-processing steps are available at https://github.com/neurogenomics/EnformerCelltyping. Blueprint phase 2, histone mark QTL data was downloaded from the European Genome-Phenome Archive (ID: EGAD00001005199). The full training and all validation scripts are also available at https://github.com/neurogenomics/EnformerCelltyping. A list of the training regions and cell types, the one-hot encoded DNA sequence, the trained Enformer Celltyping model's weights, the average chromatin accessibility signal of the training cell types and the SNP effect predictions on the hQTL sets for both Enformer and Enformer Celltyping are all available through figshare (https://figshare.com/account/home#/projects/159143). For the interest of the reviewer and readers, we have also made available full genome browser track visualisations for T-Cells H3K27ac, including the experimental track, Enformer Celltyping and average predictions: https://genome-euro.ucsc.edu/s/almurphy/EC_hg19. Supplementary figures and tables referenced within this chapter can be found in **Appendix A**.

5.7. Code availability

The Enformer Celltyping model architecture, training scripts and all analysis are available from https://github.com/neurogenomics/EnformerCelltyping.

5.8. Acknowledgements

I would like to thank Leo Linden, Simon Mathis, David Kelley and Jacob Schreiber for their invaluable conversations and feedback that supported this work. I would also like to thank

Alyssa Morrow for her help supplying a version of Epitome for benchmarking. This study makes use of data generated by the Blueprint Consortium. A full list of the investigators who contributed to the generation of the data is available from www.blueprint-epigenome.eu. Funding for the project was provided by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 282510 – BLUEPRINT.

5.9. Funding

This work is supported by the UK Dementia Research Institute award number UK DRI-5008 through UK DRI Ltd, principally funded by the UK Medical Research Council. N.S. also received funding from a UKRI Future Leaders Fellowship [grant number MR/T04327X/1].

6. Predicting gene expression from histone marks using chromatin deep learning models depends on histone mark function, regulatory distance and cellular states

6.1. Introduction

The work in this chapter is adapted from that set out in our publication (Alan E Murphy et al., 2024). As discussed in **Section 1.6**, numerous studies have employed machine learning to explore the relationship between epigenetics and gene expression, such as by predicting transcriptional levels based on histone mark signals. These works assume that the observed chromatin state captures the other contributing epigenetic regulators without directly experimentally observing them. Most of these efforts have focused primarily on identifying which histone marks are most predictive of gene expression, rather than investigating the broader regulatory landscape. For example, addressing this over 10 years ago, Karlić *et al.* used a linear regression model but only considered histone mark levels at promoter regions, and only tested the effect in a single cell type, CD4+ T-cells (Karlić *et al.*, 2010). González-Ramírez and colleagues identified predictive histone marks at promoters and other regulatory regions, leveraging chromatin interaction data from a Hi-C assay to link enhancers to target genes (González-Ramírez *et al.*, 2021). However, this study also only considered one cell type, mouse embryonic stem cells. Moreover, there is some circularity in the selection of training regions based on derived (from their histone mark data) promoter and enhancer locations and the model's input measuring the same histone mark levels. Finally, Wang *et al.*, investigated the relationship between histone marks and transcription but inverted the problem, predicting histone mark levels from transcription (Wang *et al.*, 2022). For transcription, they used Pro-Seq and GRO-Seq which labels RNA as it is being transcribed,

avoiding issues with RNA degradation (Core, Waterfall and Lis, 2008; Wang *et al.*, 2022). The authors used a Support Vector Regression model but again only investigated relationships in the K562 cell line.

Here, we expand on previous research by considering multiple cell types, histone marks and regulatory distances. We investigated the effects of seven histone marks on gene expression (**Table 1-1**) in eleven human cell or tissue types from the Roadmap Epigenomics Consortium (Kundaje *et al.*, 2015). We will refer to these as cell types hereafter, but note that they also contain tissue samples and cell lines. We used two neural network architectures to predict gene expression: a simple convolutional neural network considering genes promoter regions, and a recently published transformer-based, DNA interaction-aware deep learning architecture called Chromoformer (Lee, Yang and Kim, 2022) (see associated publication, Fig. 1 for architecture). Chromoformer was originally trained to predict expression using seven histone marks. Here, we adapt and retrain the model to predict based on single histone marks, and pairwise combinations of histone marks, to investigate their effect on transcription in isolation. Our work highlights how histone mark function, cellular differentiation and genomic distance to regulatory elements all collectively influence the relationship between histone modification levels and gene expression. We find that there is no universal histone mark which is consistently the most predictive of expression. We recommend that researchers consider all three of these influencing factors when determining the effect of histone mark levels on the transcriptional state of a cell in their work.

In the related field of genomic deep learning, models predict expression or epigenetic marks from DNA sequence, such as Enformer (Avsec, Agarwal, *et al.*, 2021), Borzoi (Linder *et al.*, 2023) and Sei (Chen *et al.*, 2022). There has been a recent shift away from arbitrarily benchmarking performance, to prioritising the use of these models to make new biological

discoveries (Avsec, Weilert, *et al.*, 2021; Nair *et al.*, 2023; Taskiran *et al.*, 2024). This is still lacking for models linking histone mark levels to expression. We aim to address this by outlining a framework to use these models to identify the cell type-specific functional and disease related genomic loci, leading to new biological insights and expanding the utility of these models beyond simple investigation into which histone marks are the most predictive of gene expression.

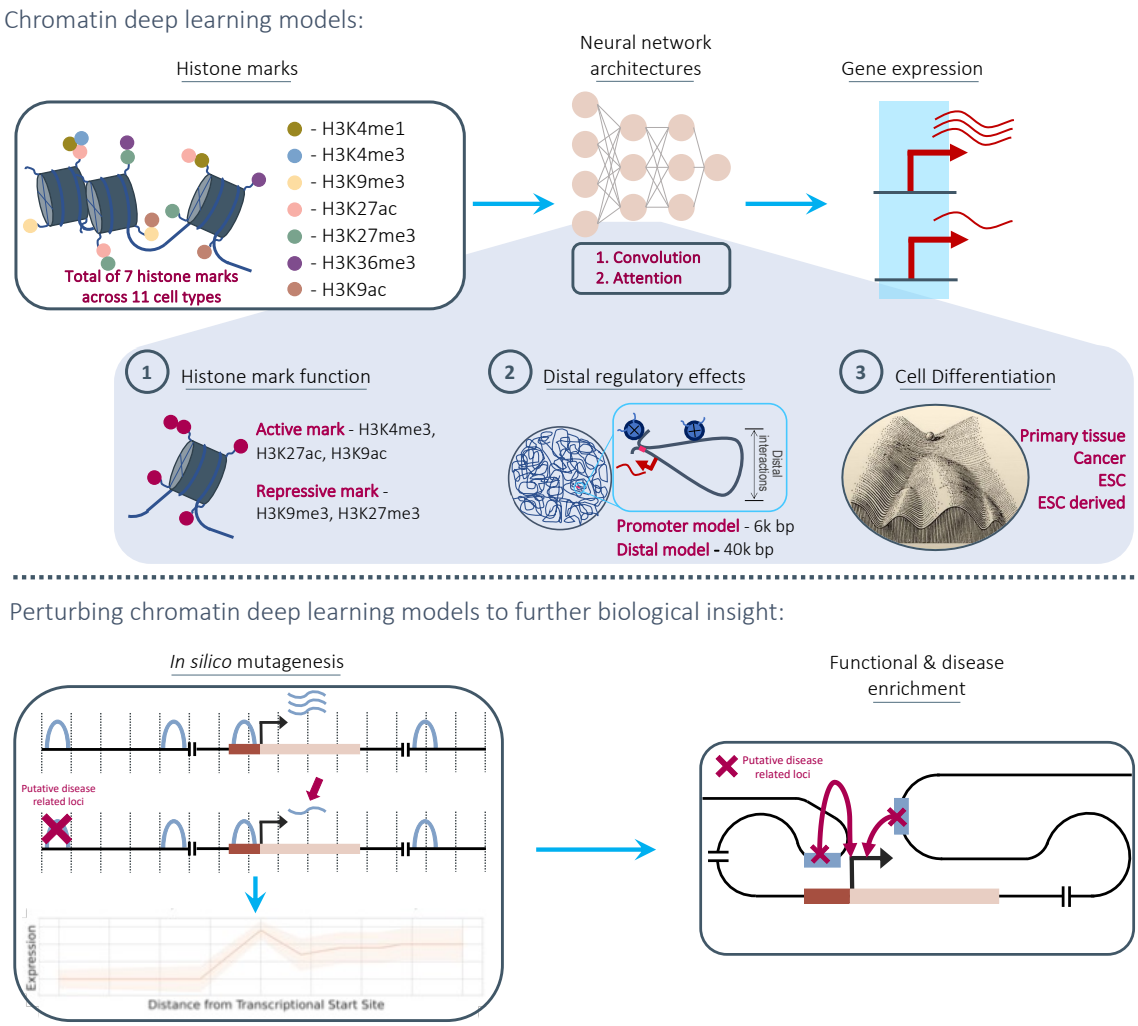


Figure 6-1: Understanding the complex relationship between histone mark activity and gene expression in silico.

Recent *in silico* prediction approaches of transcription from histone mark activity have omitted key contributing factors like cell state, histone mark function or distal effects, that impact the relationship, limiting their findings. Moreover, downstream use of these models for new biological insight is lacking. Here, we present the most comprehensive study to date - investigating seven histone marks, in eleven cell types, across a diverse range of cell states using convolutional and attention-based models at promoters and distal regulatory elements. Our work shows that histone mark function, genomic distance and cellular states collectively influence a histone mark's relationship with transcription and that no individual histone mark

is consistently the strongest predictor of gene expression. Furthermore, *in silico* histone mark perturbation assays using these chromatin deep learning models can uncover functional and disease related loci uncovering new biological insight.

6.2. Methods

6.2.1. Data collection and processing

The data for our analysis was derived from the Roadmap Epigenomics Consortium (Kundaje *et al.*, 2015) and follows the same preprocessing pipeline used by Chromoformer (Lee, Yang and Kim, 2022). We used a subset of eleven cell types from Roadmap, for which gene expression, histone mark, and 3D chromatin interactions profiles were available (**Table B-1**). Specifically, we included data from H1 embryonic stem cells (E003), H1 BMP4 derived mesendoderm (E004), H1 BMP4 derived trophoblast (E005), H1 derived mesenchymal stem cells (E006), H1 derived neuronal progenitor cultured cells (E007), HUES64 embryonic stem cells (E016), Liver (E066), Pancreatic islets (E087), A549 EtOH 0.02pct lung carcinoma (E114), GM12878 lymphoblastoid (E116) and HepG2 hepatocellular carcinoma (E118). TagAlign-formatted, ChIP-Seq read alignments for seven histone marks - H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K27ac, and H3K9ac were used. For consistency, the data was subsampled to 30 million reads and reads themselves were truncated to 36 base-pairs, reducing possible read length biases. The alignments were sorted and indexed using Sambamba (Tarasov *et al.*, 2015) v0.6 and read depths for each base-pair position were derived along the hg19 reference genome using Bedtools (Quinlan and Hall, 2010) v2.23. Note that our goal was to predict the transcriptional signal in the same experimental conditions; same experiment, same histone mark(s), same cell type while holding out subsets of chromosomes. Thus, the experimental conditions were the same for the training and test sets and we did not compare across experiments to avoid any issues with experimental variance in signal/noise ratios or sequencing depths.

Both the promoter and distal models used the averaged log2-transformed 100 base-pair binned signal with our distal model also averaging at 500 and 2,000 base-pairs to also use as

model input features. Using three different resolutions of the histone mark signal in the distal model is intended to represent prior knowledge that epigenetic regulation operates on differing scales and has been shown to improve performance of other models (Schreiber, Durham, *et al.*, 2020). Promoter-capture Hi-C 3D chromatin interaction data (Jung *et al.*, 2019) was incorporated into the distal model and mapped to the Roadmap cell types using the same approach as previously described for Chromoformer (Lee, Yang and Kim, 2022). Reads Per Kilobase of transcript, per Million mapped reads (RPKM) normalised gene expression levels from protein coding genes were downloaded from Roadmap for the matching cell types. Both the promoter and distal models predict the log2-transformed RPKM (log_2(RPKM+1)) gene expression levels and account for gene strandedness. The TSS was identified using RefSeq annotations (release v210)(O'Leary *et al.*, 2016) for each gene. Our final dataset included a total of 18,955 genes.

6.2.2. Promoter model

Our promoter model is a custom convolutional neural network, similar in architecture to DeepChrome (Singh *et al.*, 2016). The model takes in a symmetrical 6,000 base-pair genomic window averaged at 100 base-pair bins, centred on the TSS of the gene of interest. This 6,000 base-pair window was used as a lenient cut-off to include all relevant local regions. The model architecture is composed of three standard convolutional blocks. These blocks each consist of a one-dimensional convolutional layer, batch normalisation, rectified linear unit (ReLU) activation, max-pooling and dropout. This was followed by two fully-connected blocks, which have dropout (in the first block), a dense layer, ReLU activation and a final output layer with linear activation. The convolutional blocks and their sliding window converted the histone mark signal into a position-wise representation highlighting genomic loci that correlate with

expression. The fully connected blocks scaled down the size of the representation, to finally produce a single score representing the gene's RPKM. The size of each layer is provided in our github repository (https://github.com/neurogenomics/chromexpress) (Murphy, 2024).

6.2.3. Distal model

Our distal model architecture was based on the Chromoformer model (Lee, Yang and Kim, 2022). This is an attention-based model which uses cell type-specific promoter capture Hi-C data to identify interacting regions in a 40,000 base-pair genomic window centred on the TSS. This approach captures the histone mark signal both at the TSS and at putative cis-regulatory regions. The model has three independent modules at different resolutions (100, 500 and 2,000 base-pairs), producing a multi-scale representation of the histone mark landscape. Each module goes through a transformer block before being combined and passed through a full-connected block with ReLU activation and a final output layer with linear activation. The architecture of the model is discussed in more detail in the original publication (Lee, Yang and Kim, 2022).

6.2.4. Model training

The same model training approach was used for both the promoter and distal model. Model training and evaluation was based on a 4-fold cross-validation regime to give a stronger estimate of model performance. The total 18,955 genes were split into four sets, 5045, 4751, 4605, and 4554 respectively, with genes from the same chromosome in the same split to avoid data leakage (Whalen *et al.*, 2022). For every fold, one set was used as the blind test set, while the other three sets were used for model training and validation. Performance on

the test set for each fold was measured with Pearson's correlation coefficient. A separate model was trained for each histone mark, cell type and cross-validation fold combination.

The models were trained using the ADAM (Kingma and Ba, 2017) optimiser with default parameters with a batch size of 64 over a maximum of 100 epochs. An early stopping regularisation was implemented based on the model's validation loss with a patience of twelve epochs. The initial learning rate was set at 0.001 and decayed by a factor of 0.2 when the loss did not improve after a patience of three epochs. Mean squared error (MSE) was used as the loss function.

6.2.5. Histone mark levels

Histone mark occupancy was measured separately for our promoter and distal models and for each cell type, histone mark and cross-validation fold. It was measured as the average log2-transformed, 100 base-pair binned read depth of the histone mark signal. For the promoter model, this signal was taken from the 6,000 base-pairs around the TSS of each gene and for the distal model, from the full 40,000 base-pairs.

6.2.6. Gene expression state

Histone mark occupancy was measured for both active and inactive genes. A gene was defined as active or inactive based on whether its expression level was above or below the median for that cell type. This calculation was performed for each cell type independently. This approach was first implemented in DeepChrome (Singh *et al.*, 2016) and has been frequently used in the literature (Singh *et al.*, 2017; Sekhon, Singh and Qi, 2018; Kang *et al.*, 2020; Lee, Yang and Kim, 2022).

6.2.7. Correlation analysis

We measured agreement in model performance for both our promoter and distal models by matching the cross-validation fold and cell type for each model trained on a (or pair of) different histone mark(s). The Pearson correlation coefficient was used to quantify agreement. The performance across different cell types and folds for each mark (or pairs of marks) was aggregated to get the reported mean Pearson R and standard deviation. When comparing to the distal model based on all histone marks, the mean performance for each cell type - after aggregating based on the different folds - was used to report mean Pearson R and standard deviation.

6.2.8. *In silico* perturbation of histone mark activity levels

In silico histone mark perturbation was performed on the distal model trained on a single mark. Although we have trained Chromoformer with multiple histone marks as input, we chose to use it trained on a single mark for the perturbation analysis since perturbing one mark will likely affect the histone mark occupancy of other marks in the same region which would not be possible to accurately account for in the model.

We selected one active and one repressive mark which are found at both promoter and distal regulatory elements - H3K27ac and H3K27me3. Perturbation experiments were carried out on active genes for the active mark model and inactive genes for the inactive mark model (see Methods - Gene expression state), to measure the effect on expression of reducing the levels of the histone mark. The predictions from the different k-fold versions of the model were averaged, similar to the approach commonly used in sequence to expression models (Karollus, Mauermeier and Gagneur, 2023; Linder *et al.*, 2023; Sasse, Ng, A. E. Spiro, *et al.*, 2023; Alan E. Murphy *et al.*, 2024). For the promoter histone signal, the full 6,000 base pairs

around the TSS were perturbed, whereas for the distal histone signal, bins of 2,000 base pairs across the 40,000 base pair receptive field were perturbed iteratively (similar to the approach for DNA sequence perturbation used by the genomic deep learning model CRÈME (Toneyan and Koo, 2024)). The implemented perturbation levels were between 0 and 1 inclusively in 0.1 steps. The code to perform the *in silico* histone mark perturbation is available on our github repository (https://github.com/neurogenomics/chromexpress) (Murphy, 2024).

As well as averaging the predictions from the different k-fold versions, we also tested the correlation between the different folds to ensure the model is learning consistent regulatory code. Moreover, we benchmarked this concordance against Borzoi (Linder *et al.*, 2023), a genomic deep learning model with the current largest receptive field of 524,000 base pairs. Here, for a fair comparison, we only tested Borzoi's concordance in the centre 40,000 base-pairs to match Chromoformers receptive field, for the RNA predictions in the same cell types as those used in our analysis and added up to 500 random genetic variants to the sequences of 1,000 genes to match our perturbation test.

6.2.9. *In silico* perturbation enrichment in quantitative trait loci studies

The averaged *in silico* perturbation experiments on the active (H3K27ac) model for each cell type were filtered to those greater than 6,000 base-pairs upstream and 4,000 base-pairs downstream of the TSS to concentrate on distal, cell type-specific regulatory regions as opposed to promoter regions. To control for the greater effect of changes in gene bodies (**Figure 6-7e)**, which have a median length of ~25,000 base pairs (Grishkevich and Yanai, 2014) - longer than the downstream receptive field of the model - these predictions were sorted based on their predicted change in expression and split into deciles separately for upstream and downstream regions.

The fine-mapped eQTL data based on the UK Biobank population was sourced from Wang et al., 2021 (Q. S. Wang *et al.*, 2021). Causal SNPs were identified from those in LD using FINEMAP v1.3.1 (Benner *et al.*, 2016) and SuSiE v0.8.1 (Wang *et al.*, 2020). The resulting fine-mapped SNPs were filtered to those with a SuSiE causal probability (posterior inclusion probability (PIP)) > 0.9 in the tissue of interest and with a PIP <0.1 in other tissues to get just the high confidence, tissue-specific fine-mapped SNPs. The ROADMAP cell types were matched to five of the tissues used in eQTL study where the tissue assayed were identical across the two (available on our github repository: https://github.com/neurogenomics/chromexpress) (Murphy, 2024).

To test for enrichment of the fine-mapped, tissue-specific SNPs, a bootstrap sampling experiment was implemented where the proportion of SNPs found in each decile were compared against 10,000 randomly sampled regions from all deciles. P-values were derived and adjusted using FDR correction for multiple testing.

Since the distal model uses Hi-C chromatin interaction data as input, we also ran this eQTL enrichment test on the matched cell type and gene, significant promoter capture Hi-C interactions (2,000 base-pair resolution) to compare against the model's enrichment performance. To match the model's tested regions, the chromatin interaction data was filtered to just those upstream of the gene. Furthermore, we benchmarked against the regions of maximum histone mark activity for each gene up to 20,000 base-pairs upstream and downstream of the TSS, averaged at 2,000 base-pairs to match the Hi-C and model approach, and also against a proximal loci of the 6,000 base-pairs just upstream or downstream of the TSS. Scripts detailing the approach are available on our github repository (https://github.com/neurogenomics/chromexpress) (Murphy, 2024).

6.2.10. *In silico* perturbation disease enrichment

To test for disease enrichment, the top decile of the same averaged *in silico* perturbation experiments on the active model, filtered to just those upstream or downstream of the TSS, were considered. For this analysis, predictions in the liver and neuronal progenitor cells (NPCs) were used due to their potential respective relationships with liver and neuronal diseases. A third group of regions comprising the top decile across all cell types was included to look for cell type-consistent disease enrichment.

Summary statistics for GWAS for liver diseases - non-alcoholic fatty liver disease (NAFLD) (Ghodsian *et al.*, 2021) and hepatitis (Dönertaş *et al.*, 2021), glial diseases - Parkinson's (Nalls *et al.*, 2019) and Alzheimer's (Bellenguez *et al.*, 2022b) and neuronal diseases - amyotrophic lateral sclerosis (ALS) (Nicolas *et al.*, 2018), schizophrenia (Trubetskoy *et al.*, 2022), autism spectrum disorder (Grove *et al.*, 2019) and bipolar disorder (Mullins *et al.*, 2021) were downloaded from the IEU GWAS portal (Elsworth *et al.*, 2020) and the BioStudies database (Sarkans *et al.*, 2018) and were uniformly processed with MungeSumstats v1.11.3 (Murphy, Schilder and Skene, 2021) (default settings, converting build to hg19 where necessary and saving in 'LDSC' format).

We applied s-LDSC v1.0.1 (https://github.com/bulik/ldsc) to test for disease enrichment (Finucane *et al.*, 2015). Specifically, annotation files for each of the three groups of genomic loci were first created with Phase 3 of the 1000 genomes reference. Followed by the generation of LD scores with a window size of 1 cM i.e. approximately 1 million base pairs, filtering to HapMap3 SNPs to match the baseline model. Finally, the enrichment analysis was run for the GWAS summary statistics across the three groups as well as those in the baseline model whilst excluding the major histocompatibility complex (MHC) (due to the known difficulties predicting LD in this region) (Finucane *et al.*, 2015).

6.3. Results

6.3.1. Active histone marks prove most informative at promoter regions

We first focused on the performance of histone mark levels from the promoter regions of the gene of interest using our promoter model (**Figure 6-2**). As a naïve, baseline performance measure we compared performance against absolute correlation of histone mark and gene expression. Each histone mark was averaged at different local distances around the transcriptional start site (1,500, 3,000 and 6,000 base-pairs) to capture the best possible promoter correlation. Our promoter model outperformed these baselines for all histone marks, which was largely expected due to neural network's ability to capture non-linear relationships, with notably better performance in regressive marks. Moreover, the lenient inclusion of 3,000 base-pairs up and down stream of the TSS in the promoter model led to improved performance over a smaller region more focused on the putative gene promoter functional regions, i.e. 2,500 base-pairs upstream and 500 base-pairs downstream of the TSS (**Fig. B-1**).

Overall, we found H3K4me3, a mark located in active promoter regions (Schneider *et al.*, 2004), to be the best performing from our promoter model (**Figure 6-2a**). Moreover, active promoter marks H3K4me3, H3K27ac (Herrera-Uribe *et al.*, 2020) and H3K9ac (Karmodiya *et al.*, 2012) made up the three top performing marks, all with a Pearson's correlation above 0.73. The fourth best performing histone mark, H3K4me1, is found at active enhancers (Bae and Lesch, 2020). It likely performed worse than the other active marks due to the limited range of the model, which only took into consideration a gene's promoter region.

Repressive marks proved less informative with H3K9me3 (Foret *et al.*, 2014), H3K27me3 (Cao *et al.*, 2002) and H3K36me3 (Soldi *et al.*, 2017) making up the three worst performing marks. Importantly, there was high variability in performance across the histone marks with a

correlation difference of 0.25 between the best active and worst repressive mark (range of Pearson correlation coefficients: 0.52 - 0.76).

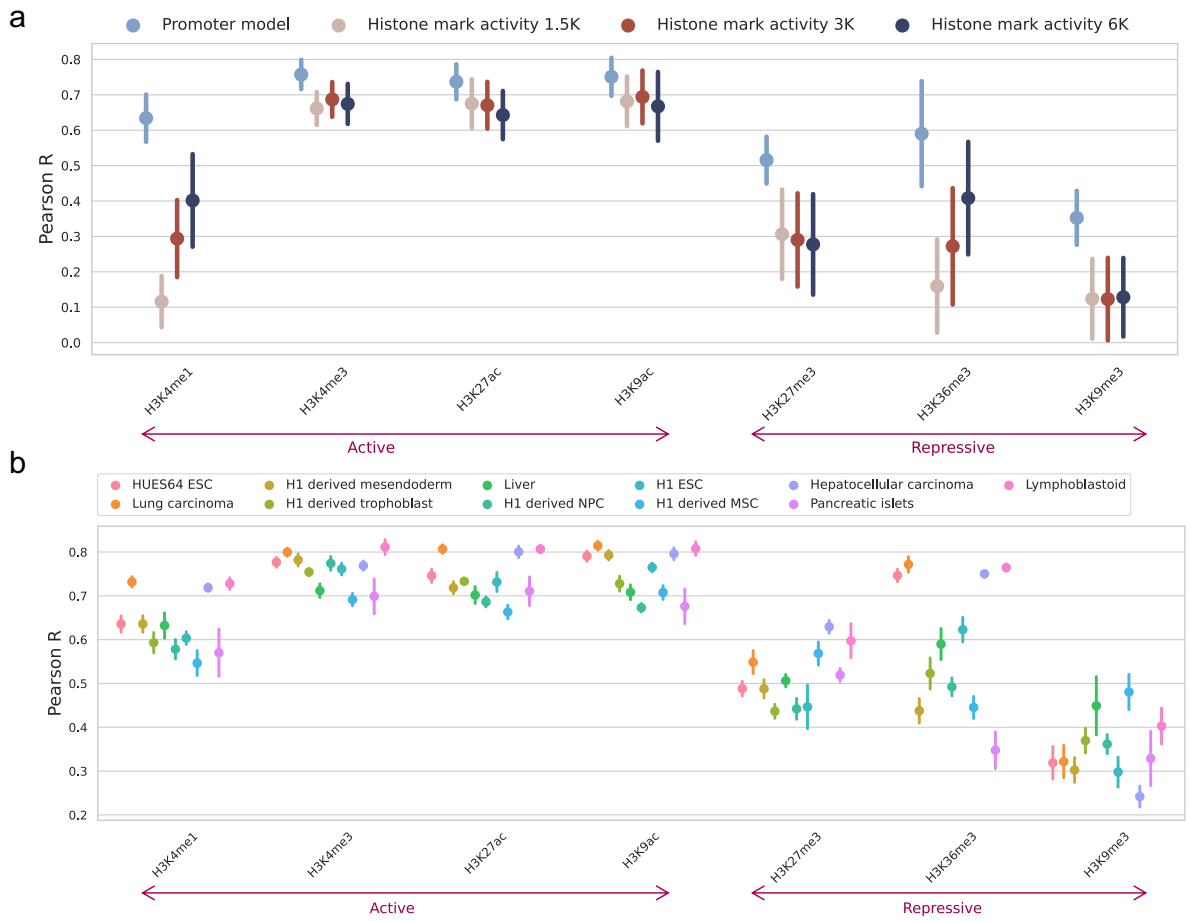


Figure 6-2 The performance of the promoter model shows the best gene expression predictions based on active histone marks H3K4me3, H3K27ac and H3K9ac.

(a) Promoter model performance (pale blue) measured by the Pearson correlation coefficient on the blind test sets across each histone mark. Also shown is the absolute correlation of each histone mark with expression at different regions around the transcriptional start site (1,500, 3,000 and 6,000 base-pairs) to act as a baseline performance measure. The whiskers represent the standard deviation across the different cell types and the 4-fold cross-validation. (b) Promoter model performance split by different cell types from Roadmap.

6.3.2. The predictive performance of active and repressive marks differ based on cell state at promoter regions

Splitting the models' performance across the different cell types highlighted histone mark groups with similar variations in their scores across cell types (**Figure 6-2b**). This is most notable for active promoter marks (H3K4me3, H3K9ac and H3K27ac). To formally evaluate this trend, we calculated the correlation between the models' predictions across the different genes, cell types, histone marks and cross-validation folds (**Figure 6-3**). One distinct group of highly correlated histone marks were apparent (highlighted in blue in **Figure 6-3a**), corresponding to the active histone marks previously observed. Interestingly, H3K9me3 showed the lowest correlation with the other histone marks, including with H3K36me3 and H3K27me3, the other repressive marks. The samples collected for Roadmap (Kundaje *et al.*, 2015) can be classified into those taken from Embryonic stem cells (ESC), cells differentiated from ESCs, adult bulk tissues and cancer cell lines (**Table B-1**). Active histone mark activity levels were significantly more predictive of expression in ESC than primary tissue whereas the opposite was noted for repressive histone mark activity levels which was more predictive in primary tissues than ESC (**Figure 6-3b**). The multi-modal performance, visible in **Figure 6-3b**, is the result of a combination of the histone mark and cell type being tested (**Fig. B-2**). These results indicated that the most accurate method by which to predict gene expression from the promoter region depends on the extent to which the cell type of interest has differentiated - active marks like H3K4me3, H3K9ac and H3K27ac were most predictive for cells at earlier stages of their differentiation process, including ESCs. In contrast, repressive marks, like H3K9me3, fared better in fully differentiated tissues or cells.

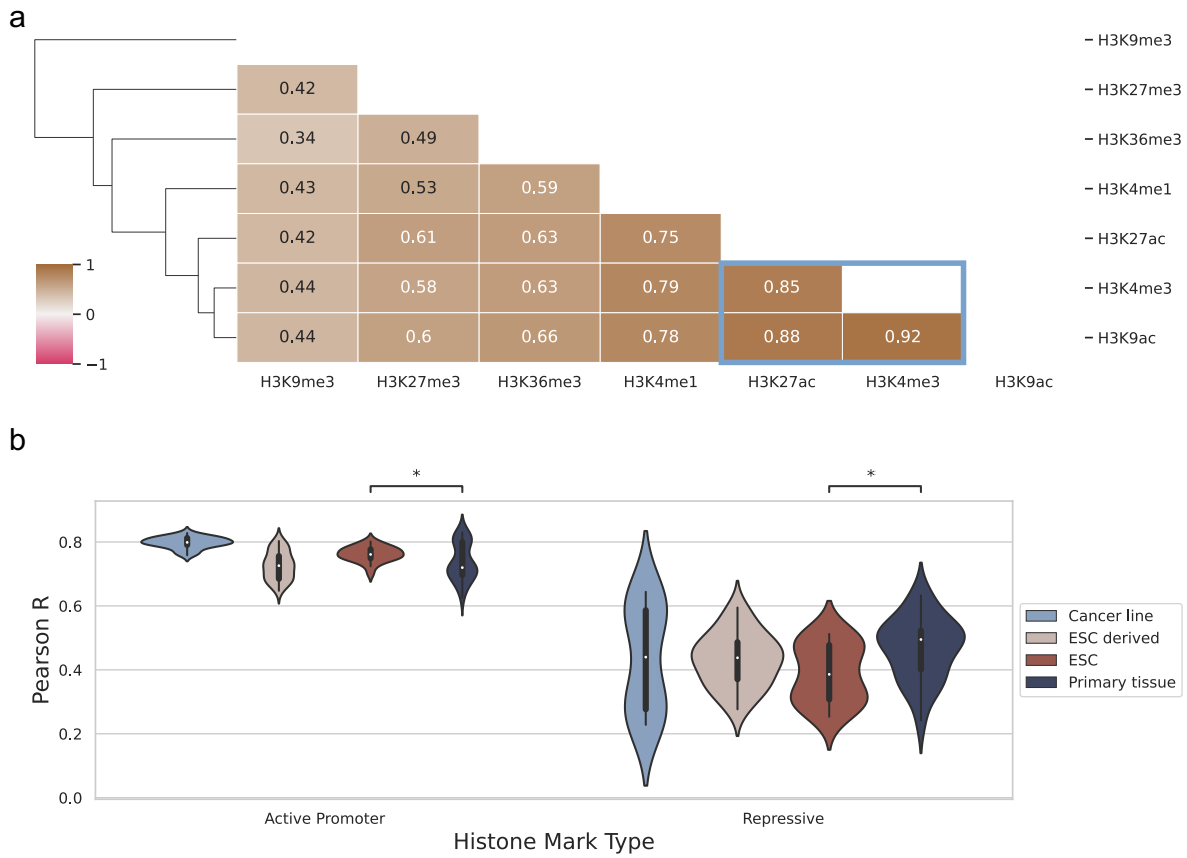


Figure 6-3 The predictive performance of the promoter model across samples clusters by histone mark function.

(a) Correlation matrix of the promoter model's predictions by the different histone marks across each gene, cell type and cross-validation fold. H3K4me3, H3K9ac and H3K27ac were characterised by high positive pairwise correlations (highlighted in blue). Bars along the y-axis show the hierarchical clustering dendrogram. (b) Violin plot of the model's performance on active and repressive histone marks measured by their Pearson correlation coefficient on the blind test sets. The cell type performance is grouped by the cell state - Embryonic stem cell (ESC), ESC derived cell, adult primary tissue or cancer cell line. A Wilcoxon rank-sum test was used to compare ESC and primary tissue performance for active and repressive marks. Significance was based on a false discovery rate (FDR) correction for multiple testing, with p-value indicator: * < 0.05.

6.3.3. Higher histone mark levels result in better predictive performance at promoter regions

We first queried whether the promoter model's performance was just a result of the quantity of histone mark signal observed, rather than learnt histone mark and cell type-specific regulatory relationships. H3K4me3 did follow this relationship with the highest levels of activity (**Fig. B-3a**), best performance and positive correlation between them (**Fig. B-3b-c**). However, on other marks, including H3K27ac, the model showed consistent performance irrespective of levels of training or test activity (**Fig. B-3b-c**). Moreover, H3K27me3 displayed a negative relationship, where more activity led to worse performance. If simply more histone mark activity led to better performance overall, a strong correlation would have been observed for all histone marks.

To investigate what was driving the difference in performance and how this related to active and repressive histone marks in the different cell states, we measured the average histone mark levels in the promoter regions for highly and lowly expressed genes (**Figure 6-4**, see **6.2 Methods** for details). We observed a strong correlation between model performance and histone mark levels for active histone marks in highly expressed genes, and conversely, for repressive marks in lowly expressed genes, indicating that the model learns the functional significance of the different histone marks (**Figure 6-4a**). For highly expressed genes (**Figure 6-4b**), repressive marks, H3K9me3, H3K36me3 and H3K27me3, had higher histone mark levels in ESCs than primary tissue (although for H3K9me3 this was not significant after multiple test correction). Conversely, active marks, H3K4me3, H3K9ac and H3K27ac, showed varying activity across primary tissue and ESCs. For lowly expressed genes (**Figure 6-4c**), histone mark levels tended to be higher in ESCs than in primary tissue across histone marks. Overall, our analysis highlighted that higher histone mark levels in a gene where the

expression status matched the function of the histone mark (active vs repressive), led to better performance of the model. RNA-Seq typically has a bias for 3' ends, which might lead to underestimation of abundance of some long transcripts (Schwanhäusser *et al.*, 2011; Tani *et al.*, 2012). To account for this, we also tested whether there was a bias in our model's predictions based on transcript lengths (**Fig. B-4**). There appeared to be no relationship, suggesting our model's performance is not biased against genes with long transcripts.

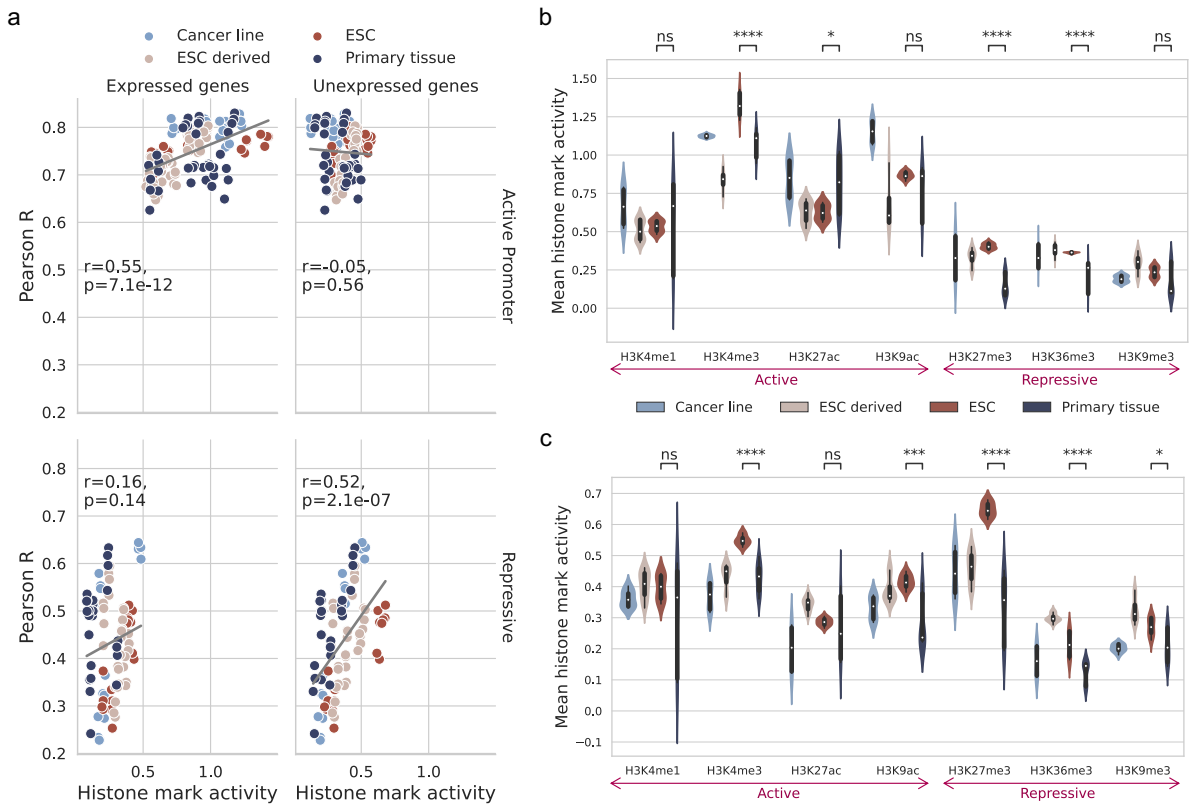


Figure 6-4 Higher histone mark levels are associated with better predictive performance in specific gene expression states.

(a) Correlation between the histone mark levels and model performance for each cell type, k-fold combination. This is split by repressive and active marks, as well as highly and lowly expressed genes. The histone mark activity is firstly averaged across the receptive field of the model then it is averaged separately for high/low expression genes for every cell for each of the 4 fold versions of the model. (b) Violin plot of histone mark levels measured by the average log2-transformed, read depth in the promoter region (6,000 base-pairs around the transcription start site) for highly expressed genes. The cell types are grouped by the cell state - Embryonic stem cell (ESC), ESC derived cell, adult primary tissue or cancer cell line and averaged at the level of cell type and k-fold. (c) Average histone mark levels for lowly

expressed genes. Significance was based on the false discovery rate (FDR) multiple test correction, with p-value indicators: **** < 1e-4, *** < 1e-3, ** < 1e-2, * < 0.05, ns >= 0.05.

6.3.4. Active marks are most predictive in the distal model

To consider histone mark levels outside of the genes' promoter regions, we next tested a model architecture with a much larger receptive field (up to 40,000 base-pairs around the TSS). We used Chromoformer (Lee, Yang and Kim, 2022), a transformer-based architecture, which accounts for distal histone mark levels, weighting important genomic regions using cell type-specific DNA interaction data. We trained this distal model on each single histone mark and benchmarked the performance across histone marks, and also against the performance when trained on all seven histone marks combined (**Figure 6-5a**). Again, we found H3K4me3, H3K9ac and H3K27ac to be the top three performing marks with very little difference in overall performance between them (mean Pearson R of 0.762, 0.757 and 0.749 respectively). All three histone modifications are active marks while only H3K27ac (Herrera-Uribe *et al.*, 2020) is found at both promoters and distal regulatory regions (enhancers).

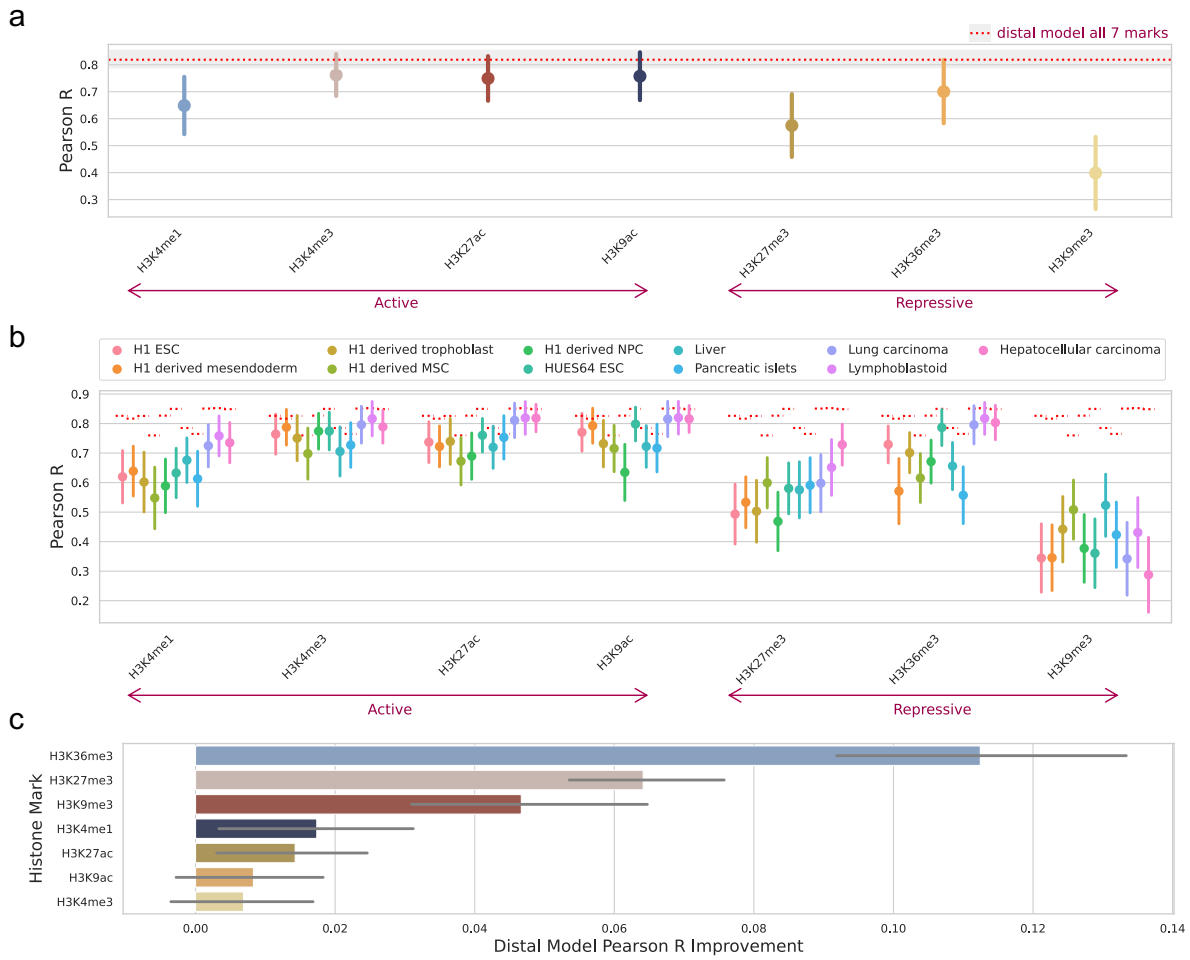


Figure 6-5 Active histone modifications performed best at predicting gene expression in the distal model.

(a) Distal model performance was measured by the Pearson correlation coefficient on the blind test sets for each histone mark. The whiskers represent the standard deviation across the different cell types and the 4-fold cross-validation. The red dashed line and shaded box shows the model's mean performance and standard deviation when trained on all seven histone marks together. (b) Predictive performance is shown split by the different cell types in Roadmap. (c) The improvement in performance for each histone mark by expanding the receptive field outside of the promoter region with the distal model. The largest increase in

performance is observed for H3K36me3. The whiskers indicate the standard deviation across cell types and folds.

6.3.5. Receptive field expansion and multi-histone mark predictions yield diminishing returns.

The performance of incorporating distal histone mark levels in a model consistently but marginally increased the Pearson correlation coefficients. The range of improvement in correlation varied from 0.01 - 0.15, despite the substantial increase in receptive field and model architecture complexity (**Figure 6-5c**). Compared to the promoter model, although the same histone mark performance ranking was observed, marks which are known to affect genomic locations outside of promoter regions showed the highest relative improvement, specifically H3K36me3, H3K27me3 and H3K27ac (**Figure 6-5c, Table 1-1**).

The relationship between histone mark type and cell state found for the promoter model, where active histone marks were more predictive in ESC and repressive in primary tissues, was similarly observed for the distal model (**Figure 6-5b, Fig. B-5a-b**). Moreover, we investigated the histone mark levels across the full 40,000 base-pair receptive field of the distal model and found the same trend as for the promoter model where the model picks up on known biology of histone mark prevalence: We observed strong correlations between model performance and histone mark levels for active histone marks around highly expressed genes, and conversely, for repressive marks around lowly expressed genes (**Fig. B-5c**).

To further interrogate the contributions of histone marks to the prediction of expression in our distal model, we benchmarked performance across pairs of histone marks with the top three performing marks, H3K4me3, H3K9ac and H3K27ac (**Figure 6-6a**). All combinations with the top three histone marks were better performing than any of the top three marks by themselves. However, this improvement was relatively small (< 0.04 mean increase in Pearson R for the best histone mark combination) and was only significant for a handful of combinations.

This result highlights that incorporating additional histone marks in the prediction led to a consistent, albeit small, boost in performance, regardless of the histone mark type (active or repressive). This became most evident when we combined pairs of the top three performing marks: All three are active marks and two are confined to promoter regions, but their combination still resulted in improved performance over the individual marks (**Figure 6-6b**). Notably though, combinations including H3K27ac, an enhancer mark, showed greater improvement over combinations including only promoter marks. Importantly, combining the top three marks with another mark showed a mean improvement of 0.043, compared to a mean improvement of 0.062 when using all marks (red dashed line in **Figure 6-6a**). This means that adding an additional five histone marks to the distal model would increase performance (Pearson R) by another 0.02, highlighting the diminishing returns of including additional histone mark information and giving insight for researchers looking to best capture gene regulation in their cell type of interest with minimal experimental work in the future.

Our analysis further highlighted the combinatorial predictive capabilities of H3K36me3 (**Figure 6-6a,c**). H3K36me3 is a repressive mark (Chantalat *et al.*, 2011) with strong distal effects on gene expression. This mark showed the largest improvement from the promoter to the distal model (**Figure 6-5c**) and when combined with the top three scoring histone marks, was its best performing pair, even when compared to combinations of the top three performing marks (**Figure 6-6a**) and far improved performance compared to the addition of the other marks (**Figure 6-6c**). Conversely, when paired with the repressive mark H3K27me3, H3K36me3 did not result in the best performing pair (**Fig. B-6**) but did still improve performance on the individual mark. This highlights the complementary information H3K36me3 provides in addition to the top three performing active promoter and enhancer marks.

We also investigated whether the performance increase for the top three marks when coupled with H3K36me3 was driven by bivalent genes. Bivalent genes are characterised by the presence of both repressive and active histone mark signals at their promoter and are known to silence developmental genes in ESCs while keeping them poised for activation (Bernstein *et al.*, 2006). However, the model performance in bivalent genes for ESCs did not notably improve over non-bivalent genes (**Fig. B-7**).

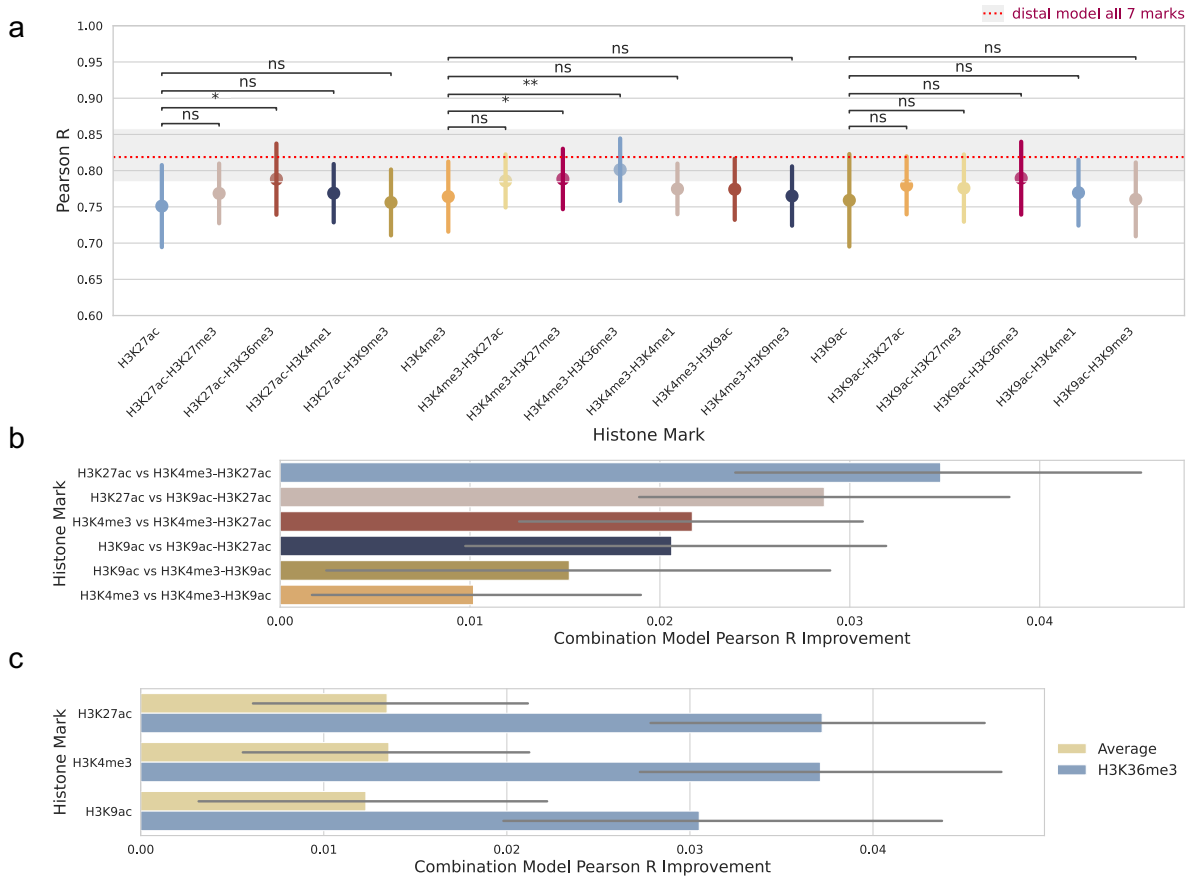


Figure 6-6 Pairwise combinatorial performance of distal model in predicting gene expression based on two histone marks.

(a) Paired distal model performance was measured by the Pearson correlation coefficient on the blind test sets across combinations of histone marks. The whiskers represent the standard deviation across the different cell types and the 4-fold cross-validation. Data is averaged at the level of cell type and k-fold. The red dashed line and shaded box shows the model's mean performance and standard deviation when trained on all seven histone marks together. Significance based on false discovery rate (FDR) multiple test correction, with p-value indicators: ** < 1e-2, * < 0.05, ns >= 0.05. (b) Performance improvement for combinations of the top three performing marks from the single histone mark distal model. (c) Performance improvement over the single histone mark distal model for combinations using H3K36me3 versus the average of the other marks.

6.3.6. *In silico* histone mark perturbation prioritises functionally relevant genomic loci and disease relevant cell types

Up to this point, our work has shown the performance of chromatin to expression deep learning models and how they encode known biological relationships between histone mark levels and expression. However, none of this highlights new information about gene regulation or disease. *In silico* perturbation enables the investigation of the effect on gene expression of varying histone mark levels in a cell type and gene-specific manner that would be impractical to undertake experimentally *in vivo*. This analysis was inspired by the recent development of systematic histone mark editing toolkits to measure the relationship with expression (Policarpi *et al.*, 2024) and follows a comparable approach of epigenomic editing but *in silico*. Our *in silico* perturbation experiments enable the systematic, high throughput quantification and comparison of the effect of histone mark activity at different genomic loci on gene expression.

We first investigated the effect of *in silico* perturbation experiments at an aggregate level, varying levels of histone mark activity, as well as distances from the TSS using the distal model trained on one active (H3K27ac) and one repressive (H3K27me3) mark (**Figure 6-7**). We used models trained on single marks to avoid issues where perturbing one type of histone mark will affect another mark's activity in the region. Here, we permuted either the entire TSS or the distal regions in 2,000 base pair bins (see **6.2.8 *In silico* perturbation of histone mark activity levels**). Furthermore, we averaged predictions across the four k-fold model versions, a standard approach *in silico* mutagenesis experiments for genomic deep learning models (Karollus, Mauermeier and Gagneur, 2023; Linder *et al.*, 2023; Sasse, Ng, A. E. Spiro, *et al.*, 2023). This step may not have been required given the notably high correlation between the different models' predictions (**Fig. B-8a**), which was on par with, if not slightly better, than

that of the genomic deep learning model Borzoi (Linder *et al.*, 2023) (**Fig. B-8b**). This indicates that models trained on histone mark levels show a similar ability to learn consistent regulatory code across differing training sets compared to that of their genomic counterparts.

For the active model, we noted a clear relationship between reducing histone mark levels and large predicted decreases in expression at the promoter (**Figure 6-7a**) while we found only minor decreases at distal regulatory regions (**Figure 6-7b**). This matches the known functional relationship between promoter and enhancer activity with expression and also the *in silico* perturbations of DNA sequence in genomic deep learning models (Karollus, Mauermeier and Gagneur, 2023). However, reducing repressive histone mark levels showed little relationship with increased expression (**Figure 6-7c-d**) which may be due to the repressive mark's relatively worse performance overall (**Figure 6-5a**). Moreover, it highlights how the lack of a repressive mark is not sufficient to confer expression to a gene, but rather it is additionally associated with the presence of an active mark which we seen through the performance of the repressive mark H3K36me3 in combination active marks (**Figure 6-6c**). The relationship between perturbation and distance and their effect on expression is more apparent when we view the predicted quantile change in expression by distance from the TSS while removing histone mark activity completely (**Figure 6-7e-f**). Here, we saw the highest predicted change near the TSS, reducing as distance to the TSS increases for both the active and repressive model. Interestingly, this reduction was not symmetrical upstream and downstream of the TSS, with downstream loci having a greater effect on expression on average. We believe this was a result of the length of the gene body (median length ~25,000 base pairs (Grishkevich and Yanai, 2014)), which would incorporate the entire downstream receptive field of the model for most genes and thus lend to greater importance for RNA-Seq predictions rather than assays of promoters like CAGE-Seq (Takahashi *et al.*, 2012) or transcription initiation like

PRO-Seq (Mahat *et al.*, 2016). Our analysis highlighted that on average, perturbations to histone mark signals in the gene body had a greater predicted effect than distal regulatory regions upstream. These results held consistently across all cell types and cell states for the active marks. However, this was not the case for repressive marks, which differed vastly by cell state (**Fig. B-9**). We intentionally chose an active (H3K27ac) and repressive (H3K27me3) mark for analysis that are known to regulate expression outside of the TSS. We confirmed that a mark that is primarily associated with the TSS (H3K4me3) has higher predicted change in expression in the TSS and lower change downstream and in the gene body (**Fig. B-10**).

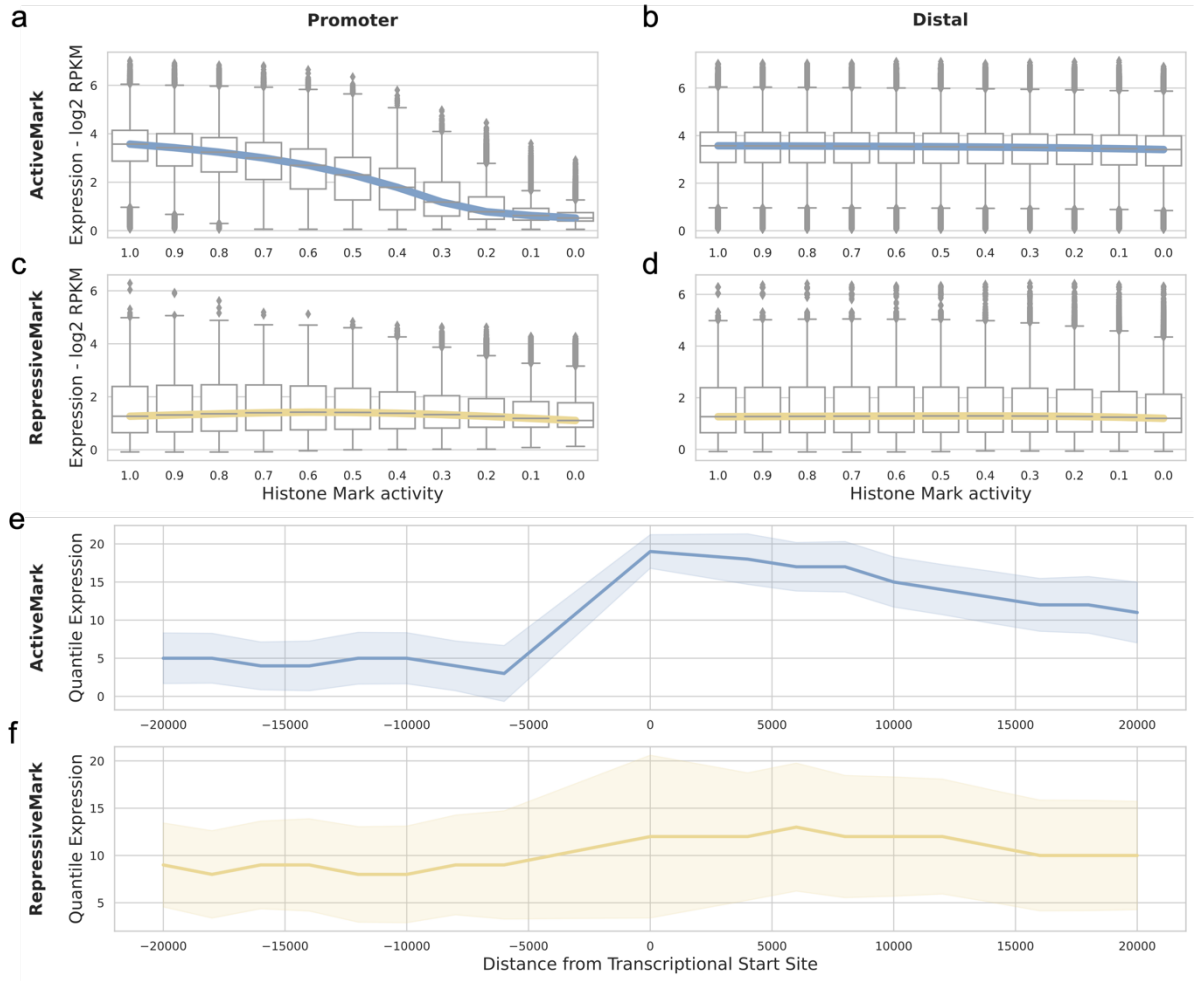


Figure 6-7 Effect of in silico histone mark perturbation on expression.

The effect on predicted expression (y-axis) of changing the proportional levels of measured histone mark activity (x-axis) for all cell types and genes, averaged across the four k-fold models. The distal model trained on a single histone mark was used to measure the effects of a perturbed active mark - H3K27ac (a,b) or a repressive mark - H3K27me3 (c,d) in 2,000 base pair bins for distal or at the promoter. (e,f) The effect of distance on expression change is shown when the histone mark activity is completely removed at a specific location for the active (e) or repressive (f) mark. The distribution of all gene expression changes in all cell types split into 20 quantile bins.

We next considered whether these *in silico* perturbation experiments could be used to gain insight into the cell type-specific regulatory function in gene expression and disease, using genetic variants to test for functional and disease enrichment (**Figure 6-8**). Given that the active model captured known biological relationships in the *in silico* perturbation, we focused on this model's perturbation experiments. Moreover, we only considered upstream predictions to capture distal regulatory regions which vary across cell types as opposed to promoter and gene body signals.

To test the model's ability to capture functional loci we used a large scale, tissue-specific, fine-mapped eQTL set based on the UK Biobank population (Sudlow *et al.*, 2015). First, we split the loci into deciles sorted based on the model's predicted change in expression (decile 10 having the largest predicted effect). We implemented a bootstrap sampling test to compare each decile to randomly sampled upstream and downstream loci for enrichment of the fine-mapped eQTLs (see **6.2.9 *In silico* perturbation enrichment in quantitative trait loci studies**). We found significant enrichment for the top decile for all but one cell types tested (**Figure 6-8a**), indicating that the model correctly predicted the loci which contributed most to the cell type-specific gene expression. Furthermore, to provide benchmarks against alternative loci prioritisation methods, we tested how much the distal model improves upon the 1. regions of highest histone mark activity, 2. Hi-C chromatin interacting loci and 3. proximal loci (<=6000 base-pairs up or down from the TSS, inspired by Wang *et al.*'s distance baselines (Wang *et al.*, 2024)) for fine-mapped eQTL enrichment in these same loci (**Fig. B-11**). With the exception of the lung tissue sample, the model's loci prioritisation approach outperformed the region of maximum histone mark activity and Hi-C in all tissues (**Fig. B-11a-b**) and matched the performance of the <=6000 base-pairs approach (**Fig. B-11c**).

We next tested whether these loci also harbour known disease related genetic variants. We used s-LDSC (Finucane *et al.*, 2015) with GWAS for eight different liver and brain diseases (Nicolas *et al.*, 2018; Grove *et al.*, 2019; Nalls *et al.*, 2019; Dönertaş *et al.*, 2021; Ghodsian *et al.*, 2021; Mullins *et al.*, 2021; Bellenguez *et al.*, 2022b; Trubetskoy *et al.*, 2022). Test regions were based on the top cell type-specific decile with the largest predicted effect for liver and NPCs, as well as the averaged top decile across all cell types of upstream and downstream loci (**Figure 6-8b**). S-LDSC measures enrichment of disease genetic variants accounting for the obscuring nature of LD (Finucane *et al.*, 2015). Here, we used the neuronal cells and liver tissue for their predicted relationship with brain and liver diseases, respectively. We used the top decile for all cell types to look for non-cell type-specific disease enrichment. While none of the expected cell type-specific significant enrichments were detected, the NAFLD GWAS was tending towards significance in liver tissue. Moreover, one significant association for bipolar disorder was detected after multiple testing correction for the non-cell type-specific disease enrichment, highlighting the functional importance to disease of regions where changes in the histone mark activity levels was predicted to have a large effect on expression across cell types.

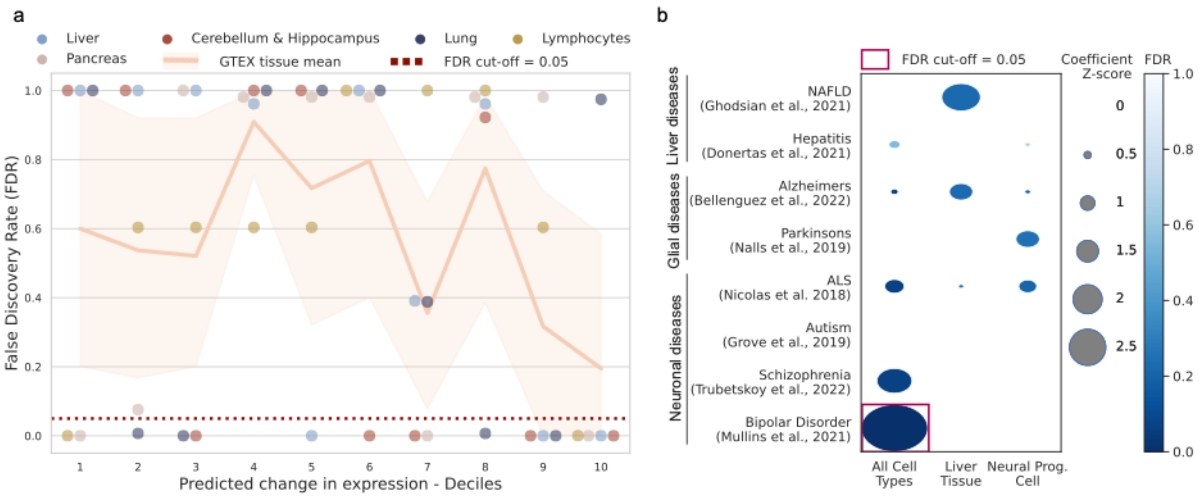


Figure 6-8 In silico histone mark perturbation experiment highlights functional and disease enrichment.

(a) Upstream and downstream *in silico* histone mark perturbation experiments from the active model were sorted into deciles based on their predicted change in expression (x-axis). Each decile was tested for enrichment of fine-mapped eQTL interactions in matched cell types and compared against bootstrap sampling random upstream loci 10,000 times to generate p-values of enrichment (y-axis). (b) False discovery rate (FDR) P-value enrichment scores (colour) and coefficient Z-score (size) for the top decile (based on their predicted change in expression) derived from all cell types (non-cell type-specific), liver tissue or neuronal progenitor cells (x-axis). Enrichment tests were conducted with s-LDSC and genetic variants relating to different diseases (y-axis). FDR adjustment was applied for all GWAS tested.

Results of the *in silico* perturbation analysis can be visualised and investigated locally by approximating the effect histone mark activity at each genomic locus has on the predicted gene expression by calculating the partial derivatives of the model with respect to the input, i.e. the gradient on the input (Hechtlinger, 2016). Using this approach, we uncovered a regulatory region upstream of *DSTN* in the lung carcinoma cell line (**Figure 6-9).** This locus was in the top decile of predicted effects on expression for this cell type from the *in silico* perturbation analysis. The region is shown to have a small effect across the three input resolutions our distal model uses prior to perturbing the H3K27ac signal, but once removed, has a large effect. This locus contains the fine-mapped SNP rs611572 from the UK Biobank population eQTL analysis in a matched cell type (Sudlow *et al.*, 2015) and *DSTN* has been noted to promote malignancy in lung adenocarcinoma. It has potential as a prognostic marker and therapeutic target (Zhang *et al.*, 2020), confirming its predicted functional importance based on our findings. Of note, we see a similar relationship when visualising the embedding attention matrix, where the weights for this region go to zero after perturbation, indicating the effect of the lack of histone mark signal (**Fig. B-12**).

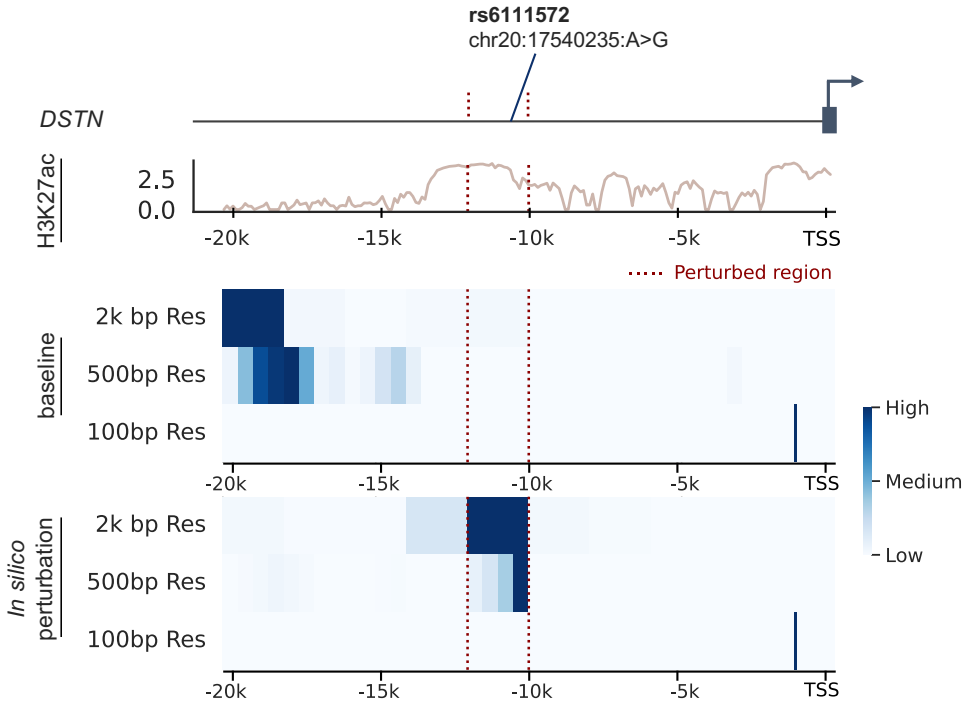


Figure 6-9 In silico histone mark perturbation visualised for DSTN in lung carcinoma cell line.

Upstream *in silico* histone mark perturbation of *DSTN* in the lung carcinoma cell line using the distal model trained on the H3K27ac signal. The experimental H3K27ac signal is shown along with an approximation of the effect each region of histone mark activity has on the prediction, calculated using the partial derivatives of the model with respect to the input, before and after ablation of the H3K27ac signal in the noted region. The fine-mapped SNP rs611572 from the UK Biobank population eQTL analysis in a matched cell type is also shown(Sudlow *et al.*, 2015). This region was chosen as the *in silico* perturbation analysis identified it in the top decile of predicted effects on expression for this cell type.

6.4. Discussion

We present the most comprehensive deep learning study of the relationship between histone mark levels and transcription to date. We considered multiple cell types and histone marks at differing receptive fields and demonstrated how the prediction of gene expression is dependent on three key contributing factors - histone mark function, regulatory distance and cellular states.

For our analysis, we used the Roadmap (Kundaje *et al.*, 2015) data repository, benefiting from the standardised experimental approach. We investigated the genome-wide activity of seven histone marks across eleven cell types. For the histone mark ChIP-Seq read alignments, we subsampled to 30 million reads to enable a fair comparison across marks. While for gene expression, we utilised RPKM values which measures the mRNA abundance of transcripts normalised by gene length, avoiding any potential within sample bias. Since model predictions were made in the same cell type as training, there was no need to standardise gene expression levels across cells (Zhao, Ye and Stanton, 2020). The use of mRNA abundance here is one limitation of the study as it may not be reflective of true cellular regulation. RNA-Seq read counts are impacted by mRNA turnover, stability and nuclear export, not solely deriving from transcriptional changes (Schwanhäusser *et al.*, 2011; Tani *et al.*, 2012). Thus, it is possible that the noted limited correlation between transcription and repressive histone marks is a result of this (**Figure 6-2a, Figure 6-5a**). For instance, particularly stable mRNAs might obscure the effects of transcriptional repression, potentially leading to underestimates of the actual repressive contribution of histone marks at these specific loci. Future work may benefit from investigating measures of transcription which avoid these issues, for example by profiling transcription initiation by assays such as precision nuclear run-on sequencing (PRO-Seq) (Mahat *et al.*, 2016).

To ensure robust benchmarking results, we repeated training of both our promoter and distal models across a 4-fold cross-validation, ensuring the test set genes were grouped by chromosomes to avoid any data leakage - where data in the training set is related to data in the test set, artificially inflating model performance (Whalen *et al.*, 2022). To run each cell type, histone mark or combination of histone marks, for each cross-validation fold using both the promoter and distal models, was a computationally intensive task. This resulted in 1,276 training and prediction iterations which were all run using an A100 80GB GPU. To avoid overfitting over such a substantial number of iterations, we automated hyperparameter tuning for both models using a learning rate decay and early stopping regime, holding out an independent validation set of genes for monitoring.

Both our promoter and distal models were developed as quantitative, regression models, predicting a gene's $\log_2$ RPKMs, which has been shown to yield better generalisation and interpretability than binary classification models (Toneyan, Tang and Koo, 2022). For the promoter model, following the same approach as past benchmarking work (Toneyan, Tang and Koo, 2022), we implemented an intentionally simple convolutional neural network architecture based a relatively small receptive field around the TSS to compare histone mark performance. On the other hand, our distal model, Chromoformer (Lee, Yang and Kim, 2022), was a transformer-based architecture accounting for distal histone mark levels in a weighted manner based on DNA interaction data. One limitation is the receptive field of our distal model which extends 40,000 base-pairs around the TSS. Although this is a large window and computationally intensive to include in such a model, it is a fraction of the known distance at which DNA interactions can occur. For example, Hi-C experiments capture cis-interactions up to 1 million base-pairs away (Lieberman-Aiden *et al.*, 2009).

The results of our promoter region analysis showed that the active marks H3K4me3 and H3K9ac were the most predictive of gene expression (**Figure 6-2**). However, their optimal performance was dependent on the cell state, performing better in ESCs whereas repressive marks like H3K9me3 performed relatively better in adult primary tissues (**Figure 6-2b**). We concluded that the stage of cell differentiation was the driving factor of performance for active and repressive marks: Active marks performed better at early stages of lineage commitment, i.e. ESCs, whereas repressive marks were more predictive in fully differentiated cells, i.e differentiated tissue. This relationship is in line with the distinctive ESC chromatin landscape, which features more accessible DNA and less heterochromatin and correspondingly, less repressive and more active histone mark activity when compared with differentiated cells (Liang and Zhang, 2013). The effect of repressive mark performance across different cell states is further substantiated by our *in silico* permutation analysis (**Fig. B-9**). We observed a consistent relationship for active marks, where perturbations to histone mark signals in the gene body had a greater predicted effect than distal regulatory regions. This pattern was only replicated for inactive marks in differentiated cell types. We hypothesise that this relates to the lack of repressive signal in the other cell states (Liang and Zhang, 2013). Furthermore, we noted that for active genes, the relationship between a histone mark's levels and performance replicated known biology: Observing a strong correlation for active histone marks in highly expressed genes, and conversely for repressive marks in lowly expressed genes (**Figure 6-4a**). This highlighted that higher histone mark levels were beneficial for the model, leading to greater predictive performance in the correct context. In relation to histone mark levels at the TSS, we also noted that for inactive genes, histone mark activity is reduced with cell differentiation (**Figure 6-4b**). The concept that cell lineage commitment leads to globally lower histone mark levels has been previously noted (Wu *et al.*, 2023).

For our distal model, active marks H3K4me3, H3K9ac and H3K27ac were the best performing (**Figure 6-5a**). Interestingly, these three marks, two of which are linked to the TSS of genes they regulate, outperformed H3K4me1 which is associated with distal enhancers (Bae and Lesch, 2020). One possible explanation for this was investigated in a recent study (Karollus, Mauermeier and Gagneur, 2023) which found that, when predicting expression from DNA sequence, a similar attention model prioritised sequences at the promoter region over distal regulatory regions. The reason being that given the multiple choice of enhancer and other regulatory regions and their relatively small influence on gene expression, a model will prioritise the information at the promoter region. We believe the same effect could have contributed to our results whereby the active promoter mark information contributed to expression to a greater degree than distal regulatory regions. This same relationship was clearly notable in our *in silico* perturbation experiments (**Figure 6-7e-f**). Moreover, given that H3K4me1 is indicative of poised rather than active enhancers (Bae and Lesch, 2020), it would presumably be less predictive of gene activity.

A key point of our findings is the marginal return in performances by: 1. Extending from an intentionally simple local promoter model to an attention based, computationally complex model which accounts for distal histone mark levels (**Figure 6-5c**), and 2. Increasing the number of histone marks included in the model (**Figure 6-6b**). We noted that understanding the cell state (undifferentiated or fully differentiated) and the gene state (highly or lowly expressed) of interest and choosing the most appropriate mark for these had a greater impact on performance than the number or receptive field of histone mark levels considered.

Comparing performance across the promoter and distal models, H3K27ac showed the biggest gain of the top three performing marks (**Figure 6-5**). This was expected given its relationship with active promoters and enhancer regions (Herrera-Uribe *et al.*, 2020). However, its

performance based solely in promoter regions was still relatively strong, which was reassuring given the mark's prevalence in complex disease research.

We also trained our distal model on pairs of histone marks, showing that the added performance of incorporating additional histone marks diminishes markedly after this point. This result could benefit researchers wishing to capture transcription in a cell type of interest from limited histone mark information. The paired analysis also highlighted the strong combinatorial performance of H3K36me3. This repressive mark was the best performing choice as a pair with any of our three top marks (**Figure 6-6a**) and was the fourth best performing mark of the single histone mark analysis (**Figure 6-5a**). H3K36me3 is a canonical mark of transcription, serving as a binding partner for histone deacetylases (HDACs) which prevent run-away transcription of RNA polymerase II (Carrozza *et al.*, 2005). H3K36me3 is generally enriched in gene bodies of mRNAs (Li, Notani and Rosenfeld, 2016), outside of the TSS, which may explain its relatively poor performance in the promoter model and why the improvement with the distal model for this mark was the highest of any mark tested (**Figure 6-5c**).

Finally, we performed an array of *in silico* histone mark level perturbation experiments, showing the relationship between distance from the TSS and a regulatory region's effect on gene expression (**Figure 6-7**). Our analysis highlighted the very high correlation for the *in silico* perturbations between the different cross-validation fold models (**Fig. B-8**). A possible advantage of histone mark deep learning models over genomic deep learning models trained on DNA sequence is that they offer an alternative to making single base pair level changes such as genetic variants which are notoriously difficult to interpret, particularly as the genomic window considered by the model increases (Toneyan, Tang and Koo, 2022; Huang *et al.*, 2023; Linder *et al.*, 2023; Sasse, Ng, A. E. Spiro, *et al.*, 2023). For example, Sasse *et al.,*

found that Enformer, a long-range genomic deep learning model, did a poor job of predicting the effect of inserting multiple single nucleotide variants based on personalised genomes, even predicting the wrong direction of effect up to 40% of the time (Sasse, Ng, A. E. Spiro, *et al.*, 2023). This work highlighted that the current training paradigm for genomic deep learning models of training across the genome for variability is insufficient to accurately capture the effect of genetic variants. Moreover, these variants are subject to LD and as of yet, it has not been proven that genomic deep learning models can accurately differentiate between causal genetic variants and those in LD. By identifying genomic loci of interest based on perturbing histone mark levels, our model captures significant enrichment of eQTLs in the most predictive regions, offering an alternative to genomic deep learning models trained on DNA. Additionally, these results based on epigenetic data further underlines the functional importance of these eQTLs (**Figure 6-8a**).

Furthermore, using these identified genomic loci, we developed a framework by which such models can be applied to test for both functional and disease enrichment in a cell type-specific manner (**Figure 6-8**). The results for the disease enrichment did not recapitulate projected relationships for the neuronal progenitor cells or liver, which could be a result of the imperfect cell type matching, the limited overlap between the disease related genetic variants and the relatively small window of upstream genomic loci considered, or the observed differences in the measured genetic effects on gene expression versus complex traits (Mostafavi *et al.*, 2023). This highlights that further work on such models is needed, hopefully increasing the receptive field and using known affected cell types, to capture an association with complex diseases. Importantly, a substantial overlap and large genomic coverage of the loci considered are key recommendations for s-LDSC analyses (Finucane *et al.*, 2015). This issue when capturing complex phenotypic enrichment is not unique to these models and is also a

challenge with genomic deep learning models as highlighted recently (Alan E. Murphy *et al.*, 2024). Non-cell type specific genomic loci predictive of gene expression were enriched for GWAS signal for bipolar disorder, indicating the importance of cell type consistent regulatory regions in complex disease. Importantly, this analysis highlights the significance of these predicted genomic loci not only in functional genomics studies (**Figure 6-8a**) but also in disease (**Figure 6-8b**). Past approaches such as ICEBERG, a pipeline that uses CUT&RUN replicates to create a combined profile of binding events for H3K4me3, have been previously used to uncover functionally relevant regulatory events (Nordin *et al.*, 2024). However, our approach shows, for the first time, how chromatin deep learning models can be perturbed to uncover genome-wide and cell type-specific functionally and disease relevant regulatory regions.

Overall, our study shows that multiple factors influence the performance of histone marks when predicting gene expression, something which had not explicitly been considered by previous work (Karlić *et al.*, 2010; González-Ramírez *et al.*, 2021; Wang *et al.*, 2022). Our findings suggest that if one wishes to investigate the TSS of genes, promoter-specific active marks H3K4me3 and H3K9ac are the best options. Beyond the promoter region, active marks H3K4me3, H3K9ac and H3K27ac are most predictive, especially in combination with the transcriptionally-associated mark H3K36me3. However, it is worth considering the cell state (differentiated or early stages of lineage commitment) and the state of the genes or interest (highly or lowly expressed) to have a better understanding of the optimally predictive histone marks. Importantly, more effort should be placed on using these models to uncover new biological insights, particularly for phenotypic and disease-based studies. Similar to genomic deep learning models, chromatin deep learning models are capable of capturing functionally relevant genomic loci.

6.5. Data and code availability

The Histone mark ChIP-Seq read alignments, RPKM gene expression profiles were downloaded from the Roadmap Epigenomics Web Portal (https://egg2.wustl.edu/roadmap/web_portal/index.html). The promoter capture Hi-C experiments were obtained from the 3DIV database (available at http://3div.kr/), specifically the tissue mnemonics H1, ME, TB, MSC, NPC, LI11, PA, LG, and GM. The UK Biobank fine-mapped eQTL data were downloaded from the supplementary material of Wang et al.'s study (Q. S. Wang *et al.*, 2021). The summary statistics were downloaded from the IEU GWAS portal (Elsworth *et al.*, 2020) (IDs: ieu-b-7, ebi-a-GCST90027158, ebi-a-GCST90027158, ebi-a-GCST005647, ebi-a-GCST90091033, ebi-a-GCST90091033, ebi-a-GCST90038627, ieu-b-5099, ieu-a-1185, ieu-b-5110) and for hepatitis, from the BioStudies database (Sarkans *et al.*, 2018) (ID: S-BSST407). All reference datasets used to run s-LDSC (Finucane *et al.*, 2015) were downloaded following the links from the source material: https://github.com/bulik/ldsc. The model architectures and all training and analysis scripts, along with scripts to download and complete all pre-processing steps on the training data (sourced from Roadmap (Kundaje *et al.*, 2015) and largely replicated from Chromoformer's scripts (Lee, Yang and Kim, 2022)) are available at https://github.com/neurogenomics/chromexpress (Murphy, 2024). The model results and weights are available for download on figshare (https://figshare.com/account/home#/projects/201105).

6.6. Funding

This work is supported by the UK Dementia Research Institute award numbers UKDRI-5008 and UKDRI-6009 through UK DRI Ltd, principally funded by the UK Medical Research Council. SJM furthermore received funding from the Edmond and Lily Safra Early Career Fellowship Program (https://www.edmondjsafra.org). NGS also received funding from a UKRI Future Leaders Fellowship [grant number MR/T04327X/1].

7. General Discussion

The expanding body of genetic research, including large-scale GWAS, have identified genetic variations primarily located in regulatory, non-coding regions of the genome as significant molecular drivers of AD (P. R. Jansen *et al.*, 2019; Bellenguez *et al.*, 2022a). Early analyses have suggested that microglia may bear much of the genetic burden of the disease (Skene and Grant, 2016; Nott *et al.*, 2019); however, the specific regulatory pathways, transcription factors, and genes involved remain largely unidentified. Fundamental advancements in computational approaches are essential for a comprehensive analysis of this genetic data, which would help illuminate the cis-regulatory code associated with AD. This PhD thesis aims to address the existing gaps in computational analyses, encompassing both the protein-coding and non-coding genome, and making contributions that extend beyond AD to broader genomic studies. Although the primary focus is on AD, the developed open-source, computational and deep learning techniques are applicable to a wider understanding of the cis-regulatory code. This final chapter summarises my main findings and how these impact, and propose future directions for, the field. The work in this chapter is adapted from that set out in our previous publications and detailed in the empirical chapters of this thesis (Murphy, Schilder and Skene, 2021; Murphy and Skene, 2022; Murphy, Fancy and Skene, 2023; Alan E Murphy *et al.*, 2024; Alan E. Murphy *et al.*, 2024).

7.1. Key findings

7.1.1. A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis

The aim of my first empirical data chapter was to benchmark differential expression approaches and give guidance on the best approach to uncover DEGs. This work was completed to inform a standardised approach for disease research. Recently, Zimmerman *et al.*, highlighted the importance of accounting for the dependence between cells from the same individual when conducting differential expression analysis on single-cell RNA-sequencing data (Zimmerman, Espeland and Langefeld, 2021). Their work proved the inadequacy of pseudoreplication approaches for such analysis—this was an important step forward that was conclusively proven by them. However, there were limitations in both their benchmarking and simulation approaches. Here, we corrected these issues, reran the author's analysis and found that pseudobulk methods outperformed mixed models. Based on these findings, we recommend the use of pseudobulk approaches for differential expression in single-cell RNA-sequencing analyses.

7.1.2. Avoiding false discoveries in single-cell RNA-Seq by revisiting the first Alzheimer's disease dataset

The aim of the second empirical data chapter, which is linked to the first, was to apply a standardised approach to processing and differential expression for scRNA-Seq studies in AD. Mathys *et al.* conducted the first snRNA-Seq study of AD (Mathys *et al.*, 2019). With bulk RNA-Seq, changes in gene expression across cell types can be lost, potentially masking the DEGs across different cell types. Through the use of single-cell techniques, the authors benefitted from increased resolution with the potential to uncover cell type-specific DEGs in AD for the

first time. However, there were shortcomings in both their data processing and quality control and their differential expression analysis. Here, we used their work as a case study, correcting these issues and using best-practice approaches to snRNA-Seq differential expression, resulting in 549 times fewer DEGs at a false discovery rate of 0.05. Our analysis gives new insight into the dysregulated genes based on the original experimental data. Overall, these findings highlight the impact of quality control and differential analysis methods on the discovery of disease-associated genes and aims to refocus the AD research field away from spuriously identified genes.

7.1.3. A tool for the standardisation and quality control of many GWAS summary statistics

Our third empirical chapter concentrates on the standardisation and quality control of genetic studies. GWAS summary statistics have popularized and accelerated genetic research. However, a lack of standardisation of the file formats used has proven problematic when running secondary analysis tools or performing meta-analysis studies. All downstream analyses of genetic data requires standardisation and quality control, something which is lacking from the broader computational biology field, hindering progress in genetic studies. To address this issue, we developed MungeSumstats, a publicly available Bioconductor R package for the standardization and quality control of GWAS summary statistics. MungeSumstats can handle the most common summary statistic formats, including VCF and producing a reformatted, standardized, tabular summary statistic file, VCF or R native data object.

7.1.4. Predicting cell type-specific epigenomic profiles accounting for distal genetic effects

The aim of our fourth empirical chapter was to develop a model to uncover the epigenomic effect of genetic variants in a cell type-specific manner. Understanding genetic variants' effects on the epigenome is crucial for interpreting GWAS results, yet profiling these effects across the non-coding genome remains challenging due to the scalability limits of experimental methods. This necessitates accurate computational models. Existing machine learning approaches, while progressively improving, are confined to the cell types they were trained on, limiting their applicability. Here, we proposed the development of a deep learning model, Enformer Celltyping, which incorporates distal effects of DNA interactions, up to 100,000 base-pairs away, to predict epigenetic signals in previously unseen cell types. We demonstrate that Enformer Celltyping's predictions of the epigenome out-perform the current best-in-class approach, have strong performance in a range of different cell types and biological regions and generalise to cell types assayed independently of the original training set. Finally, we proposed a framework to test epigenetic models' performance on genetic variant effect predictions using regulatory quantitative trait loci mapping studies, highlighting current genomic deep learning models' limitations for this task and show how, despite this, Enformer Celltyping can be used to study cell type-specific genetic enrichment of complex traits. Our work introduced a new, accurate model for cell type epigenetic imputation based on DNA and chromatin accessibility data from the cell type of interest and highlights the benefit of transfer learning for more efficient model development for deep learning in genomics.

7.1.5. Predicting gene expression from histone marks using chromatin deep learning models depends on histone mark function, regulatory distance and cellular states

Our final empirical data chapter links epigenetics to transcription utilising deep learning approaches and offering an alternative approach to uncover the cis-regulatory code *in silico*. To understand the complex relationship between histone mark activity and gene expression, recent advances have used *in silico* predictions based on large-scale machine learning models. However, these approaches have omitted key contributing factors like cell state, histone mark function or distal effects, that impact the relationship, limiting their findings. Moreover, downstream use of these models for new biological insight is lacking. Here, we presented the most comprehensive study of this relationship to date - investigating seven histone marks, in eleven cell types, across a diverse range of cell states. We used convolutional and attention-based models to predict transcription from histone mark activity at promoters and distal regulatory elements. Our work showed that histone mark function, genomic distance and cellular states collectively influence a histone mark's relationship with transcription. Moreover, we found that no individual histone mark is consistently the strongest predictor of gene expression across all genomic and cellular contexts. This highlighted the need to consider all three factors when determining the effect of histone mark activity on transcriptional state. Furthermore, we conducted *in silico* histone mark perturbation assays, uncovering functional and disease related loci and highlighting frameworks for the use of chromatin deep learning models to uncover new biological insight.

7.2. Implications of my findings

My PhD projects proposed computational approaches that addressed both AD-specific and broader challenges in the field of genomics. This thesis contributes to the field in a number of important ways, as detailed in the empirical chapters and the main aspects of which are summarised below.

7.2.1. The standardisation of analysis for cell type-specific transcriptional changes

Our works in **Chapter 2** and **Chapter 3** on scRNA-Seq analysis proposed a standardised, best-practice approach to quality control and differential expression which has already been adopted by the field. We created open-source tools for the scRNA-Seq community to encourage reproducibility and ease of re-analysis for research. This has improved the credibility of published studies. Furthermore, understanding which genes' transcriptional landscapes have changed is fundamental to understand the processes involved in neurogenerative diseases' progression like AD. Thus, we believe our work will also have an impact in AD research field.

Falling experimental and sequencing costs will lead to increased sample sizes and as a consequence, more researchers will endeavour for standardised approaches to processing and analysis. We believe this shift, along with the influence of benchmarking works including ours, will result, and to a certain extent have already resulted, in pseudoreplication approaches losing popularity in favour for more robust approaches like pseudobulk. Such approaches do not artificially inflate the number of DEGs as this is less of a necessity to uncover any, albeit possibly spurious, DEGs. This can already been seen in the largest study of scRNA-Seq in AD of 427 individuals which used pseudobulk approaches for its differential expression analysis (Mathys *et al.*, 2023) – Amongst its findings were a differentially expressed

astrocyte program that is associated with cognitive resilience to AD pathology. This work was by the same group who used pseudoreplication for the first study of scRNA-Seq for AD (Mathys *et al.*, 2019).

7.2.2. The standardisation of processing and quality control of genetic information

The fundamental issues of standardisation and quality control of genetic data has hindered the progress of any downstream analyses and potentially, led to spurious results. I believe our work in **Chapter 4,** creating MungeSumstats, addresses these issues and has already had a positive impact and uptake in the field – The tool receives around a thousand unique downloads per month on R Bioconductor (https://bioconductor.org/packages/stats/bioc/MungeSumstats/) and has been citied over 60 times. MungeSumstats along with other complementary approaches like EBI GWAS Catalogue (Buniello *et al.*, 2019), the MRC IEU OpenGWAS Catalogue (Elsworth *et al.*, 2020) and GWAS-VCF formatting (Lyon *et al.*, 2021), will accelerate future genetic discoveries by removing any barriers of usage and ensuring the quality of such data. Further work has continued to be done to standardise the field, including community driven standards for GWAS summary statistics data, dissemination and availability (Hayhurst *et al.*, 2023). However, such efforts raise the question of how the field will sustainably fund these resources beyond their initial development, given the limited focus of current academic funding mechanisms on software maintenance. Such tools are paramount for AD, and the wider research field, with the rise in related genetic studies (P. R. Jansen *et al.*, 2019; Schwartzentruber *et al.*, 2021; Bellenguez *et al.*, 2022a). For these standards to have a lasting impact, it is essential that journals enforce their implementation. In an ideal scenario, such enforcement would render tools like MungeSumstats unnecessary.

7.2.3. Predicting the cell type-specific effects of genetic variants while accounting for distal regulation

Our model Enformer Celltyping, proposed in **Chapter 5**, enables accurate cell type-specific epigenetic imputation based on DNA and chromatin accessibility data from the cell type of interest. This differs from many genomic deep learning models in the field which predict in held out chromosomes of the same cell type (Kelley *et al.*, 2018; Avsec, Agarwal, *et al.*, 2021; Avsec, Weilert, *et al.*, 2021). Moreover, we highlight how Enformer Celltyping's predictions capture cell type-specific genetic enrichment for complex traits. We believe that this work will prove of use for the field, enabling the breadth of publicly available ATAC-Seq data in public databases like ATACdb (F. Wang *et al.*, 2021), to be used to derive histone mark signals for the cell types of interest which harbour greater enrichment for the study of complex traits and diseases including AD.

7.2.4. Prioritising functional and disease relevant genomic loci in silico with deep learning by linking epigenetics to transcription.

Chapter 6 details our work to study the relationship between histone mark levels and transcription using deep learning models. We considered multiple cell types and histone marks at differing receptive fields and established three key contributing factors that gene expression is dependent on - histone mark function, regulatory distance and cellular states. I believe that this work will be of general interest for the field and give insight for researchers looking to best capture gene regulation in their cell type of interest with minimal experimental work. Furthermore, we performed an array of novel *in silico* histone mark level perturbation experiments, and developed an alternative approach to uncover cell type-specific regulatory

and disease relevant loci **(Figure 6-8).** This approach to genomic loci prioritisation has been previously underutilised by chromatin deep learning models which, given the results of our preliminary analysis using our model, should be adopted by the field to yield greater insight into the cis-regulatory code, furthering our understanding of complex disease like AD .

7.3. Future directions for the field

The field of genomics, including AD genomics research, has made significant strides in recent years. Yet, multiple avenues remain unexplored or only partially understood. Substantial advancements in computational approaches are required to propel further progress. The findings of this PhD thesis address critical gaps in the standardisation of genetic data processing, analysis of cell type-specific transcriptional changes, and the prioritization of functional genomic loci. These works set the stage for further exploration into the intricate relationship between genetic variants and disease pathology, furthering our understanding of the cis-regulatory code. Future research should build on these contributions by refining computational models, enhancing the accuracy of genomic deep learning approaches, and expanding the applicability of these techniques to other complex diseases. This will not only deepen our understanding of AD but also contribute to the broader field of genomics and precision medicine, ultimately paving the way for the development of more targeted and effective therapeutic strategies. This section will explore such future directions.

7.3.1. Uncovering the cis-regulatory code of AD

A major goal of this work was to understand the effect of genetic variants on the cis-regulatory code in a cell type-specific manner. Such as epigenetic regulation of transcription factor binding in microglia within the hippocampus, a cell type implicated in the genetic burden of AD (Skene and Grant, 2016; Nott *et al.*, 2019). Albeit, it is too early to tell that other neuronal subtypes or peripheral macrophages aren't instigated in the disease. Nonetheless, this approach holds potential for identifying transcription factors that are functionally relevant in the progression of AD.

By applying the same SLDP method developed in **Chapter 5,** with AD GWAS data (I. E. Jansen *et al.*, 2019; Wightman *et al.*, 2021; Bellenguez *et al.*, 2022a), it would be possible to systematically assess associations for all known transcription factors across multiple cell types. The identified transcription factors could then be inhibited in a cell type-specific manner as a form of targeted therapeutics as has been explored for cancer research (Bushweller, 2019).

However, such epigenetic experimental assays in these targeted brain regions and cell types have only been sparsely explored to date. To address this gap, in **Chapter 5**, we developed Enformer Celltyping to predict epigenetic signals in previously unseen cell types to circumvent the lack of experimental assays. We chose histone marks as a first attempt due to the availability of hQTL sets to validate this approach. After such, the original goal was to apply the same approach to predict transcription factor binding assays; however, we found that current genomic deep learning models can not capture the effect of genetic variants (**Figure 5-6**, see **Section 7.3.2** for a detailed discussion).

Future research aimed at elucidating the cis-regulatory code of AD should prioritise the collation of experimental data for all relevant cell types and transcription factors. This is essential for two reasons. First, our developed approach to extrapolate predictions to previously unseen cell types with Enformer Celltyping, requires the model to learn cell type representations which leads to worse predictive performance compared to models that predict in the same cell type but in held-out chromosomes. Second, as discussed in **Section 5.4**, when predicting the effect of genetic variants with Enformer Celltyping, we can not account for the genetic variants' effect on the chromatin accessibility data the model uses to infer cell type representations thus, misrepresenting the cell state for the model. Outside of amassing functional data, future work to uncover the cis-regulatory code of AD requires

substantial investments into methods to improve the efficacy of deep learning models at variant effect predictions, as further explored in **Section 7.3.2**.

7.3.2. Improving the utility of genomic deep learning models for personalised genomics

While the limitations of these works are discussed in the relevant chapters, one such limitation discussed in **Chapter 5** merits further expansion here as the implications of which, based on the growing evidence from us and others groups, necessitates a shift in the field's future direction. This relates to the prediction of genetic variant effects *in silico* with genomic deep learning models. To summarise the discussion in **Section 5.4**, we developed Enformer Celltyping for the prediction of epigenetic marks in previously unseen cell types. Enformer Celltyping achieved best-in-class performance for histone mark predictions (**Figure 5-3a**), strong predictive performance in functional genomic regulatory regions (**Figure 5-4**) and maintains accuracy even with a domain shift to isolated cell types assayed outside of ENCODE (**Figure 5-5**).

Yet despite this, Enformer Celltyping's performance in predicting the effect of genetic variants predominately did not match the effects from the QTL studies accounting for LD, using the SLDP measure of concordance (Reshef *et al.*, 2018) (**Figure 5-6a**). We believe that genomic deep learning models, including Enformer Celltyping, inherently struggle to accurately predict the effect of genetic variants based on their current training paradigm. Enformer Celltyping has to extrapolate predictions to previously unseen cell types, potentially exacerbating this problem. Therefore, we also tested Enformer on the same genetic variant effect prediction test, where it yielded comparably poor results (**Figure 5-6b**). Our finding that genomic deep learning models fail to accurately predict the effect of genetic variants on the epigenome is not the first of its kind in the field; this issue has also received growing attention in the

literature (Toneyan, Tang and Koo, 2022; Huang *et al.*, 2023; Linder *et al.*, 2023; Sasse, Ng, A. E. Spiro, *et al.*, 2023). Sasse *et al.*, highlighted how Enformer underperforms at predicting the effect of genetic variants on transcription, even predicting the incorrect direction of effect in up to 40% of individual genomes tested (Sasse, Ng, A. Spiro, *et al.*, 2023). Moreover, Huang *et al.*, demonstrated how four current state-of-art models, including Enformer, poorly capture variation in expression or chromatin accessibility based on fine-mapped chromatin accessibility QTL (caQTL) and eQTL datasets (Huang *et al.*, 2023).

We believe one potential cause of this issue is in the lack of genetic variation seen in genomic deep learning model's current training regime, leading to out-of-sample predictions on comparatively small, single base pair changes to the DNA input. Therefore, it is paramount that future work hoping to use machine learning models, to predict the effect of genetic variants, concentrate on training models on data which accounts for such. However, this is a non-trivial task for a number of reasons relating to model architecture and the genetically diverse training data. However, there are multiple immediate approaches the field could explore to circumvent this issue.

Firstly, in terms of model architecture, an approach could be to incorporate genetic variant data using fine-tuning as a second stage of training for models pretrained on the reference genome. However, refining a deep learning model on an out-of-distribution task like this can lead to catastrophic forgetting; where the pre-trained model loses the ability to accurately predict previous tasks (Aleixo *et al.*, 2023). This could cause such models to 'forget' all previously learned sequence grammar, rendering the pretraining on the reference genome, and all its associated learnt syntax, futile. Furthermore, enabling such genomic deep learning models to account for single base-pair changes, especially those with small effects, as their receptive field ever increases, is analogous to finding a needle in a haystack. Addressing this

challenge will likely require careful model architecture and training design. One potential option would be to use dual loss functions, one which accounts for inter-sample variation and another for intra-sample variation across the genome. Another option would be to use an architecture inherently designed to uncover differences in inputs like Siamese Neural Networks (Koch, Zemel and Salakhutdinov, 2015). Siamese networks have been heavily used in tasks like identifying differences and similarities between images. Traditionally, a Siamese network's architecture has identical layers for each input with shared weights across them and uses a loss function like contrastive loss based on differences in the inputs (Koch, Zemel and Salakhutdinov, 2015). This could be similarly applied to capture differences in DNA sequences across samples.

Secondly, in relation to the genetically diverse training data which itself poses some non-trivial problems. One could train a model using xQTL studies or similar data, where the genetic variance is naturally observed across populations. However, as set out in **Section 1.4**, a person's set of genomic variants will only account for just on average just 1% of their whole genome (Trost, Loureiro and Scherer, 2021). Thus, training a model to predict across different genomes at the same locus rather than the current paradigm of across loci for the same reference genome, would result in much less variance in the DNA sequence and also considerably less measured variance in the functional assay output. This depletion in variance could cause the current training approaches to fail to learn the differences in syntax and how these changes affect the cis-regulatory code. Instead models may just focus on minimising their loss functions across loci and learn an average signal for a locus as this would produce reasonable performance across batches of individuals. However, there are other approaches to incorporate greater proportions of sequence diversity, such as leveraging evolutionary diversity by incorporating predictions across multiple species (Kelley, 2020). Since many gene

regulatory mechanisms are conserved across related species, this approach can provide new, valuable training examples that are also applicable to a model's understanding of human's cis-regulatory code. Yet, it is also important to recognise that naive genomic DNA only represents a limited subset of all possible genomic DNA arrangements, as certain variations are never observed in nature due to their deleterious effects (Sasse, Chikina and Mostafavi, 2024). Thus, it may be advantageous to use synthetic sequences from perturbation experiments, like MAVEs, during training time so the model can learn a holistic representation of the sequence grammar and cis-regulatory code. Some in the synthetic biology field go as far as to suggest withholding the entire genome and instead, training on vast amounts of randomised libraries of DNA sequence to capture substantial amounts of diversity (de Boer and Taipale, 2024). This increase in diversity may help a model, for example, to learn enhancer syntax for gene regulation despite their sparsity and cell type-specific nature. Each of these options entail a trade-off between the amount of genetic diversity they include and their similarity to the functional assays used by genomic deep learning models trained in the current paradigm using the reference genome. This relationship is represented in **Figure 7-1** where low values in either represent considerable challenges for their incorporation into training regimes for genomic deep learning models. For example, MAVE datasets offer a high-throughput approach to measure the effect of thousands of genetic variants on gene regulatory sequences in one experiment. However, the output of such experiments are usually a phenotypic readout which at best is correlated with a gene's functional activity rather than expression levels which genomic deep learning sequence-to-function models often predict (Claussnitzer *et al.*, 2024).

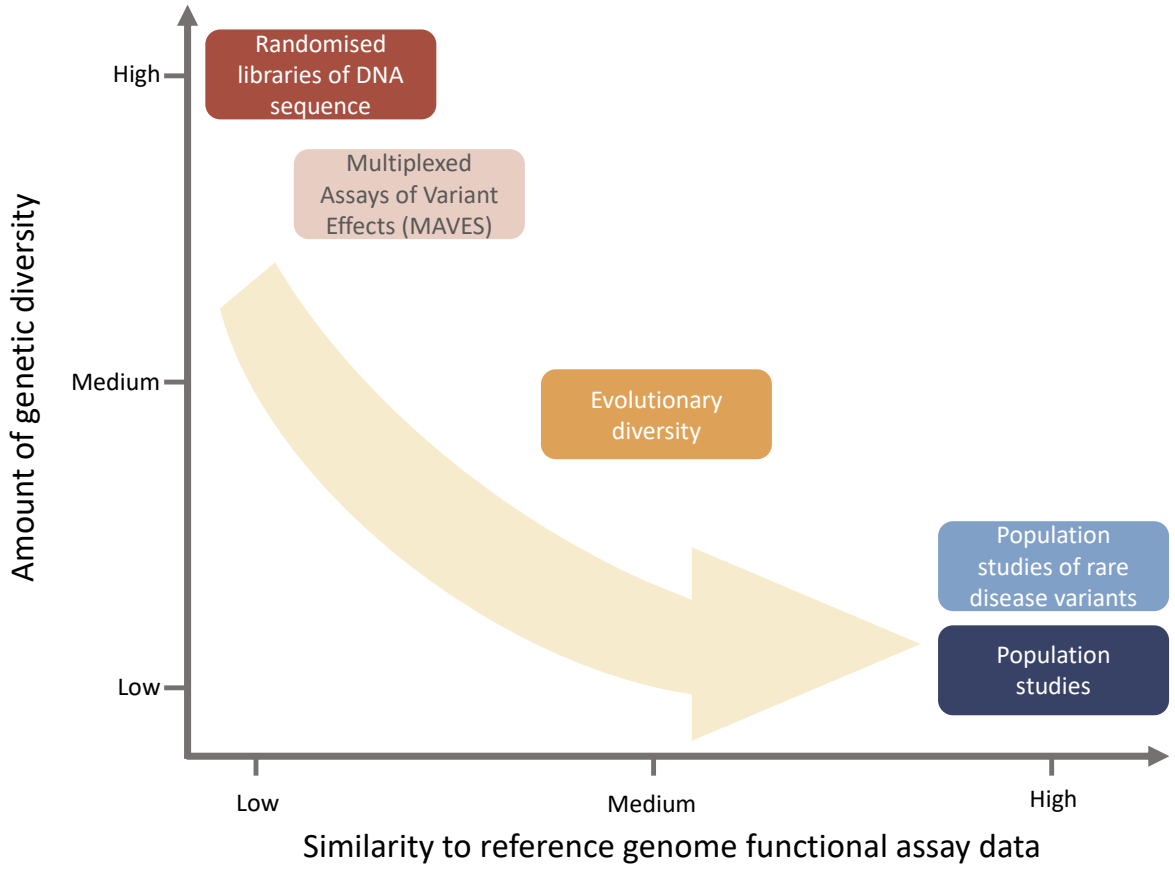


Figure 7-1: Differing genetic variant data available for incorporation to genomic deep learning model's training regime

The relationship between the amount of genetic diversity present (y-axis) and the similarity to functional assays of models trained on the reference genome (x-axis) for differing genetic variant sources is represented. The yellow arrows highlights the noted trade-off between these two attributes. A low value in either axis represent considerable challenges for their incorporation into training regimes for genomic deep learning models.

There is, however, an alternative view in the field as to why current genomic deep learning models fail to accurately predict the effect of genetic variants which is due to loss functions focusing on variation across genes or more generally genomic loci. Models like Enformer predict, for example, CAGE-Seq for numerous cell types as a multi-task prediction. The loss function used by such models penalises variation across genes. Even though multiple cell types are predicted, the model is not penalised for the variation of each gene across cell contexts. A naïve model could thus predict the average expression of a gene across cell types to minimise loss. An argument could be made that penalising a model for not identifying cell type-specific changes in expression (or other functional assay) will force it to pay attention to cell type-specific regulatory motifs like enhancer motifs which in turn would benefit the model's variant effect predictions. In the same manner as considering multiple personalised genomes at a location, the argument here is to consider multiple cellular states. Although the DNA input does not change, the output will, forcing the model to understand cell type-specific regulation. This view, although promising, would be non-trivial to incorporate and may require multiple loss functions to ensure the model maintains strong performance across genes or genomic loci or, if modelling expression, using pretrained models which capture epigenetic functional assays as a starting point as their learnt regulatory syntax is paramount e.g. chromatin accessibility as a proxy for transcription factor binding sites or regulatory histone marks like H3K27ac. It is worth noting that this has yet to be fully evaluated in the field.

To summarise, advancements to genomic deep learning models are paramount, however, this is limited by the availability of suitable, publicly available genetic variant data, advancements in deep learning techniques like training regimes and careful curation of training data to effectively incorporate genetic variant effects.

7.3.3. Benchmarking genomic deep learning models for personalised genomics

A clear first step towards this goal of improving the performance of genomic deep learning models on genetic diversity, is defining benchmark datasets to better understand the extent of their shortcomings. The use of such benchmarks have become widely adopted in the field of computer science where new models are compared across a range of datasets, to understand limitations compared to current state-of-the-art approaches. However, developing benchmarks for genomic deep learning models poses unique challenges. For example, large language models (LLMs) often use the Abstraction and Reasoning Corpus (ARC) to measure their ability to perform comprehension and abstract reasoning (Chollet, 2019). However, the functional assays used to measure genomic deep learning models' ability do not provide as straightforward a measure as this in terms of comprehending the cis-regulatory code: Each assay is subject to experimental variability within the technique such as batch effects but also variability across differing techniques resulting from technique-specific biases – for example, ChIP-Seq results differ dramatically to CUT&Tag using the same experimental conditions (Hu *et al.*, 2023).

Such benchmark approaches require standardised practices to quality control and processing for both the genetic information and the functional assay data to avoid any biases in performance measurements. Improvements to both of which have been proposed in this thesis in chapters **2**, **3**, and **4**. Some initial attempts have been proposed including that by Li *et al.*, with GV-Rep (Li *et al.*, 2024). Importantly, the GV-Rep dataset was specifically designed for benchmarking genomic deep learning models, consisting of 7 million records, reporting model performance across the differing sources, including *in vitro* MAVE experiments like gene knockouts as well as *in vivo* measurements like xQTLs. Here, we will discuss the differing

available genetic variant data sources and their suitability for genomic deep learning benchmarks and highlight those included in GV-Rep. These critiques will overlap with using such data when training genomic deep learning models, discussed in **Section 7.3.2,** but also offer unique considerations.

The goal of GV-Rep is to validate genomic language models for the classification of genetic variants in a clinical settings (Li *et al.*, 2024). For example, the records collated from ClinVar classify the pathogenicity of genetic variants; benign, likely benign, likely pathogenic or pathogenic, for differing diseases (Landrum *et al.*, 2020). Other records are from GWAS summary statistics, associating SNPs to diseases based on their p-values. However, I believe these tasks are too broad for genomic deep learning models as they lack any functional measurements or cell type/cell state contexts and should not be included in such a benchmark dataset.

Moreover, LD must be considered for relevant data sources to be included in a benchmark dataset. GV-Rep captures xQTL datasets, sourced from GTEx (Lonsdale *et al.*, 2013), but uses the reported p-values from the summary statistics without accounting for LD. This will dilute the true performance signal of models and instead, possibly rewarding models that pick up on LD structure rather than causal relationships in DNA sequence (Li *et al.*, 2024). LD could be accounted for by fine-mapping to define a set of causal SNPs (separating causal SNPs as high confidence cases from the fine-mapping) or by measuring the model predictions' concordance with the functional data is using SLDP, as proposed in this thesis (**Chapter 5**). Thus, xQTL datasets could be considered for such a benchmark and could be easily incorporated due to their similarity to the functional outputs of these models (**Figure 7-1**). However, as previously highlighted, despite considerable advancements, there is still work to

be done to improve current fine-mapping approaches (Schaid, Chen and Larson, 2018; Broekema, Bakker and Jonkers, 2020; Burgess, 2022). Moreover, approaches to define positive or negative sets from fine-mapping is non-trivial and generally results in using just the top and bottom 5% of SNPs; for example this approach was used by Avsec *et al.* (Avsec, Agarwal, *et al.*, 2021). However, excluding all but the top and bottom 5% of SNPs could bias the measured performance to the easily identifiable cases – For example, Enformer did well on fine-mapped SNPs in their work but poorly for predicting personalised genome effects (Avsec, Agarwal, *et al.*, 2021; Sasse, Ng, A. E. Spiro, *et al.*, 2023). Thus, while these datasets are of use, their inclusion such a benchmark dataset is dependent on careful processing and interpretation.

Other data sources, which were also included in GV-Rep, are *in vitro* experiments like MAVEs and CRISPRi assays which test the effect of genetic perturbations and so, avoiding LD. MAVE datasets have become more prevalent with whole databases of experiments such as MAVEdb (Esposito *et al.*, 2019). There are barriers to their immediate use however as when deposited, most experiments lack information to map their imputed genetic code to the reference genome and thus, only the immediate ~150 base-pairs are available which is far smaller than most genomic deep learning model's receptive field. There is, however, current ongoing work to map the imputed genetic code to the reference genome for all entries in MAVEdb post-hoc (Arbesfeld *et al.*, 2023). Secondly, the output of such experiments differ significantly to the assays genomic deep learning models are trained to predict (see **Figure 7-1**), so using such data to improve models is non-trivial. Finally, MPRAs, a commonly used MAVE, traditionally use a barcoded DNA sequence with a minimal promoter and a small upstream loci to create MPRA vectors and so does not capture in vivo measurements from the whole genome like cell type-specific effects. Although GV-Rep did include MPRA data, they did not address any

of these issues, thus limiting it's utility for the field. CRISPRi experiments like prime editing could capture these whole genome effects. Prime editing is a precise genome editing method that uses a modified Cas9 enzyme and a guide RNA to target and introduce specific genetic changes without causing double-strand breaks (Anzalone *et al.*, 2019). However, public matched genomic and functional data for CRISPRi experiments are limited but one exception is the Project score database of whole-genome CRISPR–Cas9 gene knockout screens which captures transcriptional changes (RNA-Seq) for over 18,000 genes in over 300 cancer cell lines (Dwane *et al.*, 2021). Overall, MAVE and CRISPRi assays are publicly available and should be included in such a benchmark for the field but careful processing and selection is required to fully utilise these resources.

There are also other *in vivo* study options for benchmark datasets not captured by GV-Rep which would avoid LD issues and are worth consideration. These include studies of rare variant associations with common diseases (RVCD) with large effects, represented in **Figure 7-1**, which render the variants detectable in small populations and unlikely to be confounded by LD (Turkmen and Lin, 2017). Or, finally, to consider individual genomes rather than genetic variants in isolation to avoid the effect of LD like the recent work previously discussed (Sasse, Ng, A. E. Spiro, *et al.*, 2023). However this last approach does not test a model's ability to differentiate between causal and tagged-SNPs. All of these population study options would require substantial investment in standardised assaying, sequencing data collection to gather larger repositories of variant effects and so are not viable as immediate candidates. Moreover the substantial investment may not be warranted given the relatively small payoff in the amount of natural variation observed across populations - approximately 1% of the person's genome will differ (Trost, Loureiro and Scherer, 2021).

In conclusion, while these differing strategies for benchmarking approaches provide valuable options, they require differing levels of processing to align with the functional outputs of genomic deep learning models. Moreover, it is essential to recognise that no single method is universally applicable, and it is likely that combining multiple of these approaches will yield the most robust insights into model performances for comprehension of the cis-regulatory code and personalised genomics. Finally, it is important to note that these benchmarks should match the cell types/states, genomic contexts or transcription/epigenetic mark that the models were trained on to give a fair measure of performance. For example, if a genomic deep learning model was trained to predict transcription factor binding of in vivo cell type A, a benchmark based of whole-genome CRISPR–Cas9 gene knockout screens measuring RNA-Seq in cell line B would not be informative. Coordinated efforts to decide on which cell types and functional measurements to use across the differing data sources will be paramount for the field. The Impact of Genomic Variation on Function (IGVF) Consortium has recently been setup to map genetic variant effects across cell types, including predictive modelling (IGVF Consortium, 2023). However, it remains to be seen whether developing such benchmarks will be within their remit.

7.3.4. An alternative approach to prioritising disease-related genomic loci

As is evident from the **Sections 7.3.2** and **7.3.3**, there are numerous open challenges to be addressed before genomic deep learning models can reach their potential of uncovering the elusive cell type and state-specific, cis-regulatory code. Thus, alternative avenues of insight to genomic deep learning models, where the goal is to interpret sequence grammar of DNA, should be considered. In **Chapter 6** of this thesis, despite little previous attention in the field,

we show how epigenetics, specifically histone mark, deep learning models can be used as an alternative approach to prioritising disease-related and functionally-relevant genomic loci.

An advantage of these models, as opposed to those focused on DNA sequence, is their ability to circumvent the challenges associated with interpreting single base pair changes, i.e. genetic variants, and LD. These challenges are notable in current models and become particularly pronounced as the genomic window considered expands (Huang *et al.*, 2023; Linder *et al.*, 2023; Sasse, Ng, A. E. Spiro, *et al.*, 2023; Toneyan and Koo, 2024), as detailed in **Section 7.3.2**. By focusing on perturbations of histone mark levels, or other epigenetic signals, to identify genomic loci of interest, these models avoid making SNP permutations or any issues associated with LD. Using such an approach, we demonstrate these model's ability to identify loci harbouring causal SNPs **(Figure 6-8)** – albeit, based on fine-mapping approaches. I believe future research should also focus on alternative approaches of applying deep learning to interpret the cis-regulatory code and believe the frameworks we created can be expanded on for this goal.

7.4. Final words

Computational biology is a paramount part of modern research and as more and more efforts are made to make experimental data available across the research community, there is increasing pressure to ensure that it is analysed in a standardised and best-practices manner. Moreover, greater research emphasis should be placed on interpreting the cell type-specific effect disease-relevant genetic variants cause and the related transcriptional and regulatory mechanisms, both for AD and the wider disease research field. Such work holds the potential to yield improved, personalised therapeutics.

My PhD thesis focused on developing computational and machine learning techniques to robustly detect the genome's cell type-specific, protein coding and non-coding effects in AD. This work spanned both AD-specific and broader advancements, addressing shortcomings in the standardisation of processing and quality control of genetic information, the standardisation of analysis of single-cell transcriptional changes in disease and the development of genomic and histone mark deep learning models to attempt to elucidate the functional role of disease relevant genomic loci and genetic variants in a cell type-specific manner. The developed open-source, computational and deep learning techniques are all broadly applicable to our comprehension of the cis-regulatory code.

In closing, I hope that the methodologies and insights presented in this thesis will not only advance our understanding of AD, leading to better treatments for patients, but also serve as valuable tools for the broader scientific community in deciphering the complex regulatory mechanisms that underlie a wide array of genetic diseases.

References

Abadi, M. *et al.* (2016) 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems'. arXiv. Available at: https://doi.org/10.48550/arXiv.1603.04467.

Aguirre, M. *et al.* (2021) 'Polygenic risk modeling with latent trait-related genetic components', *European Journal of Human Genetics*, pp. 1–11. Available at: https://doi.org/10.1038/s41431-021-00813-0.

Al-Chalabi, A. *et al.* (2010) 'An estimate of amyotrophic lateral sclerosis heritability using twin data', *Journal of Neurology, Neurosurgery, and Psychiatry*, 81(12), pp. 1324–1326. Available at: https://doi.org/10.1136/jnnp.2010.207464.

Aleixo, E.L. *et al.* (2023) 'Catastrophic Forgetting in Deep Learning: A Comprehensive Taxonomy'. arXiv. Available at: https://doi.org/10.48550/arXiv.2312.10549.

Amezquita, R.A. *et al.* (2020) 'Orchestrating single-cell analysis with Bioconductor', *Nature Methods*, 17(2), pp. 137–145. Available at: https://doi.org/10.1038/s41592-019-0654-x.

Anzalone, A.V. *et al.* (2019) 'Search-and-replace genome editing without double-strand breaks or donor DNA', *Nature*, 576(7785), pp. 149–157. Available at: https://doi.org/10.1038/s41586-019-1711-4.

Arbesfeld, J.A. *et al.* (2023) 'Mapping MAVE data for use in human genomics applications'. bioRxiv, p. 2023.06.20.545702. Available at: https://doi.org/10.1101/2023.06.20.545702.

Arnold, S.E. *et al.* (1991) 'The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with Alzheimer's disease', *Cerebral Cortex (New York, N.Y.: 1991)*, 1(1), pp. 103–116. Available at: https://doi.org/10.1093/cercor/1.1.103.

Arthur, T.D. *et al.* (2024) 'Multi-omic QTL mapping in early developmental tissues reveals phenotypic and temporal complexity of regulatory variants underlying GWAS loci'. bioRxiv, p. 2024.04.10.588874. Available at: https://doi.org/10.1101/2024.04.10.588874.

Atak, Z.K. *et al.* (2021) 'Interpretation of allele-specific chromatin accessibility using cell state–aware deep learning', *Genome Research*, 31(6), pp. 1082–1096. Available at: https://doi.org/10.1101/gr.260851.120.

Atlasi, Y. and Stunnenberg, H.G. (2017) 'The interplay of epigenetic marks during stem cell differentiation and development', *Nature Reviews Genetics*, 18(11), pp. 643–658. Available at: https://doi.org/10.1038/nrg.2017.57.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. Available at: https://doi.org/10.1038/nature15393.

Avsec, Ž. *et al.* (2018) 'Kipoi: accelerating the community exchange and reuse of predictive models for genomics'. bioRxiv, p. 375345. Available at: https://doi.org/10.1101/375345.

Avsec, Ž., Weilert, M., *et al.* (2021) 'Base-resolution models of transcription-factor binding reveal soft motif syntax', *Nature Genetics*, 53(3), pp. 354–366. Available at: https://doi.org/10.1038/s41588-021-00782-6.

Avsec, Ž., Agarwal, V., *et al.* (2021) 'Effective gene expression prediction from sequence by integrating long-range interactions', *Nature Methods*, 18(10), pp. 1196–1203. Available at: https://doi.org/10.1038/s41592-021-01252-x.

Bae, S. and Lesch, B.J. (2020) 'H3K4me1 Distribution Predicts Transcription State and Poising at Promoters', *Frontiers in Cell and Developmental Biology*, 8. Available at: https://www.frontiersin.org/articles/10.3389/fcell.2020.00289 (Accessed: 24 April 2023).

Banerji, J., Rusconi, S. and Schaffner, W. (1981) 'Expression of a β-globin gene is enhanced by remote SV40 DNA sequences', *Cell*, 27(2, Part 1), pp. 299–308. Available at: https://doi.org/10.1016/0092-8674(81)90413-X.

Bannister, A.J. and Kouzarides, T. (2011) 'Regulation of chromatin by histone modifications', *Cell Research*, 21(3), pp. 381–395. Available at: https://doi.org/10.1038/cr.2011.22.

Bartosovic, M., Kabbe, M. and Castelo-Branco, G. (2021) 'Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues', *Nature Biotechnology*, 39(7), pp. 825–835. Available at: https://doi.org/10.1038/s41587-021-00869-9.

Bassnett, S. (2009) 'On the mechanism of organelle degradation in the vertebrate lens', *Experimental Eye Research*, 88(2), pp. 133–139. Available at: https://doi.org/10.1016/j.exer.2008.08.017.

Bellenguez, C. *et al.* (2022a) 'New insights into the genetic etiology of Alzheimer's disease and related dementias', *Nature Genetics*, 54(4), pp. 412–436. Available at: https://doi.org/10.1038/s41588-022-01024-z.

Bellenguez, C. *et al.* (2022b) 'New insights into the genetic etiology of Alzheimer's disease and related dementias', *Nature Genetics*, pp. 1–25. Available at: https://doi.org/10.1038/s41588-022-01024-z.

Benner, C. *et al.* (2016) 'FINEMAP: efficient variable selection using summary data from genome-wide association studies', *Bioinformatics*, 32(10), pp. 1493–1501. Available at: https://doi.org/10.1093/bioinformatics/btw018.

Bennett, D.A. *et al.* (2006) 'Neuropathology of older persons without cognitive impairment from two community-based studies', *Neurology*, 66(12), pp. 1837–1844. Available at: https://doi.org/10.1212/01.wnl.0000219668.47116.e6.

Bernstein, B.E. *et al.* (2006) 'A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells', *Cell*, 125(2), pp. 315–326. Available at: https://doi.org/10.1016/j.cell.2006.02.041.

de Boer, C.G. and Taipale, J. (2024) 'Hold out the genome: a roadmap to solving the cis-regulatory code', *Nature*, 625(7993), pp. 41–50. Available at: https://doi.org/10.1038/s41586-023-06661-w.

Bohacek, J. and Mansuy, I.M. (2013) 'Epigenetic Inheritance of Disease and Disease Risk', *Neuropsychopharmacology*, 38(1), pp. 220–236. Available at: https://doi.org/10.1038/npp.2012.110.

Boix, C.A. *et al.* (2021) 'Regulatory genomic circuitry of human disease loci by integrative epigenomics', *Nature*, 590(7845), pp. 300–307. Available at: https://doi.org/10.1038/s41586-020-03145-z.

Bondareff, W., Mountjoy, C.Q. and Roth, M. (1981) 'Selective loss of neurones of origin of adrenergic projection to cerebral cortex (nucleus locus coeruleus) in senile dementia', *Lancet (London, England)*, 1(8223), pp. 783–784. Available at: https://doi.org/10.1016/s0140-6736(81)92657-x.

Boomsma, D., Busjahn, A. and Peltonen, L. (2002) 'Classical twin studies and beyond', *Nature Reviews Genetics*, 3(11), pp. 872–882. Available at: https://doi.org/10.1038/nrg932.

Bosselut, R. (2016) 'Pleiotropic Functions of H3K27Me3 Demethylases in Immune Cell Differentiation', *Trends in Immunology*, 37(2), pp. 102–113. Available at: https://doi.org/10.1016/j.it.2015.12.004.

Brase, L. *et al.* (2023) 'Single-nucleus RNA-sequencing of autosomal dominant Alzheimer disease and risk variant carriers', *Nature Communications*, 14(1), p. 2314. Available at: https://doi.org/10.1038/s41467-023-37437-5.

Breijyeh, Z. and Karaman, R. (2020) 'Comprehensive Review on Alzheimer's Disease: Causes and Treatment', *Molecules (Basel, Switzerland)*, 25(24), p. E5789. Available at: https://doi.org/10.3390/molecules25245789.

Broekema, R.V., Bakker, O.B. and Jonkers, I.H. (2020) 'A practical view of fine-mapping and gene prioritization in the post-genome-wide association era', *Open Biology*, 10(1), p. 190221. Available at: https://doi.org/10.1098/rsob.190221.

Brooks, M.E. *et al.* (2017) 'glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling', *The R Journal*, 9(2), pp. 378–400.

Buckner, J.H. and Ziegler, S.F. (2004) 'Regulating the immune system: the induction of regulatory T cells in the periphery', *Arthritis Res Ther*, 6(5), p. 215. Available at: https://doi.org/10.1186/ar1226.

Buenrostro, J.D. *et al.* (2013) 'Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nature Methods*, 10(12), pp. 1213–1218. Available at: https://doi.org/10.1038/nmeth.2688.

Bulik-Sullivan, B. *et al.* (2015) 'An atlas of genetic correlations across human diseases and traits', *Nature Genetics*, 47(11), pp. 1236–1241. Available at: https://doi.org/10.1038/ng.3406.

Buniello, A. *et al.* (2019) 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic Acids Research*, 47(D1), pp. D1005–D1012. Available at: https://doi.org/10.1093/nar/gky1120.

Burgess, D.J. (2022) 'Fine-mapping causal variants — why finding "the one" can be futile', *Nature Reviews Genetics*, pp. 1–1. Available at: https://doi.org/10.1038/s41576-022-00484-7.

Bushweller, J.H. (2019) 'Targeting transcription factors in cancer — from undruggable to reality', *Nature reviews. Cancer*, 19(11), pp. 611–624. Available at: https://doi.org/10.1038/s41568-019-0196-7.

Cai, Y. *et al.* (2021) 'H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions', *Nature Communications*, 12(1), p. 719. Available at: https://doi.org/10.1038/s41467-021-20940-y.

Cao, J. *et al.* (2020) 'A human cell atlas of fetal gene expression', *Science (New York, N.Y.)*, 370(6518), p. eaba7721. Available at: https://doi.org/10.1126/science.aba7721.

Cao, R. *et al.* (2002) 'Role of histone H3 lysine 27 methylation in Polycomb-group silencing', *Science (New York, N.Y.)*, 298(5595), pp. 1039–1043. Available at: https://doi.org/10.1126/science.1076997.

Carrozza, M.J. *et al.* (2005) 'Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription', *Cell*, 123(4), pp. 581–592. Available at: https://doi.org/10.1016/j.cell.2005.10.023.

Ceroni, F. *et al.* (2015) 'Quantifying cellular capacity identifies gene expression designs with reduced burden', *Nature Methods*, 12(5), pp. 415–418. Available at: https://doi.org/10.1038/nmeth.3339.

Ceroni, F. *et al.* (2018) 'Burden-driven feedback control of gene expression', *Nature Methods*, 15(5), pp. 387–393. Available at: https://doi.org/10.1038/nmeth.4635.

Chantalat, S. *et al.* (2011) 'Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin', *Genome Research*, 21(9), pp. 1426–1437. Available at: https://doi.org/10.1101/gr.118091.110.

Chen, K.M. *et al.* (2022) 'A sequence-based global map of regulatory activity for deciphering human genetics', *Nature Genetics*, 54(7), pp. 940–949. Available at: https://doi.org/10.1038/s41588-022-01102-2.

Chen, X.-F. *et al.* (2013) 'Transcriptional regulation and its misregulation in Alzheimer's disease', *Molecular Brain*, 6, p. 44. Available at: https://doi.org/10.1186/1756-6606-6-44.

Chen, Y., Lun, A.T.L. and Smyth, G.K. (2016) 'From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline', *F1000Research*, 5, p. 1438. Available at: https://doi.org/10.12688/f1000research.8987.2.

Chicco, D. and Jurman, G. (2020) 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 21(1), p. 6. Available at: https://doi.org/10.1186/s12864-019-6413-7.

Chilton, J. (2020a) *Astrocyte*. Available at: https://doi.org/10.5281/zenodo.3926048.

Chilton, J. (2020b) *Microglia resting*. Available at: https://doi.org/10.5281/zenodo.3926033.

Chilton, J. (2020c) *Pyramidal neuron 1*. Available at: https://doi.org/10.5281/zenodo.3926221.

Chollet, F. (2019) *On the Measure of Intelligence*, *arXiv.org*. Available at: https://arxiv.org/abs/1911.01547v2 (Accessed: 26 September 2024).

Chow, T.W. *et al.* (1999) 'Inheritance of Frontotemporal Dementia', *Archives of neurology*, 56(7), pp. 817–822.

Citron, B.A. *et al.* (2008) 'Transcription factor Sp1 dysregulation in Alzheimer's disease', *Journal of Neuroscience Research*, 86(11), pp. 2499–2504. Available at: https://doi.org/10.1002/jnr.21695.

Claussnitzer, M. *et al.* (2024) 'Minimum information and guidelines for reporting a multiplexed assay of variant effect', *Genome Biology*, 25(1), p. 100. Available at: https://doi.org/10.1186/s13059-024-03223-9.

Coller, H.A. (2019) 'Regulation of Cell Cycle Entry and Exit: A Single Cell Perspective', *Comprehensive Physiology*, 10(1), pp. 317–344. Available at: https://doi.org/10.1002/cphy.c190014.

Core, L.J., Waterfall, J.J. and Lis, J.T. (2008) 'Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters', *Science (New York, N.Y.)*, 322(5909), pp. 1845–1848. Available at: https://doi.org/10.1126/science.1162228.

Craig, J.M. *et al.* (2020) 'The Value of Twins for Health and Medical Research: A Third of a Century of Progress', *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies*, 23(1), pp. 8–15. Available at: https://doi.org/10.1017/thg.2020.4.

CRICK, F.H. (1958) 'On protein synthesis', *Symposia of the Society for Experimental Biology*, 12, pp. 138–163.

Crowell, H.L. *et al.* (2020) 'muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data', *Nature Communications*, 11(1), p. 6077. Available at: https://doi.org/10.1038/s41467-020-19894-4.

Cummings, J. *et al.* (2024) 'Anti-Amyloid Monoclonal Antibodies for the Treatment of Alzheimer's Disease', *BioDrugs*, 38(1), pp. 5–22. Available at: https://doi.org/10.1007/s40259-023-00633-2.

Dai, Z. *et al.* (2019) 'Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context'. arXiv. Available at: https://doi.org/10.48550/arXiv.1901.02860.

Dalla-Torre, H. *et al.* (2023) 'The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics'. bioRxiv, p. 2023.01.11.523679. Available at: https://doi.org/10.1101/2023.01.11.523679.

Das, B. and Yan, R. (2017) 'Role of BACE1 in Alzheimer's synaptic function', *Translational Neurodegeneration*, 6, p. 23. Available at: https://doi.org/10.1186/s40035-017-0093-5.

Davies, G. *et al.* (2018) 'Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function', *Nature Communications*, 9(1), p. 2098. Available at: https://doi.org/10.1038/s41467-018-04362-x.

Davies, M.N. *et al.* (2015) 'Generalised Anxiety Disorder – A Twin Study of Genetic Architecture, Genome-Wide Association and Differential Gene Expression', *PLOS ONE*, 10(8), p. e0134865. Available at: https://doi.org/10.1371/journal.pone.0134865.

Delihas, N. (2015) 'Discovery and characterization of the first non-coding RNA that regulates gene expression, micF RNA: A historical perspective', *World Journal of Biological Chemistry*, 6(4), pp. 272–280. Available at: https://doi.org/10.4331/wjbc.v6.i4.272.

Devlin, J. *et al.* (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. *NAACL-HLT 2019*, Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. Available at: https://doi.org/10.18653/v1/N19-1423.

Dönertaş, H.M. *et al.* (2021) 'Common genetic associations between age-related diseases', *Nature Aging*, 1(4), pp. 400–412. Available at: https://doi.org/10.1038/s43587-021-00051-5.

Dowle, M. and Srinivasan, A. (2021) *data.table: Extension of `data.frame`*. Available at: https://CRAN.R-project.org/package=data.table.

Dunham, I. *et al.* (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489(7414), pp. 57–74. Available at: https://doi.org/10.1038/nature11247.

Durham, T.J. *et al.* (2018) 'PREDICTD PaRallel Epigenomics Data Imputation with Cloud-based Tensor Decomposition', *Nature Communications*, 9(1), p. 1402. Available at: https://doi.org/10.1038/s41467-018-03635-9.

Dwane, L. *et al.* (2021) 'Project Score database: a resource for investigating cancer cell dependencies and prioritizing therapeutic targets', *Nucleic Acids Research*, 49(D1), pp. D1365–D1372. Available at: https://doi.org/10.1093/nar/gkaa882.

Eichler, E.E. (2019) 'Genetic Variation, Comparative Genomics, and the Diagnosis of Disease', *The New England journal of medicine*, 381(1), pp. 64–74. Available at: https://doi.org/10.1056/NEJMra1809315.

Elsworth, B. *et al.* (2020) 'The MRC IEU OpenGWAS data infrastructure'. bioRxiv, p. 2020.08.10.244293. Available at: https://doi.org/10.1101/2020.08.10.244293.

ENCODE Project Consortium (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature*, 489(7414), pp. 57–74. Available at: https://doi.org/10.1038/nature11247.

Eraslan, G. *et al.* (2019) 'Deep learning: new computational modelling techniques for genomics', *Nature Reviews Genetics*, 20(7), pp. 389–403. Available at: https://doi.org/10.1038/s41576-019-0122-6.

Erkkinen, M.G., Kim, M.-O. and Geschwind, M.D. (2018) 'Clinical Neurology and Epidemiology of the Major Neurodegenerative Diseases', *Cold Spring Harbor Perspectives in Biology*, 10(4), p. a033118. Available at: https://doi.org/10.1101/cshperspect.a033118.

Ernst, J. and Kellis, M. (2015) 'Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues', *Nature Biotechnology*, 33(4), pp. 364–376. Available at: https://doi.org/10.1038/nbt.3157.

Ernst, J. and Kellis, M. (2017) 'Chromatin-state discovery and genome annotation with ChromHMM', *Nature Protocols*, 12(12), pp. 2478–2492. Available at: https://doi.org/10.1038/nprot.2017.124.

Esposito, D. *et al.* (2019) 'MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect', *Genome Biology*, 20(1), p. 223. Available at: https://doi.org/10.1186/s13059-019-1845-6.

Faraone, S.V. and Larsson, H. (2019) 'Genetics of attention deficit hyperactivity disorder', *Molecular Psychiatry*, 24(4), pp. 562–575. Available at: https://doi.org/10.1038/s41380-018-0070-0.

Farley, E.K. *et al.* (2015) 'Suboptimization of developmental enhancers', *Science (New York, N.Y.)*, 350(6258), pp. 325–328. Available at: https://doi.org/10.1126/science.aac6948.

Fernandes, H.J.R. *et al.* (2020) 'Single-Cell Transcriptomics of Parkinson's Disease Human In Vitro Models Reveals Dopamine Neuron-Specific Stress Responses', *Cell Reports*, 33(2). Available at: https://doi.org/10.1016/j.celrep.2020.108263.

Finak, G. *et al.* (2015) 'MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data', *Genome Biology*, 16(1), p. 278. Available at: https://doi.org/10.1186/s13059-015-0844-5.

Finucane, H.K. *et al.* (2015) 'Partitioning heritability by functional annotation using genome-wide association summary statistics', *Nature Genetics*, 47(11), pp. 1228–1235. Available at: https://doi.org/10.1038/ng.3404.

Foret, M.R. *et al.* (2014) 'Molecular targets of chromatin repressive mark H3K9me3 in primate progenitor cells within adult neurogenic niches', *Frontiers in Genetics*, 5, p. 252. Available at: https://doi.org/10.3389/fgene.2014.00252.

Franzén, O., Gan, L.-M. and Björkegren, J.L.M. (2019) 'PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data', *Database*, 2019, p. baz046. Available at: https://doi.org/10.1093/database/baz046.

Fukushima, H.S., Takeda, H. and Nakamura, R. (2023) 'Incomplete erasure of histone marks during epigenetic reprogramming in medaka early development', *Genome Research*, 33(4), pp. 572–586. Available at: https://doi.org/10.1101/gr.277577.122.

Galliano, E. (2020) *neuron*. Available at: https://doi.org/10.5281/zenodo.3926535.

Gao, X. *et al.* (2022) 'Epigenetics in Alzheimer's Disease', *Frontiers in Aging Neuroscience*, 14. Available at: https://doi.org/10.3389/fnagi.2022.911635.

García, M.A. *et al.* (1996) 'Transcription Factor AP-2 Regulates Human Apolipoprotein E Gene Expression in Astrocytoma Cells', *The Journal of Neuroscience*, 16(23), pp. 7550–7556. Available at: https://doi.org/10.1523/JNEUROSCI.16-23-07550.1996.

Garg, M. *et al.* (2021) 'Meta-analysis of COVID-19 single-cell studies confirms eight key immune responses', *Scientific Reports*, 11(1), p. 20833. Available at: https://doi.org/10.1038/s41598-021-00121-z.

Gates, L.A. *et al.* (2017) 'Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation', *The Journal of Biological Chemistry*, 292(35), pp. 14456–14472. Available at: https://doi.org/10.1074/jbc.M117.802074.

Gatz, M. *et al.* (2006) 'Role of genes and environments for explaining Alzheimer disease', *Archives of General Psychiatry*, 63(2), pp. 168–174. Available at: https://doi.org/10.1001/archpsyc.63.2.168.

Gerrits, E. *et al.* (2021) 'Distinct amyloid-β and tau-associated microglia profiles in Alzheimer's disease', *Acta Neuropathologica*, 141(5), pp. 681–696. Available at: https://doi.org/10.1007/s00401-021-02263-w.

Ghannay, S. *et al.* (2016) 'Word Embedding Evaluation and Combination', in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. *LREC 2016*, Portorož, Slovenia: European Language Resources Association (ELRA), pp. 300–305. Available at: https://aclanthology.org/L16-1046 (Accessed: 24 April 2023).

Ghodsian, N. *et al.* (2021) 'Electronic health record-based genome-wide meta-analysis provides insights on the genetic architecture of non-alcoholic fatty liver disease', *Cell Reports. Medicine*, 2(11), p. 100437. Available at: https://doi.org/10.1016/j.xcrm.2021.100437.

Gloudemans, M. (2021) 'mikegloudemans/gwas-download'. Available at: https://github.com/mikegloudemans/gwas-download (Accessed: 20 April 2021).

Goldman, S.M. *et al.* (2019) 'Concordance for Parkinson's disease in twins: A 20-year update', *Annals of Neurology*, 85(4), pp. 600–605. Available at: https://doi.org/10.1002/ana.25441.

González-Ramírez, M. *et al.* (2021) 'Differential contribution to gene expression prediction of histone modifications at enhancers or promoters', *PLOS Computational Biology*, 17(9), p. e1009368. Available at: https://doi.org/10.1371/journal.pcbi.1009368.

Gosai, S. *et al.* (2023) 'Machine-guided design of synthetic cell type-specific cis-regulatory elements', *bioRxiv*, p. 2023.08.08.552077. Available at: https://doi.org/10.1101/2023.08.08.552077.

Gotmare, A. *et al.* (2018) 'A Closer Look at Deep Learning Heuristics: Learning rate restarts, Warmup and Distillation'. arXiv. Available at: https://doi.org/10.48550/arXiv.1810.13243.

Grishkevich, V. and Yanai, I. (2014) 'Gene length and expression level shape genomic novelties', *Genome Research*, 24(9), p. 1497. Available at: https://doi.org/10.1101/gr.169722.113.

Grosselin, K. *et al.* (2019) 'High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer', *Nature Genetics*, 51(6), pp. 1060–1066. Available at: https://doi.org/10.1038/s41588-019-0424-9.

Grove, J. *et al.* (2019) 'Identification of common genetic risk variants for autism spectrum disorder', *Nature Genetics*, 51(3), pp. 431–444. Available at: https://doi.org/10.1038/s41588-019-0344-8.

Guan, J.-S. *et al.* (2009) 'HDAC2 negatively regulates memory formation and synaptic plasticity', *Nature*, 459(7243), pp. 55–60. Available at: https://doi.org/10.1038/nature07925.

Haass, C. *et al.* (2012) 'Trafficking and Proteolytic Processing of APP', *Cold Spring Harbor Perspectives in Medicine*, 2(5), p. a006270. Available at: https://doi.org/10.1101/cshperspect.a006270.

Hagenbeek, F.A. *et al.* (2023) 'Maximizing the value of twin studies in health and behaviour', *Nature Human Behaviour*, 7(6), pp. 849–860. Available at: https://doi.org/10.1038/s41562-023-01609-6.

Hao, Y. *et al.* (2021) 'Integrated analysis of multimodal single-cell data', *Cell*, 184(13), pp. 3573-3587.e29. Available at: https://doi.org/10.1016/j.cell.2021.04.048.

Hayhurst, J. *et al.* (2023) 'A community driven GWAS summary statistics standard'. bioRxiv, p. 2022.07.15.500230. Available at: https://doi.org/10.1101/2022.07.15.500230.

Hechtlinger, Y. (2016) 'Interpretation of Prediction Models Using the Input Gradient'. arXiv. Available at: https://doi.org/10.48550/arXiv.1611.07634.

Heinz, S. *et al.* (2010) 'Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities', *Molecular cell*, 38(4), pp. 576–589. Available at: https://doi.org/10.1016/j.molcel.2010.05.004.

Herrera-Uribe, J. *et al.* (2020) 'Changes in H3K27ac at Gene Regulatory Regions in Porcine Alveolar Macrophages Following LPS or PolyIC Exposure', *Frontiers in Genetics*, 11. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00817 (Accessed: 24 April 2023).

Heumos, L. *et al.* (2023) 'Best practices for single-cell analysis across modalities', *Nature Reviews Genetics*, 24(8), pp. 550–572. Available at: https://doi.org/10.1038/s41576-023-00586-w.

Hilker, R. *et al.* (2018) 'Heritability of Schizophrenia and Schizophrenia Spectrum Based on the Nationwide Danish Twin Register', *Biological Psychiatry*, 83(6), pp. 492–498. Available at: https://doi.org/10.1016/j.biopsych.2017.08.017.

Hnisz, D. *et al.* (2013) 'Super-enhancers in the control of cell identity and disease', *Cell*, 155(4), pp. 934–947. Available at: https://doi.org/10.1016/j.cell.2013.09.053.

Hodge, R.D. *et al.* (2019) 'Conserved cell types with divergent features in human versus mouse cortex', *Nature*, 573(7772), pp. 61–68. Available at: https://doi.org/10.1038/s41586-019-1506-7.

't Hoen, P.A.C. *et al.* (2013) 'Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories', *Nature Biotechnology*, 31(11), pp. 1015–1022. Available at: https://doi.org/10.1038/nbt.2702.

Hoffman, M.M. *et al.* (2012) 'Unsupervised pattern discovery in human chromatin structure through genomic segmentation', *Nature methods*, 9(5), pp. 473–476. Available at: https://doi.org/10.1038/nmeth.1937.

Højsgaard, S., Halekoh, U. and Yan, J. (2006) 'The R Package geepack for Generalized Estimating Equations', *Journal of Statistical Software*, 15, pp. 1–11. Available at: https://doi.org/10.18637/jss.v015.i02.

Hu, D. *et al.* (2023) 'CUT&Tag recovers up to half of ENCODE ChIP-seq peaks in modifications of H3K27'. bioRxiv, p. 2022.03.30.486382. Available at: https://doi.org/10.1101/2022.03.30.486382.

Huang, C. *et al.* (2023) 'Personal transcriptome variation is poorly explained by current genomic deep learning models', *Nature Genetics*, 55(12), pp. 2056–2059. Available at: https://doi.org/10.1038/s41588-023-01574-w.

Ibragimov, A.N., Bylino, O.V. and Shidlovskii, Y.V. (2020) 'Molecular Basis of the Function of Transcriptional Enhancers', *Cells*, 9(7), p. 1620. Available at: https://doi.org/10.3390/cells9071620.

IGVF Consortium, T. (2023) 'The Impact of Genomic Variation on Function (IGVF) Consortium', *ArXiv*, p. arXiv:2307.13708v1.

Ilicic, T. *et al.* (2016) 'Classification of low quality cells from single-cell RNA-seq data', *Genome Biology*, 17(1), p. 29. Available at: https://doi.org/10.1186/s13059-016-0888-1.

Jacob, F. and Monod, J. (1961) 'Genetic regulatory mechanisms in the synthesis of proteins', *Journal of Molecular Biology*, 3(3), pp. 318–356. Available at: https://doi.org/10.1016/S0022-2836(61)80072-7.

Jansen, I.E. *et al.* (2019) 'Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk', *Nature Genetics*, 51(3), pp. 404–413. Available at: https://doi.org/10.1038/s41588-018-0311-9.

Jansen, P.R. *et al.* (2019) 'Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways', *Nature Genetics*, 51(3), pp. 394–403. Available at: https://doi.org/10.1038/s41588-018-0333-3.

Jenuwein, T. and Allis, C.D. (2001) 'Translating the histone code', *Science (New York, N.Y.)*, 293(5532), pp. 1074–1080. Available at: https://doi.org/10.1126/science.1063127.

Jin, B., Li, Y. and Robertson, K.D. (2011) 'DNA Methylation', *Genes & Cancer*, 2(6), pp. 607–617. Available at: https://doi.org/10.1177/1947601910393957.

Johnson, D.S. *et al.* (2007) 'Genome-wide mapping of in vivo protein-DNA interactions', *Science (New York, N.Y.)*, 316(5830), pp. 1497–1502. Available at: https://doi.org/10.1126/science.1141319.

Johnson, W.E., Li, C. and Rabinovic, A. (2007) 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics*, 8(1), pp. 118–127. Available at: https://doi.org/10.1093/biostatistics/kxj037.

Jordan, B. (2002) 'Historical Background and Anticipated Developments', *Annals of the New York Academy of Sciences*, 975(1), pp. 24–32. Available at: https://doi.org/10.1111/j.1749-6632.2002.tb05938.x.

Joshi, A.A. and Struhl, K. (2005) 'Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation', *Molecular Cell*, 20(6), pp. 971–978. Available at: https://doi.org/10.1016/j.molcel.2005.11.021.

Jovic, D. *et al.* (2022) 'Single-cell RNA sequencing technologies and applications: A brief overview', *Clinical and Translational Medicine*, 12(3), p. e694. Available at: https://doi.org/10.1002/ctm2.694.

Jung, I. *et al.* (2019) 'A compendium of promoter-centered long-range chromatin interactions in the human genome', *Nature Genetics*, 51(10), pp. 1442–1449. Available at: https://doi.org/10.1038/s41588-019-0494-8.

Kaikkonen, M.U., Lam, M.T.Y. and Glass, C.K. (2011) 'Non-coding RNAs as regulators of gene expression and epigenetics', *Cardiovascular Research*, 90(3), pp. 430–440. Available at: https://doi.org/10.1093/cvr/cvr097.

Kampmann, M. (2024) 'Molecular and cellular mechanisms of selective vulnerability in neurodegenerative diseases', *Nature Reviews Neuroscience*, pp. 1–21. Available at: https://doi.org/10.1038/s41583-024-00806-0.

Kang, M. *et al.* (2020) 'Learning Cell-Type-Specific Gene Regulation Mechanisms by Multi-Attention Based Deep Learning With Regulatory Latent Space', *Frontiers in Genetics*, 11, p. 869. Available at: https://doi.org/10.3389/fgene.2020.00869.

Karlić, R. *et al.* (2010) 'Histone modification levels are predictive for gene expression', *Proceedings of the National Academy of Sciences*, 107(7), pp. 2926–2931. Available at: https://doi.org/10.1073/pnas.0909344107.

Karmodiya, K. *et al.* (2012) 'H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells', *BMC Genomics*, 13(1), p. 424. Available at: https://doi.org/10.1186/1471-2164-13-424.

Karollus, A. *et al.* (2024) 'Species-aware DNA language models capture regulatory elements and their evolution', *Genome Biology*, 25(1), p. 83. Available at: https://doi.org/10.1186/s13059-024-03221-x.

Karollus, A., Mauermeier, T. and Gagneur, J. (2023) 'Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers', *Genome Biology*, 24(1), p. 56. Available at: https://doi.org/10.1186/s13059-023-02899-9.

Keilwagen, J., Posch, S. and Grau, J. (2019) 'Accurate prediction of cell type-specific transcription factor binding', *Genome Biology*, 20, p. 9. Available at: https://doi.org/10.1186/s13059-018-1614-y.

Keinath, M.C., Prior, D.E. and Prior, T.W. (2021) 'Spinal Muscular Atrophy: Mutations, Testing, and Clinical Relevance', *The Application of Clinical Genetics*, 14, pp. 11–25. Available at: https://doi.org/10.2147/TACG.S239603.

Kelley, D.R. *et al.* (2018) 'Sequential regulatory activity prediction across chromosomes with convolutional neural networks', *Genome Research*, 28(5), pp. 739–750. Available at: https://doi.org/10.1101/gr.227819.117.

Kelley, D.R. (2020) 'Cross-species regulatory sequence activity prediction', *PLOS Computational Biology*, 16(7), p. e1008050. Available at: https://doi.org/10.1371/journal.pcbi.1008050.

Kelley, D.R., Snoek, J. and Rinn, J. (2016) 'Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks', *Genome Research*, p. gr.200535.115. Available at: https://doi.org/10.1101/gr.200535.115.

Kelley, D.R., Snoek, J. and Rinn, J.L. (2016) 'Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks', *Genome Research*, 26(7), pp. 990–999. Available at: https://doi.org/10.1101/gr.200535.115.

Kendler, K.S. *et al.* (2006) 'A Swedish National Twin Study of Lifetime Major Depression', *American Journal of Psychiatry*, 163(1), pp. 109–114. Available at: https://doi.org/10.1176/appi.ajp.163.1.109.

Kerimov, N. *et al.* (2021) 'A compendium of uniformly processed human gene expression and splicing quantitative trait loci', *Nature Genetics*, 53(9), pp. 1290–1299. Available at: https://doi.org/10.1038/s41588-021-00924-w.

Khan, A. and Zhang, X. (2016) 'dbSUPER: a database of super-enhancers in mouse and human genome', *Nucleic Acids Research*, 44(Database issue), pp. D164–D171. Available at: https://doi.org/10.1093/nar/gkv1002.

Khozoie, C. *et al.* (2021) 'scFlow: A Scalable and Reproducible Analysis Pipeline for Single-Cell RNA Sequencing Data'. bioRxiv, p. 2021.08.16.456499. Available at: https://doi.org/10.1101/2021.08.16.456499.

Kim, J., Basak, J.M. and Holtzman, D.M. (2009) 'The Role of Apolipoprotein E in Alzheimer's Disease', *Neuron*, 63(3), pp. 287–303. Available at: https://doi.org/10.1016/j.neuron.2009.06.026.

Kingma, D.P. and Ba, J. (2017) 'Adam: A Method for Stochastic Optimization'. arXiv. Available at: https://doi.org/10.48550/arXiv.1412.6980.

Klie, A. *et al.* (2023) 'Predictive analyses of regulatory sequences with EUGENe', *Nature Computational Science*, 3(11), pp. 946–956. Available at: https://doi.org/10.1038/s43588-023-00544-w.

Knopman, D.S. *et al.* (2021) 'Alzheimer disease', *Nature Reviews Disease Primers*, 7(1), pp. 1–21. Available at: https://doi.org/10.1038/s41572-021-00269-y.

Koch, G., Zemel, R. and Salakhutdinov, R. (2015) 'Siamese Neural Networks for One-shot Image Recognition', *ICML deep learning workshop* [Preprint].

Ku, M. *et al.* (2008) 'Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains', *PLoS Genetics*, 4(10), p. e1000242. Available at: https://doi.org/10.1371/journal.pgen.1000242.

Kumar, A. *et al.* (2019) 'ADULT NEUROGENESIS IN HUMANS: A Review of Basic Concepts, History, Current Research, and Clinical Implications', *Innovations in Clinical Neuroscience*, 16(5–6), pp. 30–37.

Kundaje, A. *et al.* (2015) 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518(7539), pp. 317–330. Available at: https://doi.org/10.1038/nature14248.

Kundu, K. *et al.* (2022) 'Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for 12 immune-mediated diseases', *Nature Genetics*, 54(3), pp. 251–262. Available at: https://doi.org/10.1038/s41588-022-01025-y.

Lal, A. *et al.* (2024) 'gReLU: A comprehensive framework for DNA sequence modeling and design'. bioRxiv, p. 2024.09.18.613778. Available at: https://doi.org/10.1101/2024.09.18.613778.

Lander, E.S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*, 409(6822), pp. 860–921. Available at: https://doi.org/10.1038/35057062.

Landrum, M.J. *et al.* (2020) 'ClinVar: improvements to accessing data', *Nucleic Acids Research*, 48(D1), pp. D835–D844. Available at: https://doi.org/10.1093/nar/gkz972.

Lareau, C.A. *et al.* (2019) 'Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility', *Nature Biotechnology*, 37(8), pp. 916–924. Available at: https://doi.org/10.1038/s41587-019-0147-6.

Larson, M.H. *et al.* (2013) 'CRISPR interference (CRISPRi) for sequence-specific control of gene expression', *Nature Protocols*, 8(11), pp. 2180–2196. Available at: https://doi.org/10.1038/nprot.2013.132.

Laws, S.M. *et al.* (2003) 'Expanding the association between the APOE gene and the risk of Alzheimer's disease: possible roles for APOE promoter polymorphisms and alterations in APOE transcription', *Journal of Neurochemistry*, 84(6), pp. 1215–1236. Available at: https://doi.org/10.1046/j.1471-4159.2003.01615.x.

Lazic, S.E. (2010) 'The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis?', *BMC Neuroscience*, 11(1), p. 5. Available at: https://doi.org/10.1186/1471-2202-11-5.

Lee, D., Yang, J. and Kim, S. (2022) 'Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer', *Nature Communications*, 13(1), p. 6678. Available at: https://doi.org/10.1038/s41467-022-34152-5.

Leeuw, C.A. de *et al.* (2015) 'MAGMA: Generalized Gene-Set Analysis of GWAS Data', *PLOS Computational Biology*, 11(4), p. e1004219. Available at: https://doi.org/10.1371/journal.pcbi.1004219.

Lettice, L.A. *et al.* (2003) 'A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly', *Human Molecular Genetics*, 12(14), pp. 1725–1735. Available at: https://doi.org/10.1093/hmg/ddg180.

Levine, M. and Davidson, E.H. (2005) 'Gene regulatory networks for development', *Proceedings of the National Academy of Sciences*, 102(14), pp. 4936–4942. Available at: https://doi.org/10.1073/pnas.0408031102.

Levine, M. and Tjian, R. (2003) 'Transcription regulation and animal diversity', *Nature*, 424(6945), pp. 147–151. Available at: https://doi.org/10.1038/nature01763.

Lewis, M. (2005) 'The lac repressor', *Comptes Rendus Biologies*, 328(6), pp. 521–548. Available at: https://doi.org/10.1016/j.crvi.2005.04.004.

Li, B. *et al.* (2023) 'SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension'. arXiv. Available at: https://doi.org/10.48550/arXiv.2307.16125.

Li, H. and Guan, Y. (2021) 'Fast decoding cell type–specific transcription factor binding landscape at single-nucleotide resolution', *Genome Research*, 31(4), pp. 721–731. Available at: https://doi.org/10.1101/gr.269613.120.

Li, J.-L. *et al.* (2006) 'Genome-wide significance for a modifier of age at neurological onset in Huntington's Disease at 6q23-24: the HD MAPS study', *BMC Medical Genetics*, 7(1), p. 71. Available at: https://doi.org/10.1186/1471-2350-7-71.

Li, W., Notani, D. and Rosenfeld, M.G. (2016) 'Enhancers as non-coding RNA transcription units: recent insights and future perspectives', *Nature Reviews Genetics*, 17(4), pp. 207–223. Available at: https://doi.org/10.1038/nrg.2016.4.

Li, Z. *et al.* (2024) 'GV-Rep: A Large-Scale Dataset for Genetic Variant Representation Learning'. arXiv. Available at: https://doi.org/10.48550/arXiv.2407.16940.

Liang, G. and Zhang, Y. (2013) 'Embryonic stem cell and induced pluripotent stem cell: an epigenetic perspective', *Cell Research*, 23(1), pp. 49–69. Available at: https://doi.org/10.1038/cr.2012.175.

Lieberman-Aiden, E. *et al.* (2009) 'Comprehensive mapping of long range interactions reveals folding principles of the human genome', *Science (New York, N.Y.)*, 326(5950), pp. 289–293. Available at: https://doi.org/10.1126/science.1181369.

Lin, H. *et al.* (2018) 'Common and Rare Coding Genetic Variation Underlying the Electrocardiographic PR Interval', *Circulation. Genomic and Precision Medicine*, 11(5), p. e002037. Available at: https://doi.org/10.1161/CIRCGEN.117.002037.

Linder, J. *et al.* (2023) 'Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation'. bioRxiv, p. 2023.08.30.555582. Available at: https://doi.org/10.1101/2023.08.30.555582.

Loh, P.-R. *et al.* (2018) 'Mixed-model association for biobank-scale datasets', *Nature Genetics*, 50(7), pp. 906–908. Available at: https://doi.org/10.1038/s41588-018-0144-6.

Lonsdale, J. *et al.* (2013) 'The Genotype-Tissue Expression (GTEx) project', *Nature Genetics*, 45(6), pp. 580–585. Available at: https://doi.org/10.1038/ng.2653.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. Available at: https://doi.org/10.1186/s13059-014-0550-8.

Lovén, J. *et al.* (2013) 'Selective inhibition of tumor oncogenes by disruption of super-enhancers', *Cell*, 153(2), pp. 320–334. Available at: https://doi.org/10.1016/j.cell.2013.03.036.

Luciano, M. *et al.* (2021) 'The influence of X chromosome variants on trait neuroticism', *Molecular Psychiatry*, 26(2), pp. 483–491. Available at: https://doi.org/10.1038/s41380-019-0388-2.

Lui, J.H. *et al.* (2021) 'Differential encoding in prefrontal cortex projection neuron classes across cognitive tasks', *Cell*, 184(2), pp. 489-506.e26. Available at: https://doi.org/10.1016/j.cell.2020.11.046.

Lun, A.T.L. *et al.* (2019) 'EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data', *Genome Biology*, 20(1), p. 63. Available at: https://doi.org/10.1186/s13059-019-1662-y.

Lynch, M. *et al.* (2016) 'Genetic drift, selection and the evolution of the mutation rate', *Nature Reviews Genetics*, 17(11), pp. 704–714. Available at: https://doi.org/10.1038/nrg.2016.104.

Lyon, M.S. *et al.* (2021) 'The variant call format provides efficient and robust storage of GWAS summary statistics', *Genome Biology*, 22(1), p. 32. Available at: https://doi.org/10.1186/s13059-020-02248-0.

Mahat, D.B. *et al.* (2016) 'Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq)', *Nature Protocols*, 11(8), pp. 1455–1476. Available at: https://doi.org/10.1038/nprot.2016.086.

Maiuri, T. *et al.* (2019) 'DNA Damage Repair in Huntington's Disease and Other Neurodegenerative Diseases', *Neurotherapeutics*, 16(4), pp. 948–956. Available at: https://doi.org/10.1007/s13311-019-00768-7.

Marian, A.J. (2014) 'Sequencing Your Genome: What Does It Mean?', *Methodist DeBakey Cardiovascular Journal*, 10(1), pp. 3–6.

Marzi, S.J. *et al.* (2018) 'A histone acetylome-wide association study of Alzheimer's disease identifies disease-associated H3K27ac differences in the entorhinal cortex', *Nature Neuroscience*, 21(11), pp. 1618–1627. Available at: https://doi.org/10.1038/s41593-018-0253-7.

Mathys, H. *et al.* (2019) 'Single-cell transcriptomic analysis of Alzheimer's disease', *Nature*, 570(7761), pp. 332–337. Available at: https://doi.org/10.1038/s41586-019-1195-2.

Mathys, H. *et al.* (2023) 'Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology', *Cell*, 186(20), pp. 4365-4385.e27. Available at: https://doi.org/10.1016/j.cell.2023.08.039.

McBrian, M.A. *et al.* (2013) 'Histone Acetylation Regulates Intracellular pH', *Molecular Cell*, 49(2), pp. 310–321. Available at: https://doi.org/10.1016/j.molcel.2012.10.025.

McCormack, M. *et al.* (2018) 'Genetic variation in CFH predicts phenytoin-induced maculopapular exanthema in European-descent patients', *Neurology*, 90(4), pp. e332–e341. Available at: https://doi.org/10.1212/WNL.0000000000004853.

McGinnis, C.S., Murrow, L.M. and Gartner, Z.J. (2019) 'DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors', *Cell Systems*, 8(4), pp. 329-337.e4. Available at: https://doi.org/10.1016/j.cels.2019.03.003.

McGuffin, P. *et al.* (2003) 'The heritability of bipolar affective disorder and the genetic relationship to unipolar depression', *Archives of General Psychiatry*, 60(5), pp. 497–502. Available at: https://doi.org/10.1001/archpsyc.60.5.497.

McInnes, L., Healy, J. and Melville, J. (2020) 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'. arXiv. Available at: https://doi.org/10.48550/arXiv.1802.03426.

McQuade, A. and Blurton-Jones, M. (2019) 'Microglia in Alzheimer's disease: Exploring how genetics and phenotype influence risk', *Journal of molecular biology*, 431(9), pp. 1805–1817. Available at: https://doi.org/10.1016/j.jmb.2019.01.045.

Meftah, S. and Gan, J. (2023) 'Alzheimer's disease as a synaptopathy: Evidence for dysfunction of synapses during disease progression', *Frontiers in Synaptic Neuroscience*, 15. Available at: https://doi.org/10.3389/fnsyn.2023.1129036.

Mehler, M.F. and Mattick, J.S. (2007) 'Noncoding RNAs and RNA editing in brain development, functional diversification, and neurological disease', *Physiological Reviews*, 87(3), pp. 799–823. Available at: https://doi.org/10.1152/physrev.00036.2006.

Mehrmohamadi, M. *et al.* (2021) 'A Comparative Overview of Epigenomic Profiling Methods', *Frontiers in Cell and Developmental Biology*, 9, p. 714687. Available at: https://doi.org/10.3389/fcell.2021.714687.

Meng, G. *et al.* (2023) 'Three-dimensional chromatin architecture datasets for aging and Alzheimer's disease', *Scientific Data*, 10, p. 51. Available at: https://doi.org/10.1038/s41597-023-01948-z.

Michán, S. *et al.* (2010) 'SIRT1 is essential for normal cognitive function and synaptic plasticity', *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 30(29), pp. 9695–9707. Available at: https://doi.org/10.1523/JNEUROSCI.0027-10.2010.

Miller, J.L. and Grant, P.A. (2013) 'The Role of DNA Methylation and Histone Modifications in Transcriptional Regulation in Humans', *Sub-cellular biochemistry*, 61, pp. 289–317. Available at: https://doi.org/10.1007/978-94-007-4525-4_13.

Min, J.L. *et al.* (2021) 'Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation', *Nature Genetics*, 53(9), pp. 1311–1321. Available at: https://doi.org/10.1038/s41588-021-00923-x.

Minnoye, L. *et al.* (2020) 'Cross-species analysis of enhancer logic using deep learning', *Genome Research*, 30(12), pp. 1815–1834. Available at: https://doi.org/10.1101/gr.260844.120.

Mitchell, P.J. and Tjian, R. (1989) 'Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins', *Science (New York, N.Y.)*, 245(4916), pp. 371–378. Available at: https://doi.org/10.1126/science.2667136.

Moore, J.E. *et al.* (2020) 'Expanded encyclopaedias of DNA elements in the human and mouse genomes', *Nature*, 583(7818), pp. 699–710. Available at: https://doi.org/10.1038/s41586-020-2493-4.

Morrow, A.K. *et al.* (2021) 'Epitome: predicting epigenetic events in novel cell types with multi-cell deep ensemble learning', *Nucleic Acids Research*, 49(19), p. e110. Available at: https://doi.org/10.1093/nar/gkab676.

Mostafavi, H. *et al.* (2023) 'Systematic differences in discovery of genetic effects on gene expression and complex traits', *Nature Genetics*, 55(11), pp. 1866–1875. Available at: https://doi.org/10.1038/s41588-023-01529-1.

Mullins, N. *et al.* (2021) 'Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology', *Nature Genetics*, 53(6), pp. 817–829. Available at: https://doi.org/10.1038/s41588-021-00857-4.

Mulvey, B., Lagunas, T. and Dougherty, J.D. (2021) 'Massively Parallel Reporter Assays: Defining Functional Psychiatric Genetic Variants across Biological Contexts.', *Biological psychiatry*, 89(1), pp. 76–89. Available at: https://doi.org/10.1016/j.biopsych.2020.06.011.

Murphy, A. (2022) 'Al-Murphy/reanalysis_scRNA_seq_benchmark: Reanalysis scRNA-seq benchmark'. Zenodo. Available at: https://doi.org/10.5281/zenodo.7356243.

Murphy, A. (2024) 'neurogenomics/chromexpress: Initial release'. Zenodo. Available at: https://doi.org/10.5281/zenodo.10940543.

Murphy, Alan E. *et al.* (2024) 'Predicting cell type-specific epigenomic profiles accounting for distal genetic effects', *Nature Communications*, 15(1), p. 9951. Available at: https://doi.org/10.1038/s41467-024-54441-5.

Murphy, Alan E *et al.* (2024) 'Predicting gene expression from histone marks using chromatin deep learning models depends on histone mark function, regulatory distance and cellular states', *Nucleic Acids Research*, p. gkae1212. Available at: https://doi.org/10.1093/nar/gkae1212.

Murphy, A.E., Fancy, N. and Skene, N. (2023) 'Avoiding false discoveries in single-cell RNA-seq by revisiting the first Alzheimer's disease dataset', *eLife*. Edited by J.-Y. An and M. Choi, 12, p. RP90214. Available at: https://doi.org/10.7554/eLife.90214.

Murphy, A.E., Schilder, B.M. and Skene, N.G. (2021) 'MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics', *Bioinformatics*, 37(23), pp. 4593–4596. Available at: https://doi.org/10.1093/bioinformatics/btab665.

Murphy, A.E. and Skene, N.G. (2022) 'A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis', *Nature Communications*, 13(1), p. 7851. Available at: https://doi.org/10.1038/s41467-022-35519-4.

Nair, S. *et al.* (2023) 'Transcription factor stoichiometry, motif affinity and syntax regulate single-cell chromatin dynamics during fibroblast reprogramming to pluripotency'. bioRxiv, p. 2023.10.04.560808. Available at: https://doi.org/10.1101/2023.10.04.560808.

Nalls, M.A. *et al.* (2019) 'Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies', *The Lancet. Neurology*, 18(12), pp. 1091–1102. Available at: https://doi.org/10.1016/S1474-4422(19)30320-5.

Narayan, P.J. *et al.* (2015) 'Increased acetyl and total histone levels in post-mortem Alzheimer's disease brain', *Neurobiology of Disease*, 74, pp. 281–294. Available at: https://doi.org/10.1016/j.nbd.2014.11.023.

Nichols, E. *et al.* (2022) 'Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019', *The Lancet Public Health*, 7(2), pp. e105–e125. Available at: https://doi.org/10.1016/S2468-2667(21)00249-8.

Nicolas, A. *et al.* (2018) 'Genome-wide Analyses Identify KIF5A as a Novel ALS Gene', *Neuron*, 97(6), pp. 1267–1288. Available at: https://doi.org/10.1016/j.neuron.2018.02.027.

Nilsberth, C. *et al.* (2001) 'The "Arctic" APP mutation (E693G) causes Alzheimer's disease by enhanced Aβ protofibril formation', *Nature Neuroscience*, 4(9), pp. 887–893. Available at: https://doi.org/10.1038/nn0901-887.

Niu, Y. *et al.* (2015) 'Oxidative stress alters global histone modification and DNA methylation', *Free Radical Biology and Medicine*, 82, pp. 22–28. Available at: https://doi.org/10.1016/j.freeradbiomed.2015.01.028.

Nordin, A. *et al.* (2024) 'Exhaustive identification of genome-wide binding events of transcriptional regulators', *Nucleic Acids Research*, p. gkae180. Available at: https://doi.org/10.1093/nar/gkae180.

Nott, A. *et al.* (2019) 'Brain cell type–specific enhancer–promoter interactome maps and disease-risk association', *Science*, 366(6469), pp. 1134–1139. Available at: https://doi.org/10.1126/science.aay0793.

Nurk, S. *et al.* (2022) 'The complete sequence of a human genome', *Science*, 376(6588), pp. 44–53. Available at: https://doi.org/10.1126/science.abj6987.

Odena, A. *et al.* (2021) 'BUSTLE: Bottom-Up Program Synthesis Through Learning-Guided Exploration'. arXiv. Available at: https://doi.org/10.48550/arXiv.2007.14381.

O'Leary, N.A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*, 44(D1), pp. D733-745. Available at: https://doi.org/10.1093/nar/gkv1189.

Padeken, J., Methot, S.P. and Gasser, S.M. (2022) 'Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance', *Nature Reviews*

Molecular Cell Biology, 23(9), pp. 623–640. Available at: https://doi.org/10.1038/s41580-022-00483-w.

Pagès, H. (2021) *BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs*. Available at: https://bioconductor.org/packages/BSgenome.

Park, P.J. (2009) 'ChIP-Seq: advantages and challenges of a maturing technology', *Nature reviews. Genetics*, 10(10), pp. 669–680. Available at: https://doi.org/10.1038/nrg2641.

Parreno, V. *et al.* (2024) 'Transient loss of Polycomb components induces an epigenetic cancer fate', *Nature* [Preprint]. Available at: https://doi.org/10.1038/s41586-024-07328-w.

Paszke, A. *et al.* (2017) 'Automatic differentiation in PyTorch'. Available at: https://openreview.net/forum?id=BJJsrmfCZ (Accessed: 15 July 2024).

Paulson, H.L. (2009) 'The Spinocerebellar Ataxias', *Journal of neuro-ophthalmology : the official journal of the North American Neuro-Ophthalmology Society*, 29(3), pp. 227–237. Available at: https://doi.org/10.1097/WNO0b013e3181b416de.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12(85), pp. 2825–2830.

Peltzer, A. *et al.* (2024) 'nf-core/scrnaseq: 2.5.1'. Zenodo. Available at: https://doi.org/10.5281/zenodo.10554425.

Pickrell, J.K. (2014) 'Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits', *The American Journal of Human Genetics*, 94(4), pp. 559–573. Available at: https://doi.org/10.1016/j.ajhg.2014.03.004.

Policarpi, C. *et al.* (2024) 'Systematic epigenome editing captures the context-dependent instructive function of chromatin modifications', *Nature Genetics*, pp. 1–13. Available at: https://doi.org/10.1038/s41588-024-01706-w.

Pott, S. and Lieb, J.D. (2015) 'What are super-enhancers?', *Nature Genetics*, 47(1), pp. 8–12. Available at: https://doi.org/10.1038/ng.3167.

Purcell, S. *et al.* (2007) 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses', *The American Journal of Human Genetics*, 81(3), pp. 559–575. Available at: https://doi.org/10.1086/519795.

Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. Available at: https://doi.org/10.1093/bioinformatics/btq033.

Rada-Iglesias, A. (2018) 'Is H3K4me1 at enhancers correlative or causative?', *Nature Genetics*, 50(1), pp. 4–5. Available at: https://doi.org/10.1038/s41588-017-0018-3.

Rao, M.S. *et al.* (2019) 'Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies', *Frontiers in Genetics*, 9, p. 636. Available at: https://doi.org/10.3389/fgene.2018.00636.

Rao, Y.L. *et al.* (2022) 'Hippocampus and its involvement in Alzheimer's disease: a review', *3 Biotech*, 12(2), p. 55. Available at: https://doi.org/10.1007/s13205-022-03123-4.

Rau, A., Marot, G. and Jaffrézic, F. (2014) 'Differential meta-analysis of RNA-seq data from multiple studies', *BMC Bioinformatics*, 15(1), p. 91. Available at: https://doi.org/10.1186/1471-2105-15-91.

Rauluseviciute, I. *et al.* (2024) 'JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles', *Nucleic Acids Research*, 52(D1), pp. D174–D182. Available at: https://doi.org/10.1093/nar/gkad1059.

Reményi, A., Schöler, H.R. and Wilmanns, M. (2004) 'Combinatorial control of gene expression', *Nature Structural & Molecular Biology*, 11(9), pp. 812–815. Available at: https://doi.org/10.1038/nsmb820.

Reshef, Y.A. *et al.* (2018) 'Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk', *Nature Genetics*, 50(10), pp. 1483–1493. Available at: https://doi.org/10.1038/s41588-018-0196-7.

Ristori, G. *et al.* (2006) 'Multiple sclerosis in twins from continental Italy and Sardinia: a nationwide study', *Annals of Neurology*, 59(1), pp. 27–34. Available at: https://doi.org/10.1002/ana.20683.

Sagai, T. *et al.* (2005) 'Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb', *Development (Cambridge, England)*, 132(4), pp. 797–803. Available at: https://doi.org/10.1242/dev.01613.

Sanz, L.A. *et al.* (2008) 'A mono-allelic bivalent chromatin domain controls tissue-specific imprinting at Grb10', *The EMBO Journal*, 27(19), pp. 2523–2532. Available at: https://doi.org/10.1038/emboj.2008.142.

Sargurupremraj, M. *et al.* (2020) 'Cerebral small vessel disease genomics and its implications across the lifespan', *Nature Communications*, 11(1), p. 6285. Available at: https://doi.org/10.1038/s41467-020-19111-2.

Sarkans, U. *et al.* (2018) 'The BioStudies database—one stop shop for all data supporting a life sciences study', *Nucleic Acids Research*, 46(D1), pp. D1266–D1270. Available at: https://doi.org/10.1093/nar/gkx965.

Sasse, A., Ng, B., Spiro, A.E., *et al.* (2023) 'Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings', *Nature Genetics*, 55(12), pp. 2060–2064. Available at: https://doi.org/10.1038/s41588-023-01524-6.

Sasse, A., Ng, B., Spiro, A., *et al.* (2023) 'How far are we from personalized gene expression prediction using sequence-to-expression deep neural networks?' bioRxiv, p. 2023.03.16.532969. Available at: https://doi.org/10.1101/2023.03.16.532969.

Sasse, A., Chikina, M. and Mostafavi, S. (2024) 'Unlocking gene regulation with sequence-to-function models', *Nature Methods*, 21(8), pp. 1374–1377. Available at: https://doi.org/10.1038/s41592-024-02331-5.

Savage, J.E. *et al.* (2018) 'Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence', *Nature Genetics*, 50(7), pp. 912–919. Available at: https://doi.org/10.1038/s41588-018-0152-6.

Schaid, D.J., Chen, W. and Larson, N.B. (2018) 'From genome-wide associations to candidate causal variants by statistical fine-mapping', *Nature reviews. Genetics*, 19(8), pp. 491–504. Available at: https://doi.org/10.1038/s41576-018-0016-z.

Schena, M. *et al.* (1995) 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray', *Science (New York, N.Y.)*, 270(5235), pp. 467–470. Available at: https://doi.org/10.1126/science.270.5235.467.

Schneider, R. *et al.* (2004) 'Histone H3 lysine 4 methylation patterns in higher eukaryotic genes', *Nature Cell Biology*, 6(1), pp. 73–77. Available at: https://doi.org/10.1038/ncb1076.

Schreiber, J., Singh, R., *et al.* (2020) 'A pitfall for machine learning methods aiming to predict across cell types', *Genome Biology*, 21(1), p. 282. Available at: https://doi.org/10.1186/s13059-020-02177-y.

Schreiber, J., Durham, T., *et al.* (2020) 'Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome', *Genome Biology*, 21(1), p. 81. Available at: https://doi.org/10.1186/s13059-020-01977-6.

Schreiber, J. *et al.* (2023) 'The ENCODE Imputation Challenge: a critical assessment of methods for cross-cell type imputation of epigenomic profiles', *Genome Biology*, 24(1), p. 79. Available at: https://doi.org/10.1186/s13059-023-02915-y.

Schreiber, J. (2024) 'jmschrei/tangermeme'. Available at: https://github.com/jmschrei/tangermeme (Accessed: 15 July 2024).

Schreiber, J., Bilmes, J. and Noble, W.S. (2020) 'Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples', *Genome Biology*, 21(1), p. 82. Available at: https://doi.org/10.1186/s13059-020-01978-5.

Schwanhäusser, B. *et al.* (2011) 'Global quantification of mammalian gene expression control', *Nature*, 473(7347), pp. 337–342. Available at: https://doi.org/10.1038/nature10098.

Schwartzentruber, J. *et al.* (2021) 'Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes', *Nature Genetics*, 53(3), pp. 392–402. Available at: https://doi.org/10.1038/s41588-020-00776-w.

Segerstolpe, Å. *et al.* (2016) 'Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes', *Cell Metabolism*, 24(4), pp. 593–607. Available at: https://doi.org/10.1016/j.cmet.2016.08.020.

Sekhon, A., Singh, R. and Qi, Y. (2018) 'DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications', *Bioinformatics (Oxford, England)*, 34(17), pp. i891–i900. Available at: https://doi.org/10.1093/bioinformatics/bty612.

Shi, S. *et al.* (2016) 'Benchmarking State-of-the-Art Deep Learning Software Tools', in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pp. 99–104. Available at: https://doi.org/10.1109/CCBD.2016.029.

Shrikumar, A. *et al.* (2020) 'Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5'. arXiv. Available at: https://doi.org/10.48550/arXiv.1811.00416.

Shrikumar, A., Greenside, P. and Kundaje, A. (2019) 'Learning Important Features Through Propagating Activation Differences'. arXiv. Available at: https://doi.org/10.48550/arXiv.1704.02685.

Singh, R. *et al.* (2016) 'DeepChrome: deep-learning for predicting gene expression from histone modifications', *Bioinformatics*, 32(17), pp. i639–i648. Available at: https://doi.org/10.1093/bioinformatics/btw427.

Singh, R. *et al.* (2017) 'Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin', *Advances in neural information processing systems*, 30, pp. 6785–6795.

Sinha, K.K. *et al.* (2023) 'Histone modifications regulate pioneer transcription factor cooperativity', *Nature*, 619(7969), pp. 378–384. Available at: https://doi.org/10.1038/s41586-023-06112-6.

Skene, N.G. and Grant, S.G.N. (2016) 'Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment', *Frontiers in Neuroscience*, 10, p. 16. Available at: https://doi.org/10.3389/fnins.2016.00016.

Slatkin, M. (2008) 'Linkage disequilibrium — understanding the evolutionary past and mapping the medical future', *Nature reviews. Genetics*, 9(6), pp. 477–485. Available at: https://doi.org/10.1038/nrg2361.

Slattery, M. *et al.* (2014) 'Absence of a simple code: how transcription factors read the genome', *Trends in Biochemical Sciences*, 39(9), pp. 381–399. Available at: https://doi.org/10.1016/j.tibs.2014.07.002.

Smith, A.M. *et al.* (2022) 'Diverse human astrocyte and microglial transcriptional responses to Alzheimer's pathology', *Acta Neuropathologica*, 143(1), pp. 75–91. Available at: https://doi.org/10.1007/s00401-021-02372-6.

Soldi, M. *et al.* (2017) 'Chromatin proteomics reveals novel combinatorial histone modification signatures that mark distinct subpopulations of macrophage enhancers', *Nucleic Acids Research*, 45(21), pp. 12195–12213. Available at: https://doi.org/10.1093/nar/gkx821.

Soneson, C. and Robinson, M.D. (2018) 'Bias, robustness and scalability in single-cell differential expression analysis', *Nature Methods*, 15(4), pp. 255–261. Available at: https://doi.org/10.1038/nmeth.4612.

Squair, J.W. *et al.* (2021) 'Confronting false discoveries in single-cell differential expression', *Nature Communications*, 12(1), p. 5692. Available at: https://doi.org/10.1038/s41467-021-25960-2.

Srinivasan, K. *et al.* (2016) 'Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses', *Nature Communications*, 7(1), p. 11295. Available at: https://doi.org/10.1038/ncomms11295.

Stankiewicz, P. and Lupski, J.R. (2010) 'Structural variation in the human genome and its role in disease', *Annual Review of Medicine*, 61, pp. 437–455. Available at: https://doi.org/10.1146/annurev-med-100708-204735.

Sudlow, C. *et al.* (2015) 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age', *PLOS Medicine*, 12(3), p. e1001779. Available at: https://doi.org/10.1371/journal.pmed.1001779.

Suomi, T. *et al.* (2017) 'ROTS: An R package for reproducibility-optimized statistical testing', *PLOS Computational Biology*, 13(5), p. e1005562. Available at: https://doi.org/10.1371/journal.pcbi.1005562.

Takahashi, H. *et al.* (2012) 'CAGE- Cap Analysis Gene Expression: a protocol for the detection of promoter and transcriptional networks', *Methods in molecular biology (Clifton, N.J.)*, 786, pp. 181–200. Available at: https://doi.org/10.1007/978-1-61779-292-2_11.

Tani, H. *et al.* (2012) 'Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals', *Genome Research*, 22(5), pp. 947–956. Available at: https://doi.org/10.1101/gr.130559.111.

Tanigawa, Y. *et al.* (2019) 'Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology', *Nature Communications*, 10(1), p. 4064. Available at: https://doi.org/10.1038/s41467-019-11953-9.

Tarasov, A. *et al.* (2015) 'Sambamba: fast processing of NGS alignment formats', *Bioinformatics*, 31(12), pp. 2032–2034. Available at: https://doi.org/10.1093/bioinformatics/btv098.

Tasic, B. *et al.* (2018) 'Shared and distinct transcriptomic cell types across neocortical areas', *Nature*, 563(7729), pp. 72–78. Available at: https://doi.org/10.1038/s41586-018-0654-5.

Taskiran, I.I. *et al.* (2024) 'Cell-type-directed design of synthetic enhancers', *Nature*, 626(7997), pp. 212–220. Available at: https://doi.org/10.1038/s41586-023-06936-2.

Telenius, H. *et al.* (1994) 'Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm', *Nature Genetics*, 6(4), pp. 409–414. Available at: https://doi.org/10.1038/ng0494-409.

Tick, B. *et al.* (2016) 'Heritability of autism spectrum disorders: a meta-analysis of twin studies', *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 57(5), pp. 585–595. Available at: https://doi.org/10.1111/jcpp.12499.

Tomlinson, B.E., Irving, D. and Blessed, G. (1981) 'Cell loss in the locus coeruleus in senile dementia of Alzheimer type', *Journal of the Neurological Sciences*, 49(3), pp. 419–428. Available at: https://doi.org/10.1016/0022-510x(81)90031-9.

Toneyan, S. and Koo, P.K. (2024) 'Interpreting Cis-Regulatory Interactions from Large-Scale Deep Neural Networks for Genomics'. bioRxiv, p. 2023.07.03.547592. Available at: https://doi.org/10.1101/2023.07.03.547592.

Toneyan, S., Tang, Z. and Koo, P.K. (2022) 'Evaluating deep learning for predicting epigenomic profiles', *Nature Machine Intelligence*, 4(12), pp. 1088–1100. Available at: https://doi.org/10.1038/s42256-022-00570-9.

Traag, V.A., Waltman, L. and van Eck, N.J. (2019) 'From Louvain to Leiden: guaranteeing well-connected communities', *Scientific Reports*, 9(1), p. 5233. Available at: https://doi.org/10.1038/s41598-019-41695-z.

Tran, H.T.N. *et al.* (2020) 'A benchmark of batch-effect correction methods for single-cell RNA sequencing data', *Genome Biology*, 21(1), p. 12. Available at: https://doi.org/10.1186/s13059-019-1850-9.

Trapnell, C. *et al.* (2014) 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nature Biotechnology*, 32(4), pp. 381–386. Available at: https://doi.org/10.1038/nbt.2859.

Trost, B., Loureiro, L.O. and Scherer, S.W. (2021) 'Discovery of genomic variation across a generation', *Human Molecular Genetics*, 30(R2), pp. R174–R186. Available at: https://doi.org/10.1093/hmg/ddab209.

Trubetskoy, V. *et al.* (2022) 'Mapping genomic loci implicates genes and synaptic biology in schizophrenia', *Nature*, 604(7906), pp. 502–508. Available at: https://doi.org/10.1038/s41586-022-04434-5.

Tsalenchuk, M., Gentleman, S.M. and Marzi, S.J. (2023) 'Linking environmental risk factors with epigenetic mechanisms in Parkinson's disease', *npj Parkinson's Disease*, 9(1), pp. 1–12. Available at: https://doi.org/10.1038/s41531-023-00568-z.

Turkmen, A. and Lin, S. (2017) 'Are rare variants really independent?', *Genetic Epidemiology*, 41(4), pp. 363–371. Available at: https://doi.org/10.1002/gepi.22039.

Vecsey, C.G. *et al.* (2007) 'Histone Deacetylase Inhibitors Enhance Memory and Synaptic Plasticity via CREB: CBP-Dependent Transcriptional Activation', *The Journal of Neuroscience*, 27(23), pp. 6128–6140. Available at: https://doi.org/10.1523/JNEUROSCI.0296-07.2007.

Verheijen, B.M., Vermulst, M. and van Leeuwen, F.W. (2018) 'Somatic mutations in neurons during aging and neurodegeneration', *Acta Neuropathologica*, 135(6), pp. 811–826. Available at: https://doi.org/10.1007/s00401-018-1850-y.

Visel, A., Rubin, E.M. and Pennacchio, L.A. (2009) 'Genomic views of distant-acting enhancers', *Nature*, 461(7261), pp. 199–205. Available at: https://doi.org/10.1038/nature08451.

Võsa, U. *et al.* (2021) 'Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression', *Nature Genetics*, 53(9), pp. 1300–1310. Available at: https://doi.org/10.1038/s41588-021-00913-z.

Vuckovic, D. *et al.* (2020) 'The Polygenic and Monogenic Basis of Blood Traits and Diseases', *Cell*, 182(5), pp. 1214-1231.e11. Available at: https://doi.org/10.1016/j.cell.2020.08.008.

Waddington, C.H. (1942) 'The epigenotype', *Endeavour*, 1, pp. 18–20.

Waddington, C.H. (1957) *The strategy of the genes : a discussion of some aspects of theoretical biology*. Allen & Unwin. Available at: https://wellcomecollection.org/works/nzwm3z65 (Accessed: 29 April 2024).

Waddington, C.H. (1968) 'Towards a theoretical biology', *Nature*, 218(5141), pp. 525–527. Available at: https://doi.org/10.1038/218525a0.

Wakhloo, D. *et al.* (2020) 'Functional hypoxia drives neuroplasticity and neurogenesis via brain erythropoietin', *Nature Communications*, 11(1), p. 1313. Available at: https://doi.org/10.1038/s41467-020-15041-1.

Wang, C. *et al.* (2018) 'Reprogramming of H3K9me3-dependent heterochromatin during mammalian embryo development', *Nature Cell Biology*, 20(5), pp. 620–631. Available at: https://doi.org/10.1038/s41556-018-0093-4.

Wang, F. *et al.* (2021) 'ATACdb: a comprehensive human chromatin accessibility database', *Nucleic Acids Research*, 49(D1), pp. D55–D64. Available at: https://doi.org/10.1093/nar/gkaa943.

Wang, G. *et al.* (2020) 'A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(5), pp. 1273–1300. Available at: https://doi.org/10.1111/rssb.12388.

Wang, L. *et al.* (2024) 'Integrating single cell expression quantitative trait loci summary statistics to understand complex trait risk genes', *Nature Communications*, 15(1), p. 4260. Available at: https://doi.org/10.1038/s41467-024-48143-1.

Wang, Q.S. *et al.* (2021) 'Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs', *Nature Communications*, 12(1), p. 3394. Available at: https://doi.org/10.1038/s41467-021-23134-8.

Wang, Z. *et al.* (2022) 'Prediction of histone post-translational modification patterns based on nascent transcription data', *Nature Genetics*, 54(3), pp. 295–305. Available at: https://doi.org/10.1038/s41588-022-01026-x.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature reviews. Genetics*, 10(1), pp. 57–63. Available at: https://doi.org/10.1038/nrg2484.

Watanabe, K. *et al.* (2019) 'A global overview of pleiotropy and genetic architecture in complex traits', *Nature Genetics*, 51(9), pp. 1339–1348. Available at: https://doi.org/10.1038/s41588-019-0481-0.

Welch, J.D. *et al.* (2019) 'Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity', *Cell*, 177(7), pp. 1873-1887.e17. Available at: https://doi.org/10.1016/j.cell.2019.05.006.

Whalen, S. *et al.* (2022) 'Navigating the pitfalls of applying machine learning in genomics', *Nature Reviews Genetics*, 23(3), pp. 169–181. Available at: https://doi.org/10.1038/s41576-021-00434-9.

Whitehouse, P.J. *et al.* (1982) 'Alzheimer's disease and senile dementia: loss of neurons in the basal forebrain', *Science (New York, N.Y.)*, 215(4537), pp. 1237–1239. Available at: https://doi.org/10.1126/science.7058341.

Wightman, D.P. *et al.* (2021) 'A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease', *Nature Genetics*, 53(9), pp. 1276–1282. Available at: https://doi.org/10.1038/s41588-021-00921-z.

Wild, C.P. (2005) 'Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology', *Cancer Epidemiology, Biomarkers & Prevention*, 14(8), pp. 1847–1850. Available at: https://doi.org/10.1158/1055-9965.EPI-05-0456.

Wray, N.R. *et al.* (2018) 'Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression', *Nature Genetics*, 50(5), pp. 668–681. Available at: https://doi.org/10.1038/s41588-018-0090-3.

Wu Ct, null and Morris, J.R. (2001) 'Genes, genetics, and epigenetics: a correspondence', *Science (New York, N.Y.)*, 293(5532), pp. 1103–1105. Available at: https://doi.org/10.1126/science.293.5532.1103.

Wu, K. *et al.* (2023) 'Dynamics of histone acetylation during human early embryogenesis', *Cell Discovery*, 9(1), pp. 1–24. Available at: https://doi.org/10.1038/s41421-022-00514-y.

Wu, Y. *et al.* (2021) 'GWAS of peptic ulcer disease implicates Helicobacter pylori infection, other gastrointestinal disorders and depression', *Nature Communications*, 12(1), p. 1146. Available at: https://doi.org/10.1038/s41467-021-21280-7.

Wu, Y. *et al.* (2024) 'SmartPlay: A Benchmark for LLMs as Intelligent Agents'. arXiv. Available at: https://doi.org/10.48550/arXiv.2310.01557.

Wysocka, J. *et al.* (2006) 'A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling', *Nature*, 442(7098), pp. 86–90. Available at: https://doi.org/10.1038/nature04815.

Xi, N.M. and Li, J.J. (2021) 'Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data', *Cell Systems*, 12(2), pp. 176-194.e6. Available at: https://doi.org/10.1016/j.cels.2020.11.008.

Yang, J. *et al.* (2011) 'GCTA: A Tool for Genome-wide Complex Trait Analysis', *American Journal of Human Genetics*, 88(1), pp. 76–82. Available at: https://doi.org/10.1016/j.ajhg.2010.11.011.

Yu, Z. *et al.* (2024) 'Genetic variation across and within individuals', *Nature Reviews Genetics*, pp. 1–15. Available at: https://doi.org/10.1038/s41576-024-00709-x.

Yue, L. *et al.* (2004) 'Divergent Effects of Peroxisome Proliferator-activated Receptor γ Agonists and Tumor Necrosis Factor α on Adipocyte ApoE Expression *', *Journal of Biological Chemistry*, 279(46), pp. 47626–47632. Available at: https://doi.org/10.1074/jbc.M408461200.

Zaret, K.S. (2020) 'Pioneer Transcription Factors Initiating Gene Network Changes', *Annual Review of Genetics*, 54, pp. 367–385. Available at: https://doi.org/10.1146/annurev-genet-030220-015007.

Zeitlinger, J. (2020) 'Seven myths of how transcription factors read the cis-regulatory code', *Current Opinion in Systems Biology*, 23, pp. 22–31. Available at: https://doi.org/10.1016/j.coisb.2020.08.002.

Zhang, F. *et al.* (2009) 'Copy Number Variation in Human Health, Disease, and Evolution', *Annual review of genomics and human genetics*, 10, pp. 451–481. Available at: https://doi.org/10.1146/annurev.genom.9.081307.164217.

Zhang, H.-J. *et al.* (2020) 'Destrin Contributes to Lung Adenocarcinoma Progression by Activating Wnt/β-Catenin Signaling Pathway', *Molecular Cancer Research*, 18(12), pp. 1789–1802. Available at: https://doi.org/10.1158/1541-7786.MCR-20-0187.

Zhang, J. *et al.* (2023) 'Genome-wide association study in Alzheimer's disease: a bibliometric and visualization analysis', *Frontiers in Aging Neuroscience*, 15, p. 1290657. Available at: https://doi.org/10.3389/fnagi.2023.1290657.

Zhang, P. *et al.* (2023) 'Deep flanking sequence engineering for efficient promoter design using DeepSEED', *Nature Communications*, 14(1), p. 6309. Available at: https://doi.org/10.1038/s41467-023-41899-y.

Zhang, X. and Song, W. (2013) 'The role of APP and BACE1 trafficking in APP processing and amyloid-β generation', *Alzheimer's Research & Therapy*, 5(5), p. 46. Available at: https://doi.org/10.1186/alzrt211.

Zhang, Y. *et al.* (2008) 'Model-based Analysis of ChIP-Seq (MACS)', *Genome Biology*, 9(9), p. R137. Available at: https://doi.org/10.1186/gb-2008-9-9-r137.

Zhao, S., Ye, Z. and Stanton, R. (2020) 'Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols', *RNA*, 26(8), pp. 903–909. Available at: https://doi.org/10.1261/rna.074922.120.

Zhao, Z. and Shilatifard, A. (2019) 'Epigenetic modifications of histones in cancer', *Genome Biology*, 20(1), p. 245. Available at: https://doi.org/10.1186/s13059-019-1870-5.

Zheng, Y. and VanDusen, N.J. (2023) 'Massively Parallel Reporter Assays for High-Throughput In Vivo Analysis of Cis-Regulatory Elements', *Journal of Cardiovascular Development and Disease*, 10(4), p. 144. Available at: https://doi.org/10.3390/jcdd10040144.

Zhou, J. and Troyanskaya, O.G. (2015) 'Predicting effects of noncoding variants with deep learning–based sequence model', *Nature Methods*, 12(10), pp. 931–934. Available at: https://doi.org/10.1038/nmeth.3547.

Zhou, Y. *et al.* (2020) 'Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease', *Nature Medicine*, 26(1), pp. 131–142. Available at: https://doi.org/10.1038/s41591-019-0695-9.

Zhu, Z. *et al.* (2016) 'Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets', *Nature Genetics*, 48(5), pp. 481–487. Available at: https://doi.org/10.1038/ng.3538.

Zimmerman, K. and Murphy, A. (2022) 'neurogenomics/hierarchicell: Reanalysis scRNA-seq benchmark'. Zenodo. Available at: https://doi.org/10.5281/zenodo.7356451.

Zimmerman, K.D., Espeland, M.A. and Langefeld, C.D. (2021) 'A practical solution to pseudoreplication bias in single-cell studies', *Nature Communications*, 12(1), p. 738. Available at: https://doi.org/10.1038/s41467-021-21038-1.

Zimmerman, K.D. and Langefeld, C.D. (2021) 'Hierarchicell: an R-package for estimating power for tests of differential expression with single-cell data', *BMC Genomics*, 22(1), p. 319. Available at: https://doi.org/10.1186/s12864-021-07635-w.

Appendices

Appendix A

The information in Appendix A relates to **Chapter 5**: **Predicting cell type-specific epigenomic profiles accounting for distal genetic effects** on page 114. The work in this appendix chapter is adapted from that set out in our previous publication (Alan E. Murphy *et al.*, 2024).

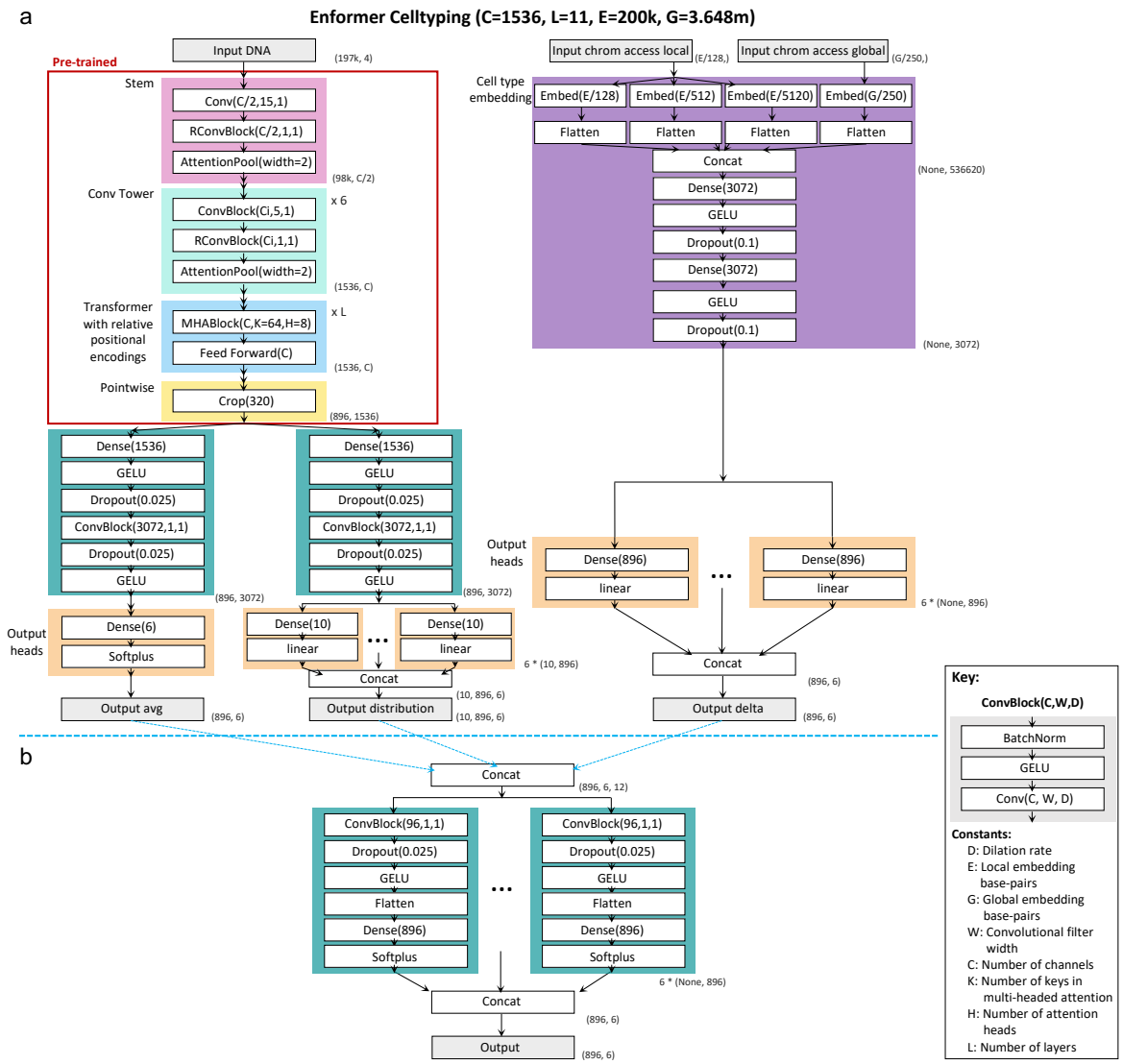


Fig. A-1 Enformer Celltyping architecture.

(a) pre-training, warm-up step: Enformer Celltyping's architecture is split into two separate submodules – The DNA half (stemming from 'Input DNA') which inputs DNA sequence and uses a pre-trained, chopped, frozen version of Enformer through a transfer learning approach to predict an average histone mark signal ('Output avg') and distribution ('Output distribution') across training cell types for the genomic region; and the celltyping half (stemming from 'Input chrom access local' and 'Input chrom access global') which embeds chromatin accessibility data from the same region and predicts the difference between the average histone mark signal and the cell type-specific signal ('Output delta'). (b) Second training step: Enformer Celltyping

combines the three outputs (‘Output avg’,‘Output distribution’ and ‘Output delta’) with subsequent layers to predict the cell type-specific histone mark signal for a given genomic region.

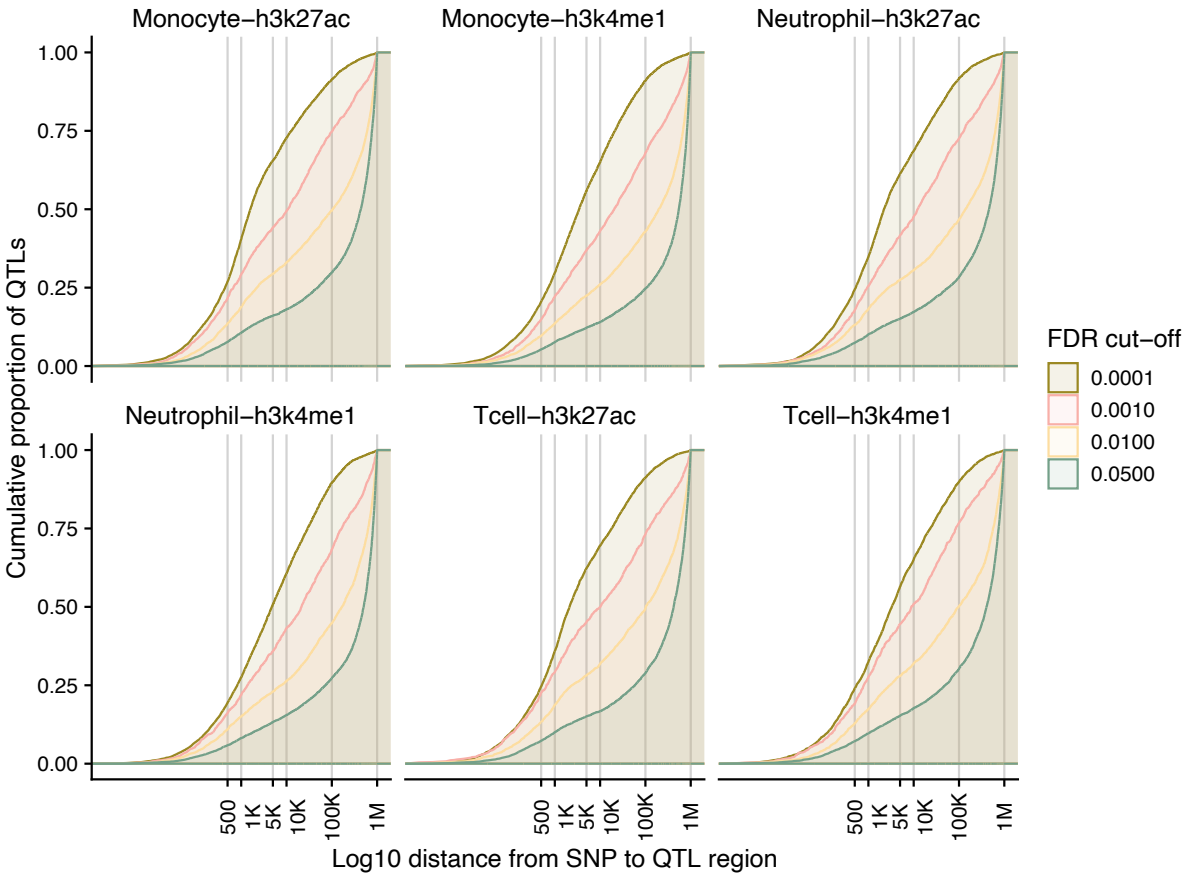


Fig. A-2 Histone mark QTL data shows that SNPs regulate peaks that are significant distances away.

The distance from the SNP to the associated histone mark peak for the six histone mark quantative trait loci (hQTL) studies from BLUEPRINT Phase 2 are shown at increasing FDR cut-offs. Here, to avoid double counting due to linkage disequilibrium (LD), SNPs were aggregated to those with the lowest p-value for each LD block based on Pickrell LD blocks. The y-axis is the cumulative proportion of interactions and the x-axis is the $\log_{10}$ distance with a grey vertical lines increasing numbers of base-pairs. At all FDR cut-offs, for all hQTL studies, we observe high proportions of interactions (between 20%-75%) at distances above 10,000 base-pairs. This goes against a null hypothesis that SNPs only regulate the epigenome by local motif disruption and highlights the need for models which can account for these distal interactions when predicting epigenetic profiles.

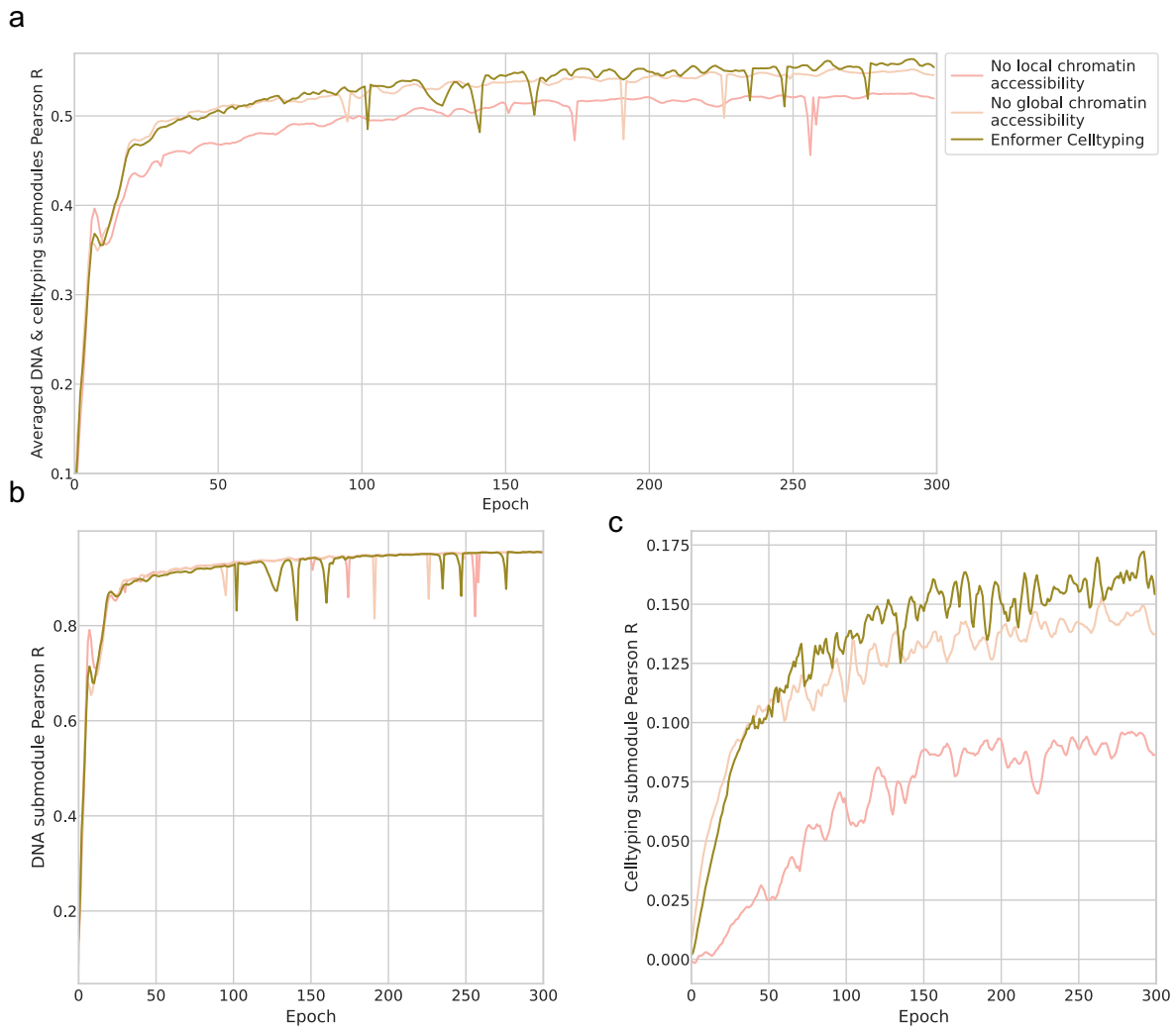


Fig. A-3 Using global and local chromatin accessibility information improves the prediction of cell type-specific histone mark signals.

We measured the performance of Enformer Celltyping on the validation set for the pre-training stage over 300 training steps (batch size 128) for both global and local chromatin accessibility information (green), no local chromatin accessibility (pink) and no global chromatin accessibility (beige). (a) Averaged Pearson R performance for both the DNA and celltyping submodules, (b) DNA submodule – note this is a negative control, these should be the same since the chromatin accessibility only effects the celltyping submodule and (c) the celltyping submodule – where this should have the largest effect. All Pearson R values are performance on the validation dataset.

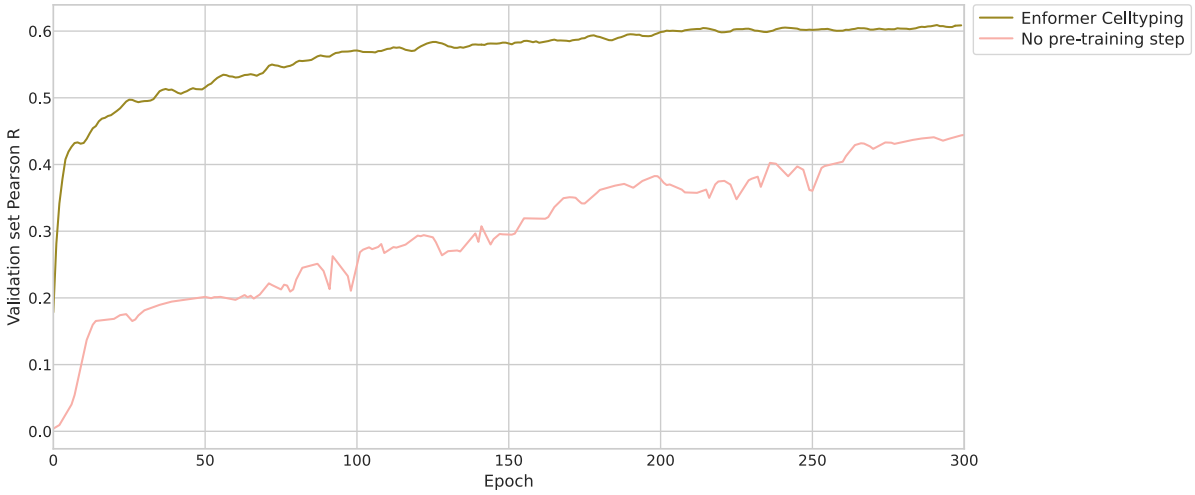


Fig. A-4 Pre-training drastically improves Enformer Celltyping's performance.

We measured the performance of Enformer Celltyping on the validation set for the full training stage over 300 training steps (batch size 128) comparing the full Enformer Celltyping model (green) against a version of Enformer Celltyping for which a pre-training stage was not performed (pink). Pearson R values are performance on the validation dataset.

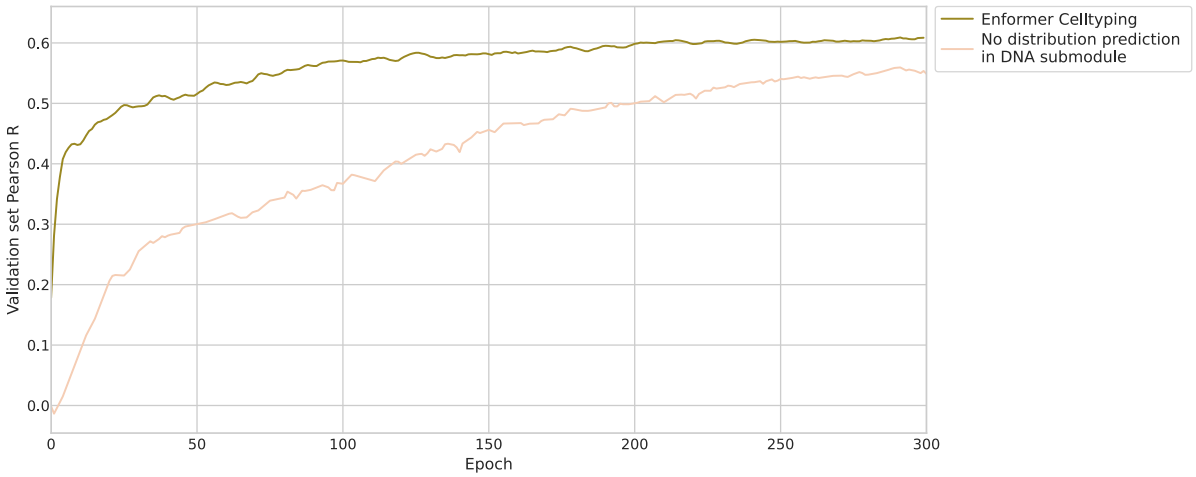


Fig. A-5 Predicting the distribution of signals as well as the average with the DNA module during pre-training improves Enformer Celltyping's performance.

We measured the performance of Enformer Celltyping on the validation set for the full training stage over 300 training steps (batch size 128) comparing the full Enformer Celltyping model (green) against a version of Enformer Celltyping for which the DNA module only predicted the average expression, as opposed to the average and distribution, in the pre-training stage (beige). Pearson R values are performance on the validation dataset.

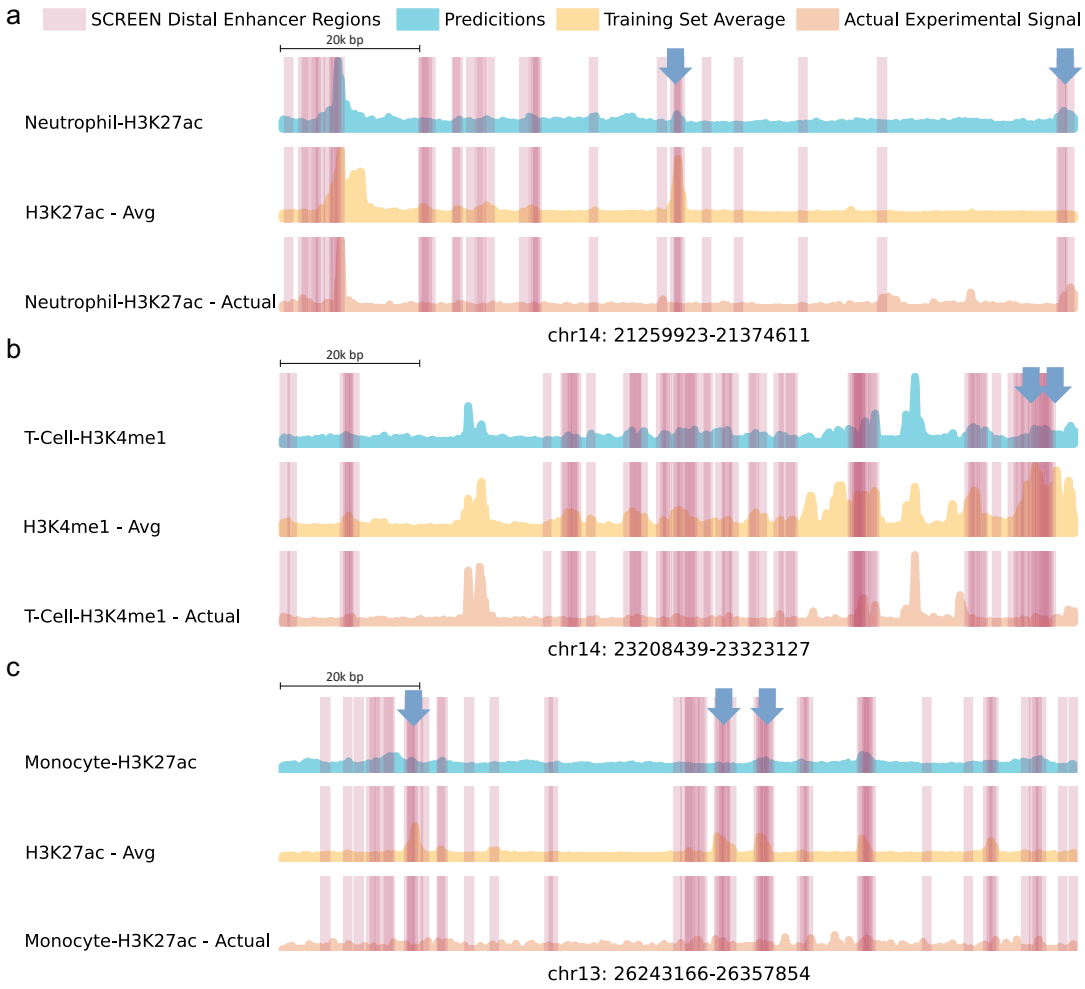


Fig. A-6 Enformer Celltyping predicts cell type-specific histone marks signals that match the experimental track.

Enformer Celltyping's H3K27ac (a,c) and H3K4me1 (b) predictions (blue), the training set average (gold) and the experimental assay signals (apricot) at differing genomic loci with SCREEN distal enhancer regions highlighted for the three blood immune cells. Navy arrows indicate true peaks in SCREEN distal enhancer regions either missed by the average but captured by Enformer Celltyping or not in the experiment or Enformer Celltyping but captured

by the average. We can see Enformer Celltyping does not simply predict the average and outperforms it, managing to capture cell type-specific signals for previously unseen cell types.

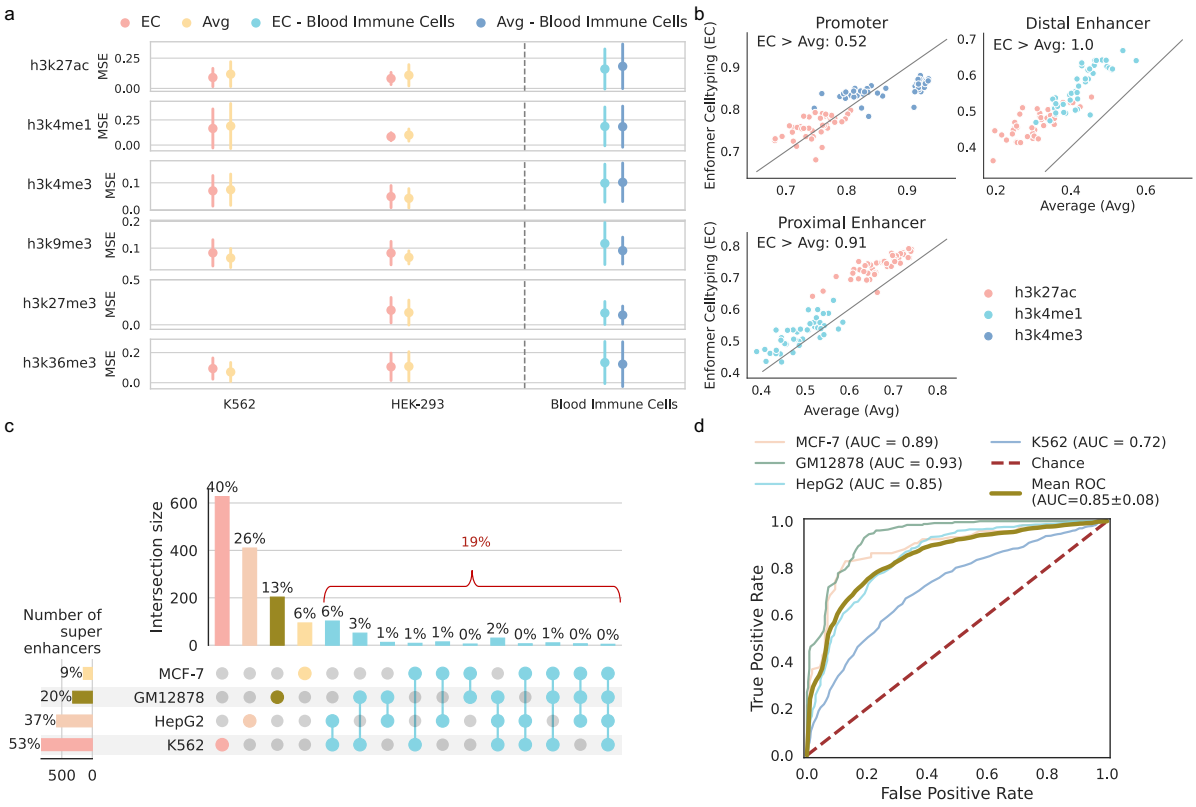


Fig. A-7 Enformer Celltyping shows superior performance on cancer cell lines and its embeddings can discriminate highly active enhancers (super-enhancers).

(a) Enformer Celltyping's (EC) mean squared error (MSE) across two held-out ENCODE cancer cell lines. The average model (Avg) performance and the average performance across the three ENCODE immune cell types (Blood Immune Cells) is shown for comparison. The results for the "Blood immune cells" is the union of prediction results for T-cells, Neutrophils and Monocytes. "Average" here refers to the MSE calculated using the average values for the histone mark, across the training cell types. Performance for H3K27me3 in K562 is not shown as it has not been assayed. Performance is based on genome-wide predictions (over 16,000 per cell type) with error bars for the standard deviation. (b) EC (y-axis) versus Avg (x-axis) Pearson R correlation performance with the experimental signal in functional genomic regions from Search Candidate cis-Regulatory Elements by ENCODE (SCREEN). Performance was measured for histone marks with known functional roles in these regions for the two cancer cell lines (K562 and HEK-293). Each point represents the Pearson R performance for each

against the true signal for each cell type-chromosome combination. The EC > Avg represents for how many cell type-chromosome values our model outperformed the average. (c) An upset plot showing the overlap of highly active enhancers, i.e. super-enhancers (SEs), across four ENCODE cancer cell lines, highlighting the cell type-specificity of the SEs, with only 19% having any overlap across the cells. (d) Performance on held-out cells for random forest classifiers trained on features derived from Enformer Celltyping, to discriminate SEs from all enhancers.

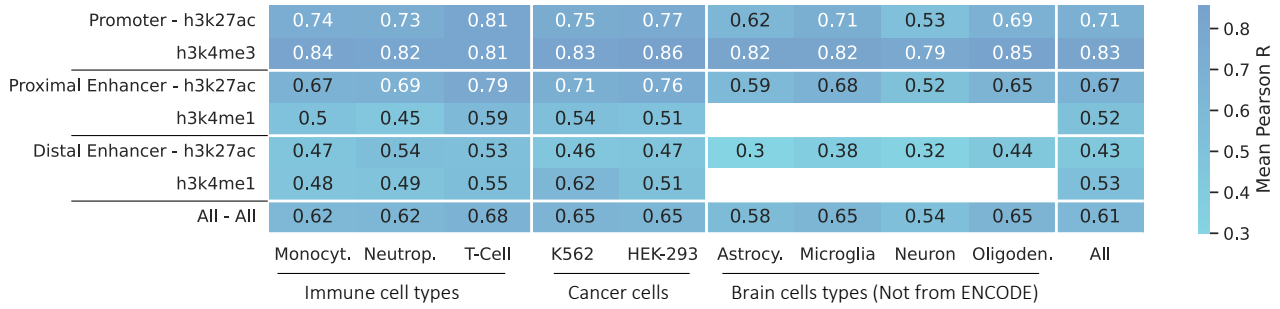


Fig. A-8 Enformer Celltyping's performance remains consistent across diverse cell types. Enformer Celltyping's Pearson R correlation performance with the experimental signal in functional genomic regions from Search Candidate cis-Regulatory Elements by ENCODE (SCREEN) (rows) across the 9 diverse cell types tested (columns). This performance is shown for the functionally relevant histone marks and the cell types are split into broad tissue groups. Monocyt. is Monocyte, Neutrop. is Neutrophil, Astrocy. is Astrocyte and Oligoden. is Oligodendrocyte. The weighted average of the three SCREEN regions for the histone marks for each cell type (last row) and for each SCREEN and histone mark combination (last column) is shown. We can see that on average, Astrocytes had the worst performance which is consistent with the results in **Figure 5-5b** and is likely attributable to the low sequencing depth of the associated ATAC-Seq data. Importantly however, average performance for cell types was approximately a Pearson R of 0.65, even on cells from outside of ENCODE (Microglia and Oligodendrocytes), highlighting Enformer Celltyping's consistent performance across cell types. Overall, performance dropped from proximal to distal regions for H3K27ac (0.67 to 0.43) but not H3K4me1 (0.52 to 0.53).

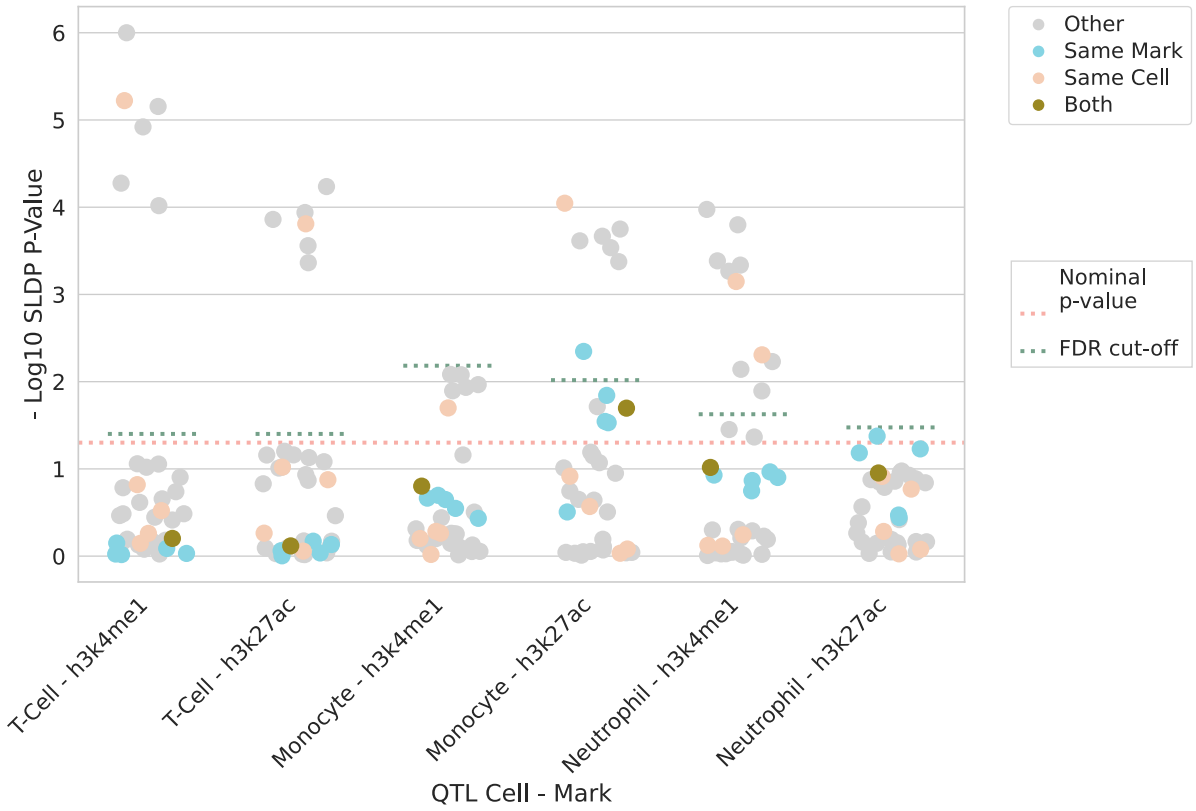


Fig. A-9 Enformer Celltyping's genetic effect predictions are not noticeably improved by filtering out QTLs associated with downregulating peaks which are missed in predictions for the major allele.

Statistical significance (y-axis) of SLDP genome-wide concordance between Enformer Celltyping's genetic variant predictions and measured hQTL effect sizes. Here, any SNP to histone mark binding hQTL entries were removed if it caused a decrease in binding and Enformer Celltyping did not predict the binding on the major allele (on the DNA sequence without the genetic variant). The x-axis depicts the six Blueprint phase 2 hQTL studies and the cell-histone mark predictions are coloured by their relationship to the hQTL study: 'Both' – both the histone mark and the cell predicted in were the same as the hQTL study, 'Same Mark' – the mark matched, 'Same Cell' – the cell matched and 'Other' – neither the cell nor mark matched. If there was perfect agreement between our model and the hQTL studies, all 'Both' entries would be significant. A nominal p-value (pink dashed line) and the cut-off after

Benjamini and Hochberg correction based on the minimum, non-significant p-value, corrected at the level of each hQTL (Turquoise dashed line) are both added.

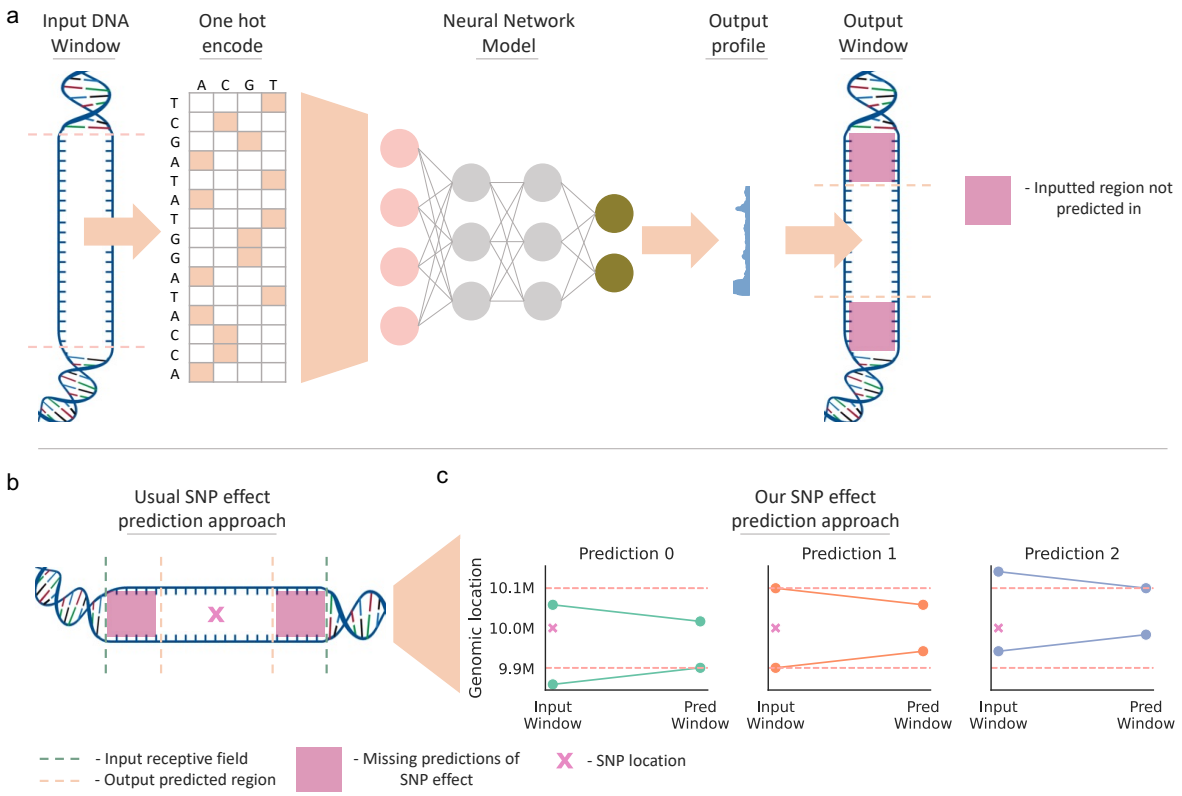


Fig. A-10 Schematic showing our strategy to capture distal effect of SNPs, by predicting each SNP multiple times, sliding the input window around the SNP location.

Genomic deep learning models tend to predict profiles for only the centre proportion of the genomic location relating to the DNA input to avoid predicting on edge regions which have less neighbouring genomic information. This funnel effect is shown in (a). Thereafter, common practice when predicting the effect of SNPs (i.e. *in silico* mutagenesis) with these models, is to centre the SNP on the input window (b). However, this approach misses distal predictions of SNP effects, not using the full receptive field of the model. We propose an alternative strategy to capture the effect of the SNP across the full receptive field by predicting each SNP multiple times, sliding the input window around the SNP location, e.g. move the relative SNP location towards the edges, or closer to the centre (c). For Enformer Celltyping, this requires three predictions (c): A prediction made with the SNP in the centre of the input window ('Prediction 1') to capture the nearby effect of the SNP and sliding the input window to make two further predictions ('Prediction 0' and 'Prediction 2') to capture distal SNP predictions.

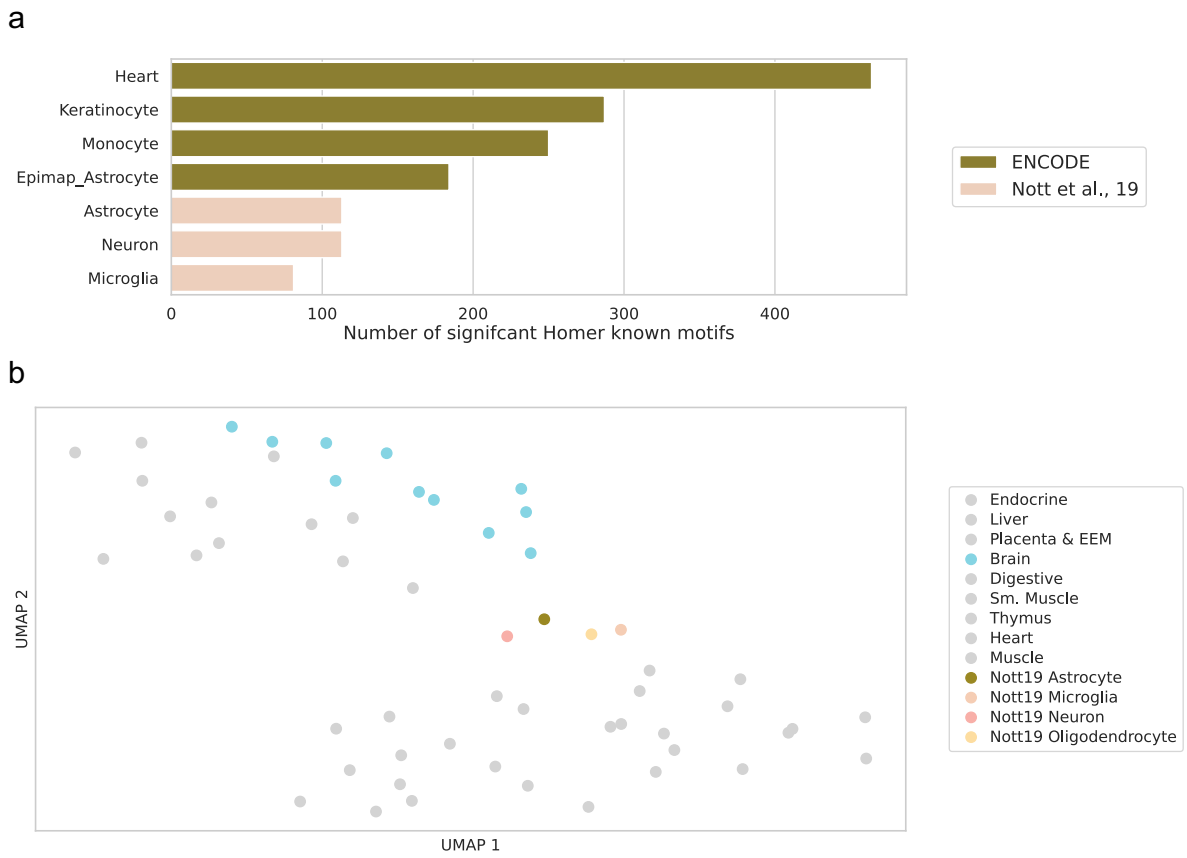


Fig. A-11 Enformer Celltyping's embedded global chromatin accessibility signal appears biased towards ENCODE samples.

(a) Gives the number of significant known motifs identified with Homer for each of the six cell types tested. The genomic regions for the motif analysis were derived from the histone mark peaks that had the highest contribution from the global chromatin accessibility signal for each cell type, identified by dropping the global signal inputted to the model and measuring the change in predicted binding. The cell types are coloured by their source. (b) A UMAP projection of Enformer Celltyping's global chromatin accessibility embeddings. Both training cell types and the four isolated cortical brain cell types from Nott et al. are included (the later are coloured). The other training cell types related to the brain (all bulk tissue, not isolated cells) are also coloured.

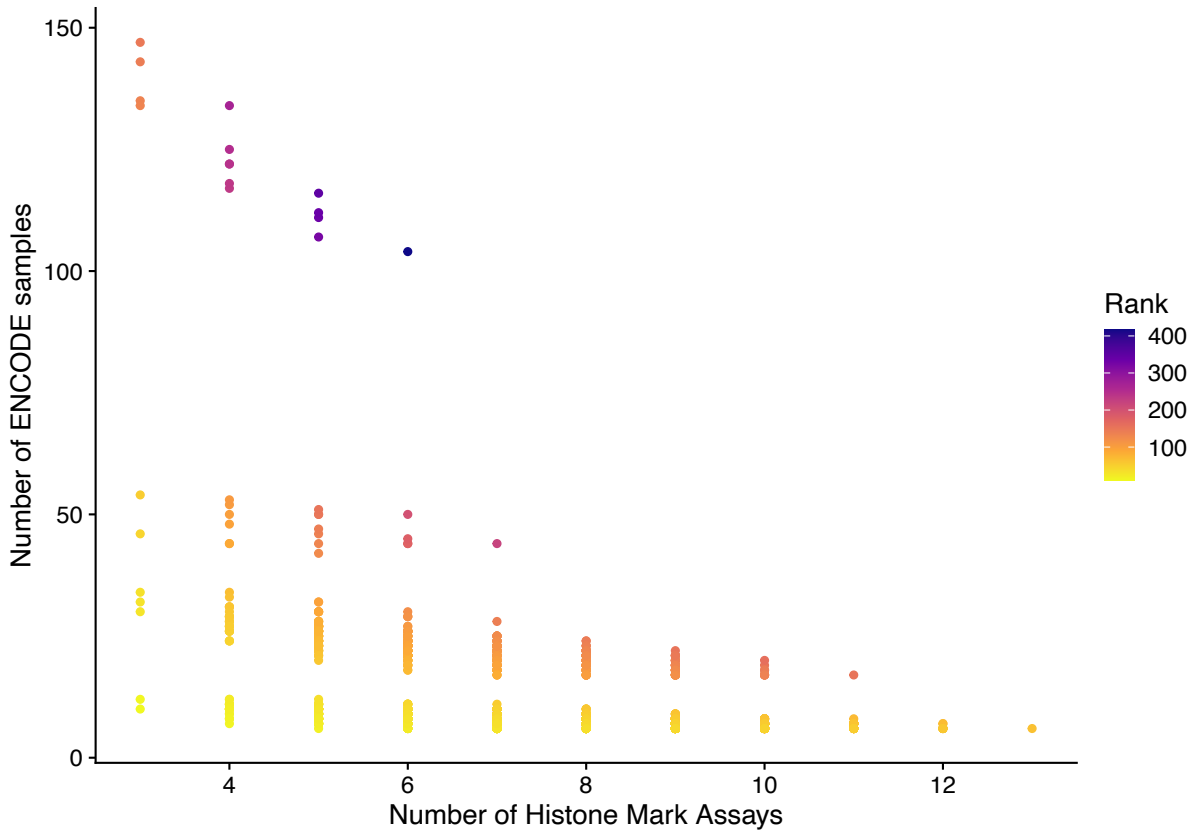


Fig. A-12 EpiMap trade-off between the number of samples and the number of assays.

The histone mark data available from EpiMap is shown with the number of samples (cell/tissue type) along the y-axis at each of the different number of histone mark assays along the x-axis. When choosing the number of assays to include in the model's predictions, the aim was to maximise both the number of assays but also the number of samples. The colour, which is the product of these numbers, of each position gives a rank which takes both into account. Note that the number of samples exclude our 3 validation cells (CD14+ monocytes, CD16+ neutrophils and naive CD4+T cells) and any related cell types.

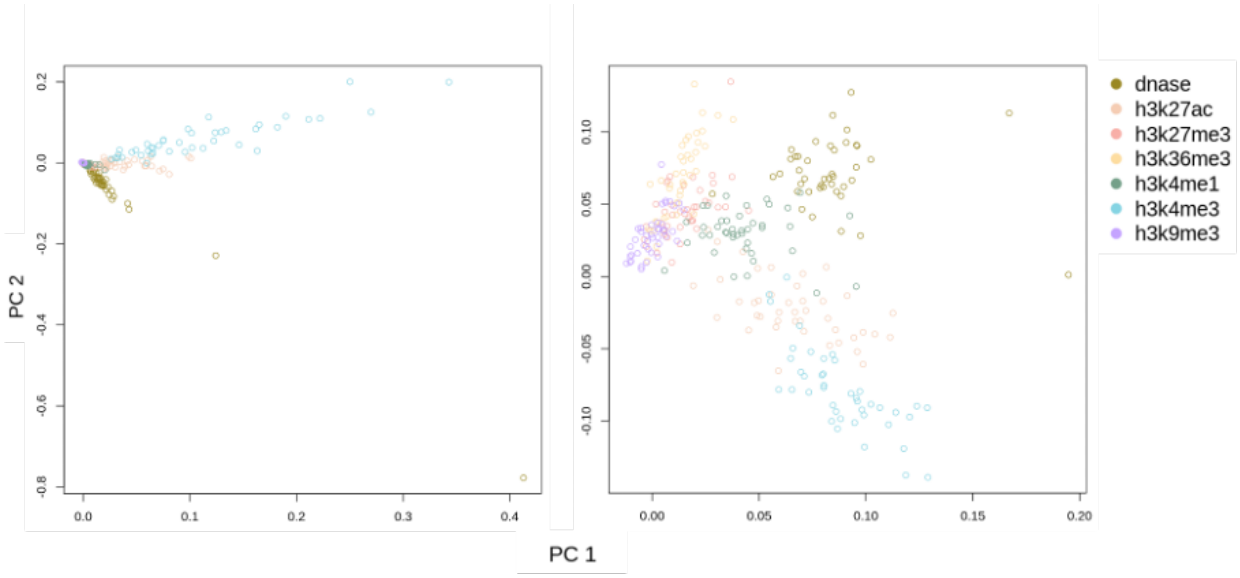


Fig. A-13 Arcsinh-transformed epigenetic signals for same marks show groupings.

Principal component analysis (PCA) plot of the first two principal components without (left) and with (right) arcsinh-transform based on the EpiMap training samples. Note that the data is not scaled here (a usual step in PCA) to replicate the data a model would be trained on. The arcsinh-transformed data shows clear separation between assay types, something which is not present with the raw data, highlighting arcsinh-transform's benefit. Note that DNase-Seq assayed were included here to show chromatin accessibility differentiation although they were not used in the model training approach.

Group	Secondary	Type	BSSID	Extended Info	MNEMONIC
Cancer	Lung	cell line	BSS00013	LUNG EPITHELIAL CARCINOMA	CA.LUEPITH.A549
Cancer	Lung	cell line	BSS00015	LUNG EPITHELIAL CARCINOMA	CA.LUEPITH.A549
Cancer	Muscle	cell line	BSS00035	MUSCLE EWING SARCOMA	CA.EWINGSC.A673
Endocrine	NA	tissue	BSS00060	ADRENAL GLAND	EDCR.EMB.ADRENAL.GLD
Brain	NA	tissue	BSS00078	ANGULAR GYRUS	BRN.AG
Heart	NA	tissue	BSS00088	ASCENDING AORTA	HRT.ASC.AORTA
Brain	NA	primary cell	BSS00089	ASTROCYTE	BRN.AST
Other	NA	tissue	BSS00146	BREAST EPITHELIUM	OTHR.BREAST.EPITH
ES-deriv	Heart	in vitro differentiated cells	BSS00171	CARDIAC MUSCLE DERIV	ESDV.CARD.MUSCLE
Brain	NA	tissue	BSS00173	CAUDATE NUCLEUS	BRN.CAUD.NUC
Brain	NA	tissue	BSS00175	CAUDATE NUCLEUS	BRN.CAUD.NUC
Placenta & EEM	NA	tissue	BSS00214	CHORIONIC VILLUS	PCTA.CHORIONIC.VILL
Placenta & EEM	NA	tissue	BSS00215	CHORIONIC VILLUS	PCTA.CHORIONIC.VILL
Placenta & EEM	NA	tissue	BSS00216	CHORIONIC VILLUS	PCTA.CHORIONIC.VILL
Placenta & EEM	NA	tissue	BSS00217	CHORIONIC VILLUS	PCTA.CHORIONIC.VILL
Brain	NA	tissue	BSS00220	CINGULATE GYRUS	BRN.CING.GYR
Digestive	NA	tissue	BSS00227	COLON MUCOSA	GI.COLON.MUC
Digestive	NA	tissue	BSS00228	COLON MUCOSA	GI.COLON.MUC
HSC & B-cell	NA	primary cell	BSS00231	CD34 CMP	HSC.CD34.CMP
HSC & B-cell	NA	primary cell	BSS00232	CD34 CMP	HSC.CD34.CMP
Cancer	Blood & T-cell	cell line	BSS00267	ACUTE LYMPHOBLASTIC LEUKEMIA	CA.ALL.DND41
Digestive	NA	tissue	BSS00270	DUODENUM MUCOSA	GI.DUOD.MUC
ES-deriv	NA	in vitro differentiated cells	BSS00273	ECTODERMAL DERIV	ESDV.ECTODERM
ES-deriv	NA	in vitro differentiated cells	BSS00287	ENDODERMAL DERIV	ESDV.ENDODERM
Endothelial	Placenta & EEM	primary cell	BSS00296	UMBILICAL VEIN ENDOTHELIAL CELL	EDTH.UBM.VEIN.ENDOTH
Digestive	NA	tissue	BSS00316	ESOPHAGUS	GI.ESOPH
Digestive	NA	tissue	BSS00318	ESOPHAGUS	GI.ESOPH
Stromal	Lung	primary cell	BSS00341	LUNG FIBROBLAST	STRM.LUNG.FIBR
Stromal	NA	primary cell	BSS00353	FORESKIN FIBROBLAST	STRM.FRSKN.FIBR
Epithelial	NA	primary cell	BSS00354	FORESKIN KERATINOCYTE	EPTH.FORESKIN.KRTN
Other	NA	primary cell	BSS00368	FORESKIN MELANOCYTE	OTHR.FRSKN.MELANOCYTE
Muscle	NA	tissue	BSS00379	GASTROCNEMIUS MEDIALIS	MUSC.GASTR.MED
Lymphoblastoid	NA	cell line	BSS00439	LYMPHOBLASTOID CELL LINE	LCL.LCL.GM12878
Stromal	NA	cell line	BSS00476	SKIN FIBROBLAST	STRM.SKIN.FIBR
ESC	NA	cell line	BSS00478	ESC	ESCH1HESC
ESC	NA	cell line	BSS00484	ESC	ESCH9
Cancer	Digestive	cell line	BSS00492	COLORECTAL ADENOCARCINOMA	CA.CRC.HCT116
Heart	NA	tissue	BSS00507	HEART LEFT VENTRICLE	HRT.LEFT.VENTR

Heart	NA	tissue	BSS00512	HEART LEFT VENTRICLE	HRT.LEFT.VENTR
Heart	NA	tissue	BSS00513	HEART LEFT VENTRICLE	HRT.LEFT.VENTR
Heart	NA	tissue	BSS00525	HEART RIGHT VENTRICLE	HRT.RIGHT.VENTR
Cancer	Reproductive	cell line	BSS00529	CERVIX ADENOCARCINOMA	CA.CVX.HELAS3
Cancer	Liver	cell line	BSS00558	HEPATOCELLULAR CARCINOMA	CA.HCC.HEPG2
ESC	NA	cell line	BSS00715	ESC	ESCHUES48
ESC	NA	cell line	BSS00716	ESC	ESCHUES6
ESC	NA	cell line	BSS00717	ESC	ESCHUES64
Stromal	Lung	cell line	BSS00720	LUNG FIBROBLAST	STRM.EMB.LUNG.FIBR
iPSC	NA	cell line	BSS00731	iPSC	IPSC.DF19.11
iPSC	NA	cell line	BSS00737	iPSC	IPSC.18a
iPSC	NA	cell line	BSS00739	iPSC	IPSC.20b
Cancer	HSC & B-cell	cell line	BSS01065	B CELL LYMPHOMA	CA.BCL.KARPAS422
Epithelial	NA	primary cell	BSS01068	KERATINOCYTE	EPTH.KRTN
Digestive	NA	tissue	BSS01119	LARGE INTESTINE	GI.EMB.LARGE.INT
Brain	NA	tissue	BSS01124	HIPPOCAMPUS	BRN.HPC
Brain	NA	tissue	BSS01125	HIPPOCAMPUS	BRN.HPC
Brain	NA	tissue	BSS01126	HIPPOCAMPUS	BRN.HPC
Liver	NA	tissue	BSS01159	LIVER	LVR.LIVER
Liver	NA	tissue	BSS01168	LIVER	LVR.LIVER
Lung	NA	tissue	BSS01190	LUNG	LNG.LUNG
Epithelial	NA	primary cell	BSS01213	MAMMARY EPITHELIAL CELL	EPTH.MMRY.EPITH
Cancer	Other	cell line	BSS01226	MAMMARY GLAND ADENOCARCINOMA	CA.BRCA.MCF7
Brain	NA	tissue	BSS01272	MIDDLE FRONTAL AREA	BRN.MID.FTL
Cancer	HSC & B-cell	cell line	BSS01274	MYELOMA	CA.MYELOMA.MM1S
Digestive	NA	tissue	BSS01282	RECTUM MUCOSA	GI.RECT.MUC
Digestive	NA	tissue	BSS01283	RECTUM MUCOSA	GI.RECT.MUC
Sm. Muscle	Digestive	tissue	BSS01287	DUODENUM MUSCLE	SMTH.DUOD.MUSCLE
Muscle	NA	tissue	BSS01319	LEG MUSCLE	MUSC.EMB.LEG.MUSCLE
Muscle	NA	tissue	BSS01332	TRUNK MUSCLE	MUSC.EMB.TRUNK.MUSCLE
Myosat	NA	in vitro differentiated cells	BSS01344	MYOTUBE	MYO.MYOTUBE
ES-deriv	Brain	in vitro differentiated cells	BSS01371	NEURAL PROGENITOR DERIV	ESDV.NEURAL.PROG
Neurosph	NA	primary cell	BSS01377	NEUROSPHERE	NSPH.NEUROSPHERE
Cancer	HSC & B-cell	cell line	BSS01390	B CELL LYMPHOMA	CA.BCL.OCILY3
Bone	NA	primary cell	BSS01397	OSTEOBLAST	BONE.OSTBLAST
Endocrine	Reproductive	tissue	BSS01399	OVARY	EDCR.REPR.OVARY
Cancer	Pancreas	cell line	BSS01405	PANCREAS DUCT EPITHELIAL CARCINOMA	CA.PANC.DUCT.EPITH.PANC1
Pancreas	NA	tissue	BSS01406	PANCREAS	PANC.PANCREAS
Cancer	Reproductive	cell line	BSS01414	PROSTATE ADENOCARCINOMA	CA.PRAD.PC3

Cancer	Lung	cell line	BSS01415	LUNG ADENOCARCINOMA	CA.LUAD.PC9
Blood & T-cell	NA	primary cell	BSS01423	MONONUCLEAR CELL	BLD.MONONUC
Placenta & EEM	NA	tissue	BSS01431	PLACENTA	PCTA.PLACENTA
Placenta & EEM	NA	tissue	BSS01438	PLACENTA	PCTA.PLACENTA
Placenta & EEM	NA	tissue	BSS01441	PLACENTA	PCTA.PLACENTA
Placenta & EEM	NA	tissue	BSS01446	PLACENTA	PCTA.PLACENTA
Muscle	NA	tissue	BSS01463	PSOAS MUSCLE	MUSC.PSOAS.MUSCLE
Digestive	NA	tissue	BSS01545	SIGMOID COLON	GI.SIG.COLON
Digestive	NA	tissue	BSS01548	SIGMOID COLON	GI.SIG.COLON
Cancer	Brain	cell line	BSS01562	NEUROBLASTOMA	CA.NEUROBM.SKNDSH
Myosat	NA	primary cell	BSS01574	SKELETAL MUSCLE MYOBLAST	MYO.SKLT.MYOBLAST
Muscle	NA	tissue	BSS01578	SKELETAL MUSCLE	MUSC.SKELETAL.MUSCLE
Digestive	NA	tissue	BSS01588	SMALL INTESTINE	GI.SMALL.INT
Digestive	NA	tissue	BSS01601	SMALL INTESTINE	GI.EMB.SMALL.INT
ES-deriv	Sm. Muscle	in vitro differentiated cells	BSS01612	SMOOTH MUSCLE DERIV	ESDV.SM.MUSCLE
Spleen	NA	tissue	BSS01631	SPLEEN	SPLN.SPLEEN
Digestive	NA	tissue	BSS01639	STOMACH	GI.STOMACH
Digestive	NA	tissue	BSS01651	STOMACH	GI.STOMACH
Sm. Muscle	Digestive	tissue	BSS01659	STOMACH MUSCLE	SMTH.STOM.MUSCLE
Adipose	NA	tissue	BSS01667	ADIPOSE TISSUE	ADIP.TISSUE
Brain	NA	tissue	BSS01676	SUBSTANTIA NIGRA	BRN.SUB.NG
Brain	NA	tissue	BSS01712	TEMPORAL LOBE	BRN.TMP.LB
Brain	NA	tissue	BSS01714	TEMPORAL LOBE	BRN.TMP.LB
Heart	NA	tissue	BSS01814	THORACIC AORTA	HRT.THRCIC.AORTA
Thymus	NA	tissue	BSS01820	THYMUS	THYM.EMB.THYMUS
Thymus	NA	tissue	BSS01825	THYMUS	THYM.THYMUS
ES-deriv	Placenta & EEM	in vitro differentiated cells	BSS01857	TROPHOBLAST DERIV	ESDV.TROPHOBLAST

Table A-1 Training samples and EpiMap IDs

Appendix B

The information in Appendix B relates to **Chapter 6**: **Predicting gene expression from histone marks using chromatin deep learning models depends on histone mark function, regulatory distance and cellular states** on page 179. The work in this appendix chapter is adapted from that set out in our previous publication (Alan E Murphy *et al.*, 2024).

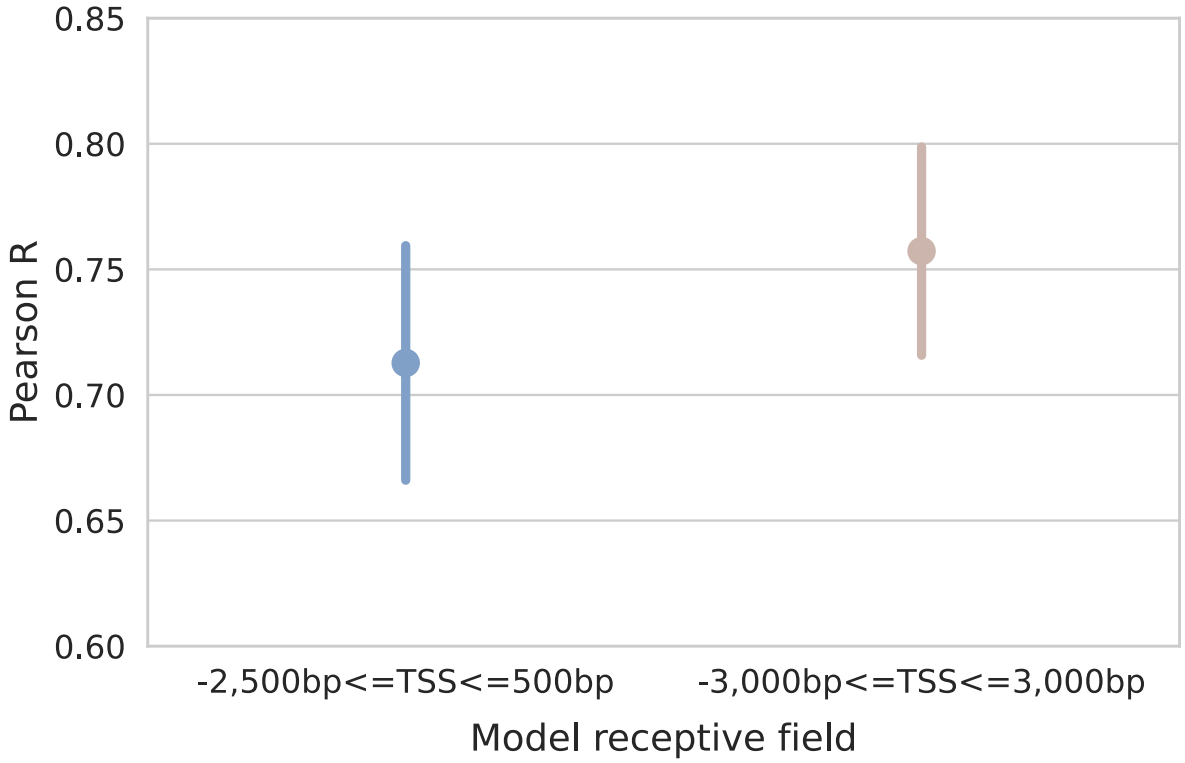


Fig. B-1 The performance of the promoter model is improved by including a larger region around the transcriptional start site (TSS) rather than a more focused, functional region. Promoter model performance (brown) with a receptive field of 6,000 base-pairs symmetrical around the TSS and a smaller receptive field model, with the same architecture, including known functional regions in the promoter (blue). Performance (y-axis) is the Pearson correlation coefficient on the blind test sets for H3K4me3, the promoter model's top performing mark. The whiskers represent the standard deviation across the different cell types and the 4-fold cross-validation.

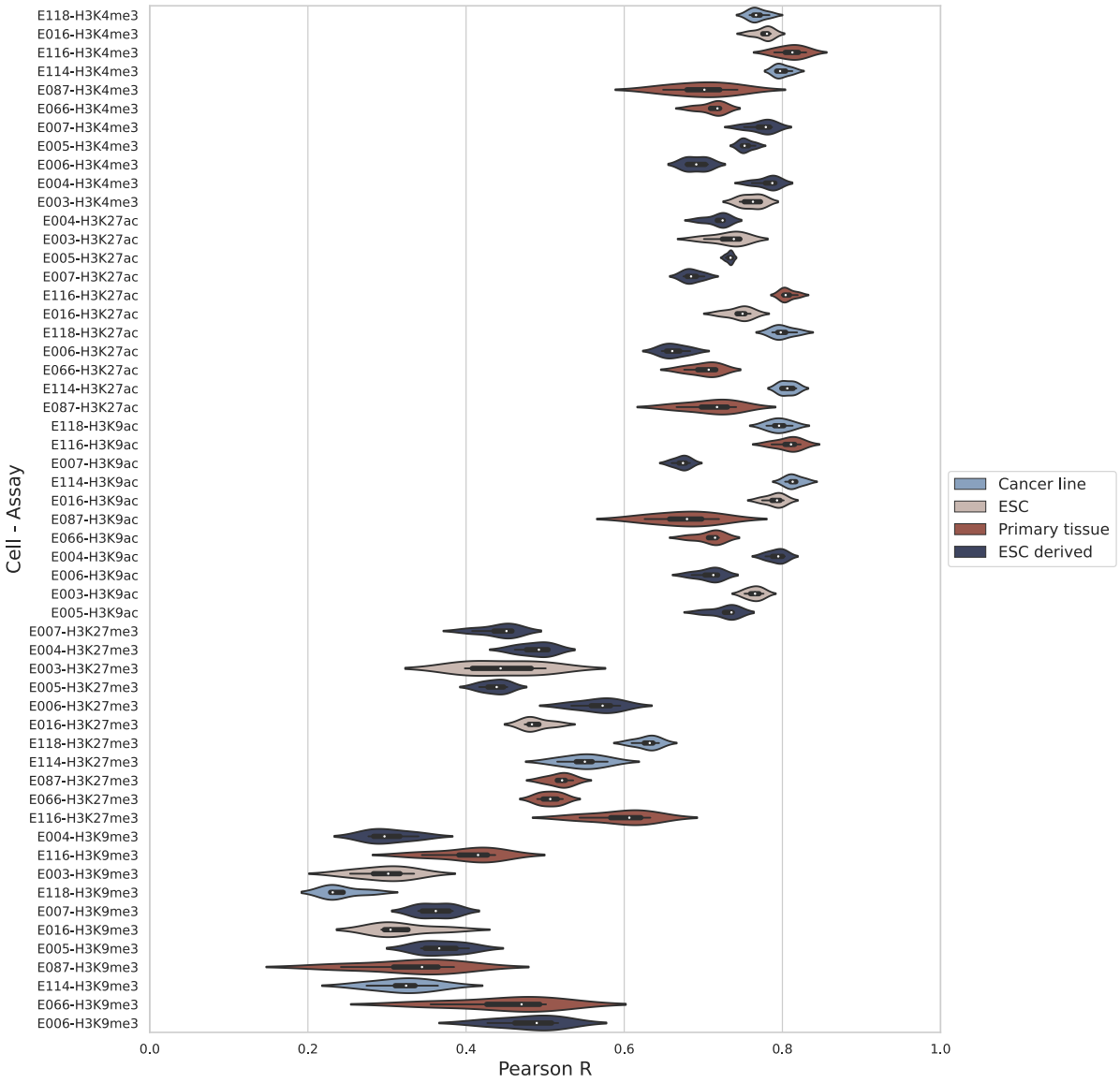


Fig. B-2 Performance of promoter model by cell type and histone mark.

Violin plot of the model's performance on histone marks measured by Pearson correlation coefficient on the blind test sets. Here, the cell type performance is coloured by the cell state - Embryonic stem cell (ESC), ESC derived cell, adult primary tissue or cancer cell line. Splitting the performance by cell type and histone mark removes the multimodal distribution noted in **Figure *6-2*b**.

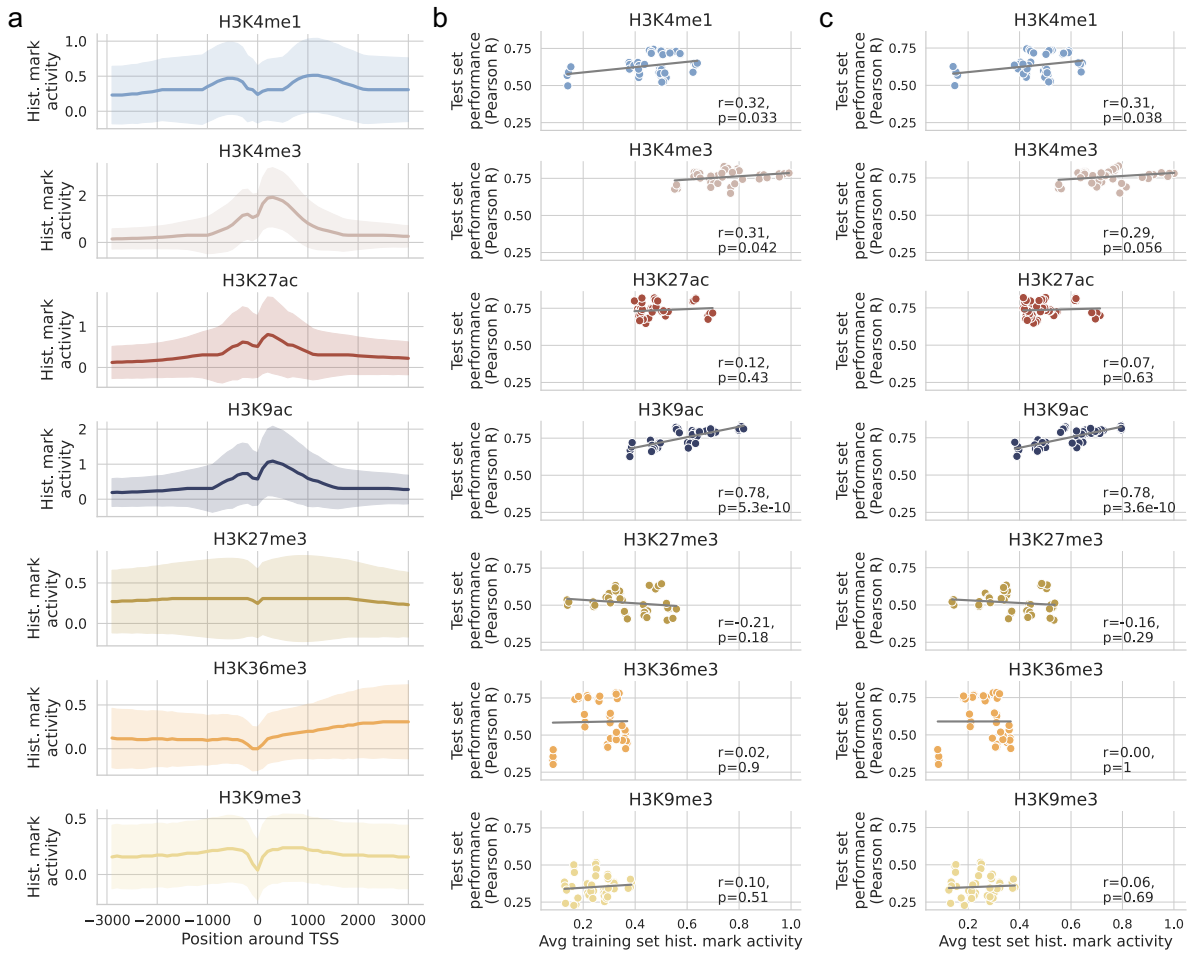


Fig. B-3 Histone mark activity and relation to performance of promoter model.

(a) Average histone mark activity for the 6,000 base-pairs around the transcriptional start site of the training genes across all cell types and folds (free y-axis). Relationship between promoter model's performance and the mean histone mark activity in the training set (b) and the test set (c). Since there isn't positive relationships for all marks, b-c highlight how the promoter model's performance was not just a result of the quantity of histone mark signal observed, rather that it learnt histone mark and cell type-specific regulatory relationships.

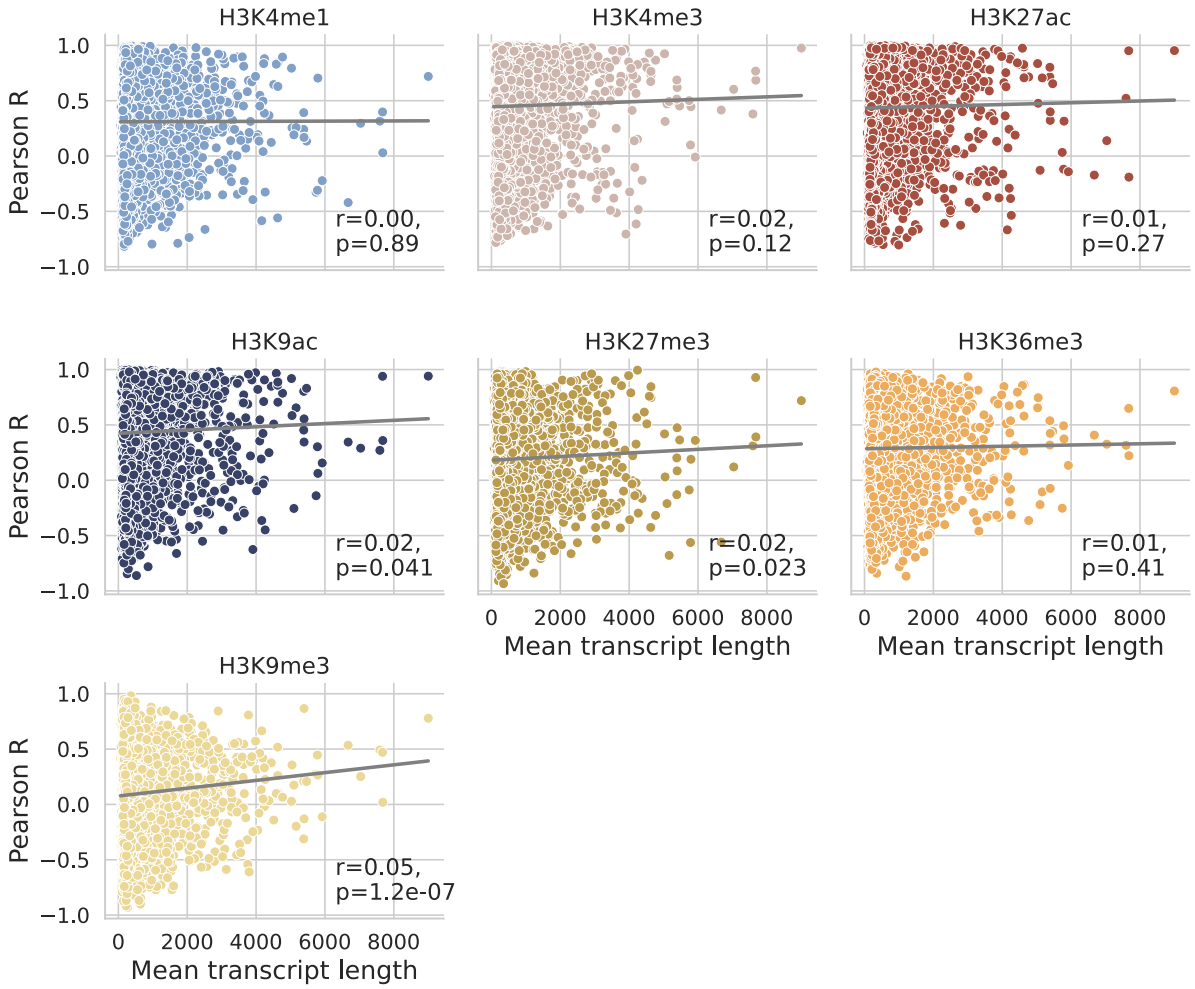


Fig. B-4 Performance of promoter model by transcript length.

Correlation between the model performance (y-axis) and the average transcript length (x-axis , in base-pairs) of all exons for each gene across all cell types and k-fold. This is split by histone marks. The average transcript length of each gene was taken from Ensembl.

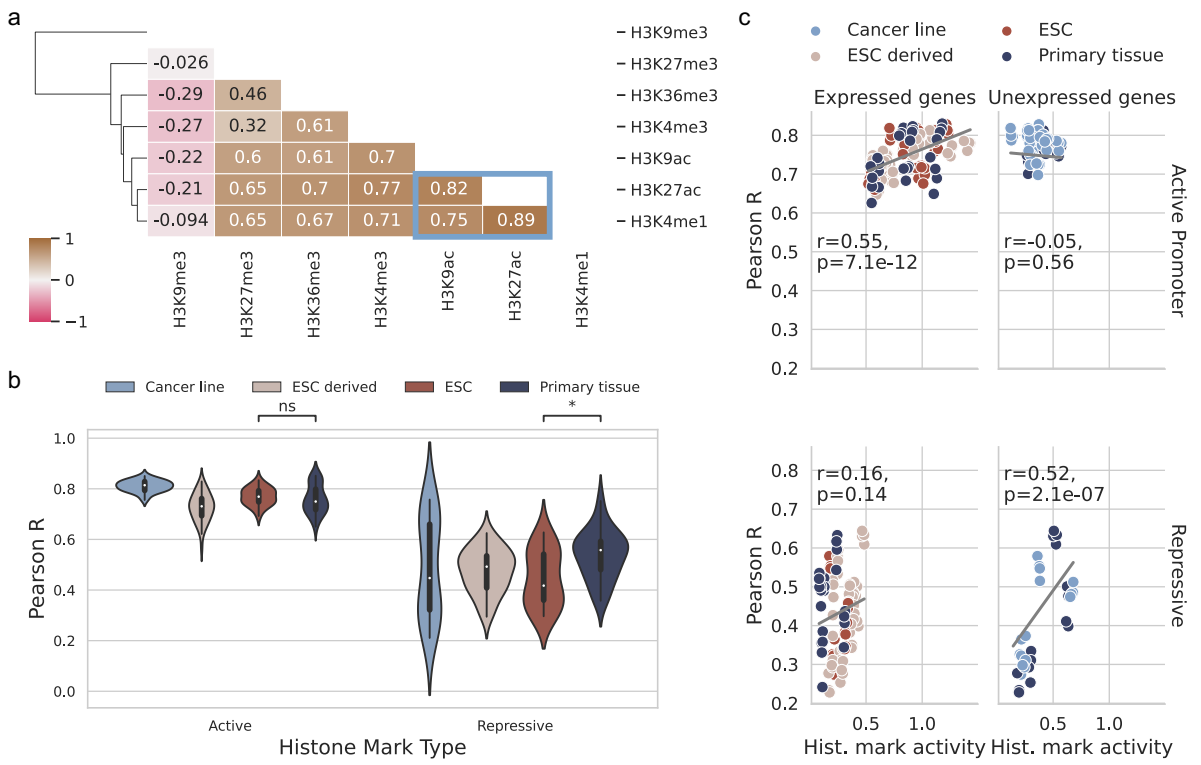


Fig. B-5 Distal model's correlation of histone marks' function and performance by cell state.

(a) Correlation matrix of the distal model's performance by the different histone marks across each cell type and cross-validation k-fold. The distinct group of active marks from the promoter model, highlighted in blue, was also replicated for the distal model. Bars along the y-axis correspond to the hierarchical clustering dendrogram. (b) Violin plot of performance measured by Pearson correlation coefficient on the blind test sets grouped by cell state and histone mark levels. Significance based on Mann-Whitney U-Test with false discovery rate multiple test correction where p-value indicators: * < 0.05, ns >= 0.05. (c) Correlation between the histone mark levels (average log2-transformed, read depth in the full 40,000 base-pairs around the transcriptional start site) and model performance for each cell type, k-fold combination. Split by repressive and active marks and highly and lowly expressed genes.

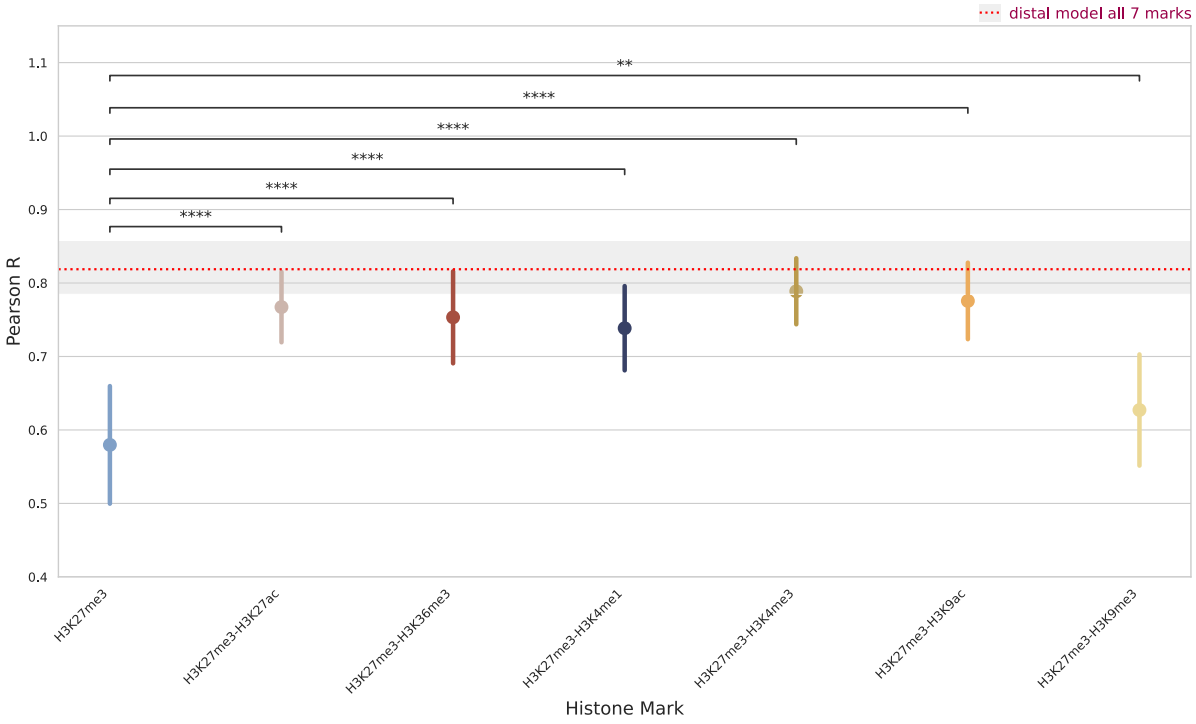


Fig. B-6 Performance of distal model in gene expression predictions from two histone mark pairs with H3K27me3.

Performance for combinations of all pairs with H3K27me3, a repressive mark. The range of values represent the standard deviation across the different cell types and the 4-fold cross-validation. Data is averaged at the level of cell type and k-fold. The red dashed line and shaded box shows the model's mean performance and standard deviation when trained on all seven histone marks together. Significance based on false discovery rate (FDR) multiple test correction where p-value indicators: **** < 1e-4, *** < 1e-3 and ** < 1e-2.

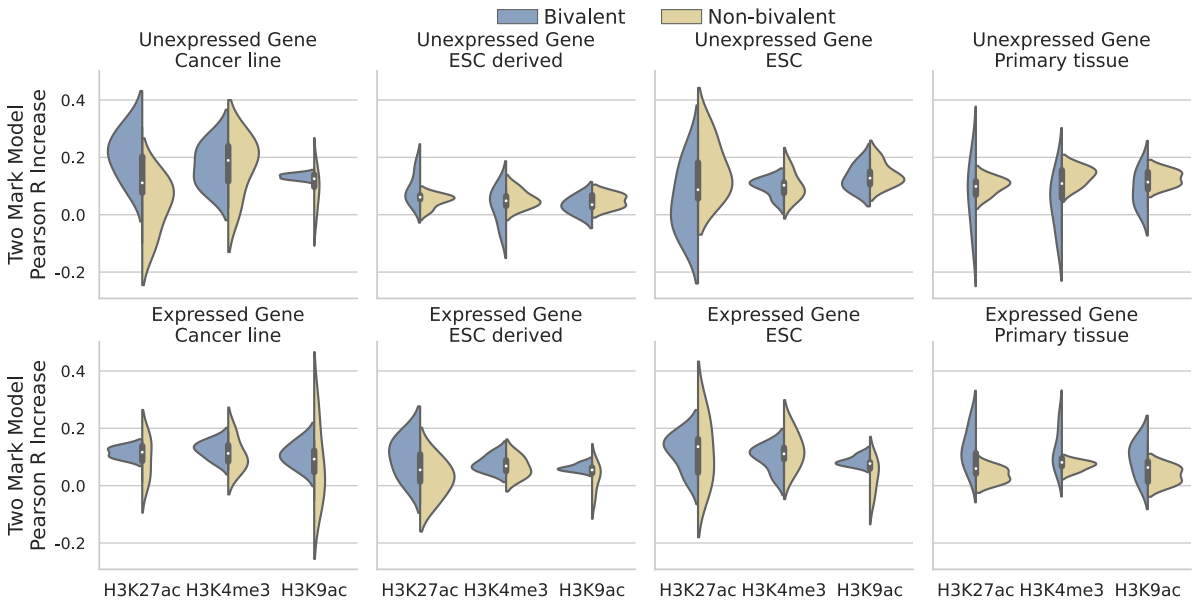


Fig. B-7 Performance improvement of distal model from two histone mark pairs with H3K27me3 in bivalent genes.

Performance improvements for combination distal model over the single mark distal model for the top three, activating histone marks with H3K36me3. The range of values represent the standard deviation across the different cell types and the 4-fold cross-validation. Data is averaged at the level of cell type and k-fold. Performance improvement is split by mark, cell type group and bivalent and non-bivalent genes. Bivalent genes are defined as those with both an active histone mark signal and repressive - H3K36me3 signal in the gene promoter region (both above the median for the cell type). Plot is also split by highly and lowly expressed genes defined in the same way as for the histone mark level, above or below the median for the cell type.

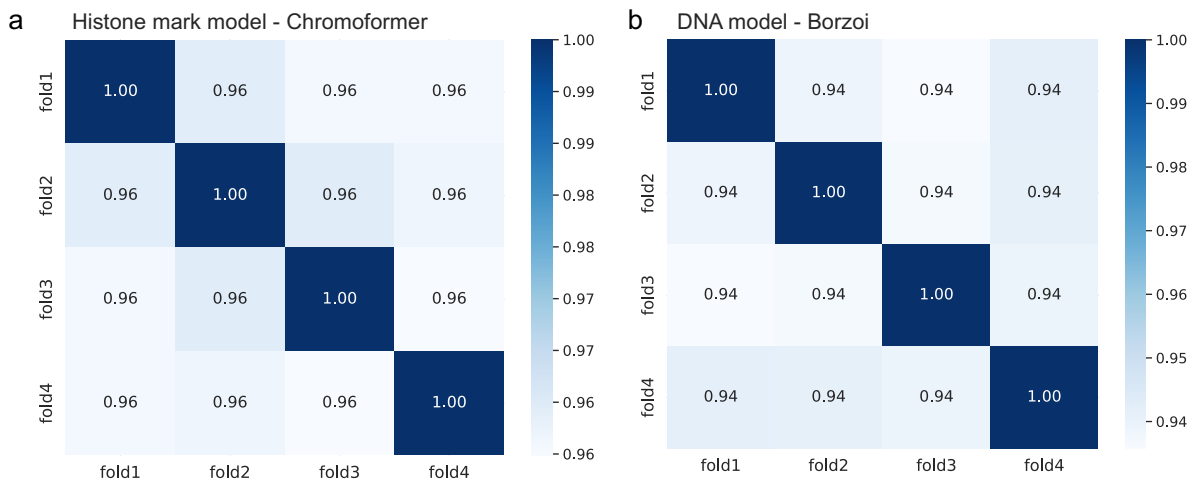


Fig. B-8 Correlation across 4 fold cross-validation models in the *in silico* perturbation analysis.

(a) Correlation for Chromoformer, the histone mark model on matched genes and histone mark perturbation loci and level across the 4-fold, cross-validation models. All models showed high concordance (Pearson R=0.96). (b) Correlation for Borzoi, a DNA model on matched genes and *in silico* mutagenesis levels across the 4-fold, cross-validation models. The same cell types and receptive field were inspected for correlation to chromoformer to make for a fair comparison. All models also showed high concordance (Pearson R=0.94) but slightly less than the histone mark model approach.

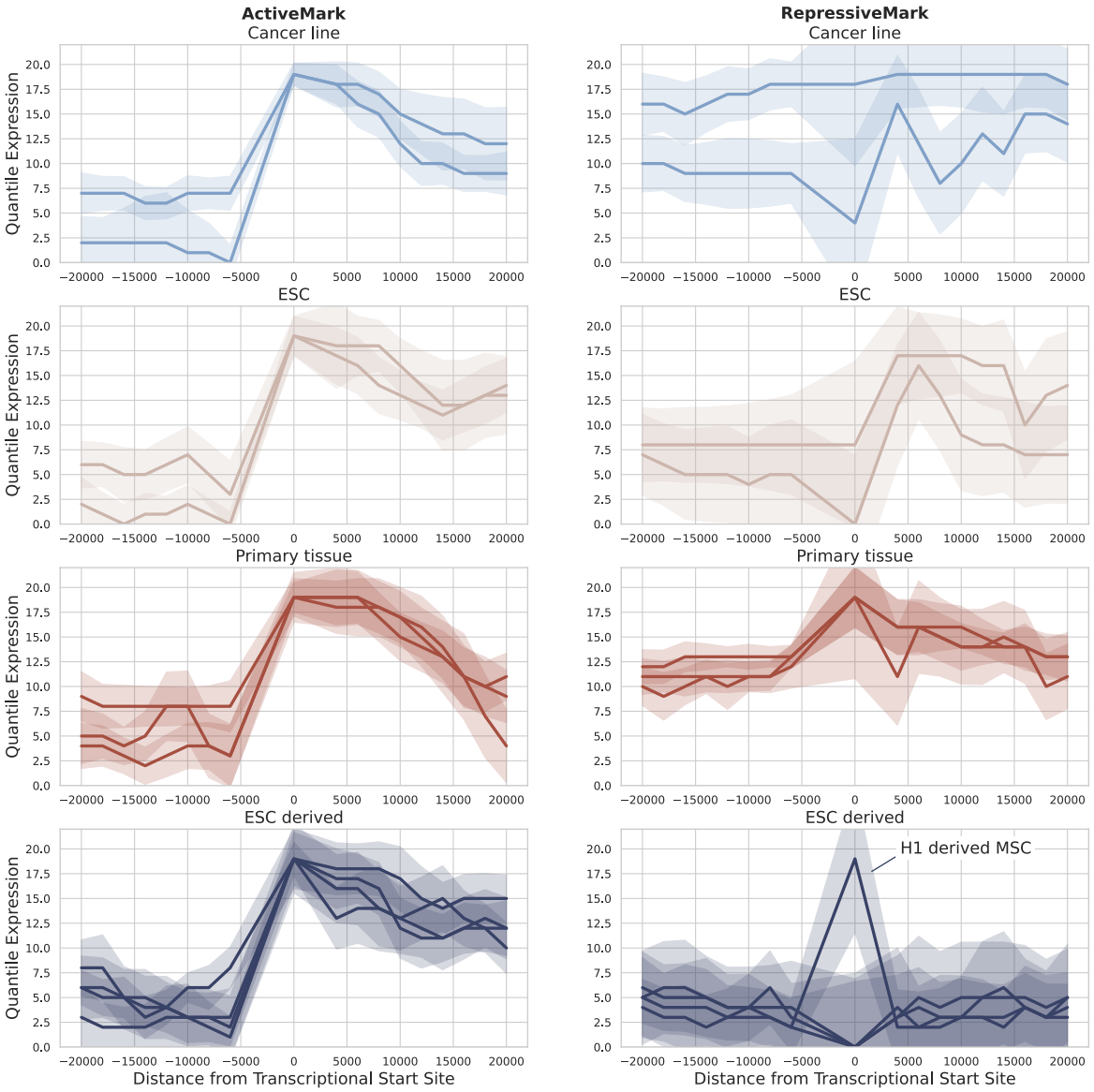


Fig. B-9 Effect of *in silico* histone mark perturbation on expression by cell state.

The effect on quantile predicted expression (y-axis) of changing the proportional levels of measured histone mark activity (x-axis) for all cell types and genes, averaged across the four k-fold models. The distal model trained on a single histone mark was used to measure the effects of a perturbed active mark - H3K27ac or a repressive mark - H3K27me3. The effect of distance on expression change is shown when the histone mark activity is completely removed at a specific locus. The distribution of all gene expression changes in all folds is split into 20 quantile bins where each line represents a cell type, grouped into cell states. H1 derived MSC is highlighted for its unique response in the repressive mark, ESC derived group.

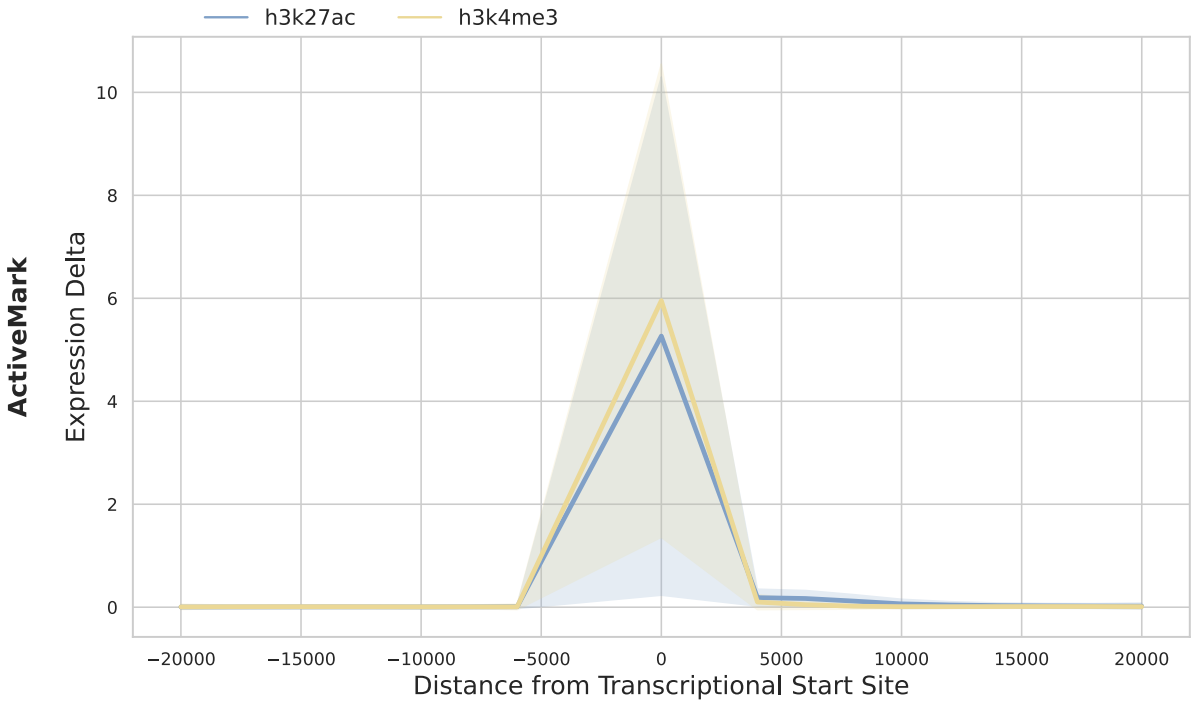


Fig. B-10 Effect of *in silico* histone mark perturbation on expression by histone mark.

The effect on predicted expression (y-axis) of changing the proportional levels of measured histone mark activity (x-axis) for all cell types and genes, averaged across the four k-fold models. Two different active histone marks were tested, a mark with known distal regulatory effects - H3K27ac, and a mark with local regulatory effects to the TSS - H3K4me3. The effect of distance on expression change is shown when the histone mark activity is completely removed at a specific locus. H3K4me3 has higher predicted change in expression in the TSS where H3K27ac has higher downstream and in the gene body.

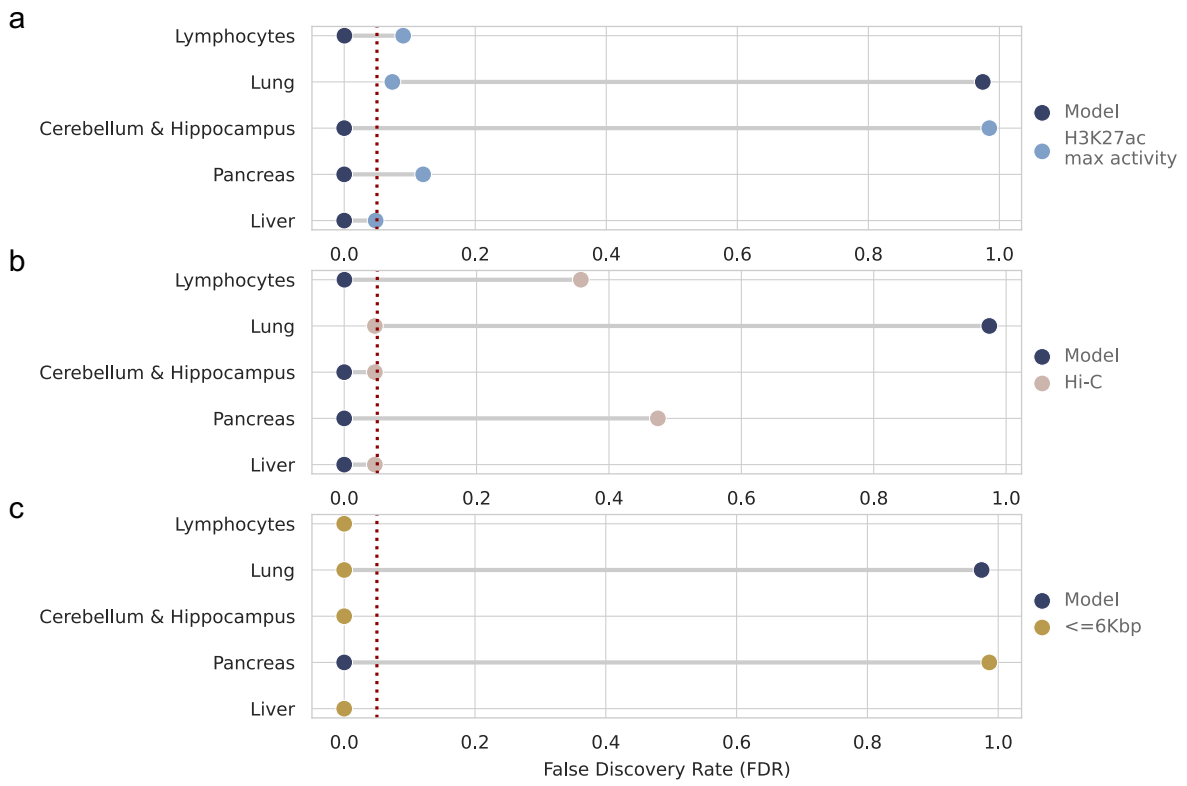


Fig. B-11 Improvement of model over alternative loci prioritisation methods for fine-mapped eQTL enrichment.

Upstream and downstream *in silico* histone mark perturbation experiments from the active model (trained on H3K27ac) were sorted into deciles based on their predicted change in expression and the top decile used; 'Model', to compare against (a) the cell type-specific region of max histone mark activity; 'H3K27ac max activity', (b) cell type-specific Hi-C data upstream of matching genes; 'Hi-C' and (c) upstream and downstream up to 6,000 base-pairs away '<=6Kbp'. All were tested for enrichment of fine-mapped eQTL interactions in matched cell types (y-axis) and compared against bootstrap sampling random upstream loci 10,000 times to generate p-values of enrichment (x-axis). Lower p-values for the model indicate greater enrichment found in more tests against randomly selected loci than the alternative method. For Hi-C Lung (b), <=6Kbp Cerebellum & Hippocampus and Lymphocytes (c), both the model and alternative method obtained equivalent p-values. Note, no fine-mapped eQTL SNPs were located <=6Kbp in the Pancreas so no dot is shown for the alternative model in (c).

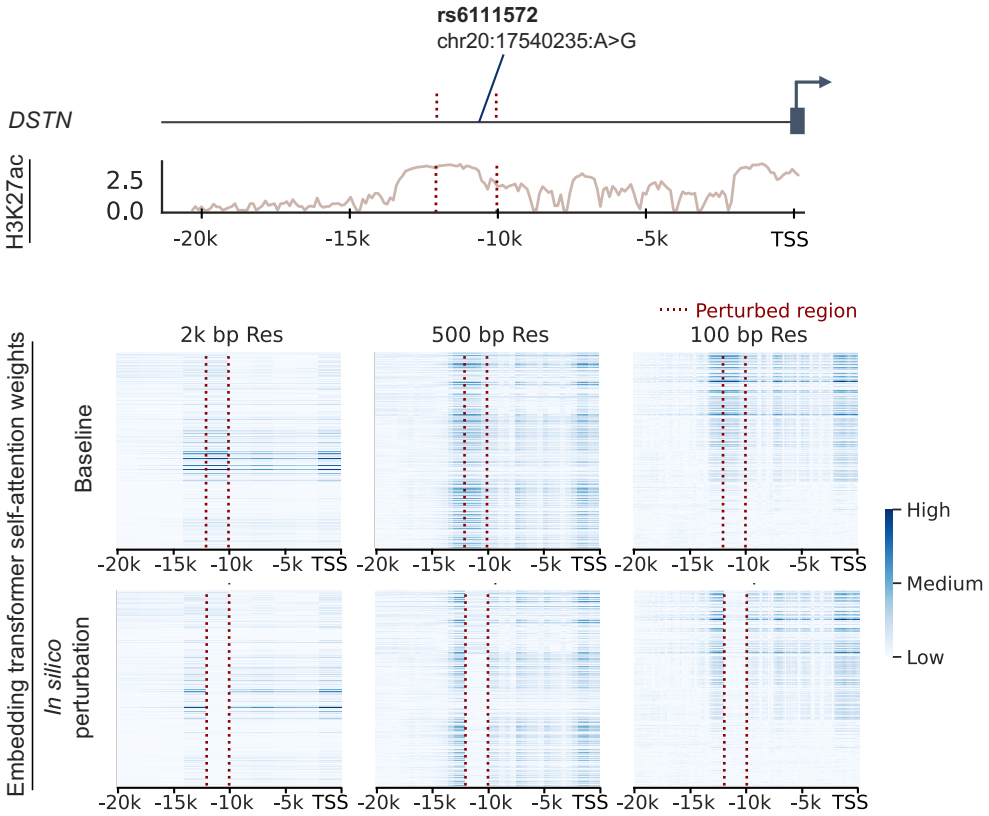


Fig. B-12 Embedding self-attention weights for the *in silico* histone mark perturbation of *DSTN* in a lung carcinoma cell line.

Upstream *in silico* histone mark perturbation of *DSTN* in the lung carcinoma cell line using the distal model trained on the H3K27ac signal. The experimental H3K27ac signal is shown along with the embedding self-attention weights for the distal model, before and after ablation of the H3K27ac signal in the highlighted region. The fine-mapped SNP rs611572 from the UK Biobank population eQTL analysis in a matched cell type is also shown. This region was chosen as the *in silico* perturbation analysis identified it in the top decile of predicted effects on expression for this cell type.

Epigenome ID (EID)	Shorthand name	GROUP	Epigenome Mnemonic	ANATOMY	TYPE
E118	Hepatocellular carcinoma	ENCODE2012	LIV.HEPG2.CNCR	LIVER	CellLine
E116	Lymphoblastoid	ENCODE2012	BLD.GM12878	BLOOD	PrimaryCulture
E114	Lung carcinoma	ENCODE2012	LNG.A549.ETOH002.CNCR	LUNG	CellLine
E087	Pancreatic islets	Other	PANC.ISLT	PANCREAS	PrimaryTissue
E066	Liver	Other	LIV.ADLT	LIVER	PrimaryTissue
E016	HUES64 ESC	ESC	ESC.HUES64	ESC	PrimaryCulture
E007	H1 derived NPC	ES-deriv	ESDR.H1.NEUR.PROG	ESC_DERIVED	ESCDerived
E006	H1 derived MSC	ES-deriv	ESDR.H1.MSC	ESC_DERIVED	ESCDerived
E005	H1 derived trophoblast	ES-deriv	ESDR.H1.BMP4.TROP	ESC_DERIVED	ESCDerived
E004	H1 derived mesendoderm	ES-deriv	ESDR.H1.BMP4.MESO	ESC_DERIVED	ESCDerived
E003	H1 ESC	ESC	ESC.H1	ESC	PrimaryCulture

Table B-1 Roadmap information on cell types.

IDs and metadata for cell types taken from Roadmap.