Exploring Mammalian Chromatin Architecture and its Impact on Gene Regulation

A thesis presented for the degree of

Doctor of Philosophy

Helen Sarah Long

Trinity term 2022

Hertford College, University of Oxford

Supervisors: Professor Cecilia Lindgren, Dr Ann-Marie Mallon,

and Dr Michelle Simon

MRC Harwell Institute

Nuffield Department of Medicine

Abstract

Mammalian chromatin architecture appears hierarchically organised and is composed of chromosome territories, compartments, topologically associating domains (TADs), loops and nucleosomes. The study of chromatin architecture has recently become prevalent due to the advancement of chromosome conformation capture (3C) and microscopy-based techniques. Many studies have predicted that TADs display important functional characteristics such as forming tissue and species-invariant regulatory domains, promoting co-regulation and insulating genes from aberrant regulation. Even with the increasing availability of data, the relationship between chromatin structures such as TADs, gene expression and function is still not fully understood. In this thesis, I use publicly available Hi-C data to rigorously assess this relationship.

To further investigate the characteristics of TADs I compiled a re-annotated TAD dataset consisting of 13 human and 20 mouse genomes, covering 8 broad tissue/cell type groups. Using this TAD dataset as a foundation, I investigated the functional features of TADs in three main areas: tissue and species specificity, functional annotations, and phenotypic characteristics. Comparative analyses between TADs and randomly placed TADs revealed that there is some evidence that TADs are both tissue- and species-specific, genes which singly occupy a TAD are highly constrained and enriched for developmental functions, and genes sharing a CTCF TAD are not more functionally similar than can be explained by their linear proximity alone. Lastly, using adipose TADs as well as chromatin loops, I investigate the aetiology of genetic variants associated with fat distribution.

The comprehensive analyses outlined suggest that contrary to common claims in the literature, TADs display a significant degree of both tissue and species specificity and play a

limited role in promoting gene co-expression. These results contribute to understanding the relationship between 3D mammalian chromatin structure and gene regulation which is crucial for dissecting gene regulatory programs.

Word count: ~54,000 (excl. references and supplementary)

Acknowledgements

Firstly, I would like to thank my supervisors Professor Cecilia Lindgren, Dr Ann-Marie Mallon and Dr Michelle Simon, for giving me the opportunity to undertake this project within the BioComputing group at Harwell and the Lindgren group at the BDI. Thank you for all your support and guidance during my DPhil. I have learnt so much about how to be a good scientist from you all. Special thanks must go to Michelle who has provided me with unwavering mentorship and support throughout.

Thank you also to all the members of the Lindgren and BioComputing groups past and present for their support and friendship. Particular thanks go to George Powell, Hugh Morgan, Simon Greenaway, Richard Reeves, Saskia Reibe-Pal and Christoffer Nellaker for their help and feedback. I would also like to thank my fellow DPhil students at MRC Harwell and the Lindgren group for being such a supportive community. I extend my sincere thanks to Adam Huffman who provided me with invaluable cluster computing support over the years.

I have been very fortunate to have been able to collaborate with both the Cox group and Nolan groups at MRC Harwell in the course of my DPhil. I would particularly like to thank Roger Cox, Rebecca Dumbell, Louisa Zolkiewski, Patrick Nolan and Gareth Banks for making these collaborations so enjoyable.

I extend a huge thanks to my friends who have provided me with support and advice throughout, particularly Sophie McManus, Sophie Roper, Natalia Zajac and Anna Parsons, as well as my friends in Hertford MCR.

I would also like to thank my parents Katherine and Keith Long, my brother Stephen Long and my grandparents Joan and Ken Mettam, and Joan and Arthur Long for their support and encouragement throughout my education and this DPhil. Not forgetting our dogs Dusty and Isla who never failed to brighten a hard day.

Lastly, my special thanks go to Christopher Han for always believing in me and supporting me in everything I do.

Research outputs arising from this DPhil

Publications:

Long, Helen, Simon Greenaway, George Powell, Ann-Marie Mallon, Cecilia M. Lindgren, and Michelle M. Simon. 2022. "Making Sense of the Linear Genome, Gene Function and TADs." Epigenetics & Chromatin 2022 15:1 15 (1): 1–19. https://doi.org/10.1186/S13072-022-00436-9.

Long, Helen, Richard Reeves, and Michelle M. Simon. 2022. "Mouse Genomic and Cellular Annotations." Mammalian Genome 33 (1): 19–30. https://doi.org/10.1007/S00335-021-09936-7.

Powell, George, **Helen Long**, Louisa Zolkiewski, Rebecca Dumbell, Ann Marie Mallon, Cecilia M. Lindgren, and Michelle M. Simon. 2022. "Modelling the Genetic Aetiology of Complex Disease: Human-Mouse Conservation of Noncoding Features and Disease-Associated Loci." Biology Letters 18 (3). https://doi.org/10.1098/RSBL.2021.0630.

Laber, Samantha, Sara Forcisi, Liz Bentley, Julia Petzold, Franco Moritz, Kirill S. Smirnov, Loubna Al Sadat, Iain Williamson, Sophie Strobel, Thomas Agnew, Shahana Sengupta, Tom Nicol, Harald Grallert, Margit Heier, Julius Honecker, Joffrey Mianne, Lydia Teboul, Rebecca Dumbell, **Helen Long**, et al. 2021. "Linking the FTO Obesity Rs1421085 Variant Circuitry to Cellular, Metabolic, and Organismal Phenotypes in Vivo." Science Advances 7 (30). https://doi.org/10.1126/SCIADV.ABG0108.

Conference talks:

Making sense of the linear genome, gene function and TADs- Genome Organization & Nuclear Function, Cold Spring Harbor Laboratory, April 2020

Poster prizes:

Outstanding research poster prize- 33rd International Mammalian Genome Conference, September 2019

Contents

Chapter 1 : Introduction 1
1.1. Gene regulation 1
1.2. Hierarchy of chromatin architecture in the mammalian genome 2
1.2.1. Methods of investigating chromatin architecture 3
1.2.2. Chromosome territories 7
1.2.3. Compartments 7
1.2.4. TADs and sub-TADs 9
1.2.5. Loops 12
1.2.6. Nucleosome density 14
1.3. Mechanisms of chromatin architecture formation 15
1.3.1. Phase separation drives the formation of compartments 15
1.3.2. Loop extrusion is required for the formation of TADs 16
1.3.3. The interplay between both mechanisms 18
1.4. The relationship between chromatin architecture and function 19
1.4.1. Regulation of gene expression 19
1.4.2. Replication timing 25
1.4.3. Importance of chromatin architecture in disease 26
1.5. Chromatin architecture across evolution 28
1.6. Thesis aims 31
Chapter 2 : Conservation of TADs between tissues and species 32
2.1. Introduction 32
2.1.1. Aims 37
2.2. Methods 38
2.2.1. Hi-C datasets 38
2.2.2. Processing of Hi-C datasets 42
2.2.3. TAD calling 43
2.2.4. Random down-sampling of Hi-C contacts 44
2.2.5. TAD chromosome coverage 45
2.2.6. TAD and TAD boundary comparison method 45

2.2.7. Comparing TADs and TAD boundaries between species 48
2.2.8. Per TAD tissue and species conservation score 48
2.3. Results 50
2.3.1. Assessment of resolution and quality of selected Hi-C data 50
2.3.2. The effect of bin size, TAD calling parameters and the number of Hi-C contacts on TADs detected by Arrowhead and TopDom 53
2.3.3. Comparison of TADs called by Arrowhead and TopDom in each dataset 60
2.3.4. Similarity of TADs and TAD boundaries between tissues 63
2.3.5. Similarity of TADs and TAD boundaries between human and mouse 69
2.3.6. Per TAD species and tissue conservation 74
2.4. Discussion 76
Chapter 3 : Assessing the relationship between TADs, expression and gene function 82
3.1. Introduction 82
3.1.1. Aims 84
3.2. Methods 85
3.2.1. Hi-C data processing 85
3.2.2. TAD calling and gene overlap 85
3.2.3. TAD randomisation 87
3.2.4. Compartment annotation 89
3.2.5. Paralogous genes 89
3.2.6. Olfactory genes 90
3.2.7. Selective gene constraint score 90
3.2.8. Functional enrichment 91
3.2.9. Gene expression correlation 92
3.2.10. GO Semantic similarity 93
3.2.11. KEGG pathway annotation 94
3.2.12. PPI annotation 94
3.2.13. Distance permutation testing 94
3.2.14. Effect size 95
3.3. Results 96
3.3.1. Assessment of Hi-C data quality and TAD caller parameter selection 96

3.3.2. General features of ESC and cortical neuron TADs 100
3.3.3. Benchmarking of TAD randomisation strategy.....104
3.3.4. Features of CTCF and non-CTCF TADs 113
3.3.5. Increased distance between pairs of genes sharing a TAD 117
3.3.6. The relationship between gene paralogous families and TADs 119
3.3.7. Evolutionary constraint of genes singly or co-occupying a TAD.....124
3.3.8. Expression correlation between pairs of genes in CTCF TADs 136
3.3.9. Similarity of functional annotations between pairs of genes in CTCF TADs....139
3.4. Discussion.....145
Chapter 4 : Using chromatin architecture to investigate the function of GWAS variants associated with obesity/fat distribution.....149
4.1. Introduction.....149
4.1.1. Aims.....154
4.2. Methods 155
4.2.1. Hi-C data processing and TAD calling 155
4.2.2. TBX15-WARS2 locus genes 156
4.2.3. Candidate SNPs.....156
4.2.4. Genotype-Tissue Expression (GTEx) project.....157
4.2.5. Compartment annotation.....157
4.2.6. Promoter/enhancer capture Hi-C data processing.....157
4.2.7. Published and recalled loop comparison.....160
4.2.8. Overlapping loops with SNPs and gene TSSs.....160
4.2.9. Human vs mouse TAD loop comparison score 161
4.2.10. GWAS and international mouse phenotyping consortium (IMPC) TADs.....163
4.2.11. Genome visualisation.....164
4.3. Results 165
4.3.1. Human TBX15-WARS2 locus and its chromatin architecture.....165
4.3.2. Characterisation of candidate causal obesity/fat distribution associated SNPs in the TBX15-WARS2 locus.....170
4.3.3. TADs and compartments in this locus are largely conserved in mouse 173
4.3.4. Chromatin looping overlying candidate SNPs in the TBX15-WARS2 locus.....177

4.3.5. Global human vs mouse regulatory circuitry conservation score 201
4.4. Discussion 210
Chapter 5 : Discussion 217
5.1. TADs as gene regulatory domains? 218
5.1.1. Tissue and species specificity 218
5.1.2. Promoting gene co-regulation 219
5.1.3. Regulatory insulation 221
5.2. TADs and loops in variant to function analysis 221
5.3. Future directions 223
5.4. Conclusions 228
Chapter 6 : References 229
Chapter 7 : Supplementary 275

Table of figures

Figure 1.1: Hierarchy of chromatin architecture 3

Figure 1.2: Schematic overview of Hi-C 6

Figure 1.3: Analysis of TAD caller similarity (measure of concordance) from Zufferey et al. .. 11

Figure 1.4: Schematic illustrating how chromosome territories, compartments, TADs, sub-TADs and loops appear in Hi-C maps 14

Figure 1.5: Phase separation and loop extrusion 17

Figure 1.6: Models of enhancer-promoter communication 24

Figure 2.1: TAD and boundary similarity score 46

Figure 2.2: Hi-C dataset resolution 52

Figure 2.3: Number and size of autosomal TADs at 10 kb, 25 kb or 50 kb bin size 55

Figure 2.4: Number and size of autosomal TADs at different parameter values for Arrowhead and TopDom 57

Figure 2.5: Map resolution, number and size of autosomal TADs detected after randomly down-sampling Hi-C contacts 59

Figure 2.6: Comparison of autosomal TADs and boundaries detected by either Arrowhead or TopDom 62

Figure 2.7: Autosomal TAD comparison between all pairs of datasets within species 65

Figure 2.8: Autosomal TAD boundary comparison between all pairs of datasets 68

Figure 2.9: Distribution of autosomal TAD similarity scores between pairs of datasets in the same tissue group vs different tissue groups. 71

Figure 2.10: Distribution of autosomal TAD boundary similarity scores between pairs of datasets in the same tissue group vs different tissue groups 73

Figure 2.11: Per autosomal TAD tissue and species conservation 75

Figure 3.1: ESC TADs detected with TopDom and different values of the “w” parameter 99

Figure 3.2: Size distribution of TopDom autosomal TADs in ESC (top) and cortical neuron (bottom) 100

Figure 3.3: General features of autosomal ESC and cortical neuron TADs detected with Arrowhead or TopDom 102

Figure 3.4: TAD randomisation and functional comparison strategy 106

Figure 3.5: TADs and random TADs 108

Figure 3.6: Comparison of my randomisation method with previously published methods (autosomal TADs) 110

Figure 3.7: Comparison of the overlap structure of random ESC TADs produced by my randomisation method and previously published methods (autosomal TADs).....112

Figure 3.8: CTCF ChIP-seq binding at autosomal TAD boundaries. 115

Figure 3.9: Median distance between genes in TADs vs random TADs (autosomal).....118

Figure 3.10: Proportion of paralogous gene pairs in TADs vs proportion in random TADs (autosomal).....120

Figure 3.11: Size of autosomal TADs which contain a pair of paralogs.....122

Figure 3.12: Average constraint of genes sharing an autosomal TAD 124

Figure 3.13: Average constraint of genes in random TADs and random genome TADs (autosomal).....126

Figure 3.14: Average constraint of genes sharing autosomal CTCF TADs or non-CTCF TADs 128

Figure 3.15: Biological process GO functional enrichment of genes singly, double or triply occupying a TAD.....130

Figure 3.16: Biological process GO term enrichment of genes singly, doubly or triply occupying autosomal TADs compared to random TADs 132

Figure 3.17: Comparison of genes singly occupying autosomal TADs in ESC and cortical neuron.....134

Figure 3.18: Expression correlation between pairs of genes sharing an autosomal CTCF TAD 137

Figure 3.19: Pairwise GO similarity scores against genomic distance.....140

Figure 3.20: Pairwise functional similarity against genomic distance.....142

Figure 3.21: Functional similarity of pairs of genes sharing an autosomal CTCF TAD 143

Figure 4.1: Human TBX15-WARS2 locus.....166

Figure 4.2: Features of candidate SNPs.....172

Figure 4.3: Mouse TADs and compartments in Tbx15-Wars2 locus 176

Figure 4.4: Schematic outlining the definition of potential regulatory SNP loops.....178

Figure 4.5: Published potential regulatory SNP loops in HWA in the TBX15-WARS2 locus..182

Figure 4.6: Published potential regulatory SNP loops in human MSC differentiating towards adipocytes in the TBX15-WARS2 locus 184

Figure 4.7: Published potential regulatory SNP loops in mouse adipogenesis in the Tbx15-Wars2 locus.....185

Figure 4.8: Proportion of reads in each HiCUP QC category per dataset.....188

Figure 4.9: Global comparison of published and recalled loops 190

Figure 4.10: Recalled potential regulatory SNP loops in SGBS in the TBX15-WARS2 locus..192

Figure 4.11: Recalled potential regulatory SNP loops in HWA in the TBX15-WARS2 locus..193

Figure 4.12: Recalled potential regulatory SNP loops in human MSC differentiating towards adipocytes in the TBX15-WARS2 locus.....196

Figure 4.13: Recalled potential regulatory SNP loops in mouse pre-adipocytes differentiating towards adipocytes in the Tbx15-Wars2 locus.....198

Figure 4.14: Schematic outlining the calculation of the global human vs mouse regulatory circuitry conservation score.....202

Figure 4.15: Number of human loops, number of human loops conserved in mouse, and the proportion of conserved loops per TAD 206

Supplementary figures:

Chapter 2: Supplementary Figure 1: Number of detected autosomal TADs vs number of Hi-C contacts in all datasets.....275

Chapter 2: Supplementary Figure 2: Similarity score between Arrowhead and TopDom TADs (top) and boundaries (bottom) detected in each dataset.....276

Chapter 2: Supplementary Figure 3: Percentage chromosome coverage of Arrowhead or TopDom autosomal TADs 277

Chapter 2: Supplementary Figure 4: Autosomal TAD comparison between all pairs of datasets, excluding those from the same data repository. 278

Chapter 2: Supplementary Figure 5: Autosomal TAD boundary comparison between all pairs of datasets, excluding those from the same data repository.....280

Chapter 2: Supplementary Figure 6: Autosomal TAD comparison between all pairs of datasets, excluding those with a common first or last author 284

Chapter 2: Supplementary Figure 7: Autosomal TAD boundary comparison between all pairs of datasets, excluding those with a common first or last author.....286

Chapter 2: Supplementary Figure 8: Number of autosomal TADs before and after liftOver between hg19 and mm10. 287

Chapter 2: Supplementary Figure 9: Heat maps indicating pairwise autosomal TAD similarity scores between datasets 288

Chapter 2: Supplementary Figure 10: Heat maps indicating pairwise TAD autosomal boundary similarity scores between datasets.....290

Chapter 2: Supplementary Figure 11: Distribution of autosomal TADs and TAD boundaries similarity scores between pairs of datasets in the same tissue group vs different tissue groups, excluding those from the same data repository.....291

Chapter 2: Supplementary Figure 12: Distribution of autosomal TADs and TAD boundaries similarity scores between pairs of datasets in the same tissue group vs different tissue groups, excluding those with a common first or last author.....292

Chapter 3: Supplementary Figure 1: Median value of measures with each added set of random autosomal TADs 293

Chapter 3: Supplementary Figure 2: Cortical neuron TADs detected with TopDom and different values of the "w" parameter 294

Chapter 3: Supplementary Figure 3: Comparison of the overlap structure of random cortical neuron TADs produced by my randomisation method and previously published methods (autosomal TADs).....295

Chapter 3: Supplementary Figure 4: Proportion of paralogous gene pairs in TADs vs proportion in random TADs (autosomal). Equivalent figure to Figure 3.9 but showing the full distribution of proportions for random TADs and random genome TADs.....297

Chapter 4: Supplementary Figure 1: Published loops in SGBS in the TBX15-WARS2 locus .303

Chapter 4: Supplementary Figure 2: Published loops in HWA in the TBX15-WARS2 locus .304

Chapter 4: Supplementary Figure 3: Published loops in human MSC differentiating towards adipocytes in the TBX15-WARS2 locus. 305

Chapter 4: Supplementary Figure 4: Published loops in mouse pre-adipocytes differentiating towards adipocytes in the Tbx15-Wars2 locus.....306

Chapter 4: Supplementary Figure 5: Recalled loops in SGBS in the TBX15-WARS2 locus. ..307

Chapter 4: Supplementary Figure 6: Recalled loops in HWA in the TBX15-WARS2 locus. ..308

Chapter 4: Supplementary Figure 7: Recalled loops in Human MSC differentiating towards adipocytes in the TBX15-WARS2 locus. 309

Chapter 4: Supplementary Figure 8: Recalled loops in mouse pre-adipocytes differentiating towards adipocytes in the Tbx15-Wars2 locus.....310

Chapter 4: Supplementary Figure 9: Number of loops in recalled datasets and percentage also detected in published loops. 311

Chapter 4: Supplementary Figure 10: Correlation between the total number of human loops and the proportion of conserved loops per TAD. 312

Table of tables

Table 2.1: Human and mouse selected Hi-C datasets. 39

Table 2.2: Table outlining the datasets included in each tissue group 47

Table 3.1: Statistics relating to the Hi-C data selected....97

Table 3.2: Comparison of autosomal TADs between cell types and TAD callers 104

Table 3.3: Compartment identity of CTCF and non-CTCF autosomal TADs 117

Table 3.4: Percentage of CTCF TADs (positive) and non-CTCF TADs (negative) in the A or B compartments for 0 gene TADs, 1 gene TAD, 2 gene TADs and 3 gene autosomal TADs. ...136

Table 4.1: Default Chicago settings for 6-cutter and 4-cutter enzymes. 160

Table 4.2: Genes in the TBX15-WARS2 locus. 169

Table 4.3: Statistics relating to the Mouse adipose Hi-C data 174

Table 4.4: Features of selected published human and mouse PCHi-C and ECHi-C data....180

Table 4.5: SNP-gene interactions occurring in multiple analysed looping datasets 200

Table 4.6: Top ten WHRadjBMI GWAS TADs with highest number of conserved loops between human and mouse 208

Supplementary Tables:

Chapter 2: Supplementary Table 1: First and last authors of publications associated with selected datasets. 281

Chapter 3: Supplementary Table 1: Comparison of autosomal CTCF and non CTCF TADs ..296

Chapter 3: Supplementary Table 2: Effect sizes corresponding to the p-values displayed in Figure 12: Average constraint of genes sharing an autosomal TAD. 298

Chapter 3: Supplementary Table 3: Effect sizes corresponding to the p-values displayed for autosomal CTCF TADs in Figure 14: Average constraint of genes sharing an autosomal CTCF TADs or non-CTCF TADs. 299

Chapter 3: Supplementary Table 4: Effect sizes corresponding to the p-values displayed for autosomal non-CTCF TADs in Figure 14: Average constraint of genes sharing an autosomal CTCF TADs or non-CTCF TADs. 300

Chapter 4: Supplementary Table 1: Method used to generate the bait map required to run ChiCAGO for each dataset used. 301

Chapter 4: Supplementary Table 2: GTEx adipose eQTLS involving SNPs from the credible set defined in Liu et al....302

Abbreviations

3C	Chromosome Conformation Capture
4C	Chromosome Conformation Capture-on-Chip
5C	Chromosome Conformation Capture Carbon Copy
BMI	Body mass index
BP	Biological process
CC	Cellular component
ChiP-seq	Chromatin immunoprecipitation sequencing
CN	Cortical neuron
CTCF	CCCTC-binding factor
EBI	European Bioinformatics Institute
ECHi-C	Enhancer capture Hi-C
eQTL	Expression quantitative trait loci
ESC	Embryonic stem cell
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GAM	Genome Architecture Mapping
GEO	Gene Expression Omnibus
GO	Gene ontology
GOSemSim	GO semantic similarity R package
GTEx project	Genotype-Tissue Expression project
GWAS	Genome-wide association study
HMM	Hidden Markov Model
LD	Linkage disequilibrium
MAPQ	Mapping quality
MF	Molecular function
MSC	Mesenchymal stem cells
NCBI SRA	The National Centre for Biotechnology Information Sequence Read Archive
NGS	Next-generation sequencing
NPC	Neural progenitor cells
PCHi-C	Promoter capture Hi-C

PCR	Polymerase chain reaction
PGC	Primordial germ cells
PPI	Protein-protein interactions
QC	Quality control
RPKM	Reads Per Kilobase of transcript per Million reads mapped
RPM	Reads Per Million Reads
SGBS	Simpson-Golabi-Behmel syndrome
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
SPRITE	Split-pool Recognition of Interactions by Tag Extension
TAD	Topologically associating domain
t-SNE	Stochastic neighbour embedding analysis
TSS	Transcription start site
VC	Vanilla coverage
WHR	Waist-hip ratio
WHRadjBMI	Waist-hip ratio adjusted for BMI
ZT	Zeitgeber time

Chapter 1: Introduction

1.1. Gene regulation

The human body is estimated to be made up of ~37 trillion cells, with 200-500 currently defined major cell types (Bianconi et al. 2013; Vickaryous and Hall 2006). Incredibly, the same set of genetic instructions is found in almost all of these cells. Deciphering these instructions and understanding how they can give rise to the vast diversity of cell types making up multicellular organisms, each with different regulatory programs and complex cellular functions, has been a critical objective in biology. It is now known that this is achieved by the complex spatiotemporal regulation of gene expression.

Regulation of gene expression occurs through two main mechanisms. The first is through *trans*-regulatory elements, such as protein transcription factors which act by binding to the DNA. The second is through regions of the genome called *cis*-regulatory elements which are able to control the spatiotemporal expression patterns of nearby genes (Wittkopp and Kalay 2011). *Cis*-regulatory elements are commonly thought of as being split into three classes: promoters, enhancers, and boundary elements (also known as insulators). However, there is growing evidence that enhancers and promoters may not always be distinct entities (reviewed in (Oudelaar and Higgs 2020; Andersson and Sandelin 2019)). Furthermore, other classes of *cis*-regulatory elements such as silencers may exist; however, they are less well studied (Pang and Snyder 2020; Ngan et al. 2020). Promoters are short sequences located directly upstream of the transcription start site (TSS), where RNA polymerase II and transcription factors bind and gene transcription begins (Haberle and Stark 2018). Enhancers

are areas of DNA sequence which can activate the expression of genes over long distances (Panigrahi and O'Malley 2021). In contrast, boundary elements have been shown to block both the spread of heterochromatin and euchromatin as well as the impact of enhancers on genes. CCCTC-binding factor (CTCF) is frequently found at mammalian boundary elements (Oudelaar and Higgs 2020; S. Kim, Yu, and Kaang 2015; Wendt et al. 2008).

Despite our tendency to visualise the genome as a linear structure, each nucleus actually contains ~2 m of DNA (Piovesan et al. 2019). Therefore, in actuality, the genome must be highly folded and compacted in order to fit within the cell nucleus. This is achieved by the condensation of individual chromosomes in the cell nucleus into a structure known as chromatin. Chromatin can condense the mammalian genome by ~10,000-20,000-fold (Belmont 2002; G. Li, Sudlow, and Belmont 1998), and the 3D chromatin architecture generated can have functional implications (Misteli 2020; Oudelaar and Higgs 2020). In this thesis I investigate 3D chromatin architecture and its impact on mammalian gene regulation, with particular focus on topologically associating domains (TADs) and loops.

1.2. Hierarchy of chromatin architecture in the mammalian genome

In the mammalian genome, chromatin folds into an intricate hierarchy of 3D structures, creating non-random interactions between sections of DNA which have functional significance in understanding processes such as gene regulation and DNA replication (Bonev and Cavalli 2016; Oudelaar and Higgs 2020). The chromatin architecture hierarchy is traditionally thought of as being composed of chromosomes which occupy their own territories at the highest scale, then at progressively smaller scales: compartments, TADs, loops and nucleosome density (Figure 1.1) (M. Yu and Ren 2017).

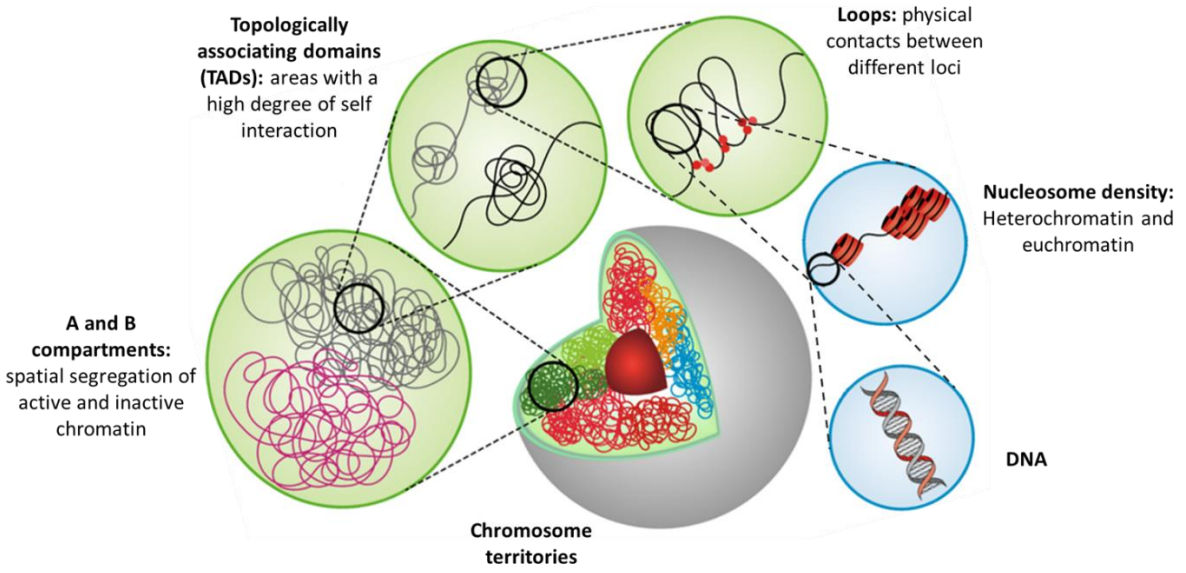


Figure 1.1: Hierarchy of chromatin architecture. *Schematic illustrating chromatin architecture in the cell nucleus at each level of the hierarchy: chromosome territories, A and B compartments, TADs, loops, and nucleosome density* [1].

1.2.1. Methods of investigating chromatin architecture

The 3D chromatin architecture of the genome has traditionally been investigated using microscopy-based techniques such as fluorescent in situ hybridisation and electron microscopy (Langer-Safer, Levine, and Ward 1982; Woodcock and Horowitz 1997). However, these techniques are limited to the study of a small number of cells and loci. The development of Chromosome Conformation Capture (3C techniques) allowed a revolution in the study of chromatin architecture by permitting its investigation at higher resolution (reviewed in (de Wit and de Laat 2012)). The original 3C method developed by Dekker et al. allowed the study

[1] Adapted by alteration of figure labels by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Plants, Doğan, Ezgi Süheyla, and Chang Liu. 2018. "Three-Dimensional Chromatin Packing and Positioning of Plant Genomes." Nature Plants 2018 4:8 4 (8): 521–29. https://doi.org/10.1038/s41477-018-0199-5.

of the frequency of chromatin contacts between any two genomic loci of interest (Dekker et al. 2002), and is often described as a 'one vs one' method (de Wit and de Laat 2012). The 3C method involves the cross-linking of the two target loci together if they are close together in 3D space/interacting, digesting the DNA using a restriction enzyme and then re-ligating cross-linked restriction fragments together. The resulting ligation products, which represent the interactions between the target loci, can then be quantified by polymerase chain reaction (PCR) (Dekker et al. 2002; de Wit and de Laat 2012). Since its development, the original 3C technique has been extended by the introduction of microarrays and next-generation sequencing (NGS) to allow the detection of interactions in increasing detail (de Wit and de Laat 2012). One such extension was developed by Simonis et al. in the form of "Chromosome Conformation Capture-on-Chip" or 4C, which is capable of detecting all interactions of a single genomic locus in an unbiased fashion i.e. "one vs all" (Simonis et al. 2006). This was then further extended by Dostie et al. who developed "Chromosome Conformation Capture Carbon Copy" or 5C, i.e. "many vs many" (Dostie et al. 2006). However, it was not until the advent of Hi-C that interactions could be captured in an unbiased fashion genome-wide i.e. "all vs all" (Lieberman-Aiden et al. 2009) (Figure 1.2).

The Hi-C protocol differs from the 3C protocol by the addition of biotin labelling of the cross-linked fragments before ligation. This allows the ligation products to be purified by biotin pull-down before sequencing (Lieberman-Aiden et al. 2009; de Wit and de Laat 2012). A significant drawback of Hi-C is the enormous sequencing depth required to capture interactions genome-wide. For this reason, there is often a trade-off between the level of sequencing and the resolution at which the Hi-C data can be analysed. The resolution of Hi-C datasets is controlled by binning the data. Reducing the resolution by binning at a larger bin size, reduces the number of possible genomic interactions and therefore increases the signal to noise ratio

(Lajoie, Dekker, and Kaplan 2015). A 10-fold increase in resolution requires a 100-fold increase in sequencing depth (de Wit and de Laat 2012).

Despite the advances enabled by the 3C based methods described above, they have a number of limitations. One of these is that they assess chromatin conformation across a population of cells, meaning that they cannot be used to predict the conformation of chromatin in individual cells (de Wit and de Laat 2012). In order to address this, single cell or single nuclei Hi-C methods have been developed (Nagano et al. 2013; Ramani et al. 2017; Nagano et al. 2017; Flyamer et al. 2017); however, data generated by these methods is inherently sparse. Within the genome it is likely that interactions occur between multiple loci; however, 3C methods are only able to capture interactions in a pairwise fashion. Methods like Tri-C, which are capable of detecting interactions between three or more loci simultaneously, as well as providing single nuclei level information, have been developed to try to address this (Oudelaar et al. 2018). Other limitations of 3C methods include the use of digestion and proximity ligation, which can introduce bias and variation due to restriction enzyme choice, digestion efficiency and specificity, and ligation variability (reviewed in (O'Sullivan et al. 2013)). Split-pool Recognition of Interactions by Tag Extension (SPRITE) was developed to address issues introduced by the proximity ligation step of 3C methods, as well as allowing the detection of multiple interactions simultaneously (Quinodoz et al. 2018). In SPRITE, DNA, RNA and protein are cross-linked, nuclei are isolated and chromatin is fragmented. After fragmentation, interacting molecules within a complex are barcoded (using a split-pool strategy) and sequenced. Interactions are then identified by sequences with the same barcode (Fiorillo et al. 2021; Quinodoz et al. 2018). Finally, Genome Architecture Mapping (GAM) was developed to address many of these issues as well as allowing chromatin compaction and nuclear location to be assayed simultaneously (Beagrie et al. 2017). GAM

assesses the proximity of genetic loci in cryosection slices taken through the nucleus, however, this is likely to be very labour intensive, limiting the number of cells which can be analysed.

Due to the limitations of approaches outlined here, a combination of different approaches, along with microscopy based techniques is likely to be essential to the advancement of the study of chromatin architecture (de Wit and de Laat 2012). However, Hi-C remains a very popular assay, as reflected by the large amount of Hi-C data which is currently in the public domain. Throughout my thesis I have primarily utilised Hi-C data and its derivatives to investigate chromatin architecture.

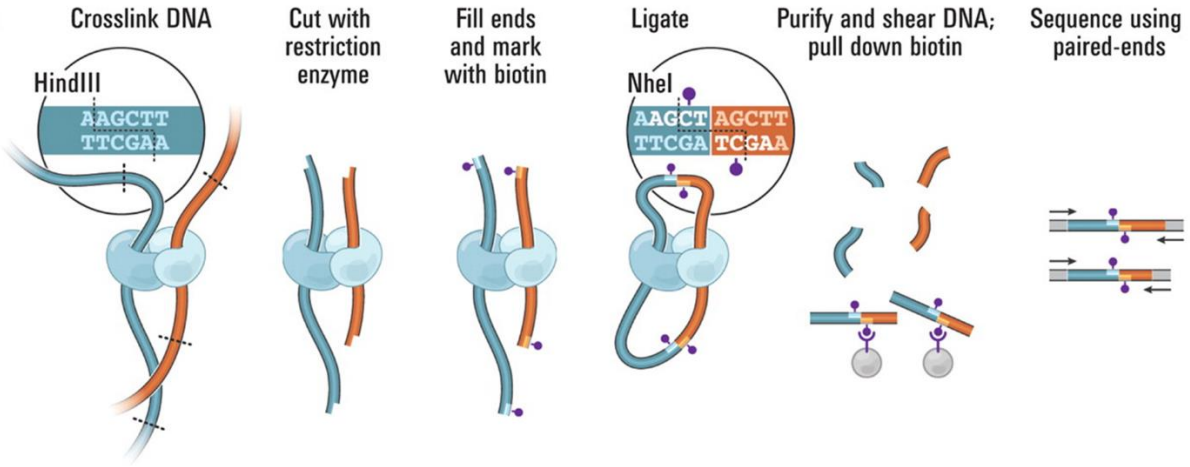


Figure 1.2: Schematic overview of Hi-C. *DNA which is in close physical proximity/interacting is cross-linked together. The DNA is then digested using a restriction enzyme and ends are filled and marked with biotin. The ends are then linked together, sheared and purified by pull down of biotin containing*

[2] Taken from from: Lieberman-Aiden et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science (New York, N.Y.) 326 (5950): 289–93. https://doi.org/10.1126/science.1181369. Reprinted with permission from AAAS.

fragments. These fragments can then be sequenced using paired end sequencing to generate a library in which each read pair maps to interacting regions [2].

1.2.2. Chromosome territories

At the highest level of 3D chromatin organisation, chromosomes occupy their own territories in the nucleus. Chromosome territories were first detected using in situ hybridisation (Cremer, Cremer, Schneider, et al. 1982; Cremer, Cremer, Baumann, et al. 1982; Lichter et al. 1988; Pinkel et al. 1988), but have also since been detected using Hi-C data (Figure 1.4B) (Lieberman-Aiden et al. 2009). The presence of chromosome territories means that interactions within the same chromosome (intra-chromosomal contacts) are more prevalent than interactions between chromosomes (inter-chromosomal contacts) in the 3D architecture of the genome (Lieberman-Aiden et al. 2009; Lajoie, Dekker, and Kaplan 2015). Furthermore, several studies have demonstrated that chromosome territories organise non-randomly within the nucleus. Small, gene rich chromosomes are more likely to co-localise at the centre of the nucleus, whereas gene poor chromosomes are more likely to locate at the nuclear periphery (Lieberman-Aiden et al. 2009; S. Boyle et al. 2001; Tanabe et al. 2002).

1.2.3. Compartments

Using Hi-C data to assess chromatin interactions at the megabase scale within chromosomes, Lieberman-Aiden et al. found that loci separate into two groups, known as compartments, with a higher degree of interaction between regions in the same compartment rather than between compartments (Lieberman-Aiden et al. 2009). The two compartments are referred to as the A and B compartments. They are visible as a plaid pattern in Hi-C maps (Figure 1.4)

and are typically defined by the eigenvector corresponding to the first principle component of the Pearson correlation matrix of the Hi-C map (Lieberman-Aiden et al. 2009; Miura et al. 2018). Regions in the A compartment are enriched for euchromatin, whereas the B compartment is enriched for heterochromatin (Lieberman-Aiden et al. 2009). This enrichment led to the A and B compartments being described as the “active” and “inactive” compartments respectively (reviewed in (Hildebrand and Dekker 2020)).

The spatial segregation of euchromatin and heterochromatin within the nucleus is also supported by microscopy studies. These studies noted that euchromatin is often localised in the centre of the nucleus whereas heterochromatin is often located at the nuclear periphery, and nucleoli (reviewed in (Solovei, Thanisch, and Feodorova 2016)).

A recent study using higher resolution Hi-C data and unsupervised clustering algorithms rather than principal component analysis indicated that the two compartments originally identified can be further divided into 6 sub-compartments (2 in compartment A and 4 in compartment B). Each of these sub-compartments was enriched for a distinct pattern of chromatin marks (Rao et al. 2014). This suggests that the compartmental structure of the genome is more complex than initially recognised. This finding has been corroborated by other studies, one finding evidence for three compartments: a high activity cluster, a low activity centromere proximal cluster and a low activity centromere distal cluster (Yaffe and Tanay 2011). Additionally, another study noted that there are at least 3 separate chromatin compartments in human cells, consisting of one active and two inactive compartments (Nichols and Corces 2021). It is likely that sub-compartments are formed by the clustering of regions with similar epigenetic profiles and therefore may correlate with nuclear regions

identified by microscopy e.g. nuclear speckles, lamin-associating domains etc. (Hildebrand and Dekker 2020).

As expected due to compartments being highly correlated with genome activity, they have been shown to exhibit a large degree of variation between different cell types and tissues. During stem cell differentiation, 36% of the genome was found to switch compartment in at least one of five lineages. Genes which switched from the A compartment to the B compartment during differentiation showed reduced gene expression, whereas genes which switched from the B compartment to the A compartment showed increased expression and were enriched for lineage-specific genes (Dixon et al. 2015). When comparing A and B compartments across 21 primary tissues and cell lines, it was found that 59.6% of the genome was dynamically compartmentalised (Schmitt et al. 2016). It has been hypothesised that compartment changes alter the accessibility of the genome to transcription factors (Dixon et al. 2015).

1.2.4. TADs and sub-TADs

At the kilobase level, chromosomes are further organised into structures known as TADs (Nora et al. 2012; Dixon et al. 2012). TADs are genomic regions which display a high level of self-interaction and a low degree of interaction with regions outside of the TAD, irrespective of distance. The boundaries of TADs are enriched for CTCF binding sites, histone modifications associated with active promoters and gene bodies, housekeeping genes, tRNA genes, TSSs, and short interspersed nuclear elements (SINEs) (Dixon et al. 2012). TADs are also thought to co-localise genes and their regulatory elements, while simultaneously insulating genes from aberrant regulation by regulatory elements beyond the TAD (Dixon, Gorkin, and Ren 2016;

Panigrahi and O'Malley 2021). It is widely thought that TADs promote gene co-regulation and form "gene regulatory domains" (Dixon, Gorkin, and Ren 2016).

Several studies have indicated that TADs are highly tissue invariant (Dixon et al. 2012; Nora et al. 2012; Dixon et al. 2015; Rao et al. 2014; Schmitt et al. 2016), unlike compartments which have been shown to vary extensively in different tissues or cell types (Dixon et al. 2015; Schmitt et al. 2016). Dixon et al. demonstrated by analysing compartment switching in stem cell differentiation that regions of the genome tend to switch compartment in the unit of whole TADs (Dixon et al. 2015).

TADs are visible in Hi-C maps as squares along the diagonal (Figure 1.4) and are commonly annotated algorithmically. There are currently a huge range of published algorithms for the annotation of TADs which vary substantially in the methods used. A recent benchmarking paper by Zufferey et al. classified the methodology employed by 22 TAD callers into the categories: linear Score (e.g. Directionality index, Insulation score, Arrowhead, TopDom); clustering (e.g. ClusterTAD, ICFinder); statistical model (e.g. HiCseg, TADbit); and network features (e.g. 3DNetMod, MrTADFinder) (Zufferey et al. 2018) [3]. However, the reproducibility of the TADs detected by different algorithms is severely limited (Zufferey et al. 2018; Dali and Blanchette 2017; Forcato et al. 2017) (reviewed in (de Wit 2020)). After calling TADs at 10 kb, Zufferey et al. compared the similarity (measure of concordance) of detected TADs between all pairs of 18 TAD callers. By using stochastic neighbour embedding analysis (t-SNE), they found that TAD callers cluster into three groups, with increased TAD similarity within groups compared to between groups (Figure 1.3). Interestingly, TAD callers did not group by

[3] Directionality index (Dixon et al. 2012), Insulation score (Crane et al. 2015), Arrowhead (Durand, Shamim, et al. 2016; Rao et al. 2014), TopDom (Shin et al. 2016), ClusterTAD (Oluwadare and Cheng 2017), ICFinder (Haddad, Vaillant, and Jost 2017), HiCseg (Lévy-Leduc et al. 2014), TADbit (Serra et al. 2017), 3DNetMod (Norton et al. 2018) and MrTADFinder (K. K. Yan, Lou, and Gerstein 2017)

algorithmic method and instead formed groups which also corresponded to the overall number of TADs detected by each TAD caller. The large difference between TAD callers has led to the recommendation that multiple TAD callers should be employed in order to make sure that findings are robust to the choice of TAD caller (Eres and Gilad 2021; Long et al. 2022). The analysis provided by Zufferey et al. is likely to be extremely useful in selecting TAD callers for this purpose; it allows researchers to select TAD callers that capture the largest degree of TAD caller attributable variation (Zufferey et al. 2018) (Figure 1.3).

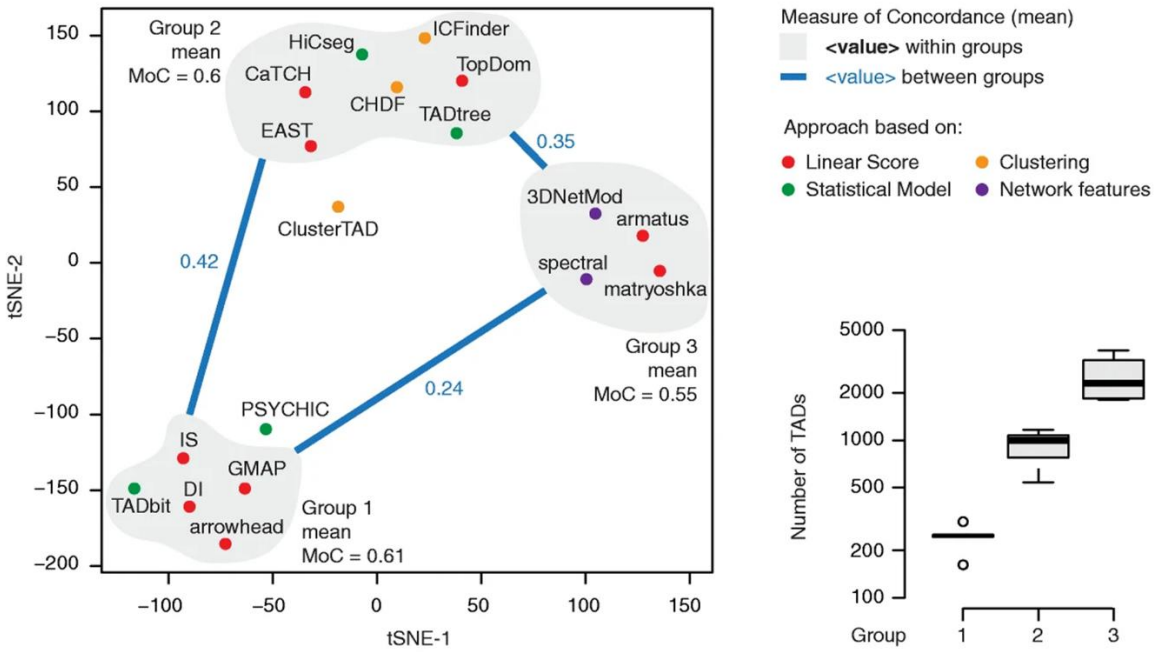


Figure 1.3: Analysis of TAD caller similarity (measure of concordance) from Zufferey et al. *Stochastic neighbour embedding analysis of pairwise between TAD caller measures of concordance reveals three groups of TAD callers, with higher within group similarity than between group similarity (mean measure of concordance between and within groups shown as annotations). Points representing each*

[4] Taken from from: Zufferey, Marie, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. 2018. "Comparison of Computational Methods for the Identification of Topologically Associating Domains." Genome Biology 19 (1): 217. https://doi.org/10.1186/s13059-018-1596-9. *Licensed under CC BY 4.0.*

TAD caller are coloured based on algorithmic method. The box plot shows the distribution of the number of TADs detected in each of the three groups [4].

Since Hi-C data is generated using a large number of cells, the resulting detected contacts represent an ensemble of contacts across all cells profiled. It is therefore possible that whilst TADs are detectable in Hi-C maps and cells on average, they are not commonly present in individual cells. Recent advances have allowed the investigation of TADs at a single cell or single nuclei level by using single cell or single nuclei Hi-C. These investigations have revealed that TADs and other chromatin structures are present on average but the actual structure of chromatin at each locus displays stochasticity within individual cells (Flyamer et al. 2017).

In higher resolution data, a further structure which is smaller than TADs and often nests within them, is commonly referred to as a sub-TAD (Phillips-Cremins et al. 2013). Sub-TADs, while similar to TADs, are characterised as having weaker insulation at their boundaries and are more likely to vary across tissues/cell types (Beagan and Phillips-Cremins 2020). However, further research is required to establish whether these structures are functionally and/or structurally distinct from TADs (reviewed in (Beagan and Phillips-Cremins 2020; Dixon, Gorkin, and Ren 2016)).

1.2.5. Loops

Individual loci on the same chromosome which are far apart in the linear sequence of the genome but are closer together in 3D space than intervening loci are commonly described as loops (Kadauke and Blobel 2009). At their smallest scale, loops are thought to occur between genes and their regulatory elements (discussed further below), as well as occurring between

enhancers, integrating their signal to form super-enhancers (Ing-Simmons et al. 2015; Madsen et al. 2020). Loops can be observed in Hi-C maps as pixels with a higher degree of interaction than surrounding pixels appearing as “dots” of high interaction frequency in the Hi-C map (Figure 1.4), which are commonly referred to in the literature as “peaks” (Rao et al. 2014). These “peaks” are frequently found in the corners of TAD or sub-TAD annotations (also described as “corner dots” in the review (Beagan and Phillips-Cremins 2020)), suggesting that the TAD is formed by a loop and is often described as a “loop domain” (Rao et al. 2014).

Loops can be detected in Hi-C data by using algorithms which detect regions with a higher than expected degree of interaction. Benchmarking of different algorithms for detecting loop interactions revealed a low degree of reproducibility between algorithms (Forcato et al. 2017). Alternatively, interactions involving specific regions of interest can be profiled by adding a capture step to the Hi-C protocol e.g. promoter capture Hi-C (PCHi-C) which is enriched for interactions involving promoters (Schoenfelder et al. 2018) or, less commonly, enhancer capture Hi-C (ECHi-C) which is enriched for interactions involving enhancers (Madsen et al. 2020).

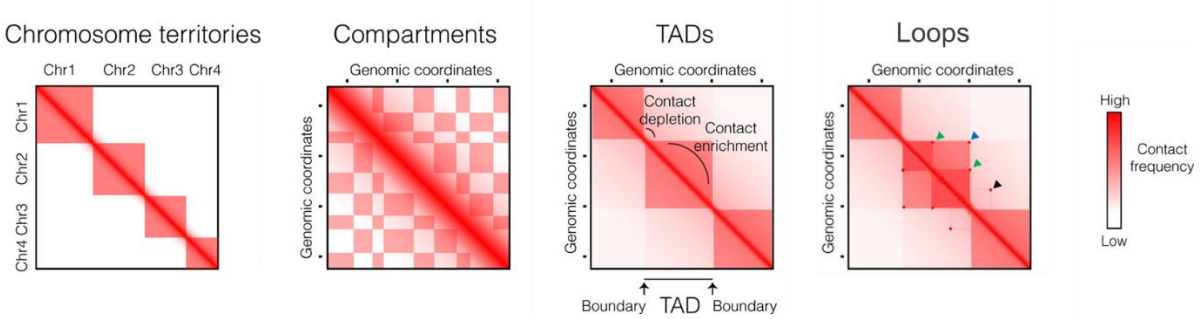


Figure 1.4: Schematic illustrating how chromosome territories, compartments, TADs, sub-TADs and loops appear in Hi-C maps. *Loops are indicated by arrow heads, the blue arrow head indicates a loop at the corner of a TAD forming a "loop domain" and green arrow heads indicate loops at the corner of sub-TADs* [5].

1.2.6. Nucleosome density

At the lowest level of the chromatin hierarchy, DNA is packaged into a chromatin fibre, consisting of nucleosomes which are formed by 146 bp regions of DNA wrapped 1.65x around octamers of core histones (reviewed in (McGinty and Tan 2015)). Euchromatin and heterochromatin differ in the level of compaction of the nucleosomes. Euchromatic regions are loosely packed and are typically enriched for genes and transcriptional activity. On the other hand, heterochromatic regions are densely packed and enriched for transcriptionally inactive genes. Heterochromatin can be further divided into constitutive heterochromatin and facultative heterochromatin. Constitutive heterochromatin forms at the centromere and telomeres of chromosomes as well as at transposons and repetitive sequences. Constitutive heterochromatin stably represses these regions preventing transcription or recombination, whereas facultative heterochromatin forms in protein-coding regions and can be dynamically

[5] Adapted by the addition of the loops panel from: Szabo, Quentin, Frédéric Bantignies, and Giacomo Cavalli. 2019. "Principles of Genome Folding into Topologically Associating Domains." Science Advances 5 (4): eaaw1668. https://doi.org/10.1126/sciadv.aaw1668. © The Authors, some rights reserved; exclusive licensee AAAS. Distributed under a CC BY-NC 4.0 License (http://creativecommons.org/licenses/by-nc/4.0/)

regulated (Murakami 2013; Saksouk, Simboeck, and Déjardin 2015; Morrison and Thakur 2021). As described above, euchromatin and heterochromatin spatially segregate in the nucleus to form compartments.

1.3. Mechanisms of chromatin architecture formation

Despite the seemingly hierarchical organisation of mammalian chromatin architecture, current evidence suggests that they are formed by two antagonistic mechanisms: phase separation and loop extrusion (Nuebler et al. 2018).

1.3.1. Phase separation drives the formation of compartments

Phase separation refers to the mechanism by which block copolymers, made up of alternating monomers (blocks) of different types, fold according to the differing affinities of the block types so that blocks of the same type cluster together in 3D space. In the genome, chromatin in different epigenetic states e.g. euchromatic or heterochromatic, have different affinities for each other and so are thought to phase separate to form the A and B compartments (Figure 1.5A) (Nuebler et al. 2018) (reviewed by (Hildebrand and Dekker 2020; Mirny, Imakaev, and Abdennur 2019)). The homotypic attraction of heterochromatin is well studied and sometimes described as the primary driver of compartment formation by phase separation (Mirny, Imakaev, and Abdennur 2019; Strom et al. 2017). Whilst it is true that the attractive forces between heterochromatic regions tend to create denser, phase-separated regions, attractive forces between euchromatic regions capable of driving phase separation are also thought to exist. Phase separation of euchromatin/the A compartment is likely to be

driven by certain transcriptional regulators, RNA Polymerase II at regions such as super-enhancers, and nuclear speckles (nuclear domains which form a cluster of active genes with splicing factors and mRNA processing proteins) (Sabari et al. 2018; Cho et al. 2018; Feric et al. 2016; Hnisz et al. 2017).

It has also been proposed that phase separation constitutes an alternative mechanism to looping in enhancer-promoter communication. In this model, enhancers are brought into close physical proximity (without direct interaction) by their mutual association with a phase-separated collection of transcription factors (Sabari et al. 2018; Hnisz et al. 2017; Shrinivas et al. 2019).

1.3.2. Loop extrusion is required for the formation of TADs

On the other hand, TADs and loops are thought to be formed by the process of active (ATP dependent) loop extrusion (Sanborn et al. 2015; Fudenberg et al. 2016; Vian et al. 2018). In loop extrusion, a loop extruding factor binds to the chromatin fibre and actively extrudes it as a loop. The loop then increases in size, until the loop extruding factor either releases the chromatin fibre or meets extrusion barriers which stall the process (Figure 1.5A) (Nuebler et al. 2018; Sanborn et al. 2015; Xi and Beer 2021). The process of loop extrusion is also thought to facilitate additional interactions between regions in the loop, explaining the high contact frequency of regions within the same TAD (Haarhuis et al. 2017). The loop extruding factor primarily proposed to be involved in the formation of TADs is cohesin (Schwarzer et al. 2017; Rao et al. 2017). Cohesin is loaded onto the chromatin fibre by the protein NIPBL and unloaded by WAPL (Kueng et al. 2006; Gandhi, Gillespie, and Hirano 2006; Ciosk et al. 2000). Cohesin can extrude chromatin into a loop until it meets the CTCF-bound boundary elements.

CTCF is thought to be the primary extrusion barrier in TAD formation. It has been noted that TAD boundaries are enriched for CTCF binding in a convergent orientation (Rao et al. 2014). This suggests that the ability for CTCF to act as a stable extrusion barrier is directional i.e. it can stall extrusion if the loop extruding factor is traversing towards it in one direction but not the other (reviewed in (Hansen 2020)).

Loop extrusion is thought to be a dynamic process, with loops constantly forming and dissociating (Hansen et al. 2017). Extrusion barriers formed by CTCF have also been shown to exhibit a level of "permeability" i.e. by fluctuating occupancy at binding sites, thus sometimes allowing the skipping of boundaries (Hansen et al. 2017). These features of loop extrusion can explain the frequently observed nesting of TADs and sub-TADs in Hi-C maps (which represent an average across cells) without these structures being hierarchically present in individual cells (Figure 1.5B) (Mirny, Imakaev, and Abdennur 2019; Hansen 2020).

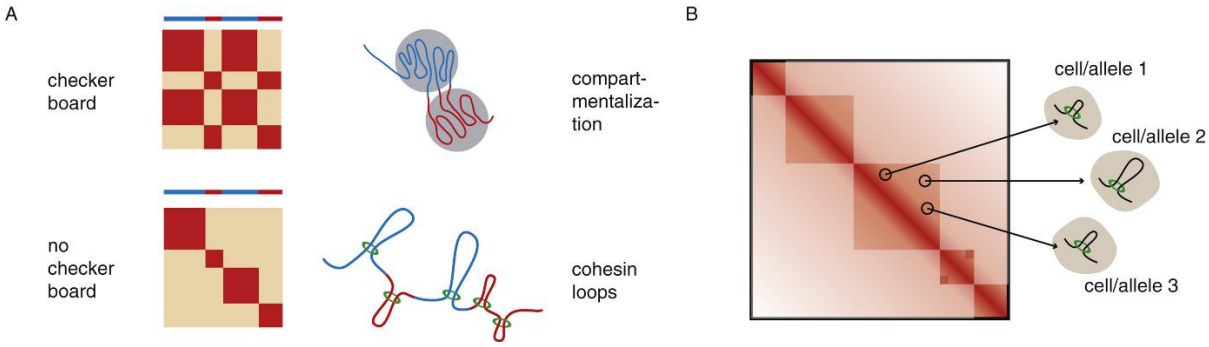


Figure 1.5: Phase separation and loop extrusion*. A) Compartments are thought to be formed by phase separation. The appearance of compartments in the Hi-C map is shown next to a schematic of how they appear within the nucleus. Phase-separated regions are shown as grey circles. Compartments A*

[6] Taken from: Wit, Elzo de. 2020. "TADs as the Caller Calls Them." Journal of Molecular Biology 432 (3): 638–42. https://doi.org/10.1016/J.JMB.2019.09.026, with permission from Elsevier.

and B are coloured red or blue respectively. TADs are thought to form by loop extrusion involving cohesin. The appearance of cohesin loops in the Hi-C map is shown next to a schematic of how these structures form in the nucleus. Cohesin is represented by green rings. B) Hi-C maps show an average of contacts across a population of cells. However, in reality cohesin loops are thought to be highly dynamic and can adopt different conformations between cells and between alleles, as indicated by the schematic [6].

1.3.3. The interplay between both mechanisms

In order to investigate the interplay between these two mechanisms, several studies have assessed chromatin architecture after the depletion of cohesin. This has typically been achieved by either degrading subunits of cohesin (e.g. RAD21 and SCC1), by knockout of the cohesin loading factor NIPBL, or by knockout of MAU2 which is a cofactor of NIPBL (Rao et al. 2017; Haarhuis et al. 2017; Gassler et al. 2017; Wutz et al. 2017; Schwarzer et al. 2017). These experiments found that in the absence of normal cohesin function, TADs were lost but compartment strength increased. Furthermore, a finer compartment structure which matched the epigenetic state of the chromatin was observed (Schwarzer et al. 2017). This suggests an antagonistic relationship in which loop extrusion limits compartment formation. Further evidence supporting this model comes from knockout experiments of WAPL, the cohesin unloading factor (Haarhuis et al. 2017; Gassler et al. 2017; Wutz et al. 2017). Increased cohesin stability after WAPL knockout/depletion was associated with increased loop length and reduced compartmentalisation (Haarhuis et al. 2017; Gassler et al. 2017; Wutz et al. 2017) (reviewed in (Mirny, Imakaev, and Abdennur 2019; Hildebrand and Dekker 2020)).

As described above, TADs are commonly algorithmically defined. However, if looked at in the context of the mechanisms by which they were formed, algorithmically defined TADs likely

represent a combination of domains formed by loop extrusion, compartment domains, simultaneous loop extrusion and compartment domains, and domains formed by other mechanisms (Beagan and Phillips-Cremins 2020). There is therefore an increasing effort to separate detected "TADs" according to the likely mechanisms which formed them (suggested in (Beagan and Phillips-Cremins 2020; de Wit 2020)).

1.4. The relationship between chromatin architecture and function

Since the identification of the 3D chromatin structures outlined above, an important research objective has been to investigate how these structures relate to function, particularly gene expression, replication timing and disease.

1.4.1. Regulation of gene expression

As introduced above, compartmentalisation is highly correlated with euchromatin and heterochromatin. Therefore, it seems plausible that compartments are related to gene transcriptional activity (Lieberman-Aiden et al. 2009). However, whilst Dixon et al. noted that 36% of the genome switched compartment during embryonic stem cell (ESC) differentiation to at least one of five lineages, they found that genes which switched compartment only displayed subtle changes in expression (Dixon et al. 2015). In a complementary study, B cells were reprogrammed to create induced pluripotent stem cells. Of the genes which were transcriptionally altered during reprogramming, those which also switched compartment were more substantially upregulated or downregulated than those which stably remained in the A compartment (for B to A, or A to B switches respectively) (Stadhouders et al. 2018).

However, only 10% of “stably upregulated genes” switched compartment from B to A and 18.6% of these genes stably remained in the B compartment (Stadhouders et al. 2018). These results suggest that compartment switching does not universally alter the expression of affected genes to a large degree, and transcriptional changes do not always co-occur with compartment changes. Therefore, it is possible that it is the epigenetic state of the surrounding chromatin, rather than transcription per se that is associated with compartmentalisation (Oudelaar and Higgs 2020).

It has been proposed that TADs are the functional units of the genome, important in the regulation of gene expression. As mentioned above, TADs co-localise enhancers and their target genes whilst simultaneously insulating against regulatory interactions beyond the TAD boundaries (Dixon, Gorkin, and Ren 2016). This feature of TADs has led to the speculation that they may play a key role in promoting co-regulation of the genes within the same TAD. Key evidence supporting this came from a study by Symmons et al. in which regulatory sensors were inserted along the mouse genome to assess the action of regulatory elements. They found that the genome could be partitioned into “regulatory domains” which closely match with TADs (Symmons et al. 2014). This suggests that regulatory elements may be able to regulate all genes within the same TAD. Other complementary studies have noted high expression correlation between genes which share a TAD (Nora et al. 2012; Flavahan et al. 2016; Tarbier et al. 2020). Together these studies suggest that gene regulation in TADs may be non-specific, i.e. a regulatory element has the potential to regulate any gene within the same TAD (Dixon, Gorkin, and Ren 2016). Furthermore, Dixon et al. found that during ESC differentiation, the areas of the genome which switch compartment are typically made up of whole TADs (Dixon et al. 2015). This is likely to be due to the ability of TAD boundaries to block the spread of heterochromatin and euchromatin (Dixon, Gorkin, and Ren 2016).

Given the evidence supporting the role of TADs in facilitating correct gene expression, there have been several surprising results from studies which have sought to globally disrupt TAD structure. In one such study, the impact of chromosomal rearrangements on transcription in *Drosophila* balancer chromosomes (spanning ~75% of the genome) was assessed. Although these rearrangements altered TAD structure, only 512 genes were differentially expressed as a result (Ghavi-Helm et al. 2019). However, it is important to note that it is possible that structures annotated as TADs in *Drosophila* may actually be representative of compartments and that structures formed by loop extrusion are not present in *Drosophila* (discussed in more detail below and in the reviews (Hildebrand and Dekker 2020; Beagan and Phillips-Cremins 2020)). Furthermore, balancer chromosomes have been selected to maintain viability which may have inadvertently selected for reduced transcriptional disruption (Beagan and Phillips-Cremins 2020; Ghavi-Helm et al. 2019). As noted above, there have been several studies in which depletion of cohesin has led to the loss of TADs and loops. Interestingly, these studies also reported only modest transcriptional changes upon the loss of TADs (Rao et al. 2017; Schwarzer et al. 2017). Similarly, a study assessing the impact of structural variants in cancer found that only 14% of boundary deletions lead to transcriptional changes of greater than 2-fold (Akdemir et al. 2020). These results challenge the hypothesis that TADs are important in correct gene regulation. However, many of these studies only assessed the effect of cohesin depletion on maintaining transcriptional programs during steady state transcription. Further studies have focused on the effects of cohesin depletion during transcriptional transitions such as differentiation or response to a stimulus and found a much larger effect (reviewed in (Popay and Dixon 2022)). In one such example, Calderon et al. depleted the cohesin subunit RAD21 during neuronal maturation and found that neuronal gene expression was depleted and neuronal maturation was impaired. They found that genes transcribed in response to

neuronal activity were pervasively affected (Calderon et al. 2022). In a similar study, Cuartero et al. looked at the response of macrophages to microbial signals after cohesin depletion. They found that cohesin was required for the proper transcriptional response of macrophages including in inflammatory gene expression (Cuartero et al. 2018). These results suggest that cohesin (loop extrusion and TAD formation) is more important in establishing new transcriptional programs than maintaining them at steady state. One possibility is that it is the dynamic nature of cohesin-mediated loop extrusion, characterised by constant dissociation and reformation of loops, which is critical to establishing new transcriptional programmes (discussed in (Popay and Dixon 2022)). This is supported by the finding that the depletion of the cohesin unloading factor WAPL (which stabilises cohesin binding) affects gene expression in a similar manner to RAD21 depletion (N. Q. Liu et al. 2020).

How enhancers communicate with promoters to impact the transcription of their target genes, sometimes over long distances, has been the subject of much research, and several proposed mechanisms exist. Historically, these mechanisms have included tracking, linking, looping and conformation (Figure 1.6). In the enhancer tracking model, the enhancer recruits transcription factors that track along the DNA until the target gene promoter is reached (Moreau et al. 1981). Whereas, in the enhancer linking model, the enhancer recruits a transcription factor that then recruits other proteins, forming a chain that ultimately reaches the target promoter (Bulger and Groudine 1999). In the conformation model, changes in chromatin conformation, e.g. the spread of de-compaction initiating at the enhancer, could be passed from the enhancer to the target promoter (Dynan and Tjian 1985). However, the mechanism that presently enjoys the greatest amount of support is the looping model. In this model, proteins and transcription factors bound at the enhancer and promoter physically interact, via the intervening sequence forming a loop (Dynan and Tjian 1985) (reviewed in

(Panigrahi and O'Malley 2021; Furlong and Levine 2018; Popay and Dixon 2022)). It has been shown that forcing a loop is enough to induce gene expression at the globin locus (Deng et al. 2014, 2012). It has also been demonstrated that enhancer-promoter looping varies across differentiation and is concomitant/correlated with transcription (Bonev et al. 2017; Freire-Pritchett et al. 2017). However, enhancer-promoter loops have also been detected before transcriptional activation (Bonev et al. 2017). This suggests that in some enhancer-promoter loops, looping is not sufficient to initiate transcription and further processes are required such as binding of transcription factors (Vermunt, Zhang, and Blobel 2018). Alternatively, these loops may be related to paused transcription (Ghavi-Helm et al. 2014). Despite a large amount of support for the looping model, recent evidence has challenged its ubiquity. For example, microscopy and 5C evidence from the Shh locus during neural differentiation suggests that the neural enhancers actually move further away from the promoter upon gene activation, which is seemingly incompatible with a simple looping model (Benabdallah et al. 2019). Although it is believed that looping interactions play an important role in enhancer-promoter communication, it is likely that other mechanisms capable of facilitating enhancer-promoter communication also exist. Other possible mechanisms, which have been more recently proposed, include phase separation, enhancer control of chromatin compaction (similar to the earlier suggested conformation model) or by enhancer facilitation of the association of the target promoter with nuclear bodies/landmarks e.g. nuclear lamina, transcription factories or nuclear speckles (discussed in (Popay and Dixon 2022)).

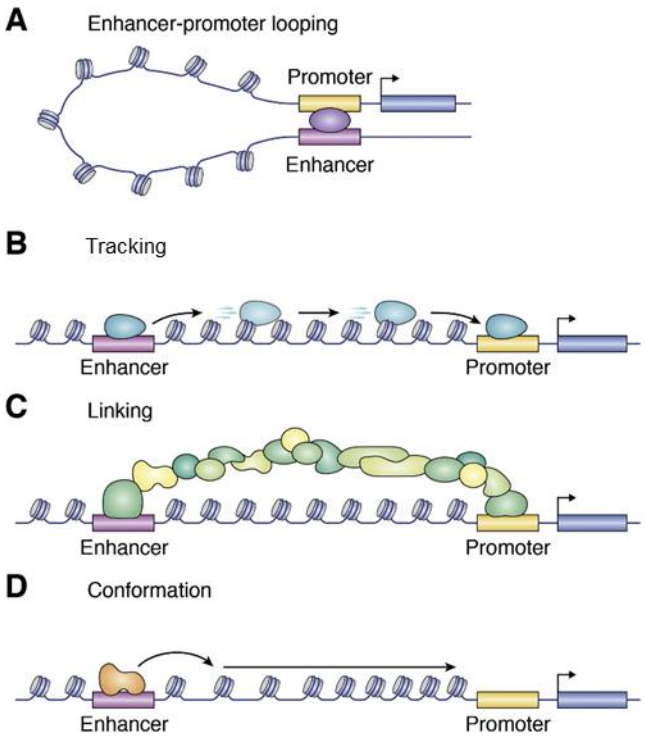


Figure 1.6: Models of enhancer-promoter communication. *A) Looping: proteins bound at the enhancer and target gene promoter physically interact forming a loop. B) Tracking: A protein bound at the enhancer tracks along the DNA until it reaches the target gene promoter. C) Linking: A protein bound at the enhancer recruits further proteins, forming a chain which eventually binds to the target gene promoter. D) Chromatin conformational changes initiating at the promoter are propagated along the DNA conveying a signal to the target promoter* [7].

[7] Adapted by a minor alteration of panel labels from: Popay, Tessa M., and Jesse R. Dixon. 2022. "Coming Full Circle: On the Origin and Evolution of the Looping Model for Enhancer–Promoter Communication." Journal of Biological Chemistry 298 (8): 102117. https://doi.org/10.1016/J.JBC.2022.102117. *Licensed under CC BY 4.0.*

1.4.2. Replication timing

Eukaryotic genomes are replicated during the S-phase of the cell cycle, ahead of mitosis. Replication takes place by the unwinding of DNA and initiation of replication at various DNA replication origins. Replication origins control the replication of individual replicons and in eukaryotic genomes clusters of replicons tend to activate simultaneously. These “replicon clusters” or “replication domains” have been shown to form stable units which remain replicated together and are activated in a conserved temporal pattern across multiple cell divisions (Jackson and Pombo 1998; Ma et al. 1998). Conservation has also been detected across evolution, for example Yaffe et al. found high conservation of replication timing between the mouse and human genomes (Yaffe et al. 2010) Whilst replication timing is stable across cell divisions it has been shown to vary between cell types (Hiratani et al. 2008; Rivera-Mulia et al. 2015).

Interestingly early and late replication domains preferentially occupy different areas of the cell nucleus, with enrichment of early replicating regions at the nuclear interior and enrichment for late replicating regions at the nuclear periphery (O’Keefe, Henderson, and Spector 1992). This bares similarity to the observed localisation of chromatin compartments suggesting a potential overlap between replication domains and 3D chromatin architecture. Studies have indicated a high correspondence between the eigenvector of Hi-C maps and replication timing (Ryba et al. 2010). Furthermore, of the 6 sub-compartments identified by Rao et al. both A sub-compartments were associated with early replication times, whilst one B sub-compartment was associated with replication in the middle of the S-phase and the other two B sub-compartments with replication at the end of the S-phase (the 6^{th} sub-compartment was only identified on chromosome 19) (Rao et al. 2014).

There has also been a reported overlap between replication domains and TADs. Pope et al. mapped the positions of “early timing transition regions”, which are regions falling between early and late replicating regions where replication timing transitions, to replication domains in mouse and human cell types. They found an “almost one-to-one” correlation between these replication domain boundaries and TAD boundaries. They argue that this suggests that TADs are stable regulatory units of replication timing (Pope et al. 2014). However, similarly to the unclear relationship between TADs and gene expression, studies deleting/depleting either cohesin (Oldach and Nieduszynski 2019) or CTCF (Sima et al. 2019) found no effect on replication timing.

1.4.3. Importance of chromatin architecture in disease

1.4.3.1. TAD disruption

Despite the inconclusive results from global disruptions of TADs, there are many examples of TAD boundary disruptions at specific loci leading to aberrant gene regulation and disease. Lupiáñez et al. found that structural variants in the WNT6/IHH/EPHA4/PAX3 locus caused distinct human limb malformations. They found that inversions, deletions and duplications misplaced a group of limb enhancers relative to a TAD boundary, allowing the formation of aberrant regulatory interactions (which would normally be blocked by the TAD boundary) (Lupiáñez et al. 2015). Similarly, a deletion in the LMNB1 locus which removes a TAD boundary results in aberrant regulation of LMNB1 causing adult-onset demyelinating leukodystrophy (Giorgio et al. 2015). Another example comes from the deletion of the CTCF and cohesin

binding sites at a TAD boundary at the mouse alpha globin gene cluster which caused the upregulation of three genes. This included *Rhbdf1*, a gene which is normally repressed by polycomb in erythroid cells, and was upregulated ~600-fold (Hanssen et al. 2017). These examples highlight the role of TAD boundaries in blocking the formation of aberrant regulatory interactions. Conversely, deletion of CTCF binding sites at the *Shh* locus leads to the loss of a preformed regulatory interaction and a 50% decrease in *Shh* expression (but no observed phenotype) (Paliou et al. 2019), showing the role of TADs in maintaining the interaction at this locus. These results demonstrate that even though the gene expression effects detected by some studies of global TAD disruption may be small, the consequences in specific loci can be severe, highlighting the importance of context in TAD disruption (reviewed in (Tena and Santos-Pereira 2021)).

1.4.3.2. Variant to function

Genome-wide association studies (GWAS) detect associations between genetic variants and phenotypes/traits. GWAS assesses the difference in allele frequency between individuals with similar ancestry but who are phenotypically different (the process of GWAS is outlined in (Uffelmann et al. 2021)). Currently, over 200,000 variants have been linked to a human complex trait via GWAS. However, understanding the mechanisms underlying these associations i.e. moving from "variant to function", has proved extremely challenging and the majority of associations are still not understood. This is largely because the majority of associated variants fall within the non-coding genome where they are likely to play a role in gene regulation but it is unclear which genes they impact (Cano-Gamez and Trynka 2020; Lappalainen and MacArthur 2021). If, as suspected, gene regulation commonly occurs via

looping between regulatory elements and their target genes, then profiling chromatin architecture will prove invaluable in moving from variant to function. Furthermore, PCHi-C and ECHi-C datasets which are enriched for chromatin contacts involving promoters or enhancers respectively are likely to be most useful (Schoenfelder et al. 2018; Madsen et al. 2020). By assessing the functional enrichments of mapped genes, Whalen et al. found that using chromatin contacts from PCHi-C to map noncoding variants to potential target genes outperforms linkage disequilibrium (LD) or selecting the nearest gene (Whalen and Pollard 2019). Similarly, Lu et al. found that Hi-C loops outperform expression quantitative trait loci (eQTLs) in identifying the target genes of neurological GWAS associations (Lu et al. 2020). Several studies have already begun using chromatin architecture information in the prioritisation of the candidate target genes of GWAS-associated variants e.g. (Javierre et al. 2016; Montefiori et al. 2018; Dryden et al. 2014; Pan et al. 2018), demonstrating its utility.

1.5. Chromatin architecture across evolution

The chromatin architecture of eukaryotes has been studied in many species. Structures, similar to TADs, which I will here term “contact domains” have been identified in a wide range of species beyond mammals e.g. *Drosophila, S. pombe*, *S. cerevisiae*, *C. elegans* and *A. thaliana* (reviewed in (Rowley and Corces 2016)). However, with the exception of *Drosophila*, CTCF is not present in these organisms (Heger, Marin, and Schierenberg 2009). While *Drosophila* does have a homolog of CTCF, evidence has suggested that unlike mammalian TADs, *Drosophila* contact domain boundaries are not strongly enriched for CTCF or Cohesin binding (Ulianov et al. 2016). Instead, *Drosophila* contact domains are highly correlated with the chromatin epigenetic state (Ulianov et al. 2016; Sexton et al. 2012). This has led to

speculation that the contact domains identified in these species may be representative of compartment domains rather than TADs. This would suggest that compartmentalisation is the primary force in chromatin organisation present in all eukaryotes, with TADs and loop extrusion representing a further mechanism of organisation present in at least mammals (suggested in the review (Rowley et al. 2017)).

Several studies have indicated that TADs are highly conserved across mammalian species (Dixon et al. 2012; Rao et al. 2014; Vietri Rudan et al. 2015). This could indicate that there is a selective pressure against TAD disruption during evolution. This assertion is supported by a study by Krefting et al. which looked at the positions of evolutionary rearrangement breakpoints (calculated across 13 vertebrate species, including human and mouse) in relation to human TAD boundaries. They found that evolutionary rearrangement breakpoints are depleted within TADs but enriched at TAD boundaries (Krefting, Andrade-Navarro, and Ibn-Salem 2018). Similarly, Lazar et al. identified an enrichment of TAD boundaries at gibbon-human breaks in synteny (Lazar et al. 2018). Further to this, Gilbertson et al. found that genes with species-specific expression in immune cells are often located near to TADs spanning human-mouse evolutionary breakpoints (Gilbertson et al. 2022), suggesting that TADs disruption contributes to species divergence. However, the relationship between TADs and genomic breakpoints is not clear-cut. Berthelot et al. identified genomic breakpoints in human, mouse, dog, cow and horse (identified by reconstructing the ancestral Boreoeutheria genome) and found that simulated breakpoints are predicted to displace TAD boundaries less frequently than is observed in real data, suggesting that there is no selective pressure against the disruption of TAD boundaries by genomic rearrangement (Berthelot et al. 2015).

Interesting insights have been generated by analysing the relationship between TADs and LD in human populations. Whalen et al. found that the boundaries of LD blocks are generally depleted at TAD boundaries. This would suggest a model in which TADs are frequently broken by recombination. Assuming that TADs constitute gene regulatory domains, this would allow natural selection to act independently on regulatory variants and their target genes (Whalen and Pollard 2019).

1.6. Thesis aims

This thesis aims to further investigate the relationship between chromatin architecture, particularly TADs and loops, and gene regulation. I aim to investigate the following questions:

1. Is there any evidence for tissues/cell type specificity of mammalian TADs and TAD boundaries? (Chapter 2).
2. How similar are TADs and TAD boundaries between human and mouse when considered across multiple tissues/cell types? (Chapter 2).
3. If TADs form “gene regulatory domains”, how similar are genes which share a TAD, in terms of paralogy, gene constraint, expression correlation, GO semantic similarity, shared pathway annotations and protein-protein interactions (PPI)? (Chapter 3).
4. Can using chromatin looping data for variant to function analysis mechanistically explain the fat distribution/obesity GWAS association of variants in the TBX15-WARS2 locus? (Chapter 4).
5. Can the comparison of human and mouse chromatin looping data during adipogenesis help to prioritise GWAS-associated loci with a high conservation of regulatory circuitry for modelling using mouse? (Chapter 4).

Chapter 2: Conservation of TADs between tissues and species

2.1. Introduction

Understanding the mechanism by which the same set of genetic instructions, existing in almost all cells, can give rise to an array of different cell types is of critical importance in biology. This cell type diversity is created by a complex process of cell differentiation in which gene expression regulation is controlled by an intricate set of regulatory elements. Regulatory elements are thought to impact the transcription of their target genes by coming into close physical proximity with the target gene promoter (Mora et al. 2015; Panigrahi and O'Malley 2021; Matharu and Ahituv 2015; Bonev and Cavalli 2016; Oudelaar and Higgs 2020). This close physical proximity is believed to be facilitated by TADs which co-localise regulatory elements and their target gene promoters within the same TAD. In addition, TADs are thought to ensure proper gene regulation by preventing aberrant interactions between genes and regulatory elements beyond the TAD boundaries (Dixon, Gorkin, and Ren 2016).

It is commonly claimed in the literature that TADs are largely tissue (Dixon et al. 2012; Nora et al. 2012; Dixon et al. 2015; Rao et al. 2014; Schmitt et al. 2016) and species (Dixon et al. 2012; Rao et al. 2014; Vietri Rudan et al. 2015) invariant. The evidence for this comes from preliminary comparisons of TADs or their boundaries across datasets from different tissues or species. Studies of the conservation of whole TADs vs TAD boundaries offer different but complementary information when considering the conservation of TADs. It is possible for both boundaries of a TAD to be conserved but for the TAD itself to have diverged, i.e. by the

gain of a new boundary between the two original boundaries which splits the TAD in two, or by alternative paring of conserved boundaries (Rao et al. 2014).

In one of the first papers describing TADs, Dixon et al. compared TAD boundaries detected in mouse ESC, mouse cortex, human ESC, and human IMR90 by comparing: mouse ESC with mouse cortex, human ESC with human IMR90, and mouse ESC with human ESC. In this study, TADs were identified by the directionality index, which quantifies upstream and downstream interaction bias at each genomic bin. TADs were then determined from the directionality index using a Hidden Markov Model (HMM). To assess which boundaries were shared between two cell types within the same species, the directionality index value from the centre of the boundary ±10 bins with a 40kb bin size was correlated between cell types; cell type-specific boundaries were then defined as boundaries only identified in one cell type that lacked significant correlation in the other. On the other hand, species-specific boundaries were defined as boundaries that did not overlap a boundary at the equivalent location in the other species. For both human and mouse samples, they found that the majority of boundaries were shared across cell types/tissues, suggesting they are cell type/tissue invariant. After comparing ESC boundaries between human and mouse, they found that 53.8% of human and 75.9% of mouse boundaries were detected in the other species, which is significantly more than would be expected at random, suggesting they are also highly conserved across species (Dixon et al. 2012). Building on this work, Rao et al. compared TADs detected using the Arrowhead TAD calling algorithm in human GM12878 B-lymphoblastoid cells with TADs detected in the human IMR90, HMEC, NHEK, K562, KBM7, HUVEC and HeLa cell lines. The conservation of TADs was determined by comparing the Euclidean distance between the TAD corner pixels. On average they found that 54% of TADs detected in one of the cell lines were also present in GM12878. They also found that 45% of TADs detected in

human GM12878 were also present in the mouse CH12-LX B- lymphoblast cell line (Rao et al. 2014). These results have been used to argue that there is substantial conservation of TADs and TAD boundaries between tissues/cell types and species. However, the assessment of tissue/cell type variance in these studies are limited to a small number of different tissues/cell types, and the assessment of inter-species variance is limited to comparisons between human and mouse.

A more comprehensive evaluation of the variance in TAD boundaries across tissues comes from a comparison of TADs detected by the insulation score method across 21 human tissues/cell types. This study found that almost 40% of TAD boundaries were detected in all 21 tissues/cell types, which is significantly more than would be expected at random. However, as boundary conservation was assessed after merging consecutive TAD boundaries into 200 kb “boundary regions”, the similarity reported is potentially inflated by the reduced resolution (Schmitt et al. 2016)

There are several major limitations of the studies which have compared TADs and TAD boundaries between different cell types, tissues and species. The largest of these relates to the generalisability of findings between studies given the fact that the compared TADs and TAD boundaries have been detected using different TAD callers. Several benchmarking studies have highlighted the extremely poor reproducibility of TADs detected by different TAD callers (Dali and Blanchette 2017; Forcato et al. 2017; Zufferey et al. 2018). TAD callers have also been shown to have varying sensitivities to sequencing depth and their own parameters. Since it is common for the datasets included in TAD comparison studies to have a large variation in sequencing depth, the sensitivity of the TAD caller to sequencing depth is an important consideration when interpreting the results. Finally, as highlighted above, TAD

comparison studies vary in the stringency of the comparison method used. It is therefore not always clear how robust findings are to the choice of TAD caller or the stringency of the TAD comparison method used.

There are also several lines of indirect evidence for TAD conservation between species (reviewed in (Eres and Gilad 2021)). Vietri Rudan et al. found that CTCF binding sites that were conserved between mouse, dog and macaque were enriched at large-scale domain borders. On the other hand, CTCF binding sites which were not conserved between species were enriched within large-scale domains. CTCF binding at TAD boundaries is thought to be involved in TAD formation; therefore, the conservation of CTCF binding sites implies that TADs themselves may also be conserved. To support this claim, they compared the correlation of Hi-C map interactions at various distances in dog, rabbit or macaque with mouse, finding a Spearman's rank correlation of 0.34-0.61 (Vietri Rudan et al. 2015; Eres and Gilad 2021). Another study found that genomic rearrangement breakpoints across 13 vertebrate species were depleted within human TADs and enriched at TAD boundaries, suggesting a tendency for TADs to remain undisrupted over evolutionary timescales (Krefting, Andrade-Navarro, and Ibn-Salem 2018). In a similar study, Lazar et al. found that breaks in synteny between humans and gibbons are enriched at TAD boundaries (Lazar et al. 2018). This suggests that there is selective pressure against the disruption of TADs in evolution. However, not all studies agree that there is a strong association between the positions of TAD boundaries and evolutionary breakpoints e.g. (Berthelot et al. 2015).

While these results point to the fact that TADs and TAD boundaries are likely to be more similar between tissues and species than would be expected if they were distributed randomly, there is still a large percentage of TADs which do vary (Sauerwald, Singhal, and

Kingsford 2020; Schmitt et al. 2016). This could be due to noise or other technical considerations. However, despite common claims to the contrary, it is possible that the TADs which vary are actually representative of tissue/species specificity in TADs.

Due to the conflicting evidence presented above, to assess tissue or species specificity in TADs, I analysed and detected TADs in 106 publicly available Hi-C samples. After merging replicates, this corresponded to 13 human and 20 mouse Hi-C maps representing eight different tissues/cell groups. These datasets were then used to compare TAD and TAD boundaries within and between tissue/cell type groups and species. As detailed above, one of the problems with robustly assessing TAD conservation is the low concordance of TADs detected with different TAD callers (Zufferey et al. 2018; Forcato et al. 2017; Dali and Blanchette 2017). To mitigate against this, I detected TADs using two popular TAD calling methods: Arrowhead (Durand, Shamim, et al. 2016; Rao et al. 2014) and TopDom (Shin et al. 2016). Arrowhead and TopDom were amongst the top three TAD callers for detecting TADs enriched for the binding of CTCF and Cohesin subunits at their boundaries. Since CTCF and Cohesin are thought to be involved in TAD formation, this indicates that Arrowhead and TopDom are amongst the best TAD callers at annotating biologically relevant TADs. Further to this, Zufferey et al. found that if TAD callers are clustered by TAD similarity, they form three groups with greater similarity within groups than between groups (Introduction: Figure 1.3). Arrowhead and TopDom fall into separate groups, so together capture a large degree of TAD caller attributable variation. In the third TAD caller group, no TAD caller ever ranked among the top scoring TADs based on any assessment criteria (Zufferey et al. 2018). Therefore, I propose that using both Arrowhead and TopDom together captures a large degree of TAD caller attributable variation in the best performing TAD callers and will allow me to assess whether findings are robust to the choice of TAD caller. I also carried out an independent

assessment of the robustness of both Arrowhead and TopDom to bin size, parameter choice and number of Hi-C contacts. Throughout this chapter, TADs is used to refer to autosomal TADs only.

2.1.1. Aims

The overall aims of this chapter are to:

- Assess the sensitivity of TAD calling with Arrowhead and TopDom to bin size, parameter choice and number of Hi-C contacts.
- Evaluate the similarity between TAD boundaries and TADs detected with Arrowhead or TopDom.
- Using human and mouse publicly available Hi-C data, assess the tissue specificity of TADs.
- Determine the level of TAD conservation between human and mouse.

2.2. Methods

2.2.1. Hi-C datasets

Publicly available Hi-C data was downloaded from the Gene Expression Omnibus (GEO) (Edgar, Domrachev, and Lash 2002; Barrett et al. 2012), the National Centre for Biotechnology Information Sequence Read Archive (NCBI SRA) (Sayers et al. 2022) or personal correspondence with the authors (Table 2.1). The selected data included Hi-C samples from: 4 human and 4 mouse ESC (Dixon et al. 2012, 2015; Freire-Pritchett et al. 2017; Y. Zhang et al. 2019; Giorgetti et al. 2016; J. Yan et al. 2018; Bonev et al. 2017); 1 human and 4 mouse neural progenitor cells (NPC) (Dixon et al. 2015; Fraser et al. 2015; Giorgetti et al. 2016; Bonev et al. 2017); 1 human and 1 mouse b-lymphoblast/lymphoblastoid cells (Rao et al. 2014); 1 human and 3 mouse livers (including mouse liver samples collected at Zeitgeber time (ZT) 10 and 22) (Leung et al. 2015; Vietri Rudan et al. 2015; Y. H. Kim et al. 2018); 1 human left ventricle (Leung et al. 2015), 1 human right ventricle (Schmitt et al. 2016), 1 human foetal heart (Bertero et al. 2019), 1 human ES derived ventricular cardiomyocytes (Y. Zhang et al. 2019); 1 human dorsolateral prefrontal cortex (Schmitt et al. 2016); 1 human mesenchymal stem cells (MSC) (Madsen et al. 2020); 1 mouse cortex (Shen et al. 2012); 1 mouse cortical neurons (Bonev et al. 2017); 5 mouse pre-adipocytes/adipocytes/adipogenesis (M. He et al. 2018; Siersbæk et al. 2017); and 1 mouse cardiac myocytes (Rosa-Garrido et al. 2017).

Table 2.1: Human and mouse selected Hi-C datasets. *Names used to refer to the dataset throughout the chapter is given in the "Dataset" column.*

Human datasets				
Dataset	**GEO or NCBI ID**	**Enzyme**	**Sample IDs**	**Reference**
ESC (1)	GEO:GSE35156	HindIII	SRX116344, SRX128221	Dixon et al. 2012
ESC (2)	GEO:GSE52457	HindIII	SRX378271, SRX378272	Dixon et al. 2015
NPC	GEO:GSE52457	HindIII	SRX378277, SRX378278	Dixon et al. 2015
Left ventricle	GEO:GSE58752	HindIII	SRX641266	Leung et al. 2015
Liver	GEO:GSE58752	HindIII	SRX641267	Leung et al. 2015
Lymphoblastoid cells (B-lymphoblastoid)	GEO:GSE63525	MboI	SRX764936, SRX764937, SRX764938, SRX764939, SRX764940, SRX764941, SRX764942, SRX764943, SRX764944, SRX764945, SRX764946, SRX764947, SRX764948, SRX764949, SRX764950, SRX764951, SRX764952, SRX764953, SRX764954, SRX764955, SRX764956, SRX764957, SRX764958, SRX764959, SRX764960, SRX764961, SRX764962,	Rao et al. 2014

			SRX764963, SRX764964	
ESC (3)	GEO:GSE86821	HindIII	SRX2159278	Freire-Pritchett et al. 2017
Dorsolateral prefrontal cortex	GEO:GSE87112	HindIII	SRX2179249	Schmitt et al. 2016
Right ventricle	GEO:GSE87112	HindIII	SRX2179261	Schmitt et al. 2016
Foetal heart	GEO:GSE106687	DNase	SRX3375355, SRX3375356	Bertero et al. 2019
ES-derived ventricular cardiomyocyte	GEO:GSE116862	MboI	SRX4377025, SRX4377026	Zhang et al. 2019
ESC (4)	GEO:GSE116862	MboI	SRX4377015, SRX4377016	Zhang et al. 2019
MSC	GEO:GSE140782	MboI	SRX7807124, SRX7807125	Madsen et al. 2020
Mouse datasets				
Dataset	**GEO or NCBI ID**	**Enzyme**	**Sample IDs**	**Reference**
Cortex	GEO:GSE34587	HindIII	SRX150196, SRX150197	Shen et al. 2012
ESC (1)	GEO:GSE35156	HindIII	SRX116341, SRX116342	Dixon et al. 2012
NPC (1)	GEO:GSE59027	HindIII	SRX644422	Fraser et al. 2015
NPC (2)	GEO:GSE59027	NcoI	SRX644423	Fraser et al. 2015
B-lymphoblasts	GEO:GSE63525	MboI	SRX765019, SRX765020, SRX765021, SRX765022, SRX765023, SRX765024, SRX765025, SRX765026, SRX765027, SRX765028,	Rao et al. 2014

			SRX765029, SRX765030, SRX765031, SRX765032, SRX765033	
Liver	GEO:GSE65126	HindIII	SRX851861, SRX851862	Vietri Rudan et al. 2015
ESC (2)	GEO:GSE72697	HindIII	SRX1182727	Giorgetti et al. 2016
NPC (3)	GEO:GSE72697	HindIII	SRX1182728	Giorgetti et al. 2016
ESC (3)	GEO:GSE74055	MboI	SRX1338543	Yan et al. 2018
3T3-L1 preadipocytes	NCBI:SRA: SRP082116	HindIII	SRX2021709	He et al. 2018
3T3-Y adipocytes	NCBI:SRA: SRP082116	HindIII	SRX2021712	He et al. 2018
Pre-adipocytes differentiating towards adipocytes, day 0	GEO:GSE95533	HindIII	SRX2599831, SRX2599832	Siersbæk et al. 2017
Pre-adipocytes differentiating towards adipocytes, 4 hours	GEO:GSE95533	HindIII	SRX2599833, SRX2599834	Siersbæk et al. 2017
Pre-adipocytes differentiating towards adipocytes, day 2	GEO:GSE95533	HindIII	SRX2599835, SRX2599836	Siersbæk et al. 2017
ESC (4)	GEO:GSE96107	DpnII	SRX2636666, SRX2636667, SRX2636668, SRX2636669	Bonev et al. 2017
NPC (4)	GEO:GSE96107	DpnII	SRX2636670, SRX2636671, SRX2636672, SRX2636673	Bonev et al. 2017

Cortical neuron	GEO:GSE96107	DpnII	SRX2636674, SRX2636675, SRX2636676, SRX2636677, SRX2636678, SRX2636679	Bonev et al. 2017
Cardiac myocytes	GEO:GSE96692	MboI	3 replicates provided by personal correspondence with the authors	Rosa-Garrido et al. 2017
Liver ZT10	GEO:GSE104129	MboI	SRX3205265, SRX3205266	Y. H. Kim et al. 2018
Liver ZT22	GEO:GSE104129	MboI	SRX3205260, SRX3205261, SRX3205262, SRX3205263, SRX3205264	Y. H. Kim et al. 2018

2.2.2. Processing of Hi-C datasets

Hi-C datasets were aligned against either the mm10 or the hg19 genome builds (excluding alternative haplotypes) and processed using the UGER version of the Juicer pipeline which was adjusted to run on Rescomp with the help of Adam Huffman and the Juicer developers (version 1.5, juicer tools version 1.7.6) (Durand, Shamim, et al. 2016). All datasets were analysed using the restriction site file corresponding to the restriction enzyme used to generate the data (Table 2.1). For data generated using DNase, no restriction site file was specified. Individual replicates were initially processed separately. To do this, Juicer was run with Juicer tools version 1.7.6. Mouse Hi-C data was processed using the option –g “mm10” and human Hi-C data was processed using custom genome build options (a custom hg19 genome build was used so that alternative haplotypes could be excluded). A split size of

90,000,000 was used for most datasets; however, for some smaller datasets a split size of 45,000,000 was used to increase run speed. Individual replicates were then merged using the mega function. This was done using juicer tools version 1.7.6 for mouse and version 1.22.01 for human (except human foetal heart where version 1.7.6 was used) using the --skip-intra-frag flag in the Juicer "pre" step to invoke the behaviour of the earlier version. The different Juicer tools version used for human was necessary due to a Juicer pipeline bug in recognising the restriction enzyme for the custom human genome build which was fixed in the later version. However, the later version was not used for the human foetal heart sample as it was not affected by the bug (it is DNase Hi-C so no restriction site file was specified in Juicer). Paired reads were removed if either end mapped to the genome with a mapping quality (MAPQ) score of less than 30. Contact matrices were normalised using the Vanilla Coverage (VC) method (Lieberman-Aiden et al. 2009). VC normalisation was used because it was compatible with both TopDom and Arrowhead TAD calling.

2.2.3. TAD calling

TADs were detected using Arrowhead and TopDom (Rao et al. 2014; Durand, Shamim, et al. 2016; Shin et al. 2016). Arrowhead TADs were detected on .hic format Hi-C maps using the ignore sparsity flag. For TopDom, dense matrices were extracted for each chromosome using the "dump" function in juicer tools; bin coordinates were then added before detection of TADs. In order to assess the impact of different bin sizes and TAD calling parameters on detected TADs, TADs were detected using Hi-C matrices binned at 10 kb, 25 kb and 50 kb using settings: m= 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000 for Arrowhead and w= 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 for TopDom. For the main analysis, Hi-C

data was binned at 50 kb and TADs were detected using the default settings for both TAD callers (m=2000 for Arrowhead TADs (Rao et al. 2014; Durand, Shamim, et al. 2016) and w=5 for TopDom TADs (Shin et al. 2016)). TADs called on the sex chromosomes were filtered in all analyses. TopDom detects “domains”, “boundaries”, and “gaps”. Similarly to Dali et al. (Dali and Blanchette 2017) I only considered “domain” annotations as TADs called by TopDom.

2.2.4. Random down-sampling of Hi-C contacts

In order to assess the sensitivity of Arrowhead and TopDom TAD calling to the number of Hi-C contacts in the Hi-C map, I synthetically down-sampled reads in a large Hi-C map. To do this I selected the Hi-C map corresponding to the mouse cortical neuron dataset (Bonev et al. 2017), as of the datasets included in this chapter, it has the largest number of Hi-C contacts (4,759,064,782 Hi-C contacts). The individual replicates in the cortical neuron dataset were analysed using Juicer, then replicates were merged using “mega” as described above. The “merged_nodups.txt” (a file containing the aligned paired Hi-C contacts, alignment positions and other information such as MAPQ scores) was filtered to remove intra-fragment reads and reads pairs where either end had a MAPQ < 30 so that only the 4,759,064,782 valid Hi-C contacts remained. The filtered merged_nodups.txt file was then split into 10 equal sized chunks. If after the creation of the chunks there were Hi-C contacts remaining, each individual Hi-C contact was allocated randomly to an existing chunk. The desired random Hi-C contact sample size was then reached by randomly sampling 1/10 of the total sample size from each chunk. The resulting random down-sampled merge_nodups.txt files contained 50 million, 100 million, 200 million, 300 million, 400 million, 500 million, 600 million, 700 million, 800 million, 900 million, 1 billion, 2 billion, 3 billion and 4 billion Hi-C contacts respectively. These files

were then used to create .hic files using the “pre” function in Juicer and then used to detect TADs using Arrowhead and TopDom as described above.

2.2.5. TAD chromosome coverage

The percentage TAD coverage of each autosomal chromosome was calculated. As TopDom TADs are non-overlapping, the chromosome coverage can be calculated as the number of bases in each chromosome which are covered by a TAD annotation, divided by the total number of bases in the chromosome. TADs detected by Arrowhead can overlap and/or nest. Therefore TADs which overlapped by greater than one base were merged and chromosome coverage was calculated as outlined above.

2.2.6. TAD and TAD boundary comparison method

In order to compare TADs and TAD boundaries between datasets, I calculated the number of shared boundaries or TADs, as well as calculating a similarity score based on the Jaccard Index. To compare TADs/boundaries between datasets A and B, TADs/boundaries were deemed to be shared if their coordinate/s were exactly the same between A and B. The similarity score was then calculated by dividing the number of shared TADs/boundaries between A and B by the total number of unique TADs/boundaries in either A or B (the intersection divided by the union) (Figure 2.1). This generates a score between 0 and 1, in which a higher score indicates a greater degree of similarity between A and B.

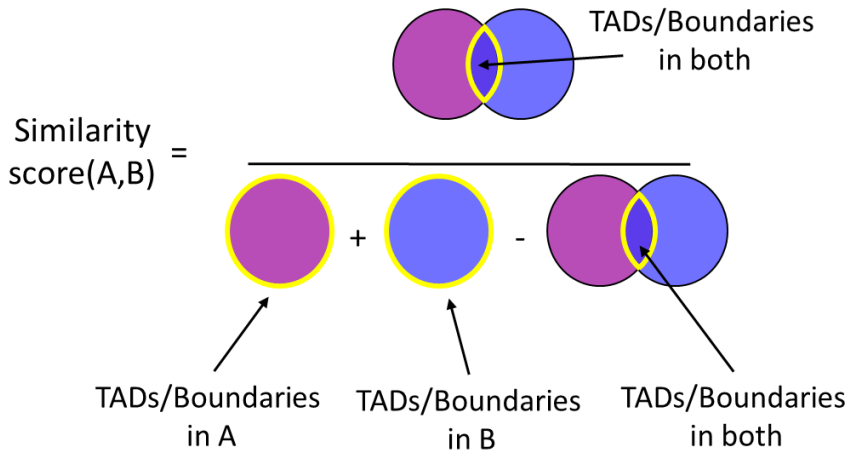


Figure 2.1: TAD and boundary similarity score. *Schematic illustrating the strategy used to calculate the similarity score (based on the Jaccard Index).*

This method was used to compare TADs/boundaries between Arrowhead and TopDom, and between datasets within and between species. In order to assess whether the detected TADs/boundaries are more similar between datasets from similar cell types/tissues, I grouped datasets into several tissue/cell type categories (henceforth referred to as 'tissue groups' for simplicity). The groups were: ESC, NPC, brain, lymphoblast, liver, adipose, and heart (Table 2.2). The scores associated with pairs of datasets from the same tissue group were then compared with the scores associated with pairs of datasets from different groups. Statistical significance was determined using the Wilcox test in R. This approach is similar to that used in Sauerwald et al. (Sauerwald, Singhal, and Kingsford 2020). In order to assess whether this analysis is affected by possible increased similarity between Hi-C datasets produced by the same lab, I assessed which pairs of datasets were published in the same data repository or published as part of papers with the same first or last authors (defined as the last author or the last corresponding authors) (Chapter 2: Supplementary Table 1). I then reran all analysis

both after excluding scores associated with pairs of datasets from the same data repository, and after excluding scores associated with pairs of datasets with at least one common first or last author.

Table 2.2: Table outlining the datasets included in each tissue group.

Tissue group	Human datasets	Mouse datasets
ESC	ESC (1), ESC (2), ESC (3), ESC (4)	ESC (1), ESC (2), ESC (3), ESC (4)
NPC	NPC	NPC (1), NPC (2), NPC (3), NPC (4)
Brain	Dorsolateral prefrontal cortex	Cortex, cortical neuron
Lymphoblast	Lymphoblastoid cells	B-lymphoblasts
Liver	Liver	Liver, liver ZT10, liver ZT22
Adipose	-	3T3-L1 preadipocytes, 3T3-Y adipocytes, pre-adipocytes differentiating towards adipocytes (day 0), pre-adipocytes differentiating towards adipocytes (4 hours), pre-adipocytes differentiating towards adipocytes (day 2)
Heart	ES derived ventricular cardiomyocyte, left ventricle, right ventricle, foetal heart	Cardiac myocytes
MSC	MSC	

2.2.7. Comparing TADs and TAD boundaries between species

To compare TADs and TAD boundaries between species, mouse coordinates (mm10) were converted to hg19 using the liftOver command line tool, with -minMatch=0.1 (Hinrichs et al. 2006). The hg19 start and end coordinates of the mouse TADs were then rounded to the nearest 50 kb bin coordinate. This was done using the formula: 50,000*rounded(coordinate/50,000) within python 3.6.2. Rounding was done using the 'decimal' package using a "rounding half to even" strategy. Therefore, coordinates which are equidistant between two 50 kb bin coordinates were rounded up or down using a "rounding half to even" strategy on coordinate/50,000. For example, the coordinate 25,000 would be rounded down to 0 (25,000/50,000=0.5 which rounds to a nearest even number of 0), whereas the coordinate 75,000 would be rounded up to 100,000 (75,000/50,000=1.5 which rounds to a nearest even number of 2). Since in this approach any mouse TADs which cannot be mapped on to hg19 are removed, human TADs which could not be successfully mapped on to mm10 were also filtered. The hg19 coordinates of human and mouse TADs were then compared as outlined above.

2.2.8. Per TAD tissue and species conservation score

Using mouse TADs converted to hg19 and rounded to the nearest 50 kb bin and human TADs after filtering those which cannot be converted to mm10 (see above), I assessed the tissue and species conservation of individual TADs. To do this, TADs from each dataset were merged and duplicated TADs were removed, creating a master set of TADs. Next, each TAD was annotated according to whether it was present in any dataset belonging to each tissue group. The total number of tissue group annotations in human or mouse was calculated for each

TAD. TADs were then further annotated as conserved between human and mouse if the TAD was present in at least one dataset in both species, regardless of tissue group. The significance of the difference between the numbers of tissue groups in TADs which are conserved in the other species vs not conserved in the other species was determined by permutation testing using the “perm” R package.

2.3. Results

2.3.1. Assessment of resolution and quality of selected Hi-C data

In order to compare TADs between tissues and species, I analysed 106 individual publicly available Hi-C samples, corresponding to 13 human and 20 mouse Hi-C maps (after merging replicates) (Methods: 2.2.1: Table 2.1). The selected datasets included a variety of tissues/cell types in both human and mouse. There was large variation in the number of sequenced reads (range 186,498,461- 8,722,587,979) and the number of Hi-C contacts per dataset (range 65,719,274- 4,759,064,782) (Figure 2.2A). This suggests that there is likely to be large variation in the dataset quality and the minimum bin size at which they can reasonably be analysed. In order to investigate this further, I calculated the *cis*/*trans* ratio to assess the noise level of each dataset. The *cis*/*trans* ratio refers to the ratio of Hi-C contacts that occur within a chromosome (*cis* contacts) to those that occur between chromosomes (*trans* contacts). High quality Hi-C experiments are expected to detect more *cis* contacts than *trans* contacts because *cis* contacts are more prevalent in the genome due to chromosome territories. Typically a ratio of ~0.67-1.5 would indicate high quality data, however, experimental noise will affect *cis* and *trans* contacts equally, resulting in a lower *cis*/*trans* ratio (Lajoie, Dekker, and Kaplan 2015). I found large variation in *cis*/*trans* ratios across datasets (range ~0.44-~12.6) (Figure 2.2B). Several datasets which suffered from a low number of sequenced reads had the highest *cis*/*trans* ratios (human MSC; mouse ESC (3); and mouse pre-adipocytes differentiating towards adipocytes at day 0, 4 hours and day 2). This suggests that whilst the level of sequencing is low in these datasets, the sequenced contacts have a high signal to noise ratio. In order to select a bin size which is appropriate for the analysis of all datasets I calculated the map resolution of each dataset. Map resolution is defined as the minimum bin

size at which 80% of bins have at least 1000 contacts and is commonly used to select the minimum bin size at which a Hi-C dataset can sensibly be analysed (Rao et al. 2014). Detected map resolutions ranged from 0.75 kb to 37.10 kb; this therefore suggests that a bin size greater than 37.10 kb would be appropriate for the analysis of all datasets (Figure 2.2C).

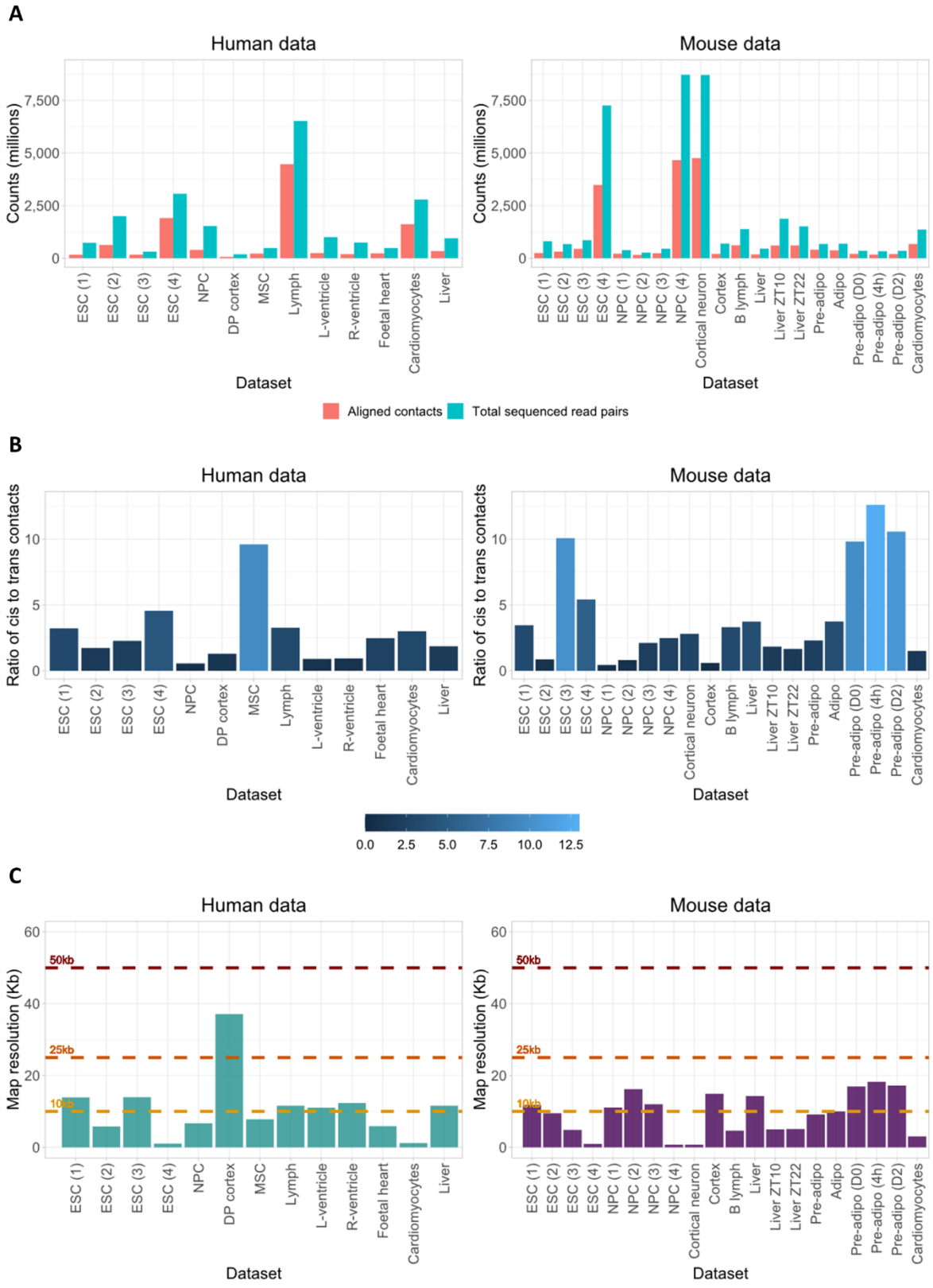


Figure 2.2: Hi-C dataset resolution. *A) The number of sequenced read pairs and the number of successfully aligned/valid Hi-C contacts per dataset. B) Ratio of cis (within the same chromosome)/trans (between chromosomes) contacts per dataset. C) The map resolution, defined as*

the minimum bin size at which 80% of bins have at least 1000 contacts (commonly used to select the minimum bin size at which a Hi-C dataset can sensibly be analysed), of each dataset. Dotted lines indicate map resolutions of 10 kb, 25 kb and 50 kb respectively.

2.3.2. The effect of bin size, TAD calling parameters and the number of Hi-C contacts on TADs detected by Arrowhead and TopDom

It has been shown that TAD callers vary in their sensitivity to bin size, their own parameters and the number of available Hi-C contacts. However, these analyses are normally restricted to a single or a smaller number of datasets than included in this chapter (Zufferey et al. 2018; Forcato et al. 2017; Dali and Blanchette 2017). I therefore sought to independently investigate the sensitivity of the number and size of TADs detected by Arrowhead and TopDom to different bin size, parameters and numbers of Hi-C contacts across all selected Hi-C datasets.

2.3.2.1. Bin size

To assess the impact of bin size on TADs, I detected TADs (using default parameters) in each dataset using a bin size of 10 kb, 25 kb and 50 kb respectively and plotted a distribution of the number and average size of Arrowhead and TopDom TADs at each bin size, for both species (Figure 2.3). For Arrowhead TADs, on average the lowest number of TADs was detected at 10 kb for both human and mouse (median number of TADs 181 and 817.5 for human and mouse respectively). This could be due to the fact that 10 kb is an inappropriately low bin size for several of the selected datasets (Figure 2.2C). The highest number of Arrowhead TADs were detected at 25 kb where a median of 1629 and 2010.5 were detected, compared to 1500 and 1414 at 50 kb for human and mouse respectively. For TopDom TADs, the number of detected

TADs decreased as bin size increased for both mouse and human (Figure 2.3A). For both TAD callers, the average size of detected TADs increased as bin size increased (Figure 2.3B). This suggests that for both TAD callers there is a tendency for the number of TADs detected to decrease and their average size to increase as bin size increases, with the exception of Arrowhead 10 kb TADs where the fewest TADs were detected. This anomaly could be because Arrowhead detects fewer TADs when data is analysed at an inappropriately low bin size for the number of Hi-C contacts. Overall, Arrowhead detected fewer but on average larger TADs than TopDom at all bin sizes.

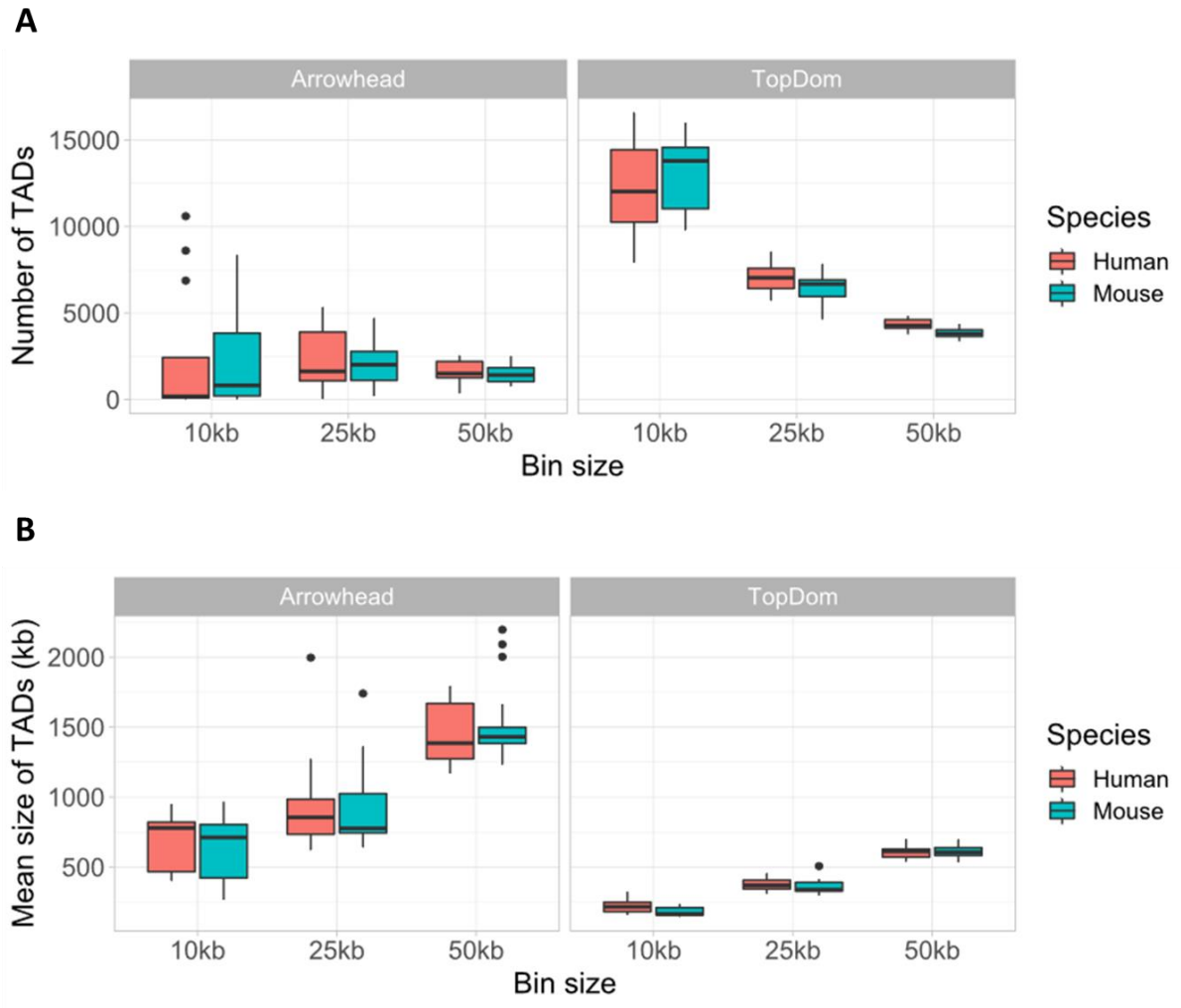


Figure 2.3: Number and size of autosomal TADs at 10 kb, 25 kb or 50 kb bin size. *A) The number of TADs detected with Arrowhead or TopDom in all selected human and mouse datasets at 10 kb, 25 kb and 50 kb split by Human and Mouse. B) The mean size of TADs detected with Arrowhead or TopDom in all selected human and mouse datasets at 10 kb, 25 kb and 50 kb split by Human and Mouse.*

2.3.2.2. Adjustable TAD calling parameters

Arrowhead and TopDom each have one adjustable parameter. Therefore, I sought to assess the sensitivity of TADs called by Arrowhead and TopDom to their respective parameters. In Arrowhead the parameter is denoted as "m" and described as the "size of the sliding window along the diagonal in which contact domains will be found". The default value of "m" is 2000

(https://github.com/aidenlab/juicer/wiki/Arrowhead). In order to assess the impact of “m” on Arrowhead TADs, I detected TADs at a 50 kb bin size with m=500-7000. For Arrowhead TADs the value of “m” had very little impact on either the number or average size of the detected TADs, indicating that detected TADs are robust to the parameter choice.

In TopDom the parameter is denoted as “w” or “window size” and is “used to compute binSignal”. Bin signal is a score which measures the average contact frequency of pairs of chromatin regions surrounding each bin. TADs are detected by TopDom as the minima of bin signal, because the average contact frequency of a bin within a TAD is expected to be greater than a bin at a TAD boundary. The default value of “w” is 5, which was optimised by the original TopDom paper for detection of TADs at 40 kb bin size (Shin et al. 2016). In order to assess the impact of “w” on TopDom TADs, I detected TADs at a 50 kb bin size w=3-20. For TopDom TADs, a larger value of “w” corresponded to the detection of fewer but on average larger TADs (Figure 2.4).

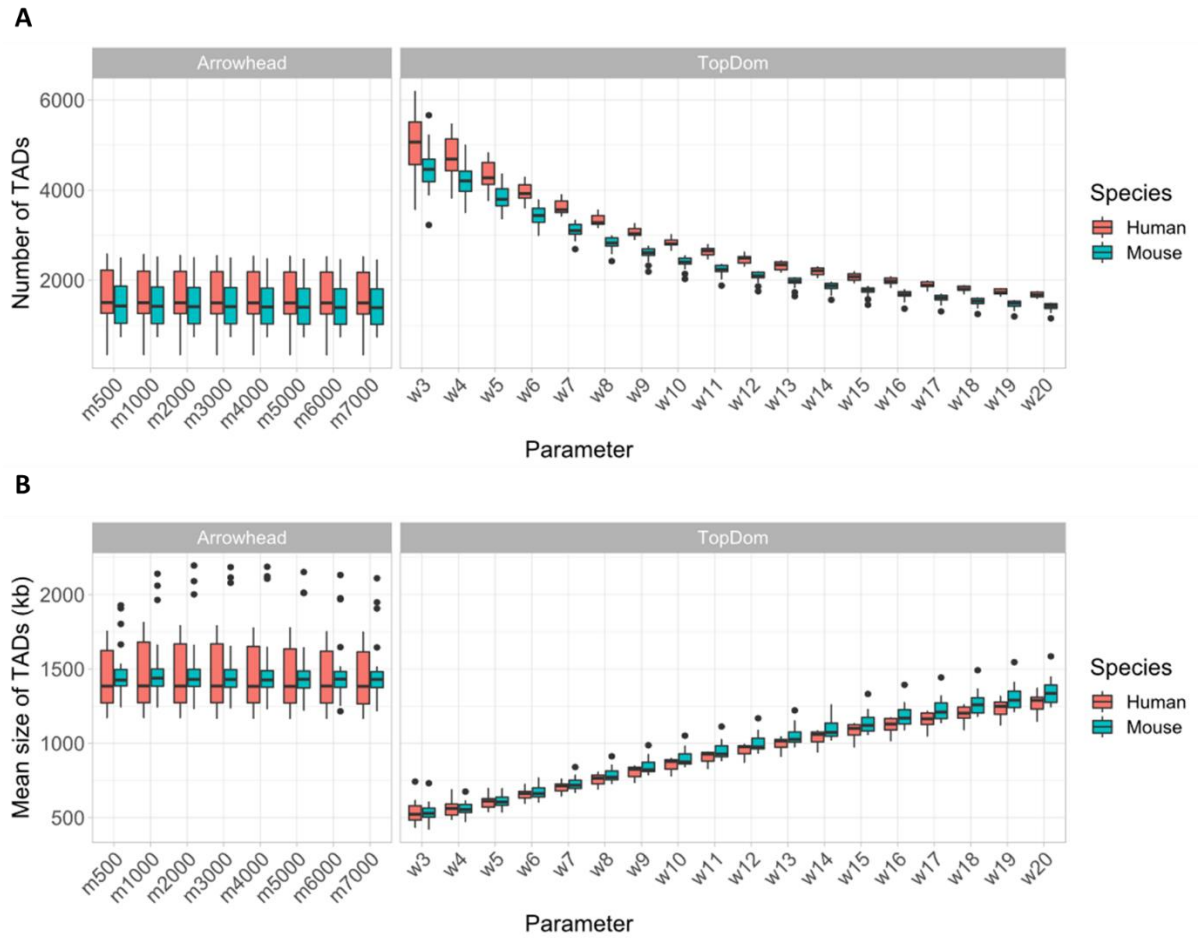


Figure 2.4: Number and size of autosomal TADs at different parameter values for Arrowhead and TopDom. *The Arrowhead parameter m is described as the "size of the sliding window along the diagonal in which contact domains will be found". The TopDom parameter w or window size, is "used to compute binSignal" in TopDom. Arrowhead parameters m=500-7000 and TopDom parameter w=3-20 were tested. A) The number of TADs detected with Arrowhead or TopDom at 50 kb with each parameter. B) The mean size of TADs detected with Arrowhead or TopDom at 50 kb with each parameter.*

2.3.2.3. Number of Hi-C contacts

Reducing the number of reads in the Hi-C map reduces the power to detect TADs and TAD callers vary in their sensitivity to this. To assess the impact of the number of Hi-C contacts on the detected TADs, I selected the dataset with the largest number of Hi-C contacts (mouse cortical neuron (Bonev et al. 2017)) and down-sampled the valid Hi-C contacts to between 50 million-4 billion Hi-C contacts (see Methods: 2.2.4). To assess how this impacted the minimum resolution at which the dataset can be sensibly analysed, I calculated the map resolution of each down-sampled dataset. Down-sampled datasets with more than 300 million reads all had a map resolution of 10 kb (Figure 2.5A). I then detected TADs at 50 kb bin size using both Arrowhead and TopDom. Arrowhead detected fewer but larger TADs than TopDom irrespective of the number of Hi-C contacts in the Hi-C map. The number and size of TopDom TADs was relatively constant across all down-sampled datasets, whereas for Arrowhead TADs, both the number and size of detected TADs increased as the number of Hi-C contacts increased (beginning to plateau at around 400-500 million Hi-C contacts) (Figure 2.5B-C). This suggests that TADs detected with TopDom are more robust to the number of Hi-C contacts in the Hi-C map than those detected with Arrowhead. This result highlights the utility of making sure conclusions are robust between TADs called with Arrowhead and TopDom.

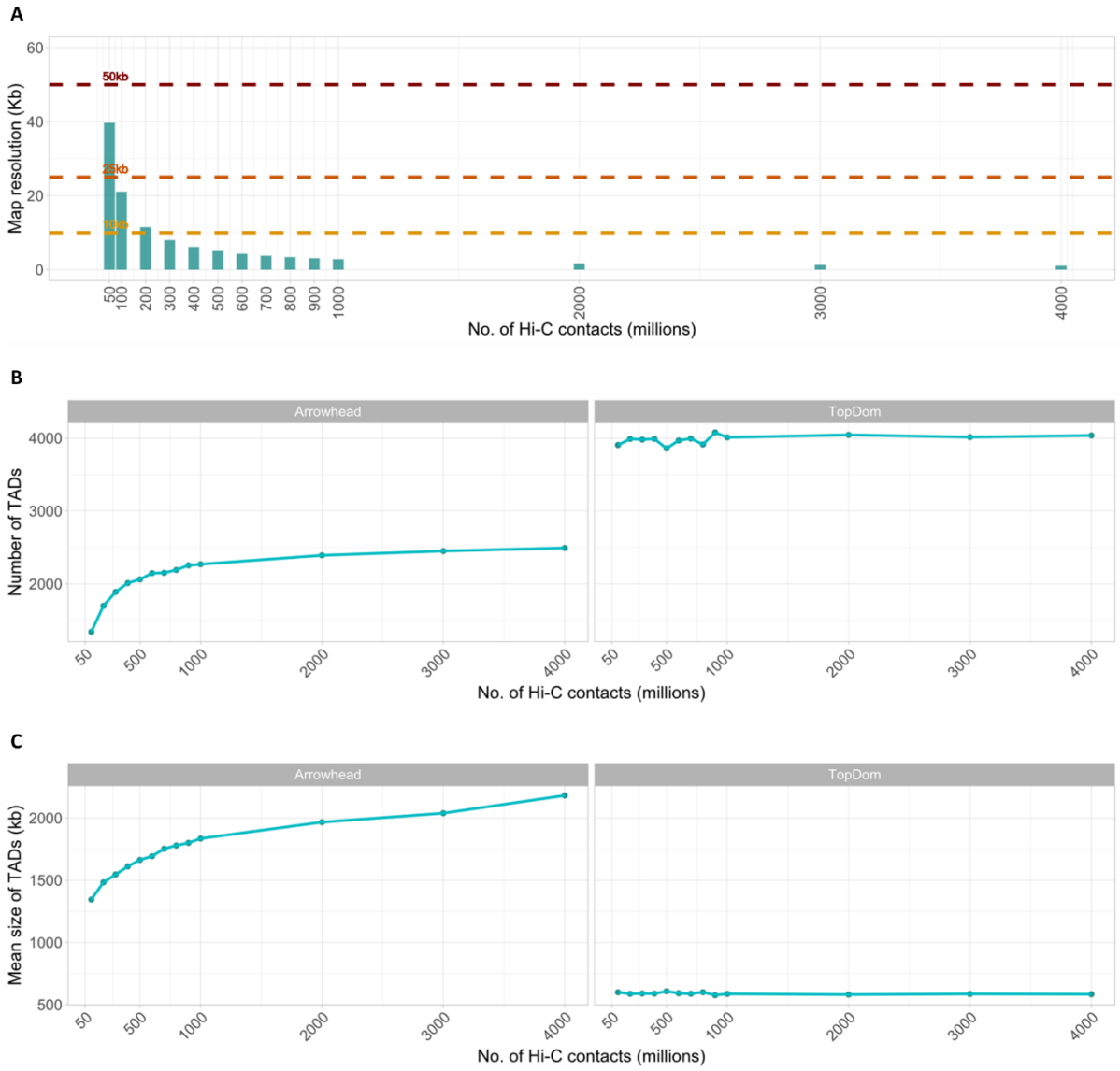


Figure 2.5: Map resolution, number and size of autosomal TADs detected after randomly down-sampling Hi-C contacts. *A) The map resolution, defined as the minimum bin size at which 80% of bins have at least 1000 contacts (commonly used to select the minimum bin size at which a Hi-C dataset can sensibly be analysed), of datasets after randomly down sampling Hi-C contacts. Dotted lines indicate map resolutions of 10 kb, 25 kb and 50 kb respectively. B) The number of TADs detected with Arrowhead or TopDom at 50 kb after randomly down sampling Hi-C contacts. C) The mean size of TADs detected with Arrowhead or TopDom at 50 kb after randomly down sampling Hi-C contacts.*

2.3.3. Comparison of TADs called by Arrowhead and TopDom in each dataset

TADs were detected in each dataset at 50 kb bin size, which was selected as it was conservatively appropriate for all datasets based on map resolution (Figure 2.2C). Default parameters were used for both Arrowhead and TopDom. This was done because, as demonstrated above, the selected Arrowhead parameter has little impact on TAD detection and the TopDom default parameter was originally optimised for a bin size close to 50 kb (Shin et al. 2016). I plotted the number of TADs detected in each dataset against the number of Hi-C contacts in each dataset. Consistent with results from the down-sampling analysis, I found that there was little relationship between the number of Hi-C contacts and the number of detected TopDom TADs. However, datasets with a greater number of Hi-C contacts also had a greater number of detected Arrowhead TADs (plateauing after ~500 million Hi-C contacts) (Chapter 2: Supplementary Figure 1). This supports my earlier finding that, unlike TopDom, Arrowhead is very sensitive to the number of Hi-C contacts, particularly in maps with less than 500 million reads.

When comparing the number of TADs or TAD boundaries detected by each TAD caller, I found a larger number of TopDom TADs or TAD boundaries than Arrowhead TADs or TAD boundaries in every dataset investigated. There was a mean of ~1598 and ~1488 Arrowhead TADs and ~4360 and ~3820 TopDom in human and mouse respectively. This equated to a mean of ~2972 and ~2753 Arrowhead TAD boundaries and ~4766 and ~4170 TopDom TAD boundaries in human and mouse respectively (Figure 2.6A-B). As expected, TADs detected with TopDom were also on average smaller than those detected with Arrowhead (mean size of Arrowhead TADs: ~1464 kb and ~1594 kb and TopDom TADs: 601 kb and 609 kb in human and mouse respectively) (Figure 2.6C). I assessed the chromosome coverage of TADs detected using

Arrowhead or TopDom. TopDom TADs had a higher percentage chromosome coverage than Arrowhead TADs. The mean coverage across all datasets and all autosomal chromosomes for Arrowhead TADs was ~52% and ~62%, whereas for TopDom it was ~89% and ~94% for human and mouse respectively (Chapter 2: Supplementary Figure 3). This is an expected result reflecting the fact that while Arrowhead detects TADs sporadically, TopDom attempts to annotate the entire chromosome.

I next compared TADs and TAD boundaries detected with Arrowhead or TopDom in order to assess the agreement between TAD callers (see Methods: 2.2.6). I found that a very low proportion of TADs or boundaries were detected by both TAD callers, but there was greater agreement between TAD callers on TAD boundaries annotation compared to whole TAD annotation. A mean of ~49 and ~41 TADs were detected by both TAD callers for human and mouse respectively (corresponding to mean similarity scores of ~0.008 and ~0.008 respectively). On the other hand, a greater number of boundaries were detected, where a mean of ~844 and ~759 TAD boundaries were detected by both TAD callers for human and mouse respectively (corresponding to mean similarity scores of ~0.118 and ~0.120 respectively) (Figure 2.6A-B, Chapter 2: Supplementary Figure 2). This suggests that TAD callers may be better at detecting boundaries than they are at assembling them into TADs. This result is supported by the findings of a TAD comparison paper which found TAD callers are better at predicting TAD boundaries than TADs (Dali and Blanchette 2017). Nevertheless, the similarity of both TADs and TAD boundaries between Arrowhead and TopDom was low, supporting previous studies which have highlighted the low concordance of TADs between different TAD callers (Dali and Blanchette 2017; Forcato et al. 2017; Zufferey et al. 2018). This highlights the importance of carrying out all analyses using both Arrowhead and TopDom, ensuring that results are robust to the TAD caller chosen.

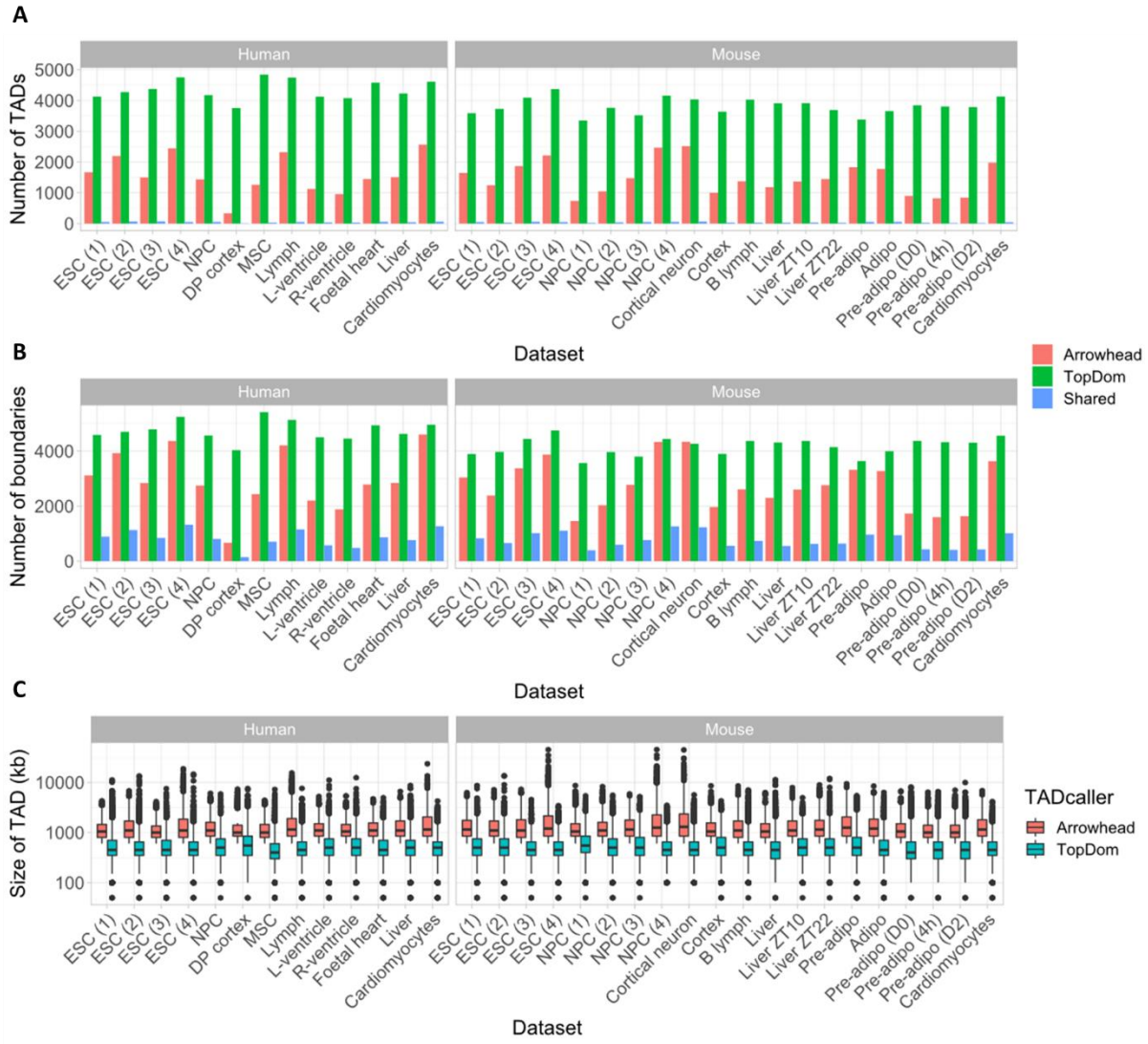


Figure 2.6: Comparison of autosomal TADs and boundaries detected by either Arrowhead or TopDom. *TADs detected at 50 kb bin size with default parameters. A) Number of TADs detected by Arrowhead (red), TopDom (green) or both (blue) per datasets. TADs were defined as shared if both boundaries had the same coordinates between datasets. B) Number of TAD boundaries detected by Arrowhead (red), TopDom (green) or both (blue) per dataset. TAD boundaries were defined as shared if they had the same coordinate between datasets. C) Distribution of TAD sizes for TADs detected by Arrowhead or TopDom per datasets (plotted on a log_{10} scale).*

2.3.4. Similarity of TADs and TAD boundaries between tissues

I calculated the similarity of TADs between all datasets within the same species (Figure 2.7). On average, the similarity scores were low for both TAD callers and species (mean similarity score: ~0.066 and ~0.058 for Arrowhead TADs and ~0.129 and ~0.120 for TopDom TADs in human and mouse respectively). This indicates that on average, a low proportion of the total detected TADs between two datasets were detected in both, pointing to a generally low similarity between TADs in different datasets. In order to assess the degree of tissue specificity, I grouped the samples into tissue groups based on tissue or cell type similarity. These groups included: ESC, NPC, brain, lymphoblast, liver, adipose, and heart (see Methods: 2.2.6). I then compared the similarity scores of pairs of datasets in the same tissue group with the similarity scores of pairs of datasets in different tissue groups (Figure 2.7). For both species and both TADs callers, TADs in pairs of datasets in the same tissue group were significantly more similar than pairs of datasets in different tissue groups (p-value <0.001). This suggests that although the similarity between different datasets is low, it is higher between datasets from similar tissues/cell types. This indicates that there is a statistically significant degree of tissue/cell type specificity in TADs. The average effect size across species and TAD callers was 0.477, indicating that the effect attributable to tissue specificity is moderate (moderate effect size defined as 0.3- 0.5). Several of the datasets analysed were published in the same paper or by the same group; it is therefore possible that these datasets could be more similar which could skew the results. To make sure that the increased similarity between datasets in the same tissue group is robust to this, I repeated the analysis both after excluding similarity scores corresponding to: datasets published in the same data repository (Chapter 2: Supplementary Figure 4) and datasets published in papers with common first or last authors (Chapter 2: Supplementary Figure 6). In both cases, pairs of datasets from the

same tissue group were significantly more similar than pairs of datasets from different tissue groups (p-value <0.001 and p-value <0.01 respectively).

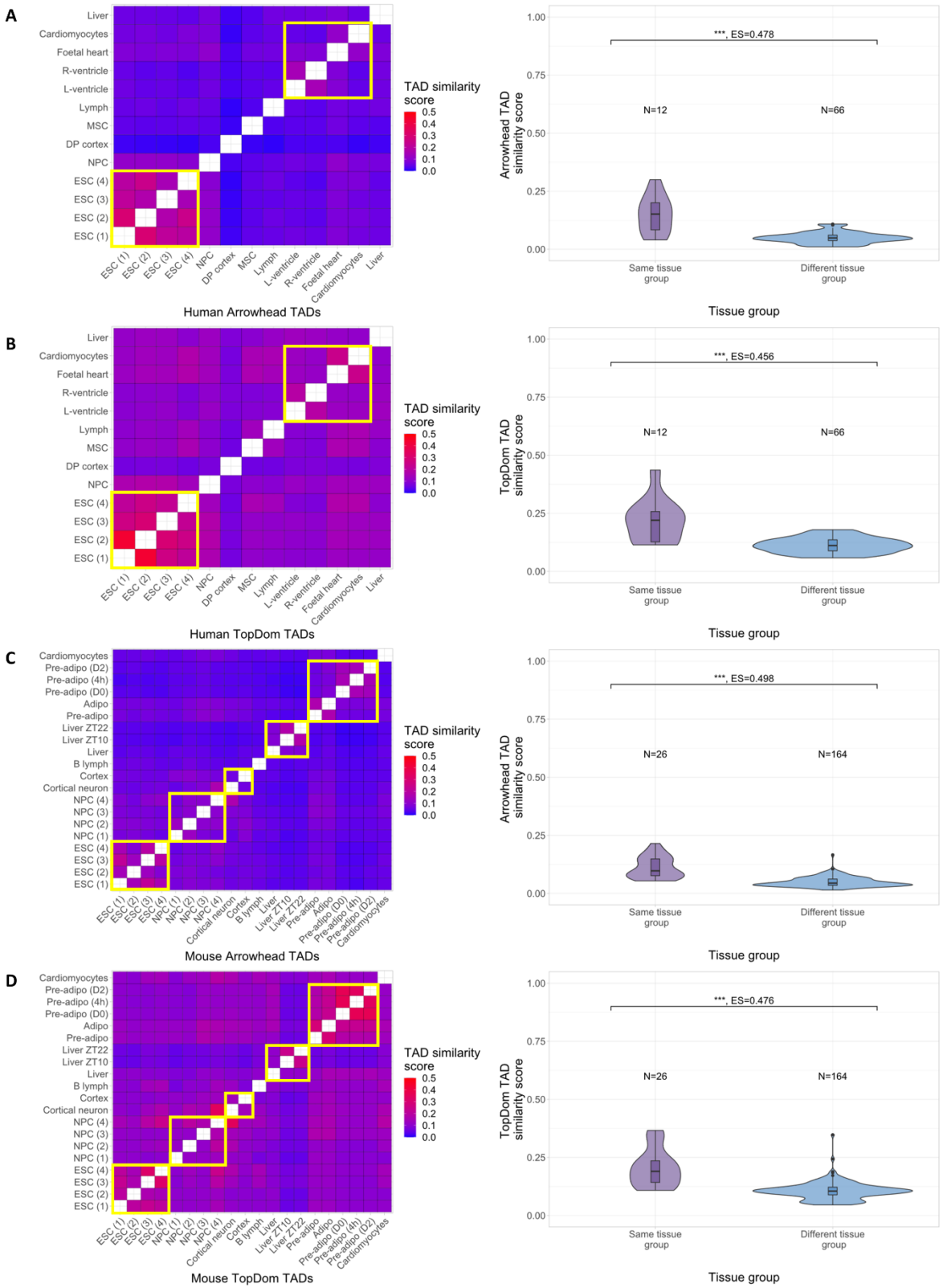


Figure 2.7: Autosomal TAD comparison between all pairs of datasets within species. *Left: heat maps indicating pairwise similarity scores between datasets. A higher similarity score indicates higher similarity. Pairs of datasets in the same tissue/cell type group are indicated by yellow squares. Right:*

*distribution of scores between pairs of datasets in the same tissue group vs different tissue groups. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across four p-values. A) Human Arrowhead TADs. B) Human TopDom TADs. C) Mouse Arrowhead TADs. D) Mouse TopDom TADs.*

To evaluate whether the tissue specificity observed in TADs is also present in TAD boundaries, I repeated the above analysis to assess the similarity of TAD boundaries between datasets. On average, boundaries were more similar between datasets than TADs (mean similarity score: ~0.195 and ~0.185 for Arrowhead TAD boundaries and ~0.342 and ~0.325 for TopDom TADs boundaries in human and mouse respectively) (Figure 2.8). This could reflect the possibility that TAD callers are better at detecting boundaries than putting them together into TADs (Dali and Blanchette 2017). Alternatively, this could suggest that TAD boundaries are relatively tissue/cell type invariant but arrange differently into tissue/cell type-specific TADs. In order to assess the degree of tissue specificity of TAD boundaries, I compared similarity scores within and between tissue groups. As was the case for TADs, I found that TAD boundaries were significantly more similar between datasets in the same tissue group compared to datasets in different tissue groups (p-value < 0.001). The average effect size across species and TAD callers was 0.41325, indicating that the effect attributable to tissue specificity is moderate. Similarly to the results for TADs, this result was robust to the exclusion of scores corresponding to datasets from the same data repository or with the same first or last authors (Chapter 2: Supplementary Figure 5 and Chapter 2: Supplementary Figure 7) (p-value <0.001 and p-value <0.01 respectively). This suggests that there is also a degree of tissue specificity in TAD boundaries, indicating that the tissue specificity detected in TADs cannot be

explained by different TADs forming between the same boundaries in different tissues/cell types.

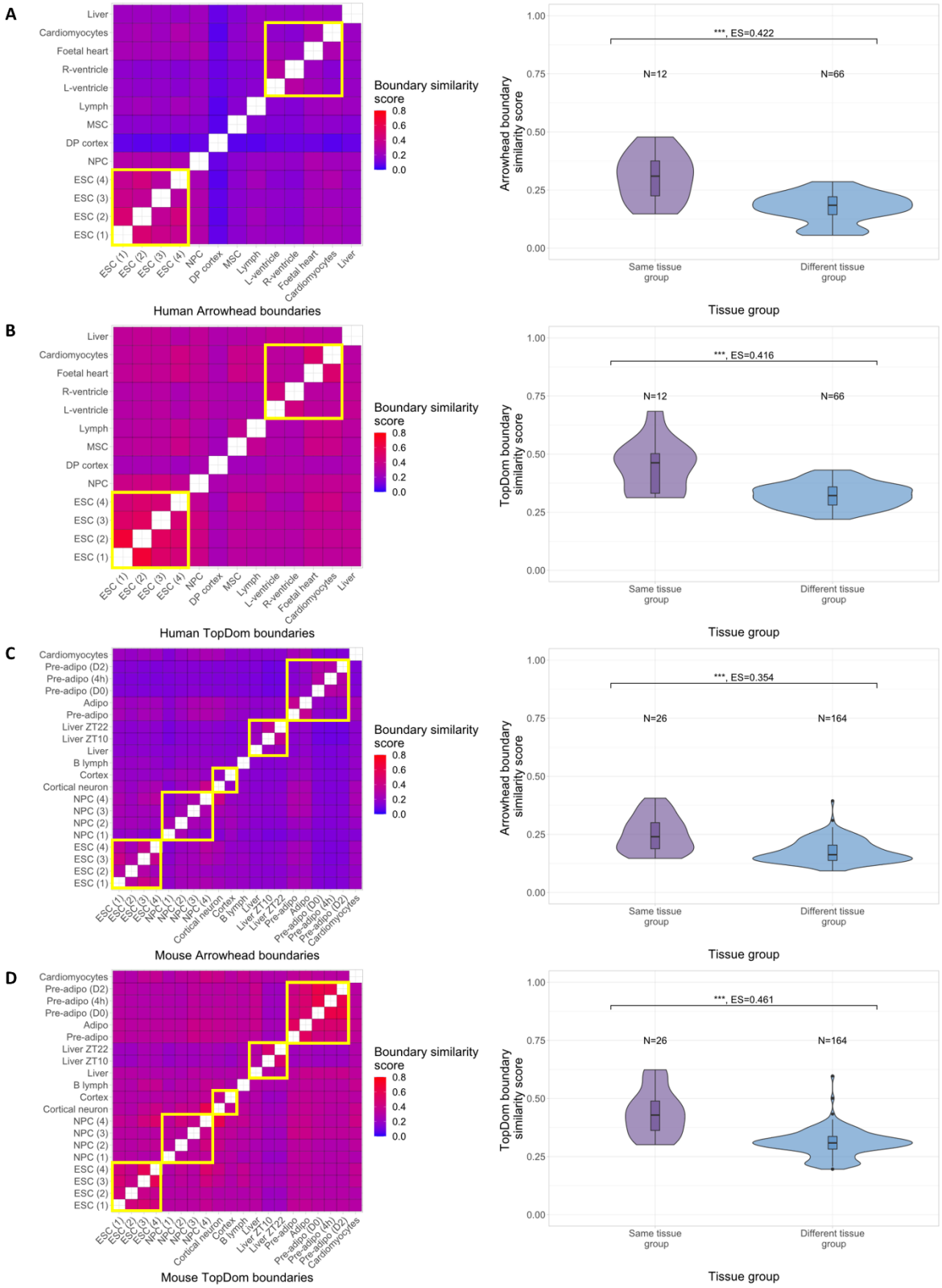


Figure 2.8: Autosomal TAD boundary comparison between all pairs of datasets. *Left: heat maps indicating pairwise similarity scores between datasets. A higher similarity score indicates higher similarity. Pairs of datasets in the same tissue/cell type group are indicated by yellow squares. Right:*

*distribution of scores between pairs of datasets in the same tissue group vs different tissue groups. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across four p-values. A) Human Arrowhead TAD boundaries. B) Human TopDom TAD boundaries. C) Mouse Arrowhead TAD boundaries. D) Mouse TopDom TAD boundaries.*

2.3.5. Similarity of TADs and TAD boundaries between human and mouse

In order to compare TADs between human and mouse tissue groups, I converted the coordinates of mouse TADs to hg19 and rounded the hg19 coordinates to the nearest 50 kb bin (see Methods: 2.2.7). This excluded a mean of ~5.3% and ~7.7% of Arrowhead and TopDom TADs respectively which could not be converted to hg19. Human TADs were also excluded if they could not (in principle) be converted to mm10, excluding a mean of ~3.0% and ~5.8% of Arrowhead and TopDom TADs respectively (Chapter 2: Supplementary Figure 8). This ensured that only TADs or TAD boundaries in syntenic regions were included. Human and hg19 mouse TADs and TAD boundaries in syntenic regions were then compared as above (Chapter 2: Supplementary Figure 9 and Chapter 2: Supplementary Figure 10).

As for the within-species tissue comparison, the average TAD similarity score between samples was low, indicating that a low proportion of the total TADs detected between two datasets are detected in both datasets (mean TAD similarity score: ~0.042 and ~0.088 for Arrowhead and TopDom respectively). For both TAD callers, pairs of samples from the same species and the same tissue group had a significantly higher similarity score than pairs of samples from the same species but different tissue groups ($p<0.001$), again suggesting that there is a significant degree of tissue specificity in TADs from the same species. Pairs of datasets from different species but the same tissue group were also on average more similar

than pairs of datasets from different species and different tissue groups for both TAD callers (mean: ~0.016 vs ~0.013 and ~0.039 vs ~0.037 for Arrowhead and TopDom respectively). This suggests that TADs from the same tissue groups display increased similarity even between species. However, this difference is only significant in TADs detected by Arrowhead ($p<0.001$) (Figure 2.9).

In order to determine the species specificity of TADs, I compared the TAD similarity scores of pairs of datasets from the same species with the TAD similarity scores of pairs of datasets from different species. For both TAD callers and when comparing pairs of datasets from the same tissue group or pairs of datasets from different tissue groups, pairs of datasets from the same species were significantly more similar than pairs of datasets from different species ($p<0.001$). The effect sizes of these comparisons ranged from 0.844- 0.863, demonstrating that this is a large effect (Figure 2.9). This shows that despite claims of high species conservation (Dixon et al. 2012; Rao et al. 2014; Vietri Rudan et al. 2015), there is a substantial difference in TADs between species indicating a high degree of species specificity.

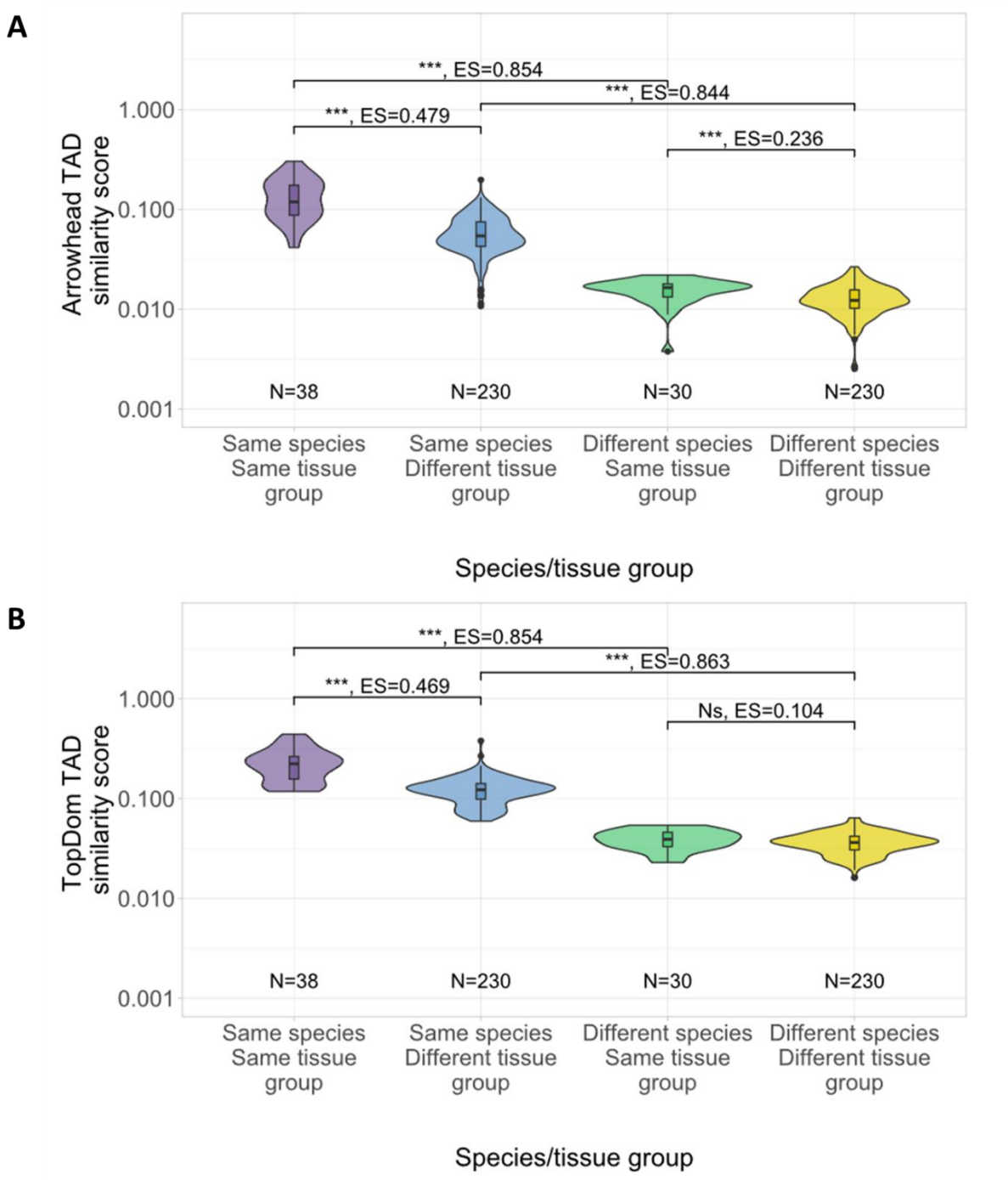


Figure 2.9: Distribution of autosomal TAD similarity scores between pairs of datasets in the same tissue group vs different tissue groups. *A higher similarity score indicates higher similarity. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across eight p-values. A) Arrowhead TADs. B) TopDom TADs.*

To assess the species specificity of TAD boundaries, I repeated the above analysis using TAD boundaries. In agreement with previous results, TAD boundaries were more similar between datasets than TADs (mean similarity score: ~0.15 and ~0.27 for Arrowhead and TopDom respectively). As seen in the comparisons of TADs, pairs of datasets from the same species and same tissue group had significantly higher similarity scores than pairs of datasets from the same species but different tissue groups ($p<0.001$). The increased similarity of pairs of datasets from the same tissue group was also observed when comparing datasets from different species. This suggests that TAD boundaries have a degree of tissue specificity even across species. However, as with TADs, this increase was only significant in TADs detected with Arrowhead (Figure 2.10).

Similarly to the results from the comparison of TADs across species, TAD boundary similarity scores were significantly higher between datasets from the same species compared to different species and the effect size was high ($p<0.001$, effect size range: 0.769- 0.867). This suggests that TAD boundaries show a high degree of species specificity.

All findings for both TADs and TAD boundaries were robust to the removal of pairs of datasets either from the same data repository or with the same first/last authors (Chapter 2: Supplementary Figure 11 and Chapter 2: Supplementary Figure 12).

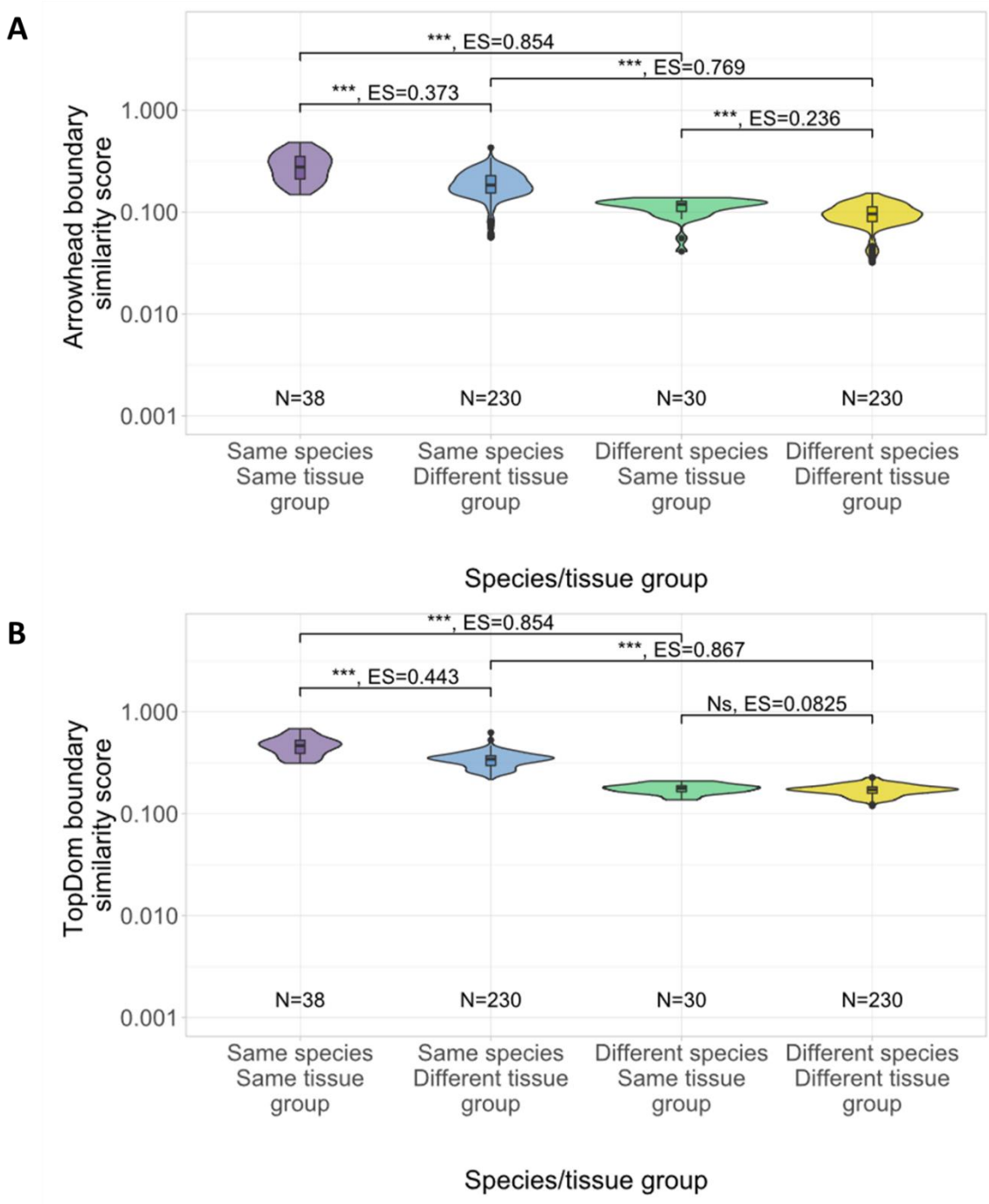


Figure 2.10: Distribution of autosomal TAD boundary similarity scores between pairs of datasets in the same tissue group vs different tissue groups. *A higher similarity score indicates higher similarity. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across eight p-values. A) Arrowhead TAD boundaries. B) TopDom TAD boundaries.*

2.3.6. Per TAD species and tissue conservation

Above, I have assessed the tissue and species conservation of TADs by comparing all TADs between pairs of datasets and calculating a score quantifying the similarity between those datasets as a whole. However, it is also of interest to investigate which individual TADs are invariant vs tissue or species-specific. To assess this I calculated species and tissue conservation of TADs on a per TAD scale. For each TAD, I calculated the number of different tissue groups in which the TAD was present, as well as whether it was conserved between species (see Methods: 2.2.8). For both TAD callers and species I found that the majority of TADs were detected in a single tissue group: 80.5% and 74.8% of Arrowhead TADs, and 67.5% and 64.7% of TopDom TADs in human and mouse respectively. This suggests that the majority of TADs are tissue/cell type-specific. I further found that TADs which were conserved between species were also present in significantly more tissue groups than those which were not conserved ($p<0.001$). This suggests that species-conserved TADs are more likely to be tissue invariant (Figure 2.11).

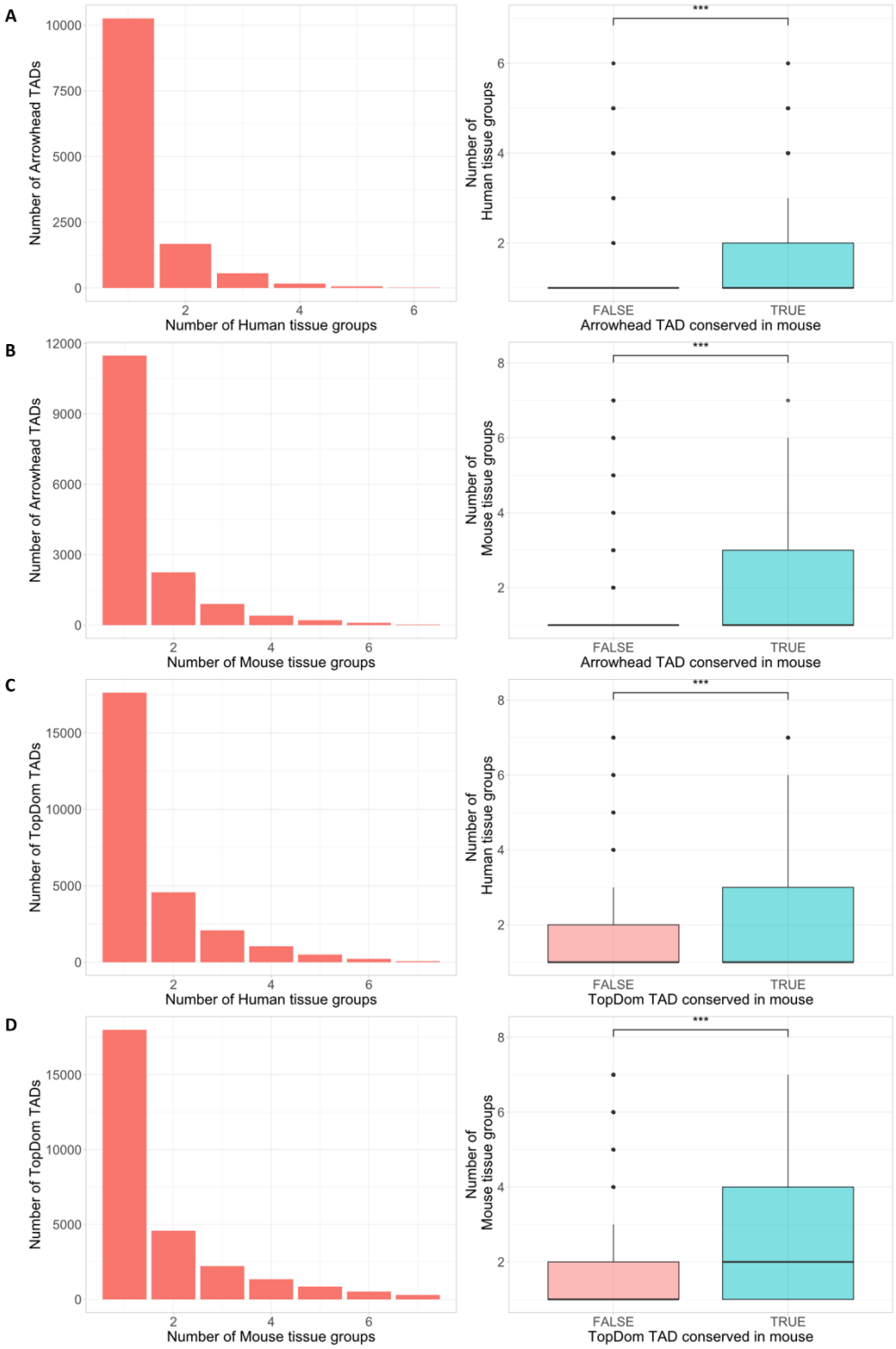


Figure 2.11: Per autosomal TAD tissue and species conservation. *Left: The number of different tissue/cell type groups for each TAD detected across all datasets. Right: As left but split by conservation in the other species. FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *. FDR correction applied across four p-values. A) Arrowhead human TADs. B) Arrowhead mouse TADs. C) TopDom human TADs. D) TopDom mouse TADs.*

2.4. Discussion

Since the first characterisation of TADs (Nora et al. 2012; Dixon et al. 2012) it has been commonly claimed that TADs are both tissue (Dixon et al. 2012; Nora et al. 2012; Dixon et al. 2015; Rao et al. 2014; Schmitt et al. 2016) and species invariant (Dixon et al. 2012; Rao et al. 2014; Vietri Rudan et al. 2015). Whilst these claims are not incompatible with a proportion of TADs showing a degree of tissue specificity, they appear to be overstated as the data underlying them often indicated that there was a substantial proportion of TADs which were not conserved between tissues or species. This is potentially problematic especially in cases where this claim is used to justify the use TADs from seemingly irrelevant tissues in analyses e.g. (Man et al. 2019; Medina-Gomez et al. 2017; Way et al. 2017). Therefore, to further investigate the degree of tissue and species conservation in TADs I compared TADs and TAD boundaries between 13 human and 20 mouse publicly available Hi-C datasets.

In the selected datasets, there was a large degree of variation in both the amount of Hi-C contacts (65,719,274- 4,759,064,782) and the *cis*/*trans* ratio (a common measure of noise in Hi-C data) (~0.44- ~12.6). This is likely to impact the bin size at which datasets can sensibly be analysed. I therefore calculated the map resolution for each dataset and selected a bin size for the detection of TADs (50 kb), which is higher than the lowest map resolution across all datasets.

Several studies have indicated that there is low concordance in the TADs detected using different TAD callers (Zufferey et al. 2018; Dali and Blanchette 2017; Forcato et al. 2017). I assessed the sensitivity of two popular TAD callers: Arrowhead (Durand, Shamim, et al. 2016; Rao et al. 2014) and TopDom (Shin et al. 2016) to bin size of the Hi-C map, adjustable parameter choice and number of Hi-C contacts. Both Arrowhead and TopDom detected fewer

but larger TADs as bin size increased. Arrowhead also detected fewer TADs when applied to data which was binned at an inappropriately low bin size. These results suggest that both TAD callers are sensitive to bin size, and particularly for Arrowhead, it is important to select an appropriate bin size based on data quality. Arrowhead and TopDom each have one adjustable parameter called “m” and “w” respectively. There was very little variation in either the number or size of Arrowhead TADs detected with different values of “m”. This suggests that detected Arrowhead TADs are very robust to the choice of “m”, therefore the default value of “m” was used throughout this chapter. For TADs detected with TopDom, as the value of “w” increased, fewer and larger TADs were detected. Since the default value of “w” was optimised for a bin size of 40 kb, I reasoned that it is likely to be appropriate for use at 50 kb. Finally, I assessed the robustness of each TAD caller to the number of Hi-C contacts in the Hi-C map. The number and size of TADs detected with TopDom remained largely constant regardless of the number of reads in the Hi-C map. On the other hand, more and larger Arrowheads TADs were detected in Hi-C maps with more Hi-C contacts. This suggests that unlike TopDom, Arrowhead is sensitive to the number of contacts captured. However, the increase in the number of detected TADs at 50 kb begins to plateau at ~400-500 million Hi-C contacts.

In each dataset analysed, more TADs and TAD boundaries were detected with TopDom than Arrowhead, TADs detected with TopDom were on average smaller than those detected with Arrowhead, and TopDom TADs had a higher chromosomal coverage than Arrowhead TADs. As expected from previous studies, the number of TADs or TAD boundaries detected by both TAD callers was low (Zufferey et al. 2018; Dali and Blanchette 2017; Forcato et al. 2017). However, the number of TAD boundaries detected by both TAD callers was higher than the number for TADs. This supports the suggestion made by Dali and Blanchette et al. that TAD

callers may be better at detecting individual boundaries than they are at putting them together as TADs (Dali and Blanchette 2017). Alternatively, if a nested hierarchy of TADs exist between available TAD boundaries, then this result could indicate that Arrowhead and TopDom are detecting TADs at different levels of the chromatin hierarchy. To mitigate against TAD caller differences and specific sensitivities, I detected TADs using both TAD callers in all analyses to ensure that conclusions were robust to the choice of TAD caller.

In order to assess the degree of tissue conservation, I compared all of the selected datasets within each species. I found that the similarity score was on average very low, suggesting that the level of TAD/TAD boundary similarity between datasets on average is low and therefore the level of tissue conservation is also low. However, as with between TAD callers, the similarity of TAD boundaries was higher than the similarity of TADs between datasets. Again, this could be attributable to TAD callers being better at detecting boundaries than TADs. Another possible explanation is that TAD boundaries are more conserved between tissues but form different TADs in a tissue-specific manner. In order to assess the level of tissue/cell type specificity, I grouped samples into tissue groups and compared the similarity of TADs in pairs of datasets from the same tissue group with pairs of datasets from different tissue groups. For both human and mouse, and for both TAD callers, I found that TADs in pairs of datasets from the same tissue group were statistically more similar than pairs of datasets from different tissue groups. To test whether the tissue specificity of TADs is driven by the different pairing of the same boundaries into TADs between different tissues, or whether the positions of TAD boundaries themselves differ between tissues, I compared TAD boundaries between tissue groups. I found that TAD boundaries are statistically more similar between pairs of datasets from the same tissue group compared to pairs of datasets from different tissue groups. Furthermore, the effect size of the difference in similarity between datasets from the

same tissue group vs different tissue groups was comparable, whether the analysis was done on TADs or TAD boundaries. This suggests that the positions of both TADs and TAD boundaries show a significant degree of tissue specificity.

The degree of conservation of TADs and TAD boundaries in syntenic regions was evaluated between human and mouse. For both TAD callers, pairs of datasets from the same species were statistically more similar than pairs of datasets from different species. This suggests that there is a large degree of species specificity in TADs and TAD boundaries. Interestingly, for both TAD callers, pairs of datasets from different species and the same tissue group were on average more similar than pairs of datasets from different species and different tissue groups, suggesting that some of the tissue specificity of TADs and TAD boundaries is conserved between species. However, this difference was only significant in TADs and TAD boundaries detected with Arrowhead.

Assessing tissue conservation of all human or mouse TADs in syntenic regions on a per TAD scale revealed that 80.5% and 74.8% of Arrowhead TADs, and 67.5% and 64.7% of TopDom TADs in human and mouse respectively, were detected in a single tissue group, indicating that the vast majority of detected TADs are tissue-specific. Additionally, TADs which were conserved between species were enriched for being detected in a larger number of tissue groups. This suggests that TADs which are conserved between species may be enriched for tissue invariance.

Overall, these results suggest that contrary to common claims (Dixon et al. 2012; Nora et al. 2012; Dixon et al. 2015; Rao et al. 2014; Schmitt et al. 2016; Vietri Rudan et al. 2015), there is evidence for a significant degree of tissue and species specificity in both TADs and TAD boundaries. The tissue specificity results outlined here support the findings of a recent study

which conducted a similar analysis on human TADs detected using the TAD caller Armatus at 100 kb bin size. They found that the overall similarity of human TADs between tissues/cell types is low, but that there is an increased similarity between samples from the same tissue/cell type (Sauerwald, Singhal, and Kingsford 2020). However, the use of Armatus to identify TADs is potentially problematic as TAD caller benchmarking papers have come to conflicting conclusions as to whether TAD boundaries detected with Armatus are enriched for CTCF binding, despite the fact that CTCF binding is thought to be required for TAD establishment. Two benchmarking papers indicated that Armatus is among the worst performing (Zufferey et al. 2018; Dali and Blanchette 2017) and one benchmarking paper indicated it is among the best performing (Forcato et al. 2017) TAD callers in detecting TADs with CTCF enrichment at their boundaries. Nevertheless, the work outlined in this chapter supports and extends the findings of Sauerwald et al. by extending the analysis to include a human vs mouse species comparison and proving that the result is robust to the choice of TAD caller as well as being detectable at a lower bin size (Sauerwald, Singhal, and Kingsford 2020). The results are also supported by another recent study which found that human TAD boundaries display “substantial variability between cell types”, but that related cell types had increased similarity. The same study also found that tissue invariant TAD boundaries had higher base pair level evolutionary constraint, which is complementary to my finding that TADs conserved between human and mouse are enriched for tissue invariance (McArthur and Capra 2021).

There are several limitations of this work, the largest of which relates to the variance in publicly available Hi-C datasets and their availability. Detected TADs (in my research and others) may have been affected by the vast difference in the number of Hi-C contacts in the included datasets (particularly those detected by Arrowhead). Whilst I have taken steps to

ensure that results are robust to technical similarities between datasets produced by the same authors, it is possible that other technical influences such as Hi-C protocol or restriction enzyme could affect results. Analysis is also limited because Hi-C samples are not available in the public domain for all cell types and/or many different species. Furthermore, many of the available datasets are generated using tissue samples; therefore, the generated Hi-C map will constitute an average of the chromatin contact profiles over all the constituent cell types. In order to improve the analysis presented here, it would be necessary to generate standardised Hi-C datasets, capturing a similar number of Hi-C contacts, for a large range of different cell types across a range of different species. The establishment of a gold standard method for detecting TADs is also essential to the future of the field.

In conclusion, my results support previous work demonstrating that there is better agreement between TAD callers in the annotation of TAD boundaries rather than TADs (Dali and Blanchette 2017). I found that contrary to common claims that TADs are tissue and species invariant, the similarity of TADs and TAD boundaries between tissues and species is low. However, there is a significant increase in the similarity of TADs and TAD boundaries between datasets from the same tissue group compared to datasets from different tissue groups. This suggests that TADs and TAD boundaries show a significant degree of tissue specificity. The increase in similarity attributable to tissue specificity is detectable when comparing TADs and TAD boundaries both within and between species. Even though this trend is present for both TAD callers, the difference is only significant for TADs detected using Arrowhead when comparing between species.

Chapter 3: Assessing the relationship between TADs, expression and gene function [8]

3.1. Introduction

It has been proposed that TADs constitute functional units in the genome, important for the correct regulation of gene expression. TADs co-localise regulatory elements with their target genes, and are thought to promote co-regulation of the genes they contain, creating "gene regulatory domains" (Dixon, Gorkin, and Ren 2016). By inserting regulatory sensors along the length of the genome, Symmons et al. found evidence that the activity of enhancers is split into regulatory domains which highly correlate with TADs (Symmons et al. 2014). This provided experimental evidence that TADs potentially facilitate enhancers to carry out "non-specific" co-regulation of all genes in the TAD (Symmons et al. 2014; Dixon, Gorkin, and Ren 2016). Simultaneously, TADs are thought to insulate genes from aberrant regulation by regulatory elements outside the TAD ('enhancer hijacking') (Dixon, Gorkin, and Ren 2016). Several examples of congenital disease have been linked to TAD boundary disruptions

[8] *The findings presented in this chapter have also been published in:* ***Long, Helen S.****, Simon Greenaway, George Powell, Ann-Marie Mallon, Cecilia M. Lindgren, and Michelle M. Simon. 2022. "Making Sense of the Linear Genome, Gene Function and TADs." Epigenetics & Chromatin 2022 15:1 15 (1): 1–19. https://doi.org/10.1186/S13072-022-00436-9.*

Figures, text and analysis that have been reused/modified from this publication are licensed under CC BY 4.0 and were authored by me (unless explicitly stated) and have been reused with the permission of all authors.

allowing enhancer hijacking, demonstrating that at least in these cases, TAD boundaries are essential for proper gene regulation (Franke et al. 2016; Lupiáñez et al. 2015; Giorgio et al. 2015). TAD boundaries are also able to block the spread of transcription, and repressive chromatin (Dixon, Gorkin, and Ren 2016). It has been observed that the unit of compartment switching in the genome tends to be whole TADs (Dixon et al. 2015). Adding to this picture, it has been suggested that genes within the same TAD have highly correlated expression patterns (Nora et al. 2012; Flavahan et al. 2016; Sarnataro, Riba, and Molina 2020; Tarbier et al. 2020). This has led to the speculation that genes which are required for specific processes may be contained within the same TAD to allow them to share regulatory programs and efficiently switch between the active and inactive compartments (suggested in the review (Ruiz-Velasco and Zaugg 2017)). Studies have already indicated that some TADs may be enriched for lineage-specific genes (Neems et al. 2016), but the global relationship between TADs and gene function is yet to be fully understood.

It has long been known that the linear order of genes in the genome is non-random with respect to gene function. Genes that are close together in linear space are more likely to have correlated expression patterns (reviewed in (Hurst, Pál, and Lercher 2004)), and share pathways and protein-protein interactions (PPI) (Thévenin et al. 2014). Genes within TADs are by definition also close together in the linear genome; therefore, the linear proximity between genes is an important confounding factor when studying the similarity of genes that share a TAD. It is also possible that the increased similarity of genes that are proximal in the linear order occurs on a similar scale to TADs. By promoting co-regulation of genes, TADs could explain the increased functional similarity between proximal genes.

Given this, I hypothesised that TADs form functional units, and therefore genes within them are more likely to share functional annotations than can be explained by linear proximity alone. In order to test my hypothesis, I utilised some of the highest quality mammalian Hi-C data currently in the public domain (Bonev et al. 2017). In order to ensure that findings were robust to the choice of TAD caller, I repeated the strategy outlined in detail in Chapter 2 and annotated TADs using two popular TAD calling algorithms; Arrowhead (Durand, Shamim, et al. 2016; Rao et al. 2014) and TopDom (Shin et al. 2016). I first assessed the relationship between TADs and gene paralogy as well as constraint. Then, focusing on annotated TADs most likely to have been formed by loop extrusion involving CTCF, I assessed the functional relatedness of non-paralogous protein-coding genes within them, using four functional annotations: expression correlation, Gene ontology (GO) semantic similarity, shared pathways and PPI.

3.1.1. Aims

The overall aims of this chapter are to:

- Develop an appropriate TAD randomisation method to serve as a 'null' TAD distribution.
- Assess the relationship of paralogous gene families and gene constraint with TADs.
- Evaluate the functional similarity of pairs of genes sharing a CTCF TAD using: expression correlation, GO semantic similarity, shared pathway annotations and PPI.

3.2. Methods

3.2.1. Hi-C data processing

In order to assess the functional relationship between genes sharing a TAD, Hi-C sequencing data for Mouse ESCs and cortical neurons was downloaded from Gene Expression Omnibus (Edgar, Domrachev, and Lash 2002; Barrett et al. 2012) (accession number GSE96107 (Bonev et al. 2017)). These Hi-C datasets represent some of the most deeply sequenced mammalian Hi-C data currently in the public domain. Hi-C data was processed as outlined in Chapter 2: Methods: 2.2.2. In brief, raw sequencing data was processed using the Juicer pipeline (Durand, Shamim, et al. 2016), aligning to mm10. Individual replicates were processed separately and then merged using the "mega" function in Juicer. Reads which aligned to the genome with a MAPQ score below 30 were excluded from the resulting contact map. The contact map was then normalised with the VC normalisation method (Lieberman-Aiden et al. 2009) (as this normalisation method was compatible with both TAD callers). The contact matrix was binned to 10 kb. Hi-C contact maps were visualised using Juicebox 1.11.08 (Durand, Robinson, et al. 2016).

3.2.2. TAD calling and gene overlap

In order to ensure conclusions were robust to the choice of TAD caller (see Chapter 2), TADs were called using both Arrowhead and TopDom. For TADs called using Arrowhead, the ignore sparsity flag and default settings were used (m=2000) (Rao et al. 2014; Durand, Shamim, et al. 2016). For TADs called with TopDom, w=20 was used (justified in Results: 3.3.1) (Shin et al. 2016). As previously outlined (Chapter 2) TopDom detects "domains", "boundaries" and

"gaps"; only "domain" annotations were considered as TADs (as previously done in (Dali and Blanchette 2017)). The smallest TADs detected with TopDom were equivalent in length to a single bin. As very small TADs are unlikely to be real, I filtered TopDom TADs which were <90 kb in length. For Arrowhead TADs the smallest TADs were 120 kb in length so no filtering was necessary.

TADs are hypothesised to be formed by loop extrusion involving CTCF binding at both TAD boundaries (Sanborn et al. 2015; Fudenberg et al. 2017). In order to investigate TADs which are most likely to have been formed by this mechanism, where indicated, TADs were split into CTCF and non-CTCF TADs. To do this, CTCF Chromatin immunoprecipitation sequencing (ChIP-seq) peaks for ESC and cortical neuron (generated alongside the Hi-C data) were downloaded from GEO (Barrett et al. 2012; Edgar, Domrachev, and Lash 2002) (accession GSE96107) (Bonev et al. 2017). ESC CTCF ChIP-seq peaks were overlapped with ESC TADs, and cortical neuron CTCF ChIP-seq peaks were overlapped with cortical neuron TADs. TADs which had a CTCF ChIP-seq peak within ±1 bin (10 kb) from both boundaries were considered "CTCF TADs"; TADs which had a CTCF ChIP-seq peak within ±1 bin of only one boundary or neither boundary were considered "non-CTCF TADs".

The Ensembl IDs of protein-coding genes were downloaded from BioMart (Ensembl release 96) (Smedley et al. 2009). Protein-coding genes were overlapped with TADs using bedtools intersect (Quinlan and Hall 2010). Genes were only assigned to a TAD if the full length of the gene fell between the TAD boundaries, meaning any genes which fell incompletely into a TAD were excluded.

3.2.3. TAD randomisation

In order to produce a null distribution of TADs for comparison, I synthesised “random TADs” by two methods, which differently control for possible confounding signals. In the first randomisation strategy which I call “random TADs”, the position of each TAD is randomised to a new position on the same chromosome, of the same size as the original TAD. Each new position is only accepted if it contains the same number of genes as the original TAD. As TopDom TADs are non-overlapping, for random TopDom TADs, the proposed new random TAD position is also prevented from overlapping any existing random TopDom TADs. For Arrowhead TADs I observed an overlap structure in which TADs are more likely to overlap in a “nested” fashion (one TAD falls fully within another). In order to approximate this overlap structure in Arrowhead random TADs, each proposed random Arrowhead TAD position was checked to see if it overlapped an existing random TAD. If it overlapped an existing TAD in a “nested” fashion it was always permitted. However, “non-nested” overlaps (one TAD partially overlaps another) were only permitted for each TAD with 10% probability. This meant that for the majority of TADs, a “non-nested” overlap was never permitted, thereby minimising this type of overlap. If a random TAD location fulfilling all criteria outlined could not be found after 10,000 attempts, the TAD was excluded from the randomisation set. TADs called with Arrowhead and TopDom were each randomised 100 times to generate 100 sets of random TADs for each TAD caller and cell type. Since during the randomisation, the algorithm randomises the position of each TAD in turn, the order of TADs was shuffled before every subsequent randomisation.

In the second strategy which I refer to as “random genome TADs”, the position of the TADs in the genome and the number of protein-coding genes within them was maintained but the

identities of the genes within each chromosome were shuffled. Again, 100 sets of random genome TADs were generated for each TAD caller and cell type I tested.

In order to test if 100 randomisations was sufficient, I plotted the median value of each measure investigated in this chapter with each added random TAD/random genome TAD set. For most measures, the median began to converge at fewer than 100 randomisations (Chapter 3: Supplementary Figure 1). Scripts required to run the TAD randomisation strategies can be found at https://github.com/MRC-Harwell/TAD_randomisation.

For statistical comparisons of TADs and random TADs or random genome TADs, each set of random TADs/random genome TADs was individually compared to real TADs (100 comparisons). The median values of the measure being tested in each randomisation were plotted (e.g. distance between genes, proportion of paralogous gene pairs, expression correlation etc.). Significance was assessed using the median p-value of the comparisons and the direction of significance was assessed using either the direction of the median effect size (for Wilcoxon test) or the difference between the measure (e.g. proportion) in real TADs and the median value of the measure in random TADs/random genome TADs (for Fishers exact test).

In order to compare the TAD randomisation strategy developed here with previously published methods, I created TAD randomisation algorithms based on the descriptions outlined in Rao et al. and Nora et al. (Rao et al. 2014; Nora et al. 2012). In brief, in the Rao et al. method, the position of each TAD was randomised to a new position of equivalent size, on the same chromosome, but prevented from overlapping any gaps in the mm10 assembly (mm10 gaps were downloaded from the UCSC genome browser (Karolchik et al. 2004)). This method was adapted to prevent overlapping when randomising TopDom TADs. As in my

randomisation method I set a cut-off of 10,000 attempts to place each TAD before it was excluded from the random TAD set. In the Nora et al. method, each TAD was randomised to a new position on the same chromosome, which contained the same number of genes, and was the same length or smaller than the original TAD. Again, I adapted this method to prevent overlapping in TopDom random TADs and used a cut off of 10,000 attempts to place a random TopDom TAD before it was excluded from the TAD set. These algorithms were then used to generate an example random TAD set for each method.

3.2.4. Compartment annotation

To determine the chromatin compartment identity of each TAD, compartments were identified as follows: reads with MAPQ below 30 were filtered, contact maps were VC normalised and binned at 50 kb, and the eigenvector function (with the –p flag) in Juicer was used (Durand, Shamim, et al. 2016). The output from the eigenvector function was then further processed as outlined in Miura et al. (Miura et al. 2018). TAD annotations were then intersected with compartment annotations (the intersection was carried out by Dr Michelle Simon).

3.2.5. Paralogous genes

To investigate whether there is a relationship between paralogous gene families and TADs, I downloaded mouse paralogous gene pairs from BioMart (Ensembl release 98) (Smedley et al. 2009). In order to assess the gene expression correlation and functional annotation similarity

between genes sharing a CTCF TAD beyond that attributable to shared ancestry, paralogous gene pairs have been excluded in these analyses.

3.2.6. Olfactory genes

The olfactory genes were also removed from the analysis of gene expression correlation and functional annotation similarity between genes sharing a CTCF TAD. This was done because the olfactory gene family has undergone a significant expansion in the mouse genome compared to the human genome. The human genome contains ~800 olfactory genes, of which <400 are functional, whereas the mouse genome contains ~1400 olfactory genes, of which <1050 are functional (Niimura 2009). Therefore, in order to increase the relevance of findings to the human genome, the olfactory genes were masked in certain analyses. To remove the olfactory genes, MGI IDs (MGI v.6.15) associated with the GO term: "olfactory receptor activity" were downloaded (Carbon et al. 2009; The Gene Ontology Consortium 2018; Ashburner et al. 2000). IDs were then converted to Ensemble IDs using BioMart (Smedley et al. 2009). The Ensemble IDs were then used to remove olfactory genes from the analysis (1133 genes in total).

3.2.7. Selective gene constraint score [9]

I assessed the relationship between genes sharing a TAD and evolutionary gene constraint. Selective gene constraint measures the level of purifying selection pressure acting on a gene

[9] The details relating to the calculation of the mouse selective constraint score were provided by Dr George Powell.

and can be used as a proxy for functional importance. The mouse selective gene constraint score was calculated by Dr George Powell (Powell et al. 2019). In brief, selective constraint was quantified as a nonsynonymous z-score by comparing the variation in 15,648 genes between 36 mouse strains (Keane et al. 2011). The constraint of each gene was defined as the deviation of the observed number of nonsynonymous variants relative to the number expected under no selection. The number expected under no selection was defined as the average rate of synonymous fixation across all mouse strains. Genes with a higher relative depletion of nonsynonymous variants are more constrained. Genes were defined by their Ensembl canonical transcript IDs and were filtered if less than 90% of their transcribed sequence had a coverage ≥10x in over 90% of mouse strains. Ensembl transcript IDs were mapped to gene IDs (Ensembl release 99). When calculating the average constraint of genes in a TAD, if a gene had multiple transcripts, all transcripts were included in the average, and if a gene had no transcripts for which a constraint score could be calculated, the gene was excluded from the average.

3.2.8. Functional enrichment

The functional enrichment of genes singly, doubly or triply occupying TADs was calculated for "biological process" GO terms using g:profiler r package (version e100_eg47_p14_021df73) (Raudvere et al. 2019). A custom background including all protein-coding mouse genes was used and electronic annotations were excluded. P values were corrected using the "g:SCS" multiple testing correction method (Reimand et al. 2007). Plots show the (max) top 25 most significant GO terms which passed a corrected p value threshold of <0.05. For visualisation, the similarity between the GO terms was calculated using the GOSemSim R package using the

"Jiang" similarity method (measure parameter) and NULL as the combine parameter (G. Yu et al. 2010). The similarity scores were then used to hierarchically cluster similar GO terms in the plot for visualisation purposes using hclust clustering.

3.2.9. Gene expression correlation

In order to assess the expression correlation between pairs of genes sharing a CTCF TAD, RNA-seq data generated alongside the Hi-C data was downloaded. 'Fragments per kilobase of transcript per million mapped reads' (FPKM) RNA-seq values from two replicates each of ESC, NPC and cortical neurons were downloaded from GEO (accession number GSE96107). Genes with an FPKM<1 were treated as having no expression. Gene expression correlation between pairs of genes was calculated across all six samples using Spearman's rank correlation coefficient.

The RNA-seq data generated alongside the Hi-C data includes three cell types. However, if TADs are cell type specific, any possible association between TADs and gene expression correlation in this data may be masked. Therefore, RNA-seq data from the most closely matching cell types/tissues to ESCs and cortical neurons which had greater than three samples were also downloaded. This included:

- RNA-seq data from mouse ESCs differentiating to primordial germ cells (PGC) were downloaded as $\log_2$('Reads per million reads' (RPM)) values from GEO (accession number GSE86903). The data consisted of three replicates each of ESC, epiblast-like cells (day 2), PGC (day 4) and PGC (day 6) (von Meyenn et al. 2016). I converted the $\log_2$(RPM) values to 'reads per kilobase of transcript per million reads mapped' (RPKM)

values for analysis. Genes with an RPKM value <1 were treated as having no expression. The original data was provided with gene names rather than gene IDs. Gene names were converted to Ensembl IDs by matching the gene names and the gene coordinates to Ensembl IDs and coordinates in Ensembl release 77 (the archive predating the data to provide more complete mapping). Gene expression correlation was calculated for pairs of genes across all samples using Spearman's rank correlation coefficient. The gene expression correlation values calculated from this data were used to assess the gene expression correlation between pairs of genes in ESC TADs.

- RNA-seq data from embryonic mouse forebrain from different stages of development were downloaded from Encode as FPKM values. Accession numbers: ENCFF302TQO, ENCFF976OLT, ENCFF895JXR, ENCFF227HKF, ENCFF340XFQ, ENCFF484AOO, ENCFF601JPN, ENCFF413BXV, ENCFF465SNB, ENCFF567AFL, ENCFF590FAC, ENCFF745ZJF, ENCFF763GXJ, ENCFF804FTJ, ENCFF816CVP and ENCFF918QNL (Davis et al. 2018; P. He et al. 2020). Genes with FPKM<1 were treated as having no expression. Gene expression correlation between pairs of genes was calculated across all samples using Spearman's rank correlation coefficient. The gene expression correlation values calculated using this data were used to assess the expression correlation between pairs of genes in cortical neuron TADs.

3.2.10. GO Semantic similarity

The GO semantic similarity between genes quantifies the similarity of their GO annotations. In order to assess the similarity of GO annotations between pairs of genes sharing a CTCF TAD, GO semantic similarity was calculated using the GOSemSim R package (G. Yu et al. 2010). The

"Jiang" similarity method was used. Where genes were associated with multiple GO terms, scores were combined using the "best match average" method. Pairs of genes which did not have any GO terms were excluded from the analysis.

3.2.11. KEGG pathway annotation

The proportion of pairs of genes sharing a CTCF TAD which also shared at least one common molecular pathway was calculated. To assess which pairs of genes shared molecular pathways, I downloaded KEGG pathways from org.Mm.eg.db (v. 3.8.2) (Carlson 2017). The proportion of gene pairs sharing a pathway was considered as a proportion of all pairs of genes (which had at least one pathway annotation each) sharing a TAD.

3.2.12. PPI annotation

In order to evaluate the proportion of pairs of genes sharing a CTCF TAD which also had a PPI, PPI data was downloaded for mm10 from string v11 (Szklarczyk et al. 2019). Interactions were filtered so that only "binding" interactions were considered. Protein IDs were mapped to Ensembl IDs using the BioMart R package (Ensemble release 98). The proportion of gene pairs with a PPI was considered as a proportion of all pairs of genes (which had at least one PPI annotation each) sharing a TAD.

3.2.13. Distance permutation testing

To assess the relationship between the distance between pairs of genes in the linear genome and their functional similarity measured by co-expression, GO semantic similarity, shared pathways and PPIs. I calculated functional similarity for each of the measures outlined between pairs of genes in the real genome binned by genomic distance. For each bin, the values in the real genome were compared to the distribution of values in 1000 random genomes by permutation testing (sum(values in the random genome ≥ value in the real genome)/1000). P values were multiple testing corrected using false discovery rate (FDR) (Benjamini and Hochberg 1995).

3.2.14. Effect size

For analyses in which statistical significance is determined by the Wilcoxon rank-sum test, effect sizes have also been calculated in order to quantify the magnitude of the difference. The effect sizes were calculated using the R package rcompanion (wilcoxonR). A positive effect size indicates that the values in TADs are greater than in random TADs/random genome TADs; a negative effect size indicates the reverse. A larger effect size is indicated by a larger score.

3.3. Results

3.3.1. Assessment of Hi-C data quality and TAD caller parameter selection

Hi-C data from mouse ESC and cortical neuron originally published by Bonev et al. (Bonev et al. 2017) were selected for this analysis as they represent some of the most deeply sequenced Hi-C data currently publicly available. The datasets contain over 7 billion and 8.5 billion sequenced read pairs, corresponding to over 3 billion and 4.5 billion valid contacts for ESC and cortical neuron respectively. In order to assess the degree of noise in the dataset, I calculated the *cis*/*trans* ratio. As previously outlined (Chapter 2: Results: 2.3.1), high quality Hi-C experiments are expected to detect more *cis* than *trans* contacts; however, experimental noise is predicted to effect *cis* and *trans* contacts equally, resulting in a lower *cis*/*trans* ratio. In both datasets I observed a high *cis*/*trans* ratio (5.42 (~84.4% of contacts *cis*) in ESC and 2.81 (~73.7% of contacts *cis*) in cortical neuron) indicating that more *cis* contacts were captured than *trans* contacts, suggesting the data is of high quality (Lajoie, Dekker, and Kaplan 2015) (Table 3.1).

In this Chapter, TADs were detected on Hi-C data binned at 10 kb (see Methods: 3.2.2). In order to establish that this was a suitable resolution for the datasets, I calculated the map resolution (minimum bin size at which 80% of bins have at least 1000 contacts) which has been adopted as a method to determine the minimum bin size at which a Hi-C data can reasonably be analysed (Rao et al. 2014). Reflecting the high quality of the selected data, the map resolution was 950 bp and 750 bp for ESC and cortical neuron respectively (Table 3.1). Therefore, analysing the data at 10 kb represents a good trade-off between maximising the fine scale resolution which can be offered by this data whilst still binning to increase the signal to noise ratio.

Table 3.1: Statistics relating to the Hi-C data selected. *Including the number of replicates, the number sequenced read pairs, the number successfully aligned/ valid Hi-C contacts, the total number of cis contacts (within chromosome contacts), the ratio of cis (within chromosome contact)/ trans (between chromosome contact) contacts and the map resolution (defined as the minimum bin size at which 80% of bins have at least 1000 contacts and commonly used to select the minimum bin size at which a Hi-C dataset can sensibly be analysed).*

Hi-C dataset	No. replicates	Total no. of sequenced read pairs	Total no. of Hi-C contacts	Total no. of cis contacts	*Cis/trans* ratio	Map resolution
ESC	4	7,260,480,082	3,483,781,581	2,940,865,222	5.42	950 bp
Cortical neuron	6	8,710,217,333	4,759,064,782	3,509,057,227	2.81	750 bp

It has been widely reported that algorithmically detected TADs can vary massively depending on the TAD caller used (Zufferey et al. 2018; Dali and Blanchette 2017; Forcato et al. 2017) (Chapter 2: Figure 2.6). As justified in detail in Chapter 2, in order to mitigate against this I used two popular TAD callers: Arrowhead and TopDom, to ensure that results were robust to the choice of TAD caller.

Arrowhead and TopDom both have an adjustable parameter which is called “m” or “w” respectively. The value of “m” in Arrowhead has a default value of 2000 which is used for detecting TADs at 5 kb or 10 kb; therefore, m=2000 was used throughout this chapter. TopDom advises that the “w” parameter is set to an integer value between 5 and 20. When a large value of the “w” parameter is used, fewer TADs are detected and they are larger in size (Chapter 2: Figure 2.4). The default value of “w” is 5; however, this was optimised to be used for Hi-C data binned at 40 kb (Shin et al. 2016). TADs called with TopDom w=5 appeared suboptimal at 10 kb binning (Figure 3.1 and Chapter 3: Supplementary Figure 2). I therefore

sought to select a more appropriate value of “w” for calling TADs at 10 kb. As there is no gold standard TAD set with which to compare, in order to establish the optimal value of “w”, I visually inspected TADs called with w=5, 10, 15 and 20 respectively in several regions of the Hi-C maps where TADs called with w=5 appeared suboptimal. From these Hi-C maps I established that TADs called with a higher value of “w” appeared more accurate for 10 kb binned Hi-C maps (Figure 3.1 and Chapter 3: Supplementary Figure 2). Therefore, w=20 was selected for the TopDom TADs called throughout this chapter. w=20 was also used to detect TopDom TADs in Eres et al. (including TADs called at 10 kb) (Eres et al. 2019).

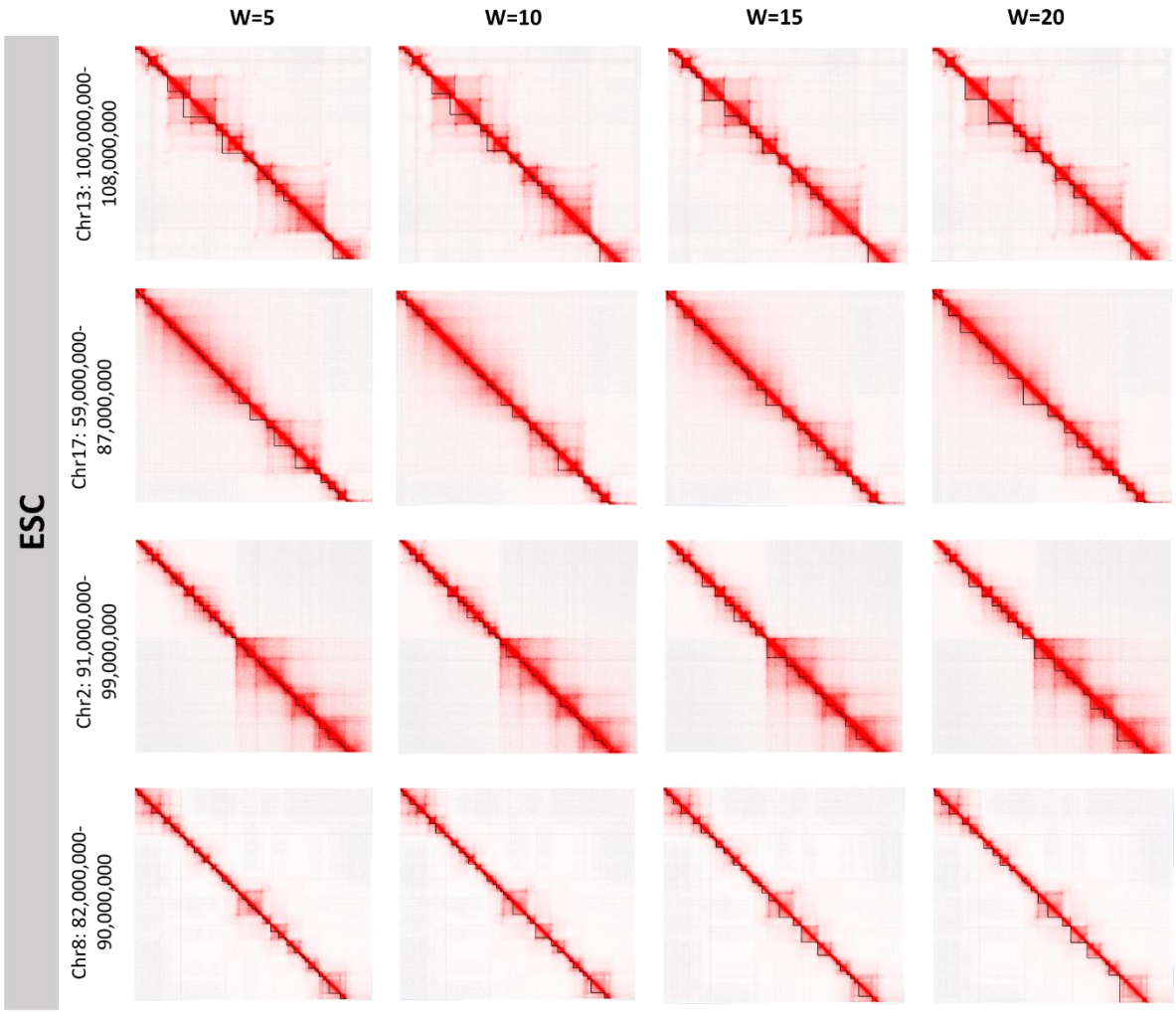


Figure 3.1: ESC TADs detected with TopDom and different values of the "w" parameter. *Hi-C maps shown for four example loci. The intensity of red indicates the degree of interaction in the Hi-C map. TADs are annotated as black triangles in lower half of the Hi-C map.*

The smallest TADs called with TopDom were 10 kb, which is the equivalent of a single genomic bin. TADs which are smaller than several bins long are unlikely to represent real TADs so TopDom TADs <90 kb were filtered from the analysis. This corresponds to the ~5% smallest TopDom autosomal TADs in each dataset (Figure 3.2). The minimum length of Arrowhead TADs was 120 kb so no filtering was necessary.

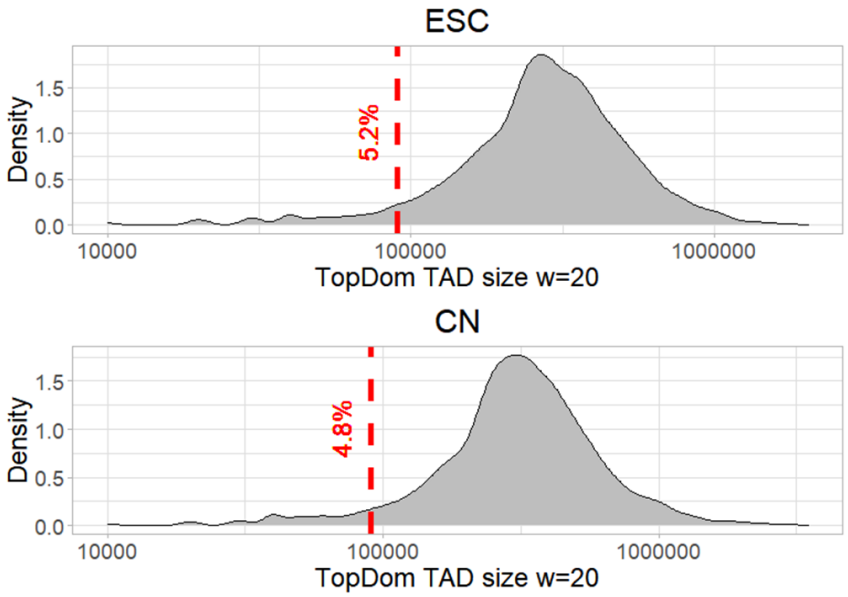


Figure 3.2: Size distribution of TopDom autosomal TADs in ESC (top) and cortical neuron (bottom). *The red dashed line indicates the position of the <90 kb size filtering threshold. The percentage indicates the percentage of TADs removed by the threshold for each dataset.*

3.3.2. General features of ESC and cortical neuron TADs

Overall I detected 8371 (median size 0.29Mb) and 5950 (median size 0.29Mb) TADs in ESCs, and 8001 (median size 0.32Mb) and 5430 (median size 0.33Mb) autosomal TADs in cortical neuron with Arrowhead and TopDom respectively. This suggests that for each TAD caller, there are more and smaller TADs in ESC compared to cortical neuron. This supports the finding using a TAD calling method based on an insulation score, from the paper which originally published the Hi-C data used in this chapter (Bonev et al. 2017) (Figure 3.3A and B). More TADs were called with Arrowhead than TopDom for both cell types but the TADs detected with Arrowhead and TopDom were of similar sizes. Throughout this chapter I focused on autosomal TADs, so unless explicitly stated, TADs refers to autosomal TADs only.

Next, I tested whether there was a correlation between the number of detected TADs on each chromosome and the size of the chromosome. As expected, there is a strong correlation

between the number of TADs and the size of chromosome for both cell types and TAD callers respectively (r ranging between 0.94- 0.98) (Figure 3.3C).

I next overlapped the detected TADs with protein-coding genes, by assigning genes to TADs if the entire length of the gene fell between the TAD boundaries. This results in a more stringent gene overlap method than the methods previously used (Nora et al. 2012; Flavahan et al. 2016; Muro, Ibn-Salem, and Andrade-Navarro 2019), which rely on assigning genes to a TAD based only on the positions of their gene start or TSS. However, the method employed in my analysis allows genes to be assigned to TADs with greater confidence and controls for the possibility that genes which span TAD boundaries may have different features. This is particularly important because recent research has indicated that TAD boundaries are not "sharp" and instead span "zones of transition" from one TAD to another (reviewed in (Chang, Ghosh, and Noordermeer 2020)). This means that it is difficult to assign a gene which spans a TAD boundary to one TAD or the other. Unexpectedly, there was no correlation between the size of TADs and the number of protein-coding genes they contained, for either TAD caller or cell type (r ranging between -0.04- 0.25). Most TADs contain relatively few genes (mean number of genes: ~2.74 and ~3.49 (Arrowhead), and ~2.15 and ~2.62 (TopDom) in ESC and cortical neuron, respectively). Deviating from the average, for each TAD caller and cell type a minority of TADs contained large numbers of genes. I investigated the gene content of some of these TADs and found that they contained large numbers of genes from paralogous gene families (e.g. olfactory, protocadherin and keratin) (Figure 3.3D). This supports previous findings which indicated that genes from functionally related families e.g. cytochrome, olfactory and protocadherin genes, commonly share a TAD, probably because of their need for a shared regulatory program (Dixon, Gorkin, and Ren 2016).

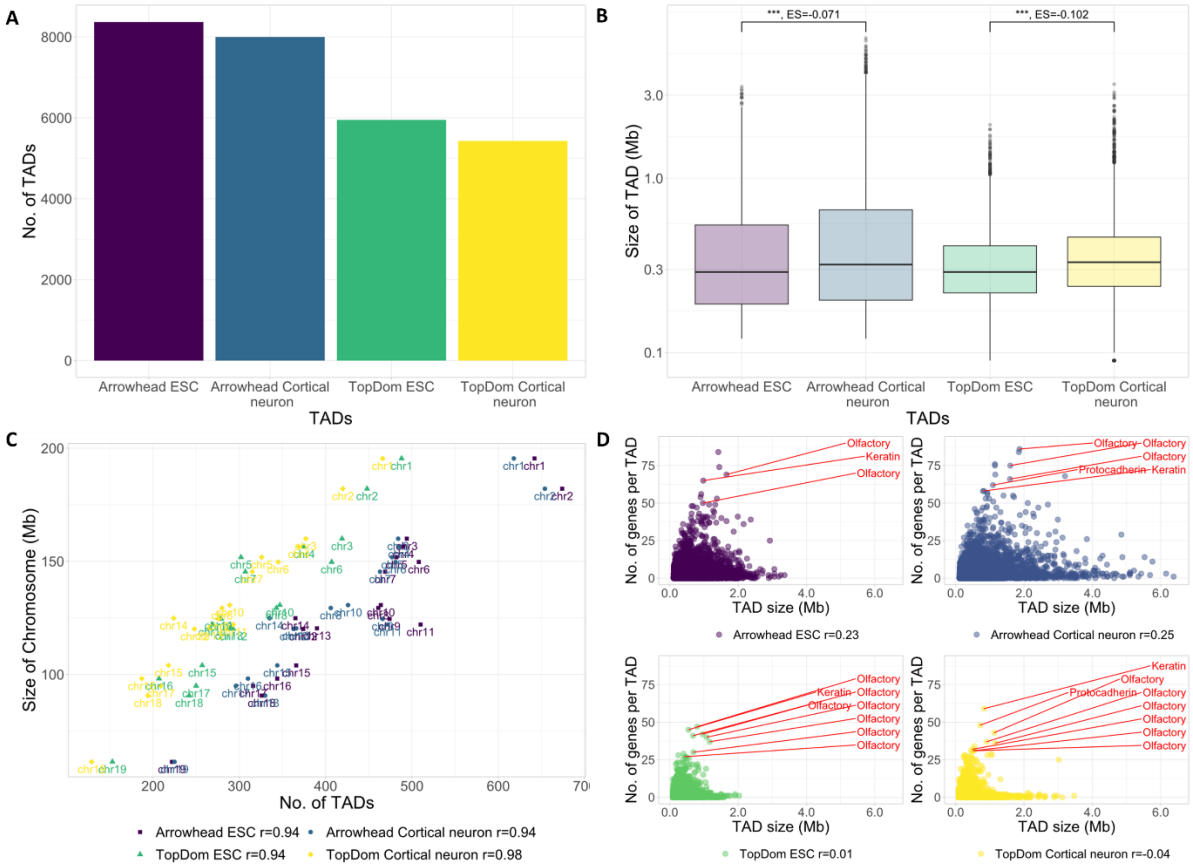


Figure 3.3: General features of autosomal ESC and cortical neuron TADs detected with Arrowhead or TopDom. *A) Number of TADs detected. B) Size of TADs detected shown in megabases on a log_{10} scale. Significant difference between ESC and cortical neuron was tested using wilcoxon test for each TAD caller, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across two p-values. C) Correlation between the size of chromosomes and the number of TADs detected, Pearson correlation coefficient (r) is indicated. D) Correlation between the number of protein-coding genes in TADs and the size of the TAD in megabases, Pearson correlation coefficient (r) is indicated. Several TADs with a large number of protein-coding genes were found to contain large numbers of genes from the same paralogous families (annotated).*

To assess the similarity between the TADs called using either Arrowhead or TopDom in ESC or cortical neuron, I calculated the percentage of TADs in each TAD set (focal) which had at least one equivalent TAD in the other TAD set (comparison). An equivalent TAD was defined as a TAD with boundaries within ±2 bins (20 kb) of the boundaries of the focal TAD. I found that between 21.38-45.5% of TADs detected by one TAD caller had an equivalent TAD detected by the other TAD caller. For Arrowhead TADs, 45.48% and 47.58% of TADs in ESC and cortical neuron respectively had an equivalent Arrowhead TAD in the other cell type. For TopDom TADs, 34.89% and 38.23% of TADs in ESC and cortical neuron respectively had an equivalent TopDom TAD in the other cell type (Table 3.2).

Table 3.2: Comparison of autosomal TADs between cell types and TAD callers. *Number (top) and percentage (bottom) of TADs in focal data set which have an equivalent TAD (both boundaries within ±2 bins (20 kb) of the boundaries of the focal TAD) in the comparison dataset. Percentages are highlighted on a colour scale from red (largest value) to white (smallest value). Cortical neuron (CN).*

Focal dataset	Comparison dataset			
	ESC Arrowhead	CN Arrowhead	ESC TopDom	CN TopDom
ESC Arrowhead	8371	3807	2707	1790
CN Arrowhead	3807	8001	1772	2404
ESC TopDom	2707	1772	5950	2076
CN TopDom	1790	2404	2076	5430
Focal dataset	**Comparison dataset**			
	ESC Arrowhead	CN Arrowhead	ESC TopDom	CN TopDom
ESC Arrowhead	100.00%	45.48%	32.34%	21.38%
CN Arrowhead	47.58%	100.00%	22.15%	30.05%
ESC TopDom	45.50%	29.78%	100.00%	34.89%
CN TopDom	32.97%	44.27%	38.23%	100.00%

3.3.3. Benchmarking of TAD randomisation strategy

To establish whether genes which share a TAD are more functionally similar than expected, I synthesised random TADs to serve as a null distribution (see Methods: 3.2.3). In brief, I adopted two strategies which allowed me to disentangle the similarity of genes sharing a TAD from the similarity of genes which can be attributed to their order and proximity in the linear genome. This is important as the order and proximity of genes in the linear genome has been shown to correlate with increased functional similarity (Hurst, Pál, and Lercher 2004; Thévenin et al. 2014). In the first strategy which I refer to as “random TADs”, each TAD was randomised to a new location on the same chromosome, which was the same size and

contained the same number of genes as the original TAD (Figure 3.4B). The placement of random TADs was also controlled to approximate the overlap structure of real TADs detected by each TAD caller. This randomisation strategy controls for the influence of the linear genome, as well as the general features of real TADs (Figure 3.5). In the second strategy, which I refer to as "random genome TADs", the position of the TADs in the genome is maintained, along with the number of genes they contain, but the identity of the genes are shuffled within each chromosome. In this randomisation strategy the impact of the linear genome is removed (Figure 3.4C). Each randomisation strategy was run 100 times for each cell type and TAD caller.

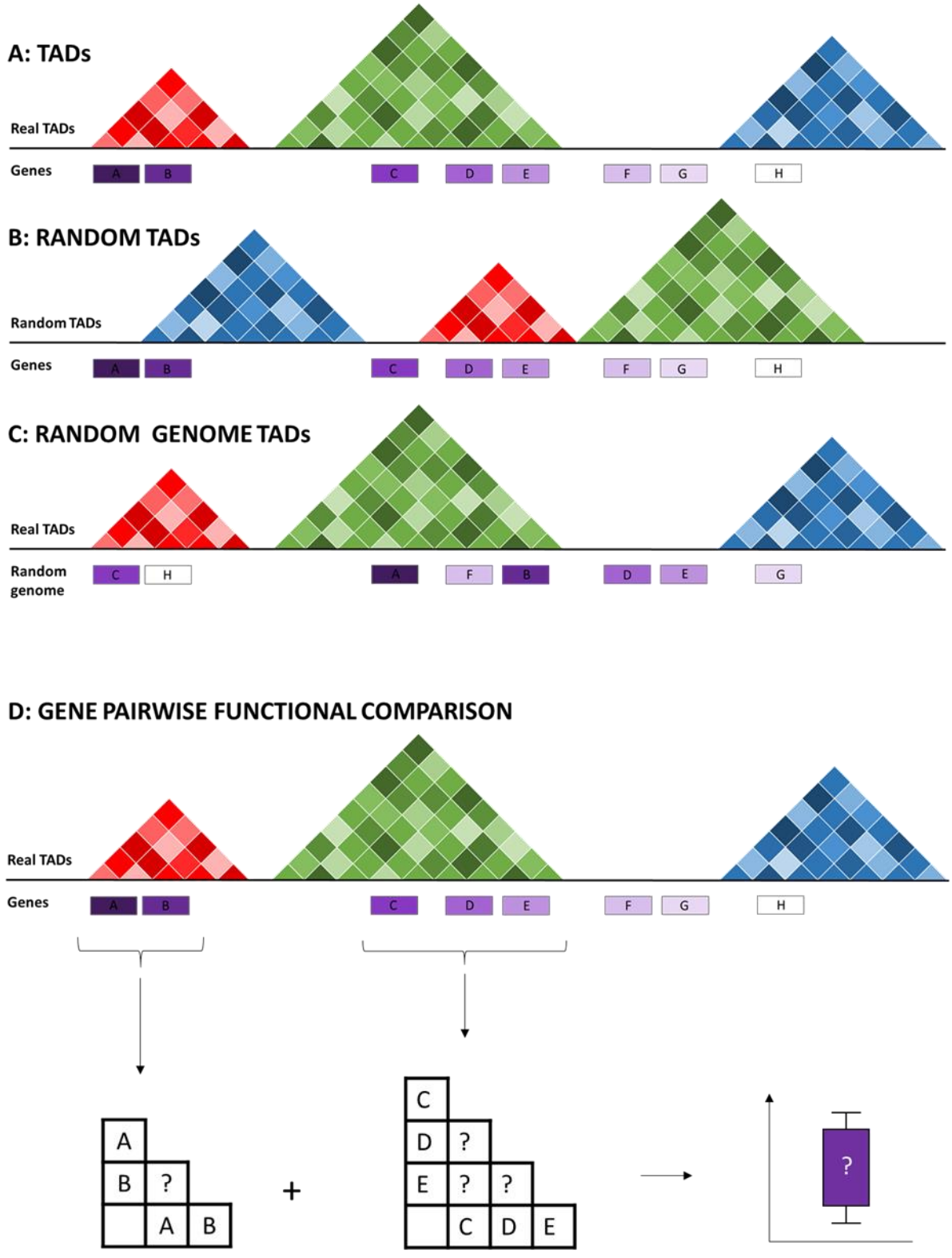


Figure 3.4: TAD randomisation and functional comparison strategy. *A) Schematic representation of real TADs. B) Random TADs were generated by randomising TADs to a new position of the same size which also contained the same number of genes as the original TAD. In this randomisation strategy the effect of linear gene order was maintained. C) Random genome TADs were generated by*

maintaining the positions of real TADs and the number of genes within them, but randomising the identity of the genes within each chromosome. In this randomisation strategy the effect of linear gene order was removed. D) In order to assess the functional similarity of genes in TADs, a pairwise approach was employed. In this approach the similarity of every possible pair of genes sharing a TAD was considered.

In order to assess the functional similarity of genes in TADs, random TADs and random genome TADs, a pairwise strategy was employed. In this strategy, the functional similarity of every possible pair of genes within the same TAD was calculated in order to create a distribution of scores (Figure 3.4D). As Arrowhead TADs, Arrowhead random TADs and Arrowhead random genome TADs can be nested or overlapping (Figure 3.5), if a pair of genes was present in multiple TADs, it was only considered once in the analysis. This means that for nested TADs (one TAD falls fully within another), pairs of genes from the largest of the nested TADs were considered, and for overlapping TADs (one TAD partially overlaps another), pairs of genes from both TADs were considered with pairs of genes from the overlapping region considered once. The distribution of scores in TADs was then compared to each of 100 random TADs and 100 random genome TADs and the median p value and median value of the functional measure was reported.

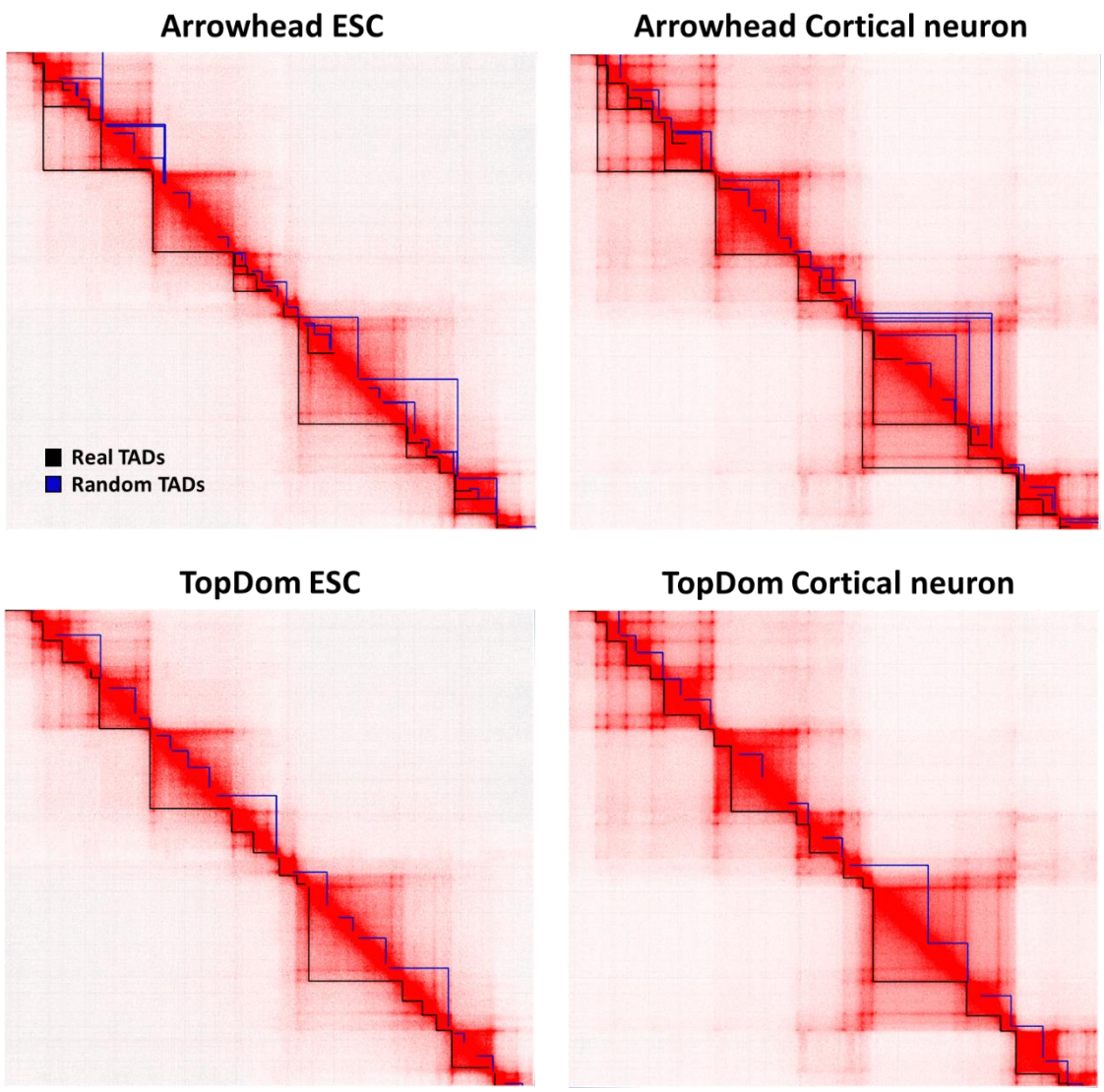


Figure 3.5: TADs and random TADs. *Hi-C map with TADs shown as triangles in black and an example set of random TADs shown as triangles in blue, for the equivalent region of chromosome 2. The intensity of red in the Hi-C map indicates the degree of interaction in the Hi-C map*

To test how well my TAD randomisation strategy captured features of real TADs compared to previously published methods, I generated random TADs by two further TAD randomisation strategies (see Methods: 3.2.3). The first strategy was as described in Rao et al.; in brief, this method randomised the position of TADs within a chromosome to a new position of the same size as long as it did not overlap a gap in the genome assembly (Rao et al. 2014). The second strategy was described in Nora et al.; in brief, TADs were randomised to new positions on the

same chromosome which contained the same number of genes and were the same size or smaller than the original TADs (Nora et al. 2012). As TopDom TADs are non-overlapping, I adapted both methods for generating TopDom random TADs to prevent TopDom random TADs from overlapping.

I first compared the distance between genes in TADs and in a single example of random TADs generated either by: the method outlined in this chapter, the Rao et al. method or the Nora et al. method. I found that irrespective of the TAD randomisation method used, the distance between genes in TADs was always significantly different to the distance between genes in random TADs (p-value < 0.001) (Figure 3.6A). However, for both TAD callers and cell types, the smallest effect size was observed for random TADs generated by my method, indicating that for this measure random TADs generated by my method are the most similar to real TADs.

I next investigated the number of genes in TADs compared to random TADs. As TADs with zero genes are not generated by the Nora et al. method, TADs with zero genes were removed from all TAD sets. There is a significant difference between the number of genes in TADs and random TADs produced by the Rao et al. method for both TAD callers and cell types (p-value< 0.001). In both my randomisation method and the Nora et al. method, TADs were randomised to new locations on the same chromosome which also contained the same number of genes as the original TAD. For this reason, the only difference between the distribution of the number of genes in TADs and random TADs produced by these methods comes from TADs which could not be placed (see Methods: 3.2.3). There is no significant difference between random Arrowhead TADs produced by either my method or the Nora et al. method, compared to real TADs indicating that for this measure, both methods produce random Arrowhead TADs

which are similar to real Arrowhead TADs. On the other hand, only random TopDom TADs produced by my method did not contain a significantly different number of genes compared to real TopDom TADs. Therefore, overall random TADs produced by my method best reflect the number of genes in real TADs (Figure 3.6B).

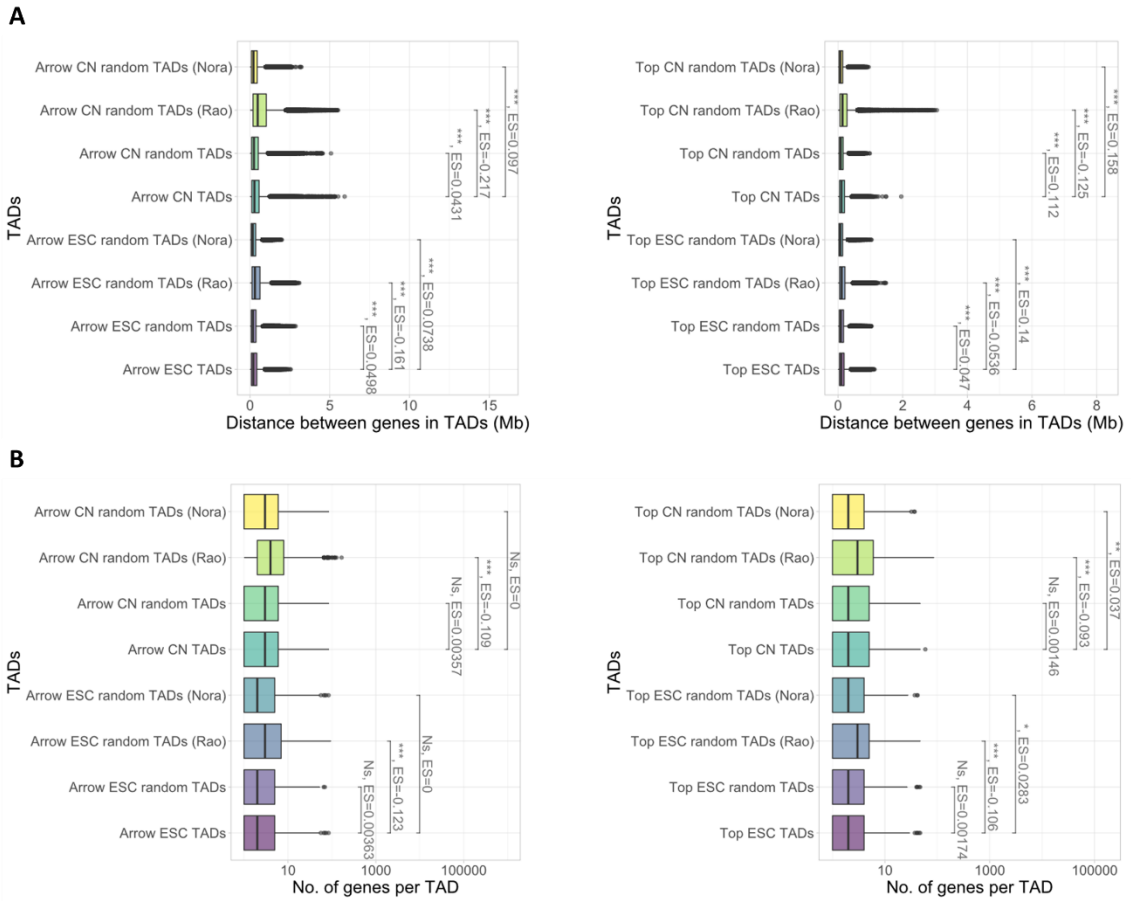


Figure 3.6: Comparison of my randomisation method with previously published methods (autosomal TADs). *A comparison of the features of TADs with an example set of random TADs produced by my method, by the method outlined in Rao et al. or by the method outlined in Nora et al. A) Distance between genes in TADs and random TADs produced by each method. Significances indicated by wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across 12 p-values. B) Number of protein-coding genes in TADs and random TADs produced by each method (log_{10} scale). TADs containing no genes were excluded. Significances indicated by wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **,*

*p<0.05 = *, NS= not significant, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across 12 p-values.*

Finally, I compared the overlap structure between real and random Arrowhead TADs produced by each method (TopDom TADs and TopDom random TADs do not overlap). Real Arrowhead TADs are characterised by a high degree of "nesting" overlaps (where one TAD falls completely within another TAD) and a depletion of "non-nesting" overlaps (one TAD partially overlaps another) (Figure 3.5, Figure 3.7 and Chapter 3: Supplementary Figure 3). I found that the overlap structure of real Arrowhead TADs was best approximated by random Arrowhead TADs generated using my method. Random Arrowhead TADs generated with either the Rao et al. or the Nora et al. method contain far more "non-nesting" overlaps than real Arrowhead TADs (Figure 3.7 and Chapter 3: Supplementary Figure 3).



Figure 3.7: Comparison of the overlap structure of random ESC TADs produced by my randomisation method and previously published methods (autosomal TADs). *Overlap structure of ESC TADs with an example set of random ESC TADs produced by my method, by the method outlined in Rao et al. or by the method outlined in Nora et al. Non-overlapping TADs were excluded. Remaining TADs were classified based on whether they took part in any "nested" (one TAD falls completely within another) or "non-nested" (one TAD partially overlaps another) overlaps.*

3.3.4. Features of CTCF and non-CTCF TADs

It has previously been shown that TAD boundaries are enriched for CTCF binding and TADs are thought to be formed by loop extrusion commonly involving CTCF (Dixon et al. 2012; Rao et al. 2014; Sanborn et al. 2015; Nuebler et al. 2018; Fudenberg et al. 2016). To assess the enrichment of CTCF at the boundaries of TADs in this analysis, I tested for the presence of CTCF ChIP-seq peaks within ±10 kb of TAD and random TAD boundaries (Figure 3.8A). I observed that ~ 62%, 59%, 28%, 35% of ESC Arrowhead TADs, ESC TopDom TADs, cortical neuron Arrowhead TADs and cortical neuron TopDom TADs respectively, had a CTCF ChIP-seq peak within in ±10 kb of both TAD boundaries. This is compared to ~ 29%, 31%, 4.5%, 4.9% of ESC Arrowhead random TADs, ESC TopDom random TADs, cortical neuron Arrowhead random TADs and cortical neuron TopDom random TADs respectively. This suggests that CTCF ChIP-seq peaks are enriched at the boundaries of TADs more than would be expected if TADs were randomly placed, which supports previous findings (Dixon et al. 2012; Rao et al. 2014; Sanborn et al. 2015; Nuebler et al. 2018; Fudenberg et al. 2016). This result also suggests that for both TAD callers, a higher proportion of ESC TADs had CTCF ChIP-seq peaks at both boundaries than cortical neuron TADs. This could perhaps suggest that there is a reduction in the number of TADs formed by loop extrusion involving CTCF over cell differentiation. I further detected 30%, 30%, 46% and 42% of TADs which had a CTCF ChIP-seq peak near only one boundary and 7.9%, 12%, 27% and 23% of TADs which did not have a ChIP-seq peak near either boundary, in ESC Arrowhead TADs, ESC TopDom TADs, cortical neuron Arrowhead TADs and cortical neuron TopDom TADs respectively. This suggests that these domains may not be formed by loop extrusion involving CTCF and therefore may represent a different class of domain. In order to assess the features of these TADs separately, I defined CTCF TADs and non-CTCF TADs. CTCF TADs are defined as TADs with ChIP-seq peaks within ±10 kb of both

TAD boundaries, whereas non-CTCF TADs are TADs which had a ChIP-seq peak within ±10 kb of only one or neither boundary (Figure 3.8B).

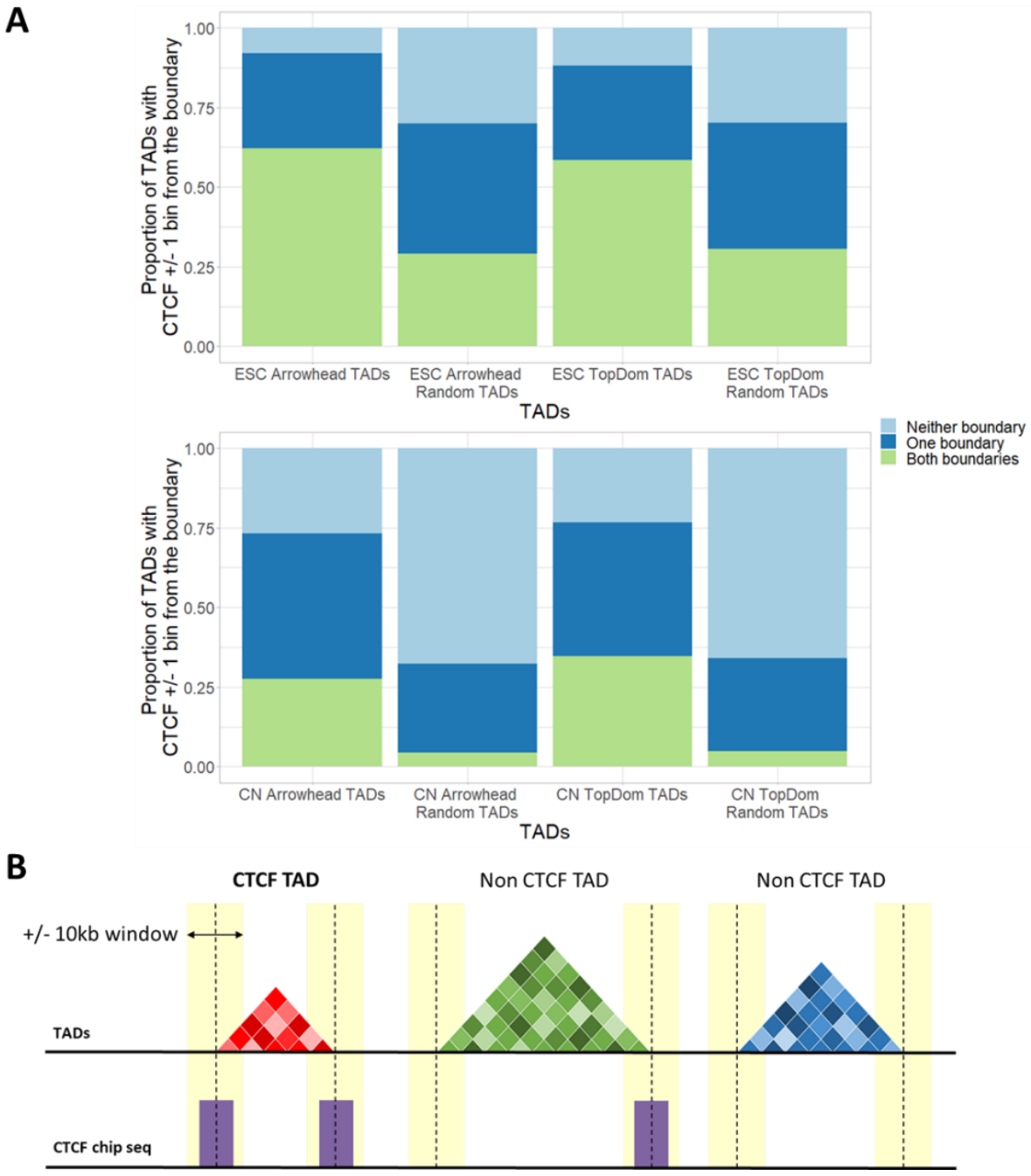


Figure 3.8: CTCF ChIP-seq binding at autosomal TAD boundaries*. A) Proportion of real and an example set of random TADs which have a CTCF ChIP-seq binding peak within ±10 kb of both TAD boundaries, only one boundary, or neither boundary, in ESC (top) and cortical neuron (CN) (bottom). B) Schematic outlining the definition of CTCF and non-CTCF TADs. CTCF TADs are defined as TADs with a CTCF ChIP-seq binding peak within ±10 kb of both TAD boundaries. Non-CTCF TADs are defined as TADs with a CTCF ChIP-seq binding peak within ±10 kb of only one or neither boundary.*

In order to assess the similarity of CTCF and non-CTCF TADs between TAD callers, CTCF TADs and non-CTCF TADs were compared. As above, an equivalent TAD was defined as a TAD with boundaries within ±2 bins (20 kb) of the boundaries of the focal TAD. I found that between 23.74%- 45.99% of CTCF TADs called by one TAD caller had an equivalent CTCF TAD called by the other TAD caller and that 15.13%- 28.72% of non-CTCF TADs called by one TAD caller had an equivalent non-CTCF TAD called by the other TAD caller, suggesting that CTCF TADs are more similar between TAD callers than non-CTCF TADs (Chapter 3: Supplementary Table 1).

To further investigate the biological context of CTCF and non-CTCF TADs, I generated A and B compartment profiles binned at 50 kb (see Methods: 3.2.4) for both the ESC and cortical neuron Hi-C datasets. These compartment annotations were intersected with TAD annotations (intersection carried out by Dr Michelle Simon) to calculate the percentage of CTCF and non-CTCF TADs which were in each compartment (Table 3.3). For both TAD callers and cell types, CTCF TADs were most commonly in the A compartment (~44%, 40%, 20% and 23% of TADs are CTCF TADs in the A compartment compared with ~18%, 18%, 7% and 12% of TAD which are CTCF TADs in the B compartment in ESC Arrowhead TADs, ESC TopDom TADs, cortical neuron Arrowhead TADs and cortical neuron TopDom TADs respectively). In contrast, ESC non-CTCF TADs were more commonly found in the B compartment (percentage of TADs which are non-CTCF in compartment A vs B in ESC Arrowhead TADs: ~11% vs 26%, ESC TopDom TADs: ~8% vs 33%) and cortical neuron non-CTCF TADs were split relatively equally between A and B compartments (cortical neuron Arrowhead TADs: ~42% vs 31% and cortical neuron TopDom TADs: ~30% vs 35%).

Table 3.3: Compartment identity of CTCF and non-CTCF autosomal TADs. *CTCF TADs are defined as TADs with a CTCF ChIP-seq binding peak within ±10 kb of both TAD boundaries. Non-CTCF TADs are defined as TADs with a CTCF ChIP-seq binding peak within ±10 kb of only one or neither boundary. Compartment annotations are derived from the eigenvector corresponding to the first principle component of the Pearson correlation matrix of the Hi-C map. Percentages are highlighted on a colour scale from red (largest value) to white (smallest value).*

Arrowhead				
TAD	**Non-CTCF**		**CTCF**	
Compartment	**A**	**B**	**A**	**B**
Cortical neuron	42%	31%	20%	7%
ESC	11%	26%	44%	18%
TopDom				
TAD	**Non-CTCF**		**CTCF**	
Compartment	**A**	**B**	**A**	**B**
Cortical neuron	30%	35%	23%	12%
ESC	8%	33%	40%	18%

3.3.5. Increased distance between pairs of genes sharing a TAD

I investigated the distance between gene start coordinates in TADs and random TADs (Figure 3.9A). For both TAD callers and cell types, I found that pairs of genes in the same TAD are significantly further apart than pairs of genes in the same random TADs (median p-value< 0.001). This indicates that genes in TADs were further apart than would be expected if TADs were randomly placed. Next, I split TADs into CTCF TADs and non-CTCF TADs and reassessed the distance between gene starts compared to random CTCF TADs and random non-CTCF TADs respectively. Similarly to the pattern observed in all TADs, for both cell types and TAD callers, pairs of genes in CTCF and non-CTCF TADs were significantly further apart than in random CTCF or non-CTCF TADs respectively (median p-value< 0.001) (Figure 3.9B and C).

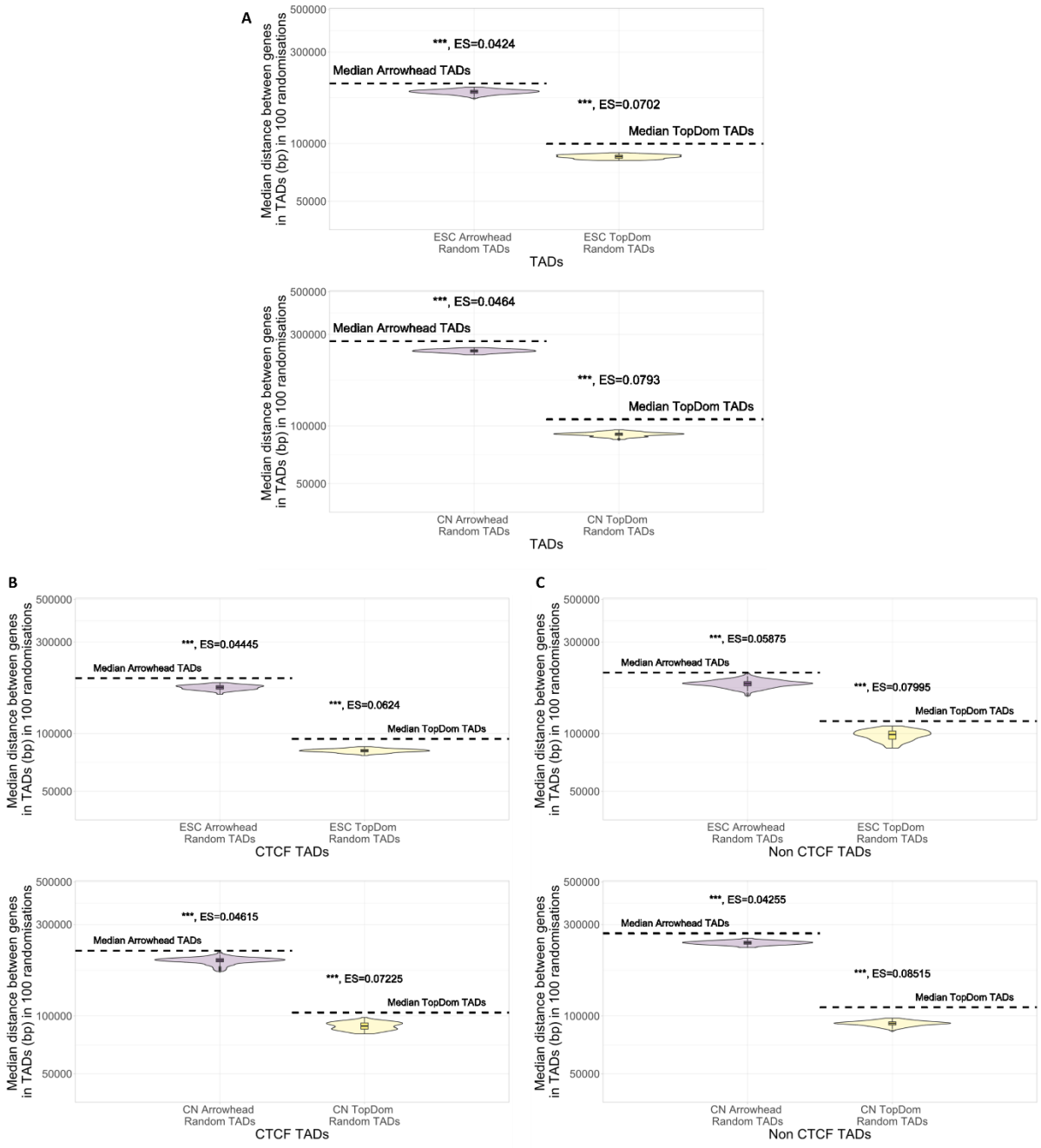


Figure 3.9: Median distance between genes in TADs vs random TADs (autosomal). *Median distance between gene start coordinates in TADs (dotted line) and the median distance between genes in 100 sets of random TADs (plotted on a log10 scale). Significance indicated by median p-value of Wilcoxon test between real TADs and each of 100 sets of random TADs. FDR correction applied across four median p-values (i.e. within A, B and C separately). FDR corrected median p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Median of effect sizes calculated using r for Wilcoxon between real and each of 100 sets of random TADs. A) All autosomal TADs. B) CTCF TADs. C) non-CTCF TADs.*

3.3.6. The relationship between gene paralogous families and TADs

In Figure 3.3D I noted several examples of TADs which contained a large number of genes from the same paralogous families. Genes from the same paralogous family are also highly functionally similar (Stamboulian et al. 2020), so genes within TADs could have increased functional similarity because of shared ancestry. To test this, I assessed the genome-wide enrichment of paralogous gene pairs in TADs, random TADs and random genome TADs. I found a significantly higher proportion of paralogous gene pairs in TADs compared to random genome TADs for both TAD callers and cell types (median p-value <0.001) (Figure 3.10A and Chapter 3: Supplementary Figure 4A). This is an expected result as paralogous gene pairs are known to cluster together in the linear genome and the influence of the linear genome is removed in random genome TADs. I next compared TADs with random TADs, again finding a significantly higher proportion of paralogous gene pairs in TADs compared to random TADs for both TAD callers and cell types (median p-value <0.001) (Figure 3.10A and Chapter 3: Supplementary Figure 4A). This result supports the findings of Ibn Salem et al. who reported an enrichment for paralogous gene pairs to fall in the same TADs (Ibn-Salem, Muro, and Andrade-Navarro 2017).

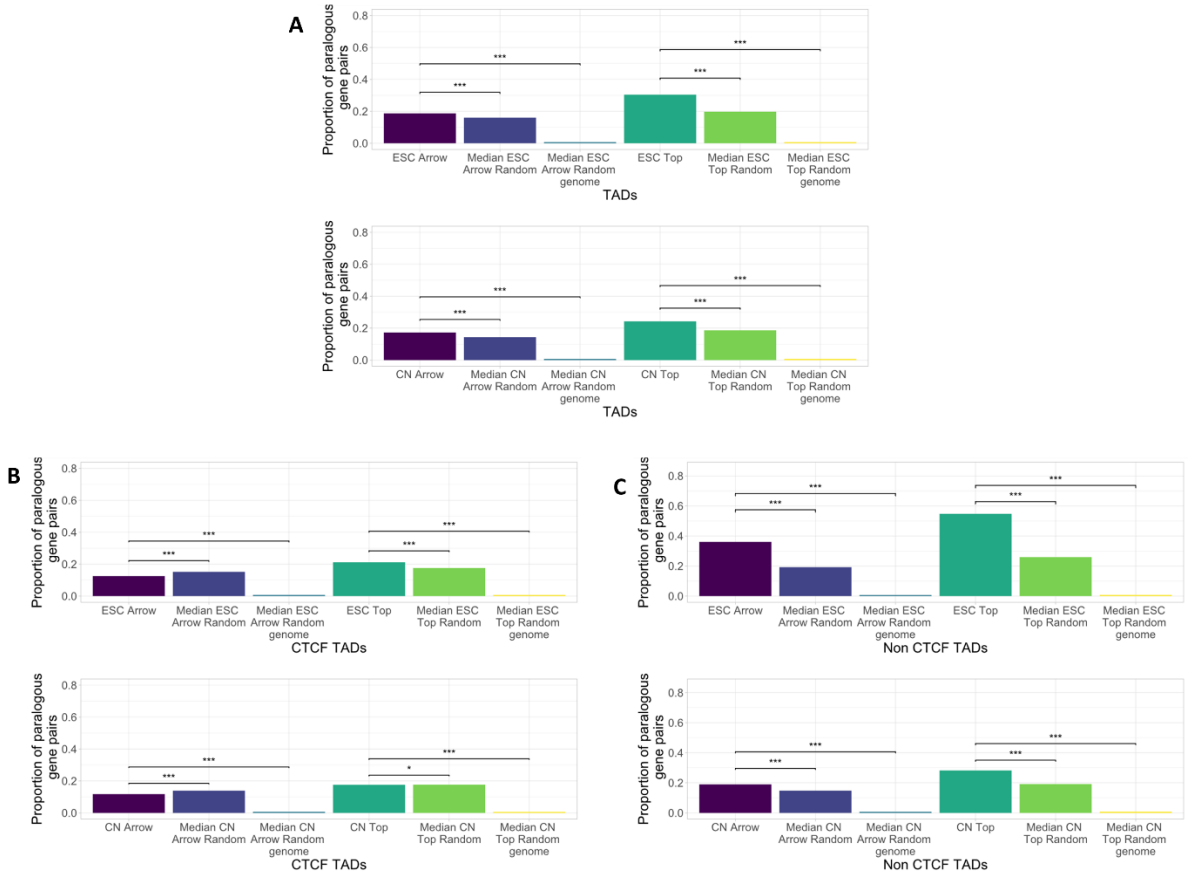


Figure 3.10: Proportion of paralogous gene pairs in TADs vs proportion in random TADs (autosomal). *Significance indicated by median p-value of Fisher's exact test between real TADs and each of 100 sets of random/random genome TADs. FDR correction applied across eight median p-values (i.e. within A, B and C separately). FDR corrected median p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *. A) Proportion in all autosomal TADs and the median proportions in 100 sets each of random TADs and random genome TADs. B) Proportion in CTCF autosomal TADs and the median proportions in 100 sets each of random CTCF TADs and random genome CTCF TADs. C) Proportion in non-CTCF autosomal TADs and the median proportions in 100 sets each of random non-CTCF TADs and random genome non-CTCF TADs.*

I investigated the features of TADs which contained at least one pair of paralogs or no pairs of paralogs. Arrowhead TADs which contained at least one pair of paralogs were larger than Arrowhead TADs with no pairs of paralogs (p-value <0.001). TopDom TADs which contained at least one pair of paralogs were also slightly larger on average than TADs containing no pairs

of paralogs, however the difference was not statistically significant (Figure 3.11A). I next compared the size of TADs containing at least one pair of paralogs or no pairs of paralogs to the expected size in random TADs and random genome TADs. For both TAD callers and cell types I found that random genome TADs containing a pair of paralogs were significantly larger than TADs containing a pair of paralogs (median p-value<0.05, Figure 3.11B). This result is likely due to the increased chance of a larger TAD containing any pairs of paralogous genes in a randomised genome. When comparing the size of TADs and random TADs, for both cell types and TAD callers, I found that random TADs containing a pair of paralogs were significantly smaller than TADs containing a pair of paralogs (median p-value<0.05, Figure 3.11B). This indicates that TADs containing pairs of paralogs are larger than would be expected if TADs were randomly placed. I next compared the size of TADs which did not contain a pair of paralogs to the expected size in random TADs or random genome TADs. I found no significant difference in any comparison, suggesting that the size of TADs which do not contain a pair of paralogs is not different to the expected size if TADs are randomly placed or the effect of linear gene order is removed (Figure 3.11B).

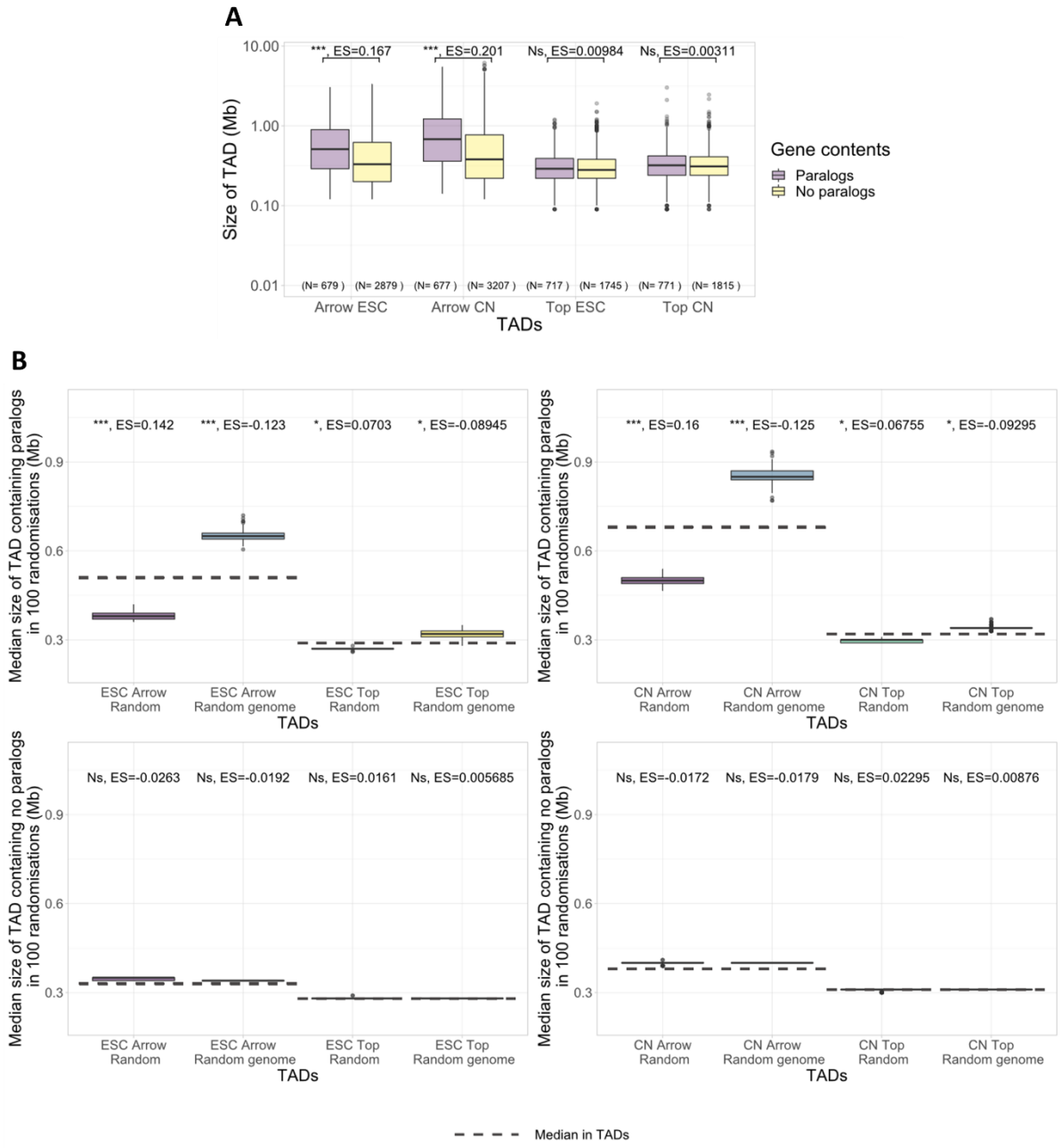


Figure 3.11: Size of autosomal TADs which contain a pair of paralogs. *A) Size of TADs containing >1 gene and a pair of paralogous genes or no pairs of paralogous genes. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, NS= not significant, ES= effect sizes calculated using r for Wilcoxon. FDR correction applied across four p-values. B) Median size of TADs (dotted line) vs 100 sets of random TADs or random genome TADs containing >1 gene and a pair of paralogous genes (top) and no pairs of paralogous genes (bottom). Significance indicated by median p-value of Wilcoxon test between real TADs and each of 100 sets of random/random genome TADs. FDR correction applied across 16 median p-values. FDR corrected median p-value: p<0.001 =*

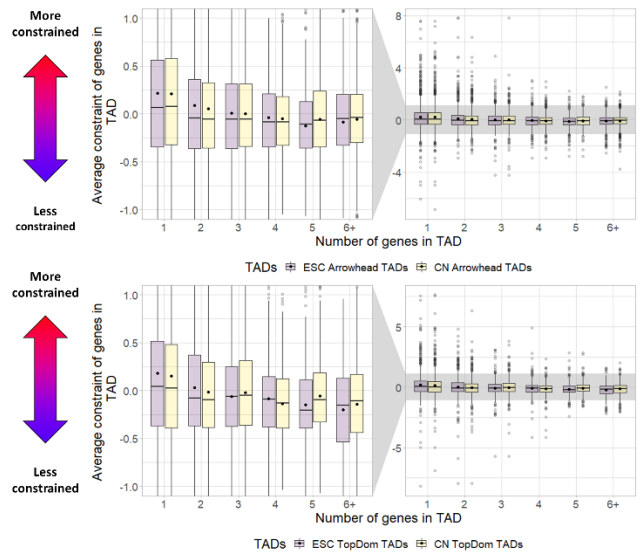
****, p<0.01 = **, p<0.05 = *, NS= not significant, ES= Median of effect sizes calculated using r for Wilcoxon between real and each of 100 sets of random TADs.*

To further investigate the relationship between TADs and paralogy, I split TADs into CTCF and non-CTCF TADs and tested their association with paralogous gene pairs separately. Similarly to the results found for all TADs, I observed a significantly higher proportion of paralogous gene pairs in non-CTCF TADs compared to both non-CTCF random TADs and non-CTCF random genome TADs (median p-value <0.001) (Figure 3.10C and Chapter 3: Supplementary Figure 4C). This suggests that paralogous gene pairs are more likely to fall within non-CTCF TADs than would be expected if they were randomly placed. However, I found that paralogous gene pairs were commonly depleted in CTCF TADs (with the exception of ESC TopDom CTCF TADs) compared to random CTCF TADs (median p-value of CTCF TADs vs random CTCF TADs <0.001 for ESC Arrowhead and cortical neuron Arrowhead and <0.05 for cortical neuron TopDom, Figure 3.10B and Chapter 3: Supplementary Figure 4B). This suggests paralogous genes pairs are commonly depleted in CTCF TADs and that TADs detected by Arrowhead and TopDom may be comprised of two functional groups of TADs. As CTCF TADs are bounded by CTCF, they are likely to have been formed by loop extrusion involving CTCF. In contrast, non-CTCF TADs may have been formed by other mechanisms and therefore may represent other types of domain e.g. compartment domains or TADs formed by loop extrusion not involving CTCF (Beagan and Phillips-Cremins 2020).

3.3.7. Evolutionary constraint of genes singly or co-occupying a TAD

To further investigate the relationship between the evolutionary forces acting on genes in TADs, I assessed the average selective constraint of genes in TADs with 1, 2, 3, 4, 5, 6+ genes. Gene selective constraint quantifies the degree of purifying selection pressure acting on a gene. Genes with greater functional importance are expected to be under stronger purifying selection; therefore, gene constraint can be used as a proxy for functional importance. I found that genes which singly occupy a TAD are significantly more constrained than the mean constraint of genes co-occupying a TAD (Figure 3.12A-B, Chapter 3: Supplementary Table 2).

A



B

No. of genes in TAD	No. of genes in TAD	ESC Arrowhead TADs P value	CN Arrowhead TADs P value	ESC TopDom TADs P value	CN TopDom TADs P value
1	2	0.001903109	4.16157E-05	0.01663705	0.01498467
1	3	0.00027194	4.16157E-05	0.00020306	0.06138803
1	4	4.96001E-05	1.39716E-06	4.4102E-05	0.00013832
1	5	5.96656E-07	4.16157E-05	1.0672E-06	0.01579016
1	6+	2.3967E-11	1.03675E-10	1.0252E-14	3.2517E-08
2	3	0.436576609	0.818985786	0.159076	0.66620451
2	4	0.161127035	0.352631112	0.03732384	0.07961412
2	5	0.007742287	0.392464738	0.00131748	0.66620451
2	6+	0.015798336	0.486836008	3.9532E-07	0.01498467
3	4	0.450552465	0.392464738	0.31568725	0.03421304
3	5	0.056191547	0.486836008	0.02909119	0.42303686
3	6+	0.213739286	0.664941756	0.00056451	0.00799969
4	5	0.213739286	0.960470859	0.159076	0.31960654
4	6+	0.917145233	0.392464738	0.03426772	0.85931841
5	6+	0.161127035	0.504822614	0.67946238	0.21868246

Figure 3.12: Average constraint of genes sharing an autosomal TAD. *A) Distribution of the mean constraint score of protein coding genes in TADs with 1, 2, 3, 4, 5 and 6+ protein coding genes. The mean of the distribution is indicated by the dot. The left panel zooms in on the grey section of the right panel. B) Table showing the FDR corrected p-value calculated with the Wilcoxon test, of the difference between groups in A. Significant p-values (p<0.05) are highlighted in red.*

Genes singly occupying TADs are also more constrained than genes singly occupying random genome TADs, as well as random TADs (with the exception of cortical neuron TopDom random TADs) (FDR corrected median p-value of genes singly occupying TADs vs genes singly occupying TADs in random TADs <0.05 or random genome TADs <0.001, Figure 3.13). This suggests that the increased constraint cannot be explained by the linear genome alone. Suggesting that genes which singly occupy TADs may be under higher selective constraint and may be more functionally important than genes which co-occupy a TAD. This in turn suggests that the protection from aberrant regulation of functionally important genes, implied by being in a private TAD, is under selective constraint.

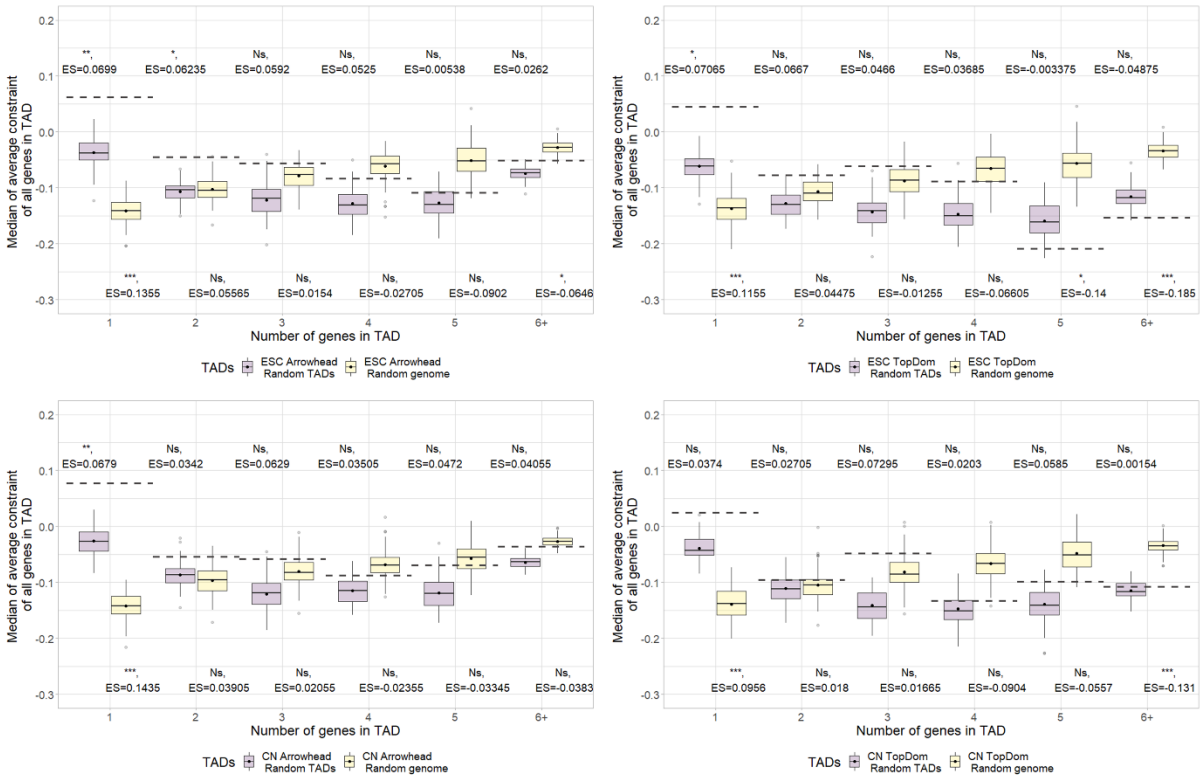
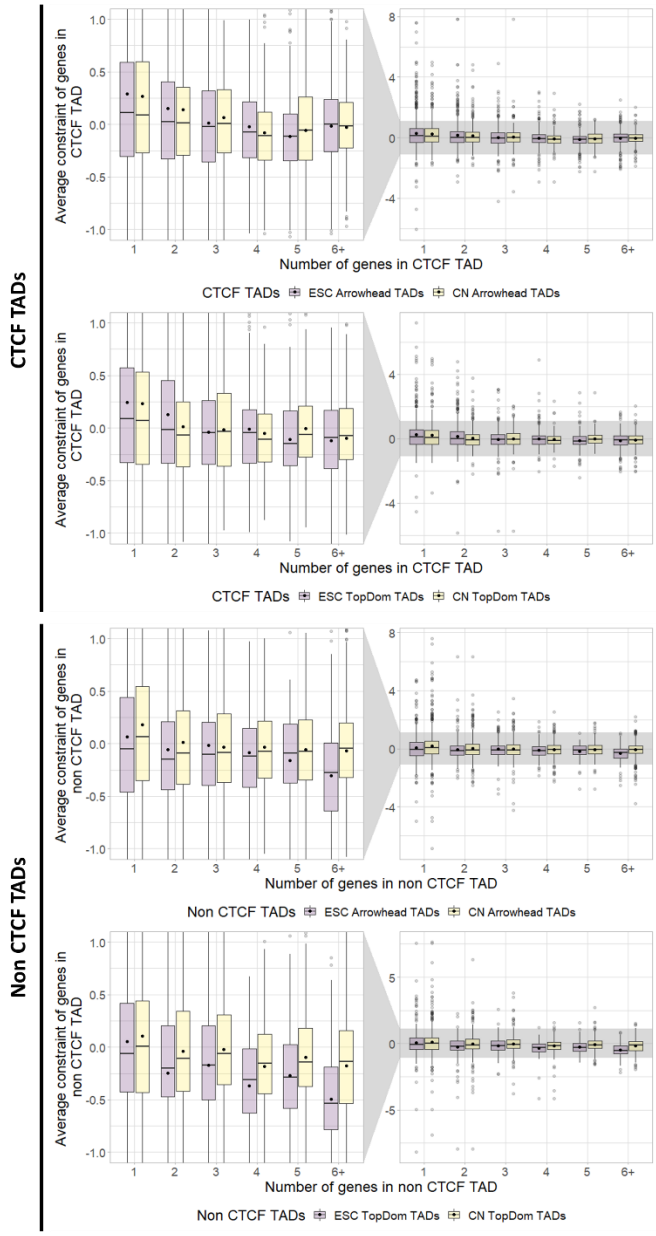


Figure 3.13: Average constraint of genes in random TADs and random genome TADs (autosomal). *The average gene constraint score was calculated for each TAD and random/random genome TAD and then binned based on whether the TAD contained 1, 2, 3, 4, 5 or 6+ protein coding genes. For TADs the median score for each bin is shown as the dotted line. For random TADs (purple) and random genome TADs (yellow), the median score for each bin across 100 sets of random/random genome TADs is shown as a distribution. The mean of the distribution is indicated by the dot. Significance is indicated by median p-values (median p-values FDR corrected individually for n=6 tests, corresponding to the number of bins) of Wilcoxon test between real TADs and each of 100 sets of random/random genome TADs: p<0.001 = ***, p<0.01 = **, p<0.05 = *, NS= not significant, ES= Median of effect sizes calculated using r for Wilcoxon between real and each of 100 sets of random TADs. P-values and effect sizes above the box plots indicate difference between random TADs and TADs and values below the boxplots indicate difference between random genome TADs and TADs.*

I then tested whether the relationship between TADs and gene constraint was observable in CTCF and non-CTCF TADs. When considering either CTCF TADs or non-CTCF TADs, as seen above, I found that generally the constraint of genes in singly occupied TADs is significantly higher relative to the average constraint of genes co-occupying a CTCF or non-CTCF TAD (Figure 3.14, Chapter 3: Supplementary Table 3 and Chapter 3: Supplementary Table 4).



No. of genes in TAD	No. of genes in TAD	ESC Arrowhead TADs P value	CN Arrowhead TADs P value	ESC TopDom TADs P value	CN TopDom TADs P value
1	2	0.004905065	0.098692601	0.108790502	0.071582203
1	3	8.25583E-06	0.066899563	0.000121037	0.169317446
1	4	5.01173E-06	0.000125136	0.000370042	0.071582203
1	5	2.96132E-08	0.020425601	3.19428E-05	0.169317446
1	6+	7.86555E-08	0.000298767	9.932E-09	0.004106037
2	3	0.086072709	0.781194238	0.043326337	0.902431719
2	4	0.031371456	0.01270249	0.066003482	0.871440638
2	5	0.000337737	0.213602465	0.002387032	0.911028586
2	6+	0.106622991	0.213602465	0.000121037	0.839269779
3	4	0.548966121	0.02629318	0.813141009	0.839269779
3	5	0.039671906	0.322803766	0.116896281	0.911028586
3	6+	0.577076225	0.336598279	0.116896281	0.644231357
4	5	0.129426566	0.425223072	0.141238298	0.871440638
4	6+	0.128549764	0.031198786	0.231679665	0.958475887
5	6+	0.001251964	0.53451915	0.567358267	0.839269779

No. of genes in TAD	No. of genes in TAD	ESC Arrowhead TADs P value	CN Arrowhead TADs P value	ESC TopDom TADs P value	CN TopDom TADs P value
1	2	0.10939455	0.000723781	0.01234586	0.185068337
1	3	0.774214108	0.000832763	0.09621708	0.332918951
1	4	0.457798142	0.002907646	0.00179673	0.004288111
1	5	0.457798142	0.002907646	0.00507668	0.087192392
1	6+	2.61593E-08	1.34629E-06	5.4988E-12	2.01104E-05
2	3	0.457798142	0.98339686	0.91226553	0.686256165
2	4	0.774214108	0.98339686	0.12837436	0.091467532
2	5	0.774214108	0.98339686	0.1773788	0.461845621
2	6+	0.000626938	0.993939667	2.5838E-06	0.011314496
3	4	0.774214108	0.98339686	0.14725395	0.052213545
3	5	0.774214108	0.98339686	0.20077935	0.332918951
3	6+	0.000110921	0.98339686	0.00014688	0.007176632
4	5	0.968511131	0.98339686	0.92160687	0.461845621
4	6+	0.004054607	0.98339686	0.03128328	0.686256165
5	6+	0.013254503	0.98339686	0.03128328	0.332918951

Figure 3.14: Average constraint of genes sharing autosomal CTCF TADs or non-CTCF TADs. *Left: Distribution of the mean constraint score of protein-coding genes in CTCF TADs (top) or non-CTCF TADs (bottom) with 1, 2, 3, 4, 5 and 6+ protein-coding genes. The mean of the distribution is indicated by the dot. The left panels of the graphs zooms in on the grey section of the right panel. Right: Tables showing the FDR corrected p-value calculated with the Wilcoxon test, of the difference between groups shown in the graphs. Significant p-values (p<0.05) are highlighted red.*

In order to assess which biological processes genes which singly occupy a TAD are involved in, I carried out a functional enrichment analysis (see Methods: 3.2.8) using biological process GO terms (Figure 3.15). I found that genes which singly occupy a TAD are highly enriched for developmental processes, genes which occupy a TAD with one other gene (double occupancy) are also enriched for developmental processes but to a lesser extent and genes which occupy a TAD with two other genes (triple occupancy) are less enriched for developmental processes still. For example, "system development" is the most significant GO term associated with singly occupied Arrowhead TADs in both ESCs and cortical neurons (p-value= 8.96×10^{-48} and 4.02×10^{-43}, respectively), but it is less significantly associated in doubly occupied or triply occupied TADs (doubly occupied: p-value= 5.31×10^{-20} and 3.11×10^{-20}, triply occupied: p-value= 9.30×10^{-06} and 1.35×10^{-06} for ESC and cortical neuron, respectively).

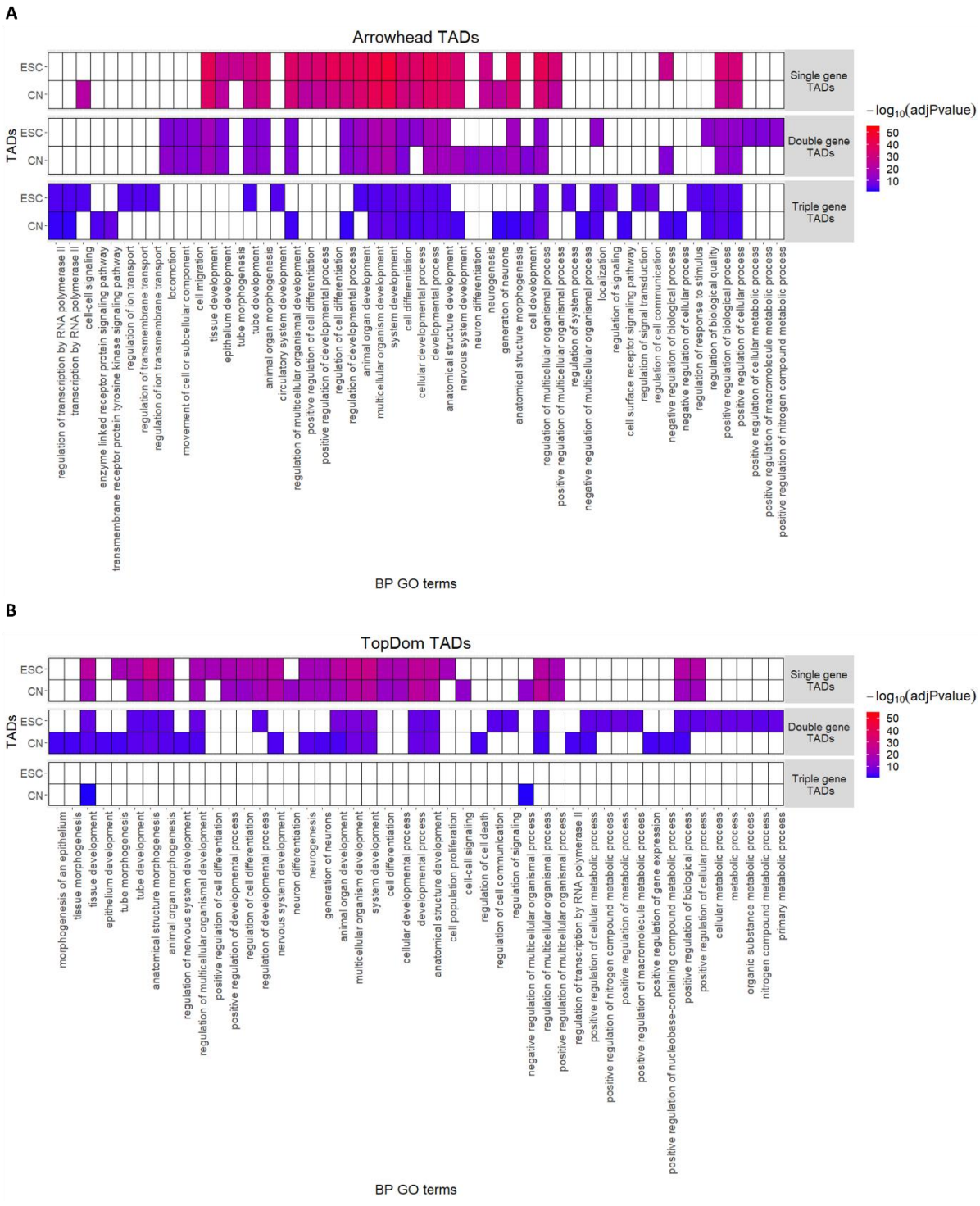


Figure 3.15: Biological process GO functional enrichment of genes singly, double or triply occupying a TAD. *TADs detected with Arrowhead (A) or TopDom (B). Only the (max) top 25 most significant GO terms passing "gSCS" corrected p-value threshold of <0.05 are shown for each gene set.*

I repeated the enrichment analysis using genes which singly, doubly and triply occupy random TADs in order to establish whether randomly placed TADs with similar features (e.g. only one gene in the length of the TAD) have a similar pattern of enrichment (Figure 3.16). I found that genes that singly occupy a random TAD are also enriched for developmental processes but to a lesser degree than TADs. This suggests that the enrichment for developmental function observed in genes that singly occupy a TAD cannot be explained by the linear genome alone.

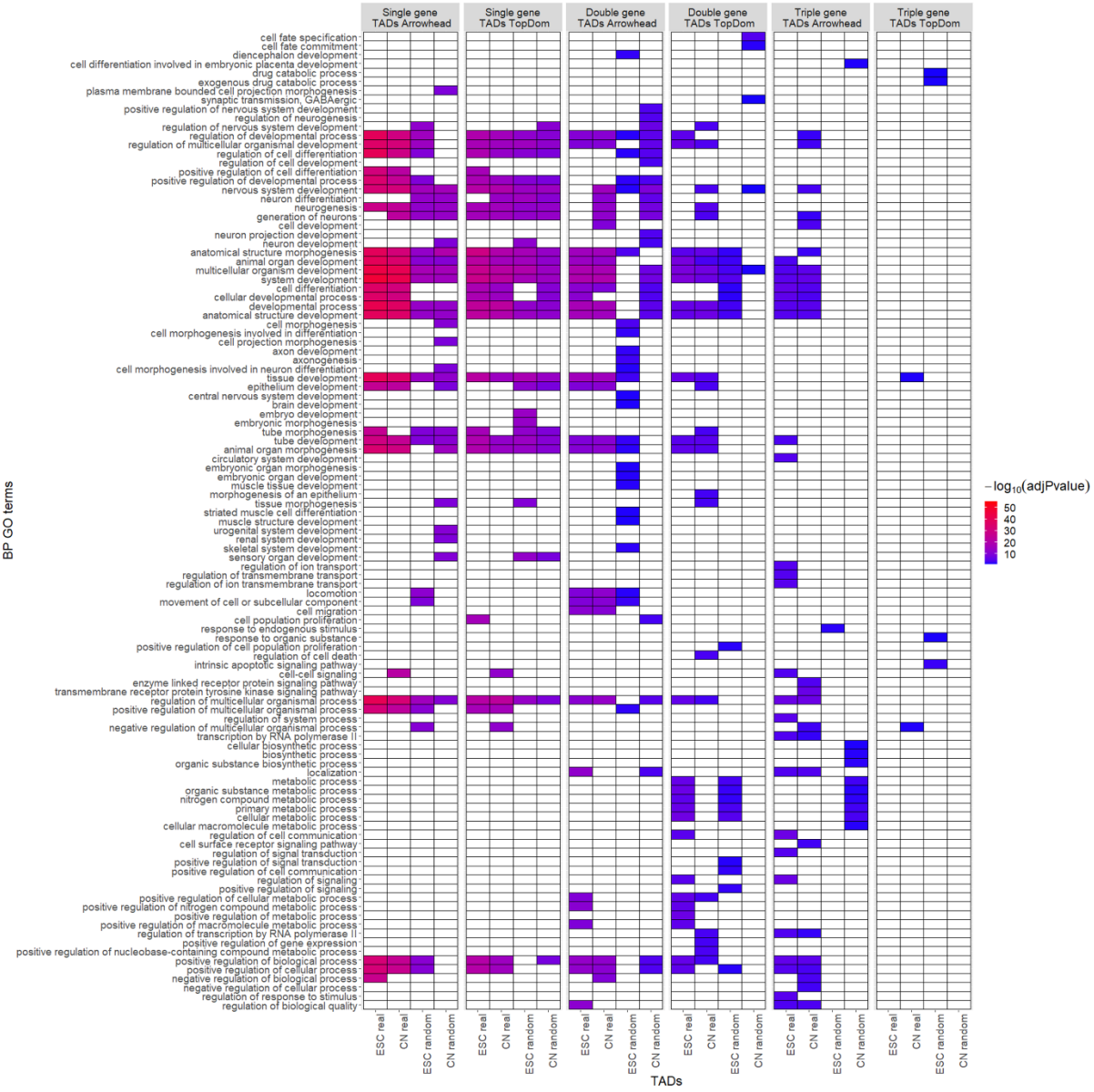


Figure 3.16: Biological process GO term enrichment of genes singly, doubly or triply occupying autosomal TADs compared to random TADs. *Enrichments shown for genes occupying Arrowhead TADs, TopDom TADs, an example set of random Arrowhead TADs and an example set of random TopDom TADs. Only the (max) top 25 most significant GO terms passing "gSCS" corrected p-value threshold of <0.05 are shown for each gene set.*

In order to further investigate the similarity of sets of genes singly occupying TADs between cell types, I compared genes singly occupying a TAD in ESC with cortical neuron. I found that for both Arrowhead and TopDom, genes which singly occupied TADs were very similar between ESC and cortical neuron (~64% and ~58% of ESC and ~68% and ~61% of cortical neuron genes singly occupying TADs were also singly occupying TADs in the other cell type for Arrowhead and TopDom respectively) (Figure 3.17A). I next tested the functional enrichment of the genes singly occupying TADs in: ESC only, cortical neuron only or both ESC and cortical neuron. In general I found enrichment for similar developmental functions between the groups (Figure 3.17B). Together, these results suggest that similar genes singly occupy TADs in both cell types.

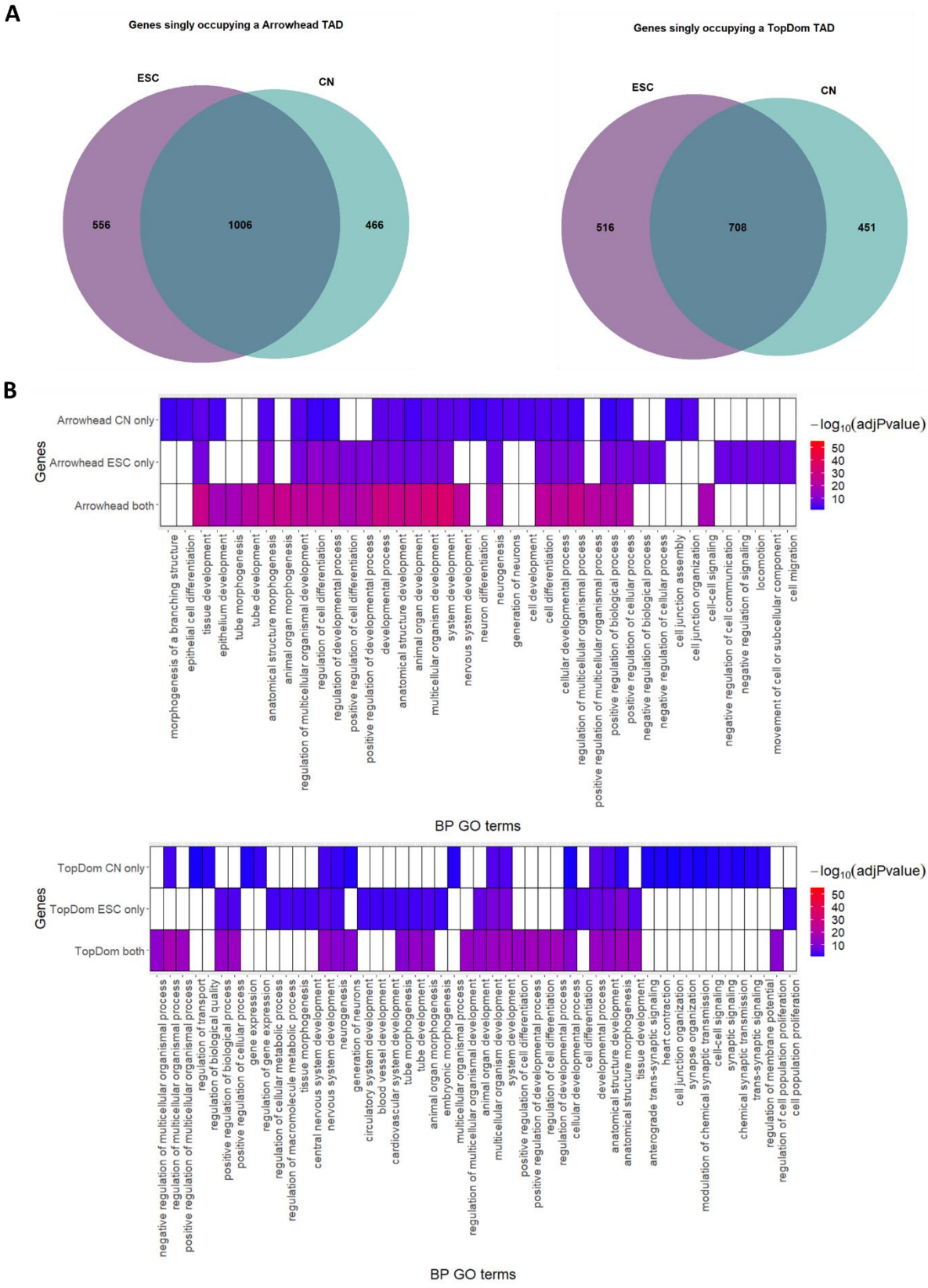


Figure 3.17: Comparison of genes singly occupying autosomal TADs in ESC and cortical neuron. *A) Venn diagrams comparing the genes singly occupying TADs in ESC and cortical neuron (CN) for Arrowhead TADs (left) and TopDom TADs (right). B) Biological process GO term enrichment of genes*

singly occupying TADs in ESC only, CN only or both ESC and CN, in Arrowhead TADs (top) TopDom TADs (bottom). Only the (max) top 25 most significant GO terms passing "gSCS" corrected p-value threshold of <0.05 are shown for each gene set.

Singly occupied TAD annotations were also intersected with compartment annotations to assess their compartment identity (intersection carried out by Dr Michelle Simon). Whether they were classed as CTCF or non-CTCF TADs was also assessed (Table 3.4). We found that ESC singly occupied TADs were most commonly CTCF TADs in the A compartment, suggesting that they are more likely to be transcriptionally active and formed by loop extrusion involving CTCF. On the other hand, cortical neuron singly occupied TADs were most commonly non-CTCF TADs with similar proportions in the A and B compartments.

Table 3.4: Percentage of CTCF TADs (positive) and non-CTCF TADs (negative) in the A or B compartments for 0 gene TADs, 1 gene TAD, 2 gene TADs and 3 gene autosomal TADs*. CTCF TADs are defined as TADs with a CTCF ChIP-seq binding peak within ±10 kb of both TAD boundaries. Non-CTCF TADs are defined as TADs with a CTCF ChIP-seq binding peak within ±10 kb of only one or neither boundary. Compartment annotations are derived from the eigenvector corresponding to the first principle component of the Pearson correlation matrix of the Hi-C map. Within each sub-table, percentages are highlighted on a colour scale from red (largest value) to white (smallest value).*

Arrowhead - 0 gene TADs					TopDom - 0 gene TADs				
CTCF TAD	negative		positive		CTCF TAD	negative		positive	
Compartment	A	B	A	B	Compartment	A	B	A	B
CN	39%	39%	15%	7%	CN	20%	51%	15%	13%
ESC	10%	43%	24%	23%	ESC	4%	58%	14%	24%
Arrowhead - 1 gene TADs					TopDom - 1 gene TADs				
CTCF TAD	negative		positive		CTCF TAD	negative		positive	
Compartment	A	B	A	B	Compartment	A	B	A	B
CN	35%	33%	22%	10%	CN	28%	35%	22%	16%
ESC	11%	24%	46%	19%	ESC	11%	26%	43%	21%
Arrowhead - 2 gene TADs					TopDom - 2 gene TADs				
CTCF	negative		positive		CTCF	negative		positive	
Compartment	A	B	A	B	Compartment	A	B	A	B
CN	43%	25%	23%	8%	CN	34%	27%	28%	11%
ESC	14%	17%	56%	14%	ESC	12%	17%	56%	15%
Arrowhead - 3 gene TADs					TopDom - 3 gene TADs				
CTCF	negative		positive		CTCF	negative		positive	
Compartment	A	B	A	B	Compartment	A	B	A	B
CN	44%	24%	23%	8%	CN	40%	26%	26%	8%
ESC	13%	13%	64%	11%	ESC	10%	13%	67%	10%

3.3.8. Expression correlation between pairs of genes in CTCF TADs

I have shown that CTCF and non-CTCF TADs are unequal in their functional relevance, with non-CTCF TADs enriched for paralogous gene pairs. I therefore next decided to focus on the functional similarity of pairs of genes in CTCF TADs. Since paralogous gene pairs are highly likely to share functional similarity and I have previously assessed their relationship with TADs (Figure 3.10 and Chapter 3: Supplementary Figure 4), I excluded all pairs of paralogous genes

and removed the olfactory genes (see Methods: 3.2.5-3.2.6) in all functional analyses. This allowed the assessment of functional similarity between genes within TADs without recent shared ancestry.

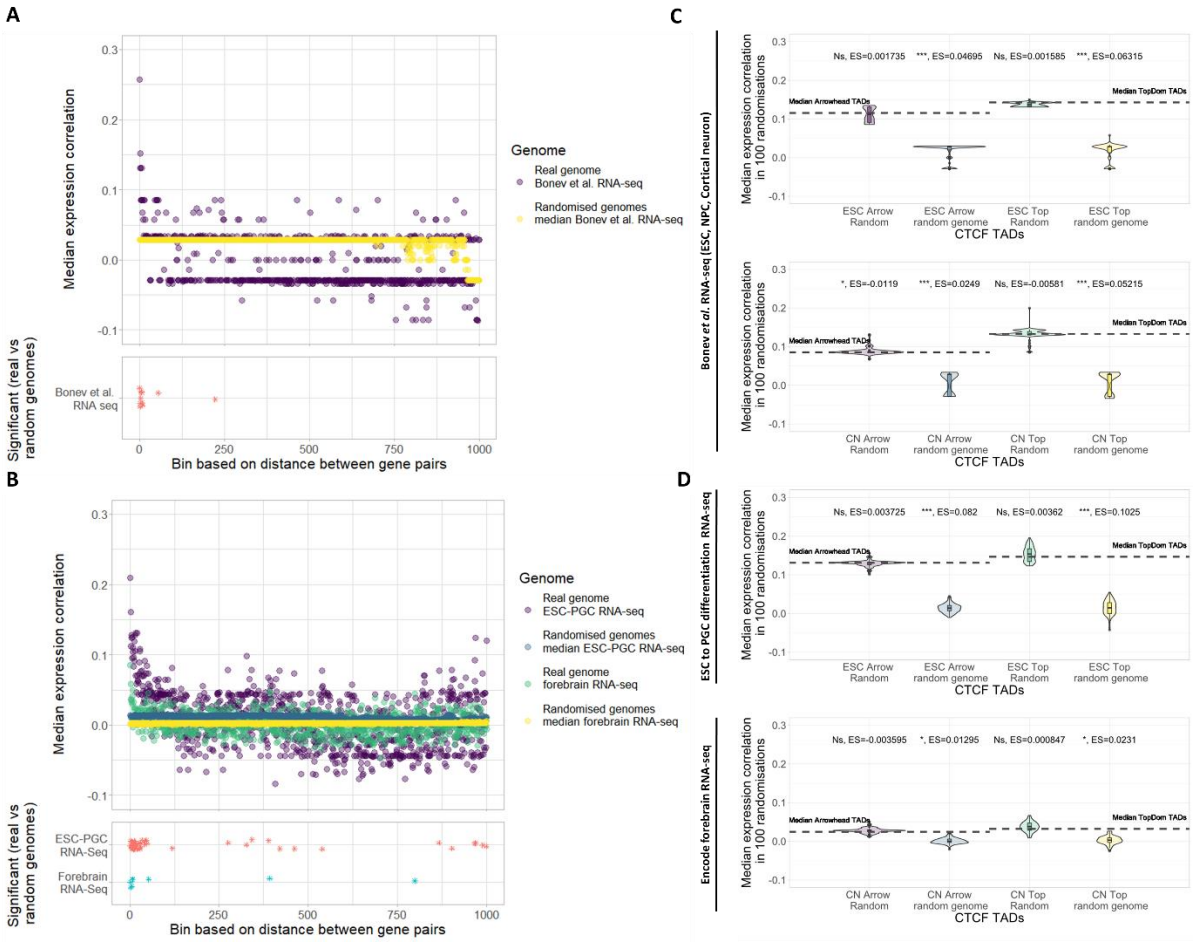


Figure 3.18: Expression correlation between pairs of genes sharing an autosomal CTCF TAD. *Olfactory genes and paralogous gene pairs have been removed in all panels. A-B) Top panel: The median expression correlation coefficient (Spearman's rank) of pairs of (autosomal) genes in the real genome binned by distance (each dot indicates one bin). Expression correlation coefficient (Spearman's rank) of pairs of (autosomal) genes in 1000 random genomes generated by shuffling the identity (but not the positions) of genes within chromosomes. Each point indicates the median of the distribution in 1000 random genomes for each bin. Paralogous gene pairs and any pairs of genes including an olfactory gene were removed from each bin in both the real and random genomes. Bottom panel: Stars indicate bins which have a significantly higher median expression correlation than expected based on*

*1000 random genomes (FDR corrected p-value<0.05). Jitter has been applied to the y axis to allow for clearer visualization of close together stars. A) Expression correlation coefficients calculated using RNA-seq from Bonev et al. (two replicates each of ESC, NPC and cortical neuron). B) Expression correlation coefficients calculated using either RNA-seq from ESC differentiating to PCG or mouse forebrain over embryonic development. C-D) Median expression correlation of pairs of genes in CTCF TADs (dotted line) and median expression correlation in 100 sets each of random CTCF TADs and random genome CTCF TADs. TADs were detected by either Arrowhead or TopDom in ESC or cortical neuron (CN). Significance indicated by median p-value of Wilcoxon test between real CTCF TADs and each of 100 sets of random/random genome CTCF TADs. FDR correction applied across eight median p-values (i.e. within C and D separately). FDR corrected median p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, NS= not significant, ES= Median of effect sizes calculated using r for Wilcoxon between real and each of 100 sets of random TADs. C) Expression correlation coefficients were calculated using RNA-seq from A. D) Expression correlation coefficients were calculated using RNA-seq from D.*

In order to assess whether pairs of genes in the same TAD have correlated expression patterns, I used FPKM/RPKM counts from RNA-seq expression data. RNA-seq generated during neural differentiation from Bonev et al. (Bonev et al. 2017) and from the most closely matching cell types/tissues to ESCs and cortical neurons which had greater than three samples (mouse ESCs differentiating to PGC (von Meyenn et al. 2016) and forebrain at different embryonic stages (P. He et al. 2020), respectively) from Encode (Davis et al. 2018) or GEO (Edgar, Domrachev, and Lash 2002) were used. Using these expression counts, I calculated Spearman's rank correlation coefficient between pairs of genes in CTCF TADs, 100 sets of random CTCF TADs, and 100 sets of random genome CTCF TADs (Figure 3.18C-D). I found that pairs of genes in CTCF TADs have a significantly higher expression correlation than pairs of genes in random genome CTCF TADs in all comparisons (median p-value<0.05). This is an expected result because randomising the genome removes the effect of linear gene proximity. However, in 7 out of 8 comparisons, I found no significant difference in expression

correlation between pairs of genes in CTCF TADs and pairs of genes in random CTCF TADs (median p-value<0.05). This suggests that contrary to the majority of other studies (Nora et al. 2012; Flavahan et al. 2016; Sarnataro, Riba, and Molina 2020; Tarbier et al. 2020), I found little evidence that pairs of genes sharing a TAD are more likely to have similar expression patterns than can be explained by their linear proximity. A study by Soler-Oliva et al. found that algorithmically identified co-expression domains in breast tissue/breast cancer tend not to coincide with TADs, which supports my findings (Soler-Oliva et al. 2017).

3.3.9. Similarity of functional annotations between pairs of genes in CTCF TADs

I next sought to assess whether pairs of genes within the same CTCF TAD are more likely to share functional annotations than pairs of genes in random CTCF TADs or random genome CTCF TADs. Here, I used molecular function GO semantic similarity scores, shared pathways, and PPI (see Methods: 3.2.10-3.2.12). GO annotations can be split into three categories: Biological processes (BP), Molecular function (MF), and Cellular component (CC) (Ashburner et al. 2000). Similarity between pairs of genes was calculated based on GO semantic similarity (see Methods: 3.2.10) using GO terms from each category separately and plotted against binned genomic distance in the real genome and 1000 random genomes. This was done for all autosomal genes both before and after the removal of paralogous gene pairs and olfactory genes (Figure 3.19). I found that for scores calculated with BP and CC, once olfactory genes and paralogous gene pairs have been removed, there is little association between GO similarity and distance. This suggests that increased similarity of these scores between genes close together in the linear genome is driven by paralogous pairs and the olfactory genes. I

therefore only assessed the GO semantic similarity of pairs of genes sharing a CTCF TAD using MF GO terms.

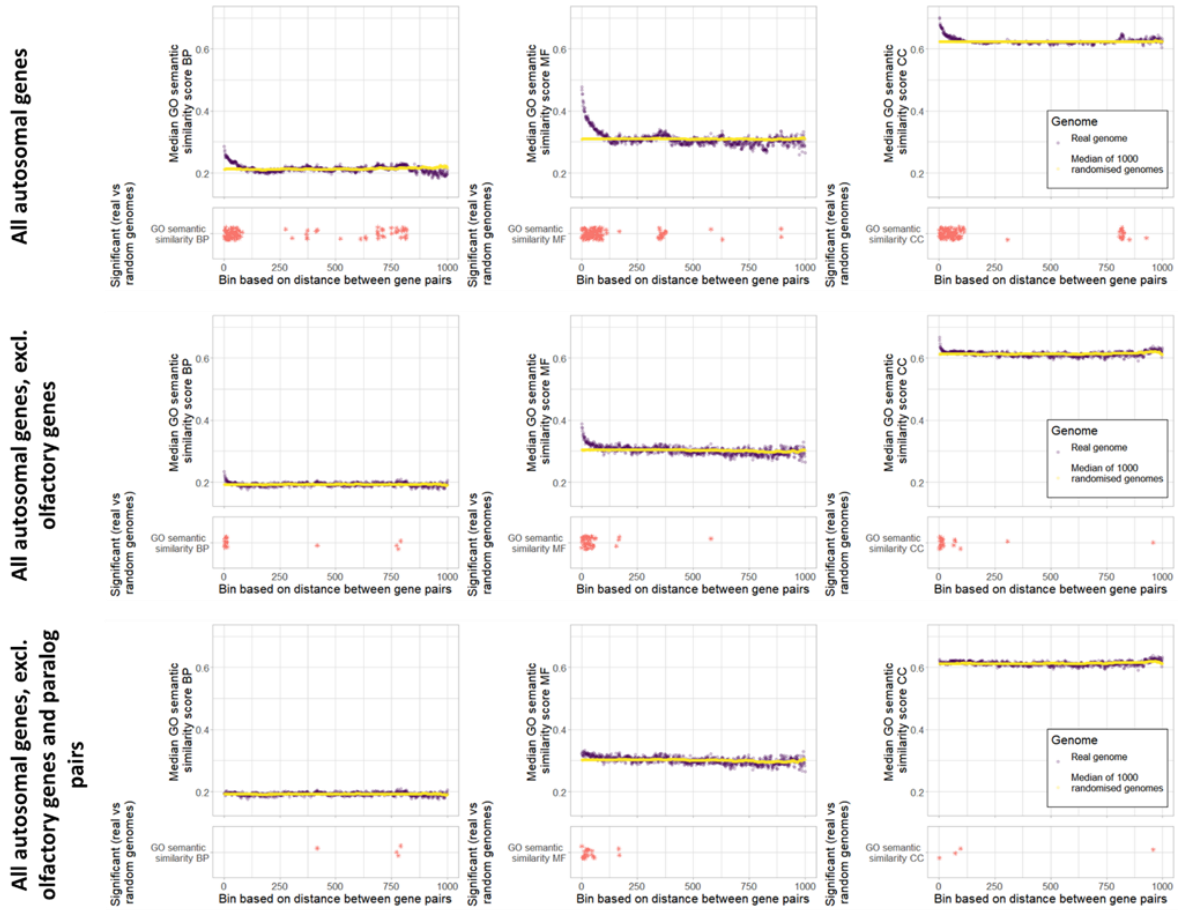


Figure 3.19: Pairwise GO similarity scores against genomic distance. *The median GO semantic similarity between pairs of (autosomal) genes in the real genome binned by distance (each dot indicates one bin). GO semantic similarity of pairs of (autosomal) genes in 1000 random genomes generated by shuffling the identity (but not the positions) of genes within chromosomes. Each point indicates the median of the distribution in 1000 random genomes for each bin. GO semantic similarity was calculated using biological process (BP) (left), molecular function (MF) (middle) or cellular component (CC) (right). Top row: all autosomal genes were included, middle row: olfactory genes were excluded and bottom row: olfactory genes and paralogous pairs of genes were excluded (from each bin in both the real and random genomes). Significance panels are included below each graph, a star indicates a significantly higher GO semantic similarity in the real genome than expected based on 1000 random genomes (FDR corrected p-value<0.05).*

I next calculated the similarity between pairs of genes sharing a CTCF TAD based on their MF GO semantic similarity score, proportion of gene pairs sharing a pathway and proportion of gene pairs sharing PPIs. In 11 out of 12 comparisons I found that pairs of genes in CTCF TADs are significantly more similar (median p-value<0.01) in terms of functional annotation than pairs of genes in random genome CTCF TADs (Figure 3.21A-C). Again, this is expected as randomising the genome removes functional similarity that can be explained by linear proximity. Using binned linear distance, I found greater similarity between pairs of genes which are in very close linear proximity than expected if genes had been randomly ordered on the chromosome (Figure 3.20A-C). I next compared the functional annotations of genes in CTCF TADs with genes in random CTCF TADs. I found that for the majority of comparisons there was no significant difference (10 out of 12 comparisons, median p-value<0.05) (Figure 3.21A-C). Pairs of genes in Arrowhead ESC CTCF TADs and Arrowhead cortical neuron CTCF TADs had significantly more similar MF GO terms than pairs of genes in random CTCF TADs (median p-value <0.01 and <0.05, respectively). A similar trend in MF GO similarity was observed for all other CTCF TADs compared to random CTCF TADs, but the difference was not significant in the other 2 comparisons (median p-value <0.05). This could indicate that pairs of genes in CTCF TADs have slightly more similar MF GO term annotations than pairs of genes in random CTCF TADs. However, perhaps this is limited to few TADs as the increase in similarity is very small and often not significant. Overall, I found that the biggest contribution to the functional similarity between pairs of genes in CTCF TADs can be attributed to their linear proximity in the genome. Once linear proximity was controlled for, I found a less consistent picture, but in the majority of comparisons, pairs of genes in CTCF TADs were no more likely to be functionally similar than if CTCF TADs had been randomly placed.

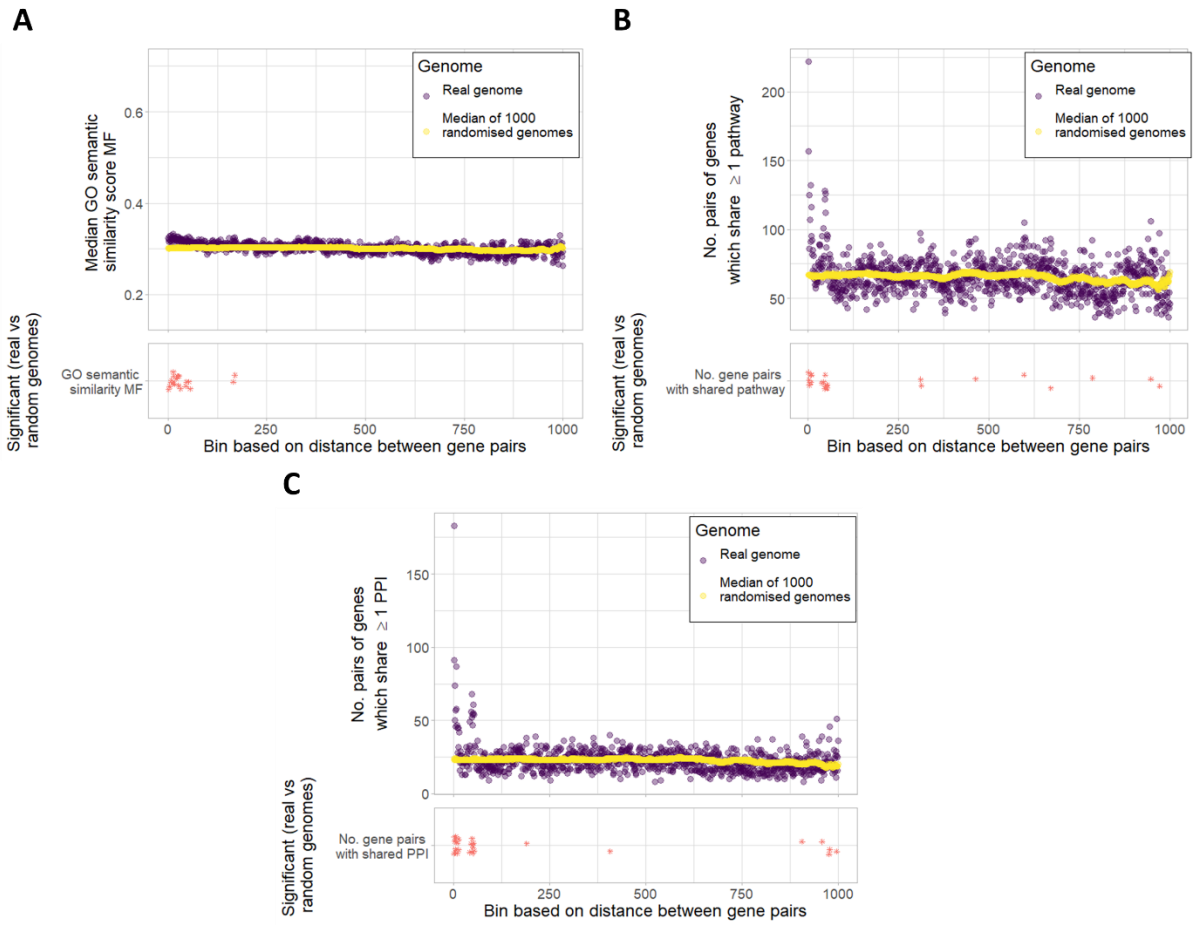


Figure 3.20: Pairwise functional similarity against genomic distance. *Olfactory genes and paralogous gene pairs have been removed in all panels. A-C) Top panel: The median GO semantic similarity between pairs of autosomal genes (A), number of pairs of autosomal genes with ≤1 shared pathway (B), or number of pairs of autosomal genes with ≤1 PPI (C) in the real genome binned by distance (each dot indicates one bin). GO semantic similarity between pairs of autosomal genes (A), number of pairs of autosomal genes with ≤1 shared pathway (B), or number of pairs of autosomal genes with ≤1 PPI in 1000 random genomes generated by shuffling the identity (but not the positions) of genes within chromosomes. Each point indicates the median of the distribution in 1000 random genomes for each bin. Paralogous gene pairs and any pairs of genes including an olfactory gene were removed from each bin in both the real and random genomes. Bottom panel: Stars indicate bins which have a significantly higher functional similarity than expected based on 1000 random genomes (FDR corrected p-value<0.05). Jitter has been applied to the y axis to allow for clearer visualization of close together stars.*

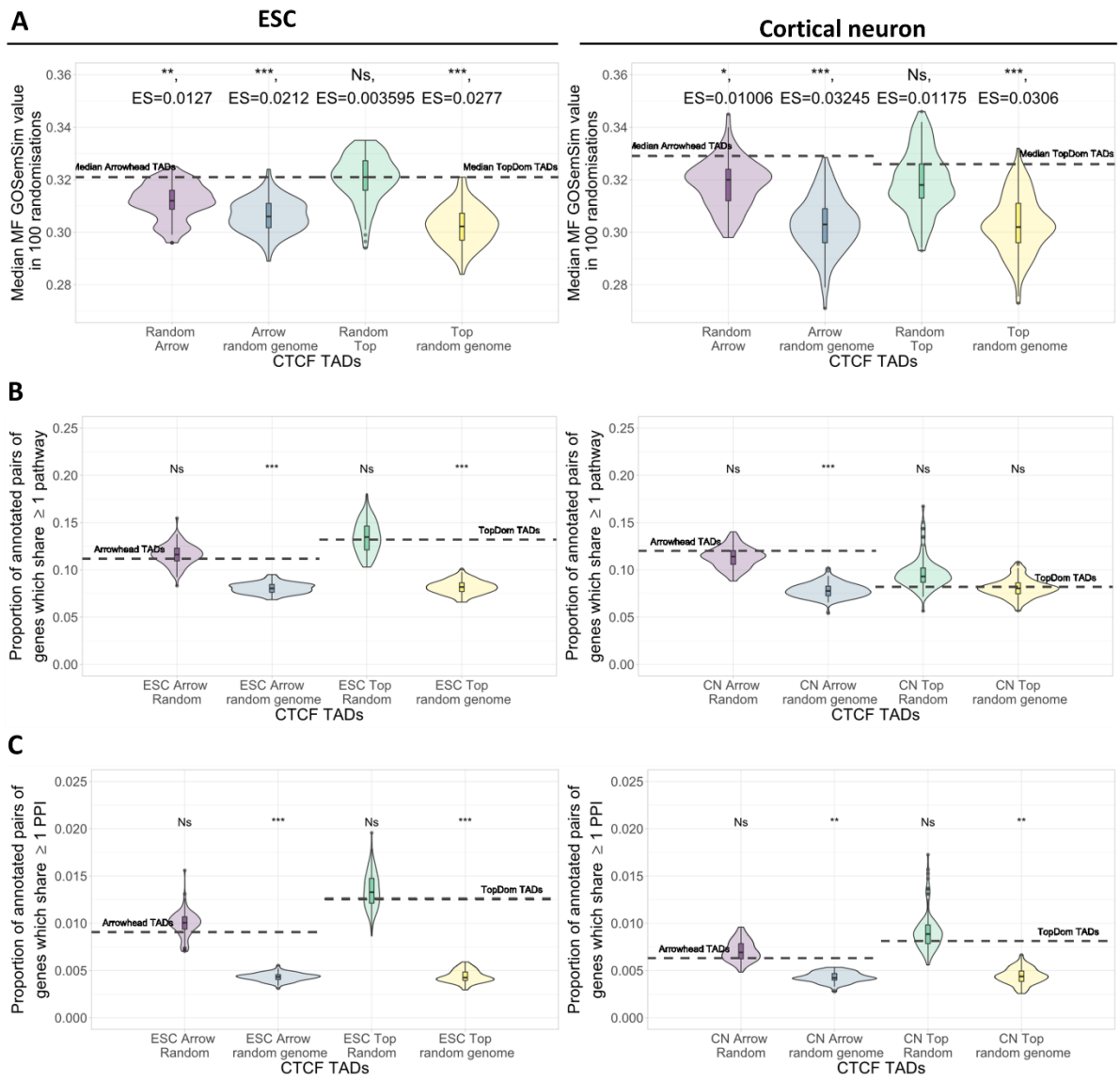


Figure 3.21: Functional similarity of pairs of genes sharing an autosomal CTCF TAD. *Olfactory genes and paralogous gene pairs have been removed in all panels. A) Median MF GO semantic similarity for pairs of genes in CTCF TADs (dotted line) compared to the distributions of median MF semantic similarity for 100 sets each of: random CTCF TADs and random genome CTCF TADs. Significance indicated by median p-value of Wilcoxon test between real CTCF TADs and each of 100 sets of random/random genome CTCF TADs. FDR correction applied across eight median p-values. B) Proportion of pairs of annotated genes sharing ≥ 1 pathway in CTCF TADs (dotted line) and the distribution of the proportion of pairs of annotated genes sharing ≥ 1 pathway in 100 sets each of: random CTCF TADs and random genome CTCF TADs. Significance indicated by median p-value of Fisher's exact test between real TADs and each of 100 sets of random TADs. FDR correction applied across eight median p-values. C) Proportion of pairs of annotated genes sharing ≥ 1 PPI in CTCF TADs (dotted line) and the distribution of the proportion of pairs of annotated genes sharing ≥ 1 PPI in 100*

*sets each of: random CTCF TADs and random genome CTCF TADs. Significance indicated by median p-value of Fisher's exact test between real TADs and each of 100 sets of random TADs. FDR correction applied across eight median p-values. D-F) CTCF TADs detected using Arrowhead and TopDom; in both ESC and cortical neuron. FDR corrected median p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, NS= not significant, ES= Median of effect sizes calculated using r for Wilcoxon between real and each of 100 sets of random TADs.*

3.4. Discussion

TADs are thought to co-localise regulatory elements and their target genes, insulate genes from off-target enhancer interactions, and block the spread of genome activation (Dixon, Gorkin, and Ren 2016). Due to these findings, I hypothesised that genes sharing a TAD would be more likely to be co-regulated. This is because in the absence of further insulation/specificity, enhancers may be able to "scan" all regulatory elements in the TAD. I hypothesised that if this is the case, one might expect genes within TADs to have higher co-expression and greater functional similarity than can be explained purely by the proximity of genes in the linear order of the genome.

In order to properly test this hypothesis, it was necessary to develop a method to create a null set of random TADs to compare against, which adequately controlled for the confounding effect of increased similarity between genes close together in the linear genome. To do this, I developed a TAD randomisation method and demonstrated that it better captures the features of real TADs than previously published methods (Figure 3.6).

I first assessed the relationship between TADs and gene paralogy, as paralogous gene pairs are likely to be highly functionally similar (Stamboulian et al. 2020). Similarly to previously described work by Ibn-Salem et al. (Ibn-Salem, Muro, and Andrade-Navarro 2017), I found that pairs of paralogous genes are more likely to fall within the same TAD than would be expected if TADs were randomly placed within chromosomes (Figure 3.10 and Chapter 3: Supplementary Figure 4). This presents a clear case in which TADs contain functionally similar genes and could reflect the need for paralogs to share regulatory elements. However, TADs can be split into CTCF TADs and non-CTCF TADs, and whilst pairs of paralogous genes are more likely to fall within non-CTCF TADs than randomly placed non-CTCF TADs, paralogous gene

pairs are commonly depleted in CTCF TADs (Figure 3.10 and Chapter 3: Supplementary Figure 4). This suggests that paralogs are depleted in domains formed by loop extrusion involving CTCF.

To further investigate the relationship between gene evolution and TADs, I analysed the average gene constraint of genes sharing TADs. I found that genes which singly occupy a TAD are statistically more constrained than genes in TADs with multiple genes. This suggests that genes in TADs on their own are less tolerant to mutation and therefore more functionally important. This is supported by Muro et al. who recently found that genes which singly occupy a TAD are more likely to be associated with disease (Muro, Ibn-Salem, and Andrade-Navarro 2019). I analysed the functional enrichment of genes singly occupying TADs and found a strong enrichment for developmental processes. A similar result was also reported in a parallel effort by Wu et al. who found that developmental genes are topologically isolated in CTCF "loop domains" in human pluripotent stem cells (Wu et al. 2021). I compared the gene sets singly occupying TADs between ESC and cortical neuron and found a high degree of similarity, suggesting that the same genes are singly occupying TADs between cell types (Figure 3.17). Together, these results indicate that there is a selective pressure for functionally important developmental genes to fall privately within TADs, providing them with strong insulation from aberrant regulation.

When considering the relationship between genes sharing a TAD and gene functional similarity, I found little global evidence for an increase in expression correlation or functional annotation similarity in genes sharing a TAD. In general, I found no difference in expression correlation between pairs of non-paralogous protein-coding genes in CTCF TADs vs random CTCF TADs. This is contrary to previous findings (Nora et al. 2012; Flavahan et al. 2016;

Sarnataro, Riba, and Molina 2020; Tarbier et al. 2020). I also found that pairs of non-paralogous protein-coding genes within CTCF TADs were largely not more similar in functional annotation than in random TADs. This suggests that globally, TADs are not associated with a higher degree of co-regulation between the genes they contain. However, there may be instances of individual TADs containing co-regulated/functionally similar genes on a local scale.

I postulate that CTCF TADs are likely to have been formed by loop extrusion involving CTCF. However, the mechanism involved in the formation of non-CTCF TADs is unclear, therefore these domains may represent other domain categories e.g. compartmental domains or TADs formed by loop extrusion not involving CTCF. The results presented here support the assertion made by Beagan et al. that it is important to separate domains formed by different mechanisms because they are likely to have different functional properties (Beagan and Phillips-Cremins 2020).

It has been widely reported that TAD calling is highly variable (Forcato et al. 2017; Dali and Blanchette 2017; Zufferey et al. 2018); in order to mitigate against this, I have included results from two TAD callers, Arrowhead and TopDom. Many of the conclusions reported here are largely consistent across the two TAD callers. Where disagreements exist, this is likely due to differences in the TAD calling methods and highlights the importance of ensuring findings are robust to the choice of TAD caller. TopDom annotates the entire genome with TADs. Therefore, if regions exist in the genome which have no TADs, TopDom will still attempt to call them; this could increase noise in TADs called by TopDom by the inclusion of false positives. On the other hand, Arrowhead calls TADs sporadically and therefore could be prone

to false negatives. In future, the field would greatly benefit from efforts to develop a gold standard TAD calling method.

A further limitation of the work outlined in this chapter is the use of only protein coding genes. This means that any functional relationships between non-coding genes sharing TADs would have been missed. It is also possible that protein coding genes which singly occupy TADs could contain other non-coding genes. In future work it would be interesting to also investigate the relationship between non-coding genes, as well as other non-coding features and TADs. As an example, one could extend the results presented above, by including genes plus their regulatory elements in the investigations of the size of genes in TADs and the distance between genes in TADs.

Together, these results suggest that TADs play a stronger role in insulating genes from aberrant regulation than promoting co-expression of genes within a TAD. I speculate that these results are more compatible with a model of TADs in which enhancers are prevented from interacting with all the genes within the same TAD by evolved enhancer-promoter specificity or further insulation in the form of tissue/cell type specific sub-TADs. I also note that genes which share a TAD are significantly further apart in linear distance than would be expected if TADs had been randomly placed. Therefore, perhaps one such mechanism by which enhancer-promoter specificity is maintained within TADs is by selection pressure for genes in TADs to be further apart in linear distance. If this is the case, disease-associated variants within a TAD may have deleterious consequences by misregulating their normal target gene, but also may acquire gain-of-function regulation of other genes within the TAD.

Chapter 4: Using chromatin architecture to investigate the function of GWAS variants associated with obesity/fat distribution

4.1. Introduction

GWAS has been pivotal in detecting associations between genomic variants and complex human disease/traits. There are currently over 200,000 variants which have been linked to a disease or trait using this method. However, it has proved challenging to move from 'variant to function' by understanding these associations in terms of their underlying mechanisms and using this knowledge to develop treatments. Over 90% of GWAS-associated variants are located in the non-coding genome and many may act via gene regulation (Cano-Gamez and Trynka 2020; Lappalainen and MacArthur 2021). Therefore, methods which assign regions of the non-coding genome with predicted regulatory function to target genes are of critical importance in GWAS variant to function efforts e.g. (Sun et al. 2022).

Regulatory elements such as enhancers are thought to regulate the expression of their target genes by coming into physical contact with the target gene promoter via chromatin looping (Mora et al. 2015; Panigrahi and O'Malley 2021; Matharu and Ahituv 2015; Bonev and Cavalli 2016; Oudelaar and Higgs 2020). This allows regulatory elements to regulate gene expression over both short and long linear genomic distances (Bonev and Cavalli 2016). In order to dissect the regulatory landscape of genes, there has been increasing interest in the generation of 3C-based datasets (Dekker et al. 2002) that are enriched for contacts involving gene promoters, allowing the genome-wide detection of promoter interacting regions which potentially have a regulatory function (PCHi-C) (Schoenfelder et al. 2018). This method has since been further

developed to allow for the enrichment of contacts involving other regulatory annotation classes such as enhancers (ECHi-C) (Madsen et al. 2020). This chapter will aim to leverage these data types in adipose to aid variant to function analysis of GWAS associations.

The global disease burden of obesity is substantial and increasing. Currently, more than one third of the global population can be classed as either obese or overweight (Hruby and Hu 2015; Dai et al. 2020; Loos and Yeo 2021). Overall obesity is commonly measured by body mass index (BMI), but this does not capture the fact that higher central (abdominal) adiposity is associated with higher metabolic disease, cardiovascular disease and mortality risk compared to higher gluteal adiposity (Shungin et al. 2015). In order to capture the effects of fat distribution (which are independent of BMI), waist hip ratio (WHR) is commonly used. WHR can be further adjusted for BMI: waist hip ratio adjusted for BMI (WHRadjBMI) (Heid et al. 2010). Using these measures, hundreds of loci have been associated by GWAS with measures of obesity or fat distribution. Further study of these associations has revealed that genes near variants associated with BMI are enriched for expression in the brain, particularly regions involved in addiction and reward. This suggests a genetic behavioural component to obesity susceptibility (Ndiaye et al. 2019; Finucane et al. 2018), whereas genes in fat distribution associated loci are enriched for expression in adipose related tissues (Shungin et al. 2015; Finucane et al. 2018). Despite these findings, variant to function studies for both obesity and fat distribution have proved challenging. The majority of detected genetic associations still require careful dissection to identify the causal variants, target genes and location of action, both in terms of tissue/cell type and developmental stage (Loos and Yeo 2021). There is significant interest in the Cox group at MRC Harwell (who use mouse models to study human obesity and type II diabetes) in carrying out variant to function analysis on obesity and fat distribution variants using mice as model organisms. In light of previous work

from the Cox group demonstrating that a mouse model with a mutation in *Wars2* showed hearing loss, reduced adiposity and reduced body weight (Agnew et al. 2018), they are particularly interested in variants in the TBX15-WARS2 locus which have been associated with human fat distribution by GWAS. The analysis presented in this chapter represents my collaboration with them, analysing chromatin architecture data over adipogenesis to aid their variant to function analysis.

An association between variants in the TBX15-WARS2 locus (index Single nucleotide polymorphism (SNP) rs984222) and fat distribution/central adiposity was initially detected via a large-scale GWAS meta-analysis by Heid et al. (Heid et al. 2010). To further investigate this association, Liu et al. leveraged varying LD patterns amongst different human populations in a multi-ethnic fine mapping approach to create a 95% credible set of SNPs (C. T. Liu et al. 2014). To do this, the posterior probability of each individual SNP driving the fat distribution association in the locus was calculated. The 95% credible set of SNPs (henceforth referred to as credible set SNPs) is made up of the SNPs with the highest posterior probability, with a cumulative posterior probability of over 95%. Using this approach, Liu et al. narrowed the 95% credible set region to 1.6 kb including rs984222, rs10923712 and rs984225, compared to the 27.654 kb region originally detected by Heid et al. rs984222 has previously been shown to be associated with the expression of TBX15 in omental fat (Heid et al. 2010), suggesting that this SNP may overlay an enhancer targeting TBX15. Recent GWAS analysis indicated that the TBX15-WARS2 locus contains multiple sexually dimorphic associations with fat distribution, identifying four signals with different sex associations (Shungin et al. 2015). This suggests that the associations with fat distribution in this locus may be more complex than originally thought.

In order to model the fat distribution association at TBX15-WARS2 in mouse, the Cox group first confirmed the conservation of enhancer activity in the mouse equivalent of the credible set region using a Luciferase-enhancer assay. This region was then deleted by the Cox group in a mouse model which was used for further investigation. RNA seq from the inguinal white adipose depot (stromal vascular fraction) revealed that very few genes were differentially expressed between mice with and without the deletion (with no differentially expressed genes in the TBX15-WARS2 locus). Body weight, body composition and metabolic phenotypes were compared between the mouse model and wild-type littermates on either high-fat or low-fat diets (manuscript in preparation). Only subtle differences were detected; there was a slight but significant difference in the rate of accumulation of body weight in male mice on a high-fat diet, but this did not translate into a difference in final body weight in these mice. There was also potentially a very subtle difference in glucose tolerance in male mice. However, no other significant phenotypic differences have so far been detected. There are several possible reasons for the lack of molecular or clear organismal phenotype in the mouse model. GWAS-associated variants typically have small effect sizes which may make differences difficult to detect in mouse studies which are limited by sample size. It is also possible that the variant requires specific environmental conditions which were not replicated. Alternatively however, it is possible that the true causal variant was not located within the credible set region or that there is insufficient conservation between mouse and human regulatory circuitry in this locus for mouse models to be useful.

Given the fat distribution association of the TBX15-WAR2 locus in human and the lack of clear phenotype detected in mouse, I sought to investigate the potential regulatory mechanisms underlying the human association, as well as to assess the suitability of the loci for modelling in mouse. To do this, I examined publicly available data to determine the chromatin

architecture of the locus. This data includes disease-relevant tissues/cell types across three levels of chromatin structure: compartments, TADs and loops. Further analysis was performed on the candidate SNPs in the locus, overlaid with loops, to prioritise candidate effector SNP-gene interactions. As noted, multi-ethnic fine mapping indicated that the credible set of SNPs associated with fat distribution in the TBX15-WARS2 locus contains only three SNPs (C. T. Liu et al. 2014). However, recent evidence of sexual dimorphism suggests that the association with fat distribution may be more complicated than originally thought (Shungin et al. 2015). There are also several GWAS studies which have found associations in this locus with BMI (Akiyama et al. 2017; Hoffmann et al. 2018; Pulit et al. 2019; Kichaev et al. 2019; Zhu et al. 2020; Sakaue et al. 2021). In order to exhaustively search for SNP-gene looping interactions which may underlie the fat distribution/obesity associations in this locus, I curated a candidate SNP set containing all index SNPs associated with WHR, BMI and WHRadjBMI in the GWAS catalog. This list was then expanded to include any SNPs in high LD with the index SNPs (proxy SNPs). This approach is similar to that used in Zhang et al. (X. Zhang et al. 2020). Since SNPs in the credible set directly relate to the mouse model created by the Cox group (rs984222, rs10923712 and rs984225), I have focused on interactions involving them in more detail. Finally, I hypothesised that if the regulatory circuitry of a locus is represented by chromatin architecture, loci with greater conservation of chromatin features between human and mouse would be more suitable for modelling in mouse. To assess this, loops detected over adipogenesis were compared between human and mouse genome-wide, in order to prioritise loci for modelling in mouse.

4.1.1. Aims

The overall aims of this chapter are to:

- Assess the conservation of TADs and compartments in disease-relevant tissues between mouse and human in the TBX15-WARS2 locus.
- Determine whether regulatory looping interactions can mechanistically explain the fat distribution/obesity association of variants in the TBX15-WARS2 locus using publicly available human PCHi-C and ECHi-C data.
- Use publicly available mouse chromatin looping data to assess whether regulatory looping in the *Tbx15-Wars2* locus is conserved between human and mouse.
- Conduct a global comparison of the human and mouse adipogenesis regulatory looping circuitry to determine the level of conservation between the two species.
- Use the results obtained above to devise a method/strategy to prioritise fat distribution/obesity associated loci for modelling in mouse.

4.2. Methods

4.2.1. Hi-C data processing and TAD calling

To assess the TAD structure at the TBX15-WARS2 locus I downloaded and analysed Hi-C data from disease-relevant tissues and cell types. As there is no human adipose-specific Hi-C data currently available in the public domain, Hi-C data from human MSC (which can differentiate into adipocytes) was downloaded from GEO (Edgar, Domrachev, and Lash 2002; Barrett et al. 2012) (accession number GSE140782 (Madsen et al. 2020)). Mouse pre-adipocyte (3T3-L1) and adipocyte (3T3-Y: 3T3-L1 after differentiation) Hi-C data was downloaded from NCBI SRA (Sayers et al. 2022) (accession number SRP082116 (M. He et al. 2018)). Mouse pre-adipocyte (3T3L1) Hi-C data after 0 hours, 4 hours or 2 days of differentiation were downloaded from GEO (accession number GSE95533 (Siersbæk et al. 2017)). Hi-C was then processed and TADs detected (as outlined in detail in Chapter 2: Methods: 2.2.2-2.2.3). In brief, raw sequencing data was processed using the Juicer pipeline, aligning to the hg19 or mm10 genome build (Durand, Shamim, et al. 2016). Individual replicates were processed separately and then merged using the "mega" option within Juicer. Reads with MAPQ < 30 were discarded and the Hi-C matrix was VC normalised (Lieberman-Aiden et al. 2009) and binned at 50 kb. TADs were detected using Arrowhead with the ignore sparsity flag (Durand, Shamim, et al. 2016; Rao et al. 2014), and TopDom (Shin et al. 2016) with default settings. As previously described, for TADs detected using TopDom, only "domain" annotations were considered. The coordinates of the largest detected human MSC TAD over the TBX15-WARS2 region were used to define the boundaries of the TBX15-WARS2 locus (chr1:118500000-120050000 detected with Arrowhead).

4.2.2. TBX15-WARS2 locus genes

The genes within the TBX15-WARS2 locus along with their biotypes, descriptions and phenotypes were downloaded from BioMart (Ensembl release 75) (Smedley et al. 2009), using the locus coordinates as filters.

4.2.3. Candidate SNPs

To curate a list of candidate causal obesity/fat distribution SNPs in the TBX15-WARS2 locus, index SNPs which have been significantly (p-value < 5×10^{-8}) associated with the mapped traits: "waist-hip ratio" (trait ID: EFO_0004343), "body mass index" (trait ID: EFO_0004340) and "BMI-adjusted waist-hip ratio" (trait ID: EFO_0007788) were downloaded from the GWAS catalog (January, 2022) (Buniello et al. 2019). Downloaded data was further filtered to remove SNPs reported in studies with more complicated trait associations e.g. cross-trait GWAS studies or longitudinal associations. Only SNP associations within the TBX15-WARS2 locus were considered (chr1:118500000-120050000). Hg19 genomic coordinates for SNPs were collected using the BioMart R package (GRCh37) (Smedley et al. 2009). In order to account for the possibility that the index SNPs identified in these studies are tagging the functional variant, I employed a strategy similar to that used in Zhang et al. (X. Zhang et al. 2020). This involved expanding the list of candidate SNPs to include any nearby 'proxy SNPs' within a 1 Mb window which are in high LD ($r^2 \geq 0.8$) with any of the index SNPs (proxy SNPs were further filtered so only SNPs within the TBX15-WARS2 locus remained). In order to identify proxy SNPs in high LD, the LDproxy tool from LDlinkR v1.1.2 was used and LD was assessed using all European populations (Myers, Chanock, and Machiela 2020). Two index SNPs were removed from the analysis at this stage because LDlink indicated that they were either not in the 1000

genome reference panel or were monoallelic in the European population. Regulomedb annotations for each SNP were downloaded from LDlinkr (A. P. Boyle et al. 2012; Myers, Chanock, and Machiela 2020) and variant consequence annotations were downloaded using the BioMart R package (Smedley et al. 2009; Durinck et al. 2009).

4.2.4. Genotype-Tissue Expression (GTEx) project

Adipose eQTLs involving SNPs in the credible set (C. T. Liu et al. 2014) (rs984222, rs10923712 and rs984225) were downloaded from the GTEx project (V8) (Ardlie et al. 2015).

4.2.5. Compartment annotation

In order to evaluate the compartment structure in the TBX15-WARS2 locus, compartment annotations were generated for all Hi-C data (chr1 for hg19 and chr3 for mm10) using the eigenvector function in the Juicer pipeline, with the –p flag (Durand, Shamim, et al. 2016). To do this, reads with MAPQ below 30 were filtered, contact maps were VC normalised and binned at 50 kb. The output was then further processed as outlined in Miura et al. (Miura et al. 2018)

4.2.6. Promoter/enhancer capture Hi-C data processing

To investigate chromatin looping in disease-relevant tissues in the TBX15-WARS2 locus, published PCHi-C and ECHi-C datasets from relevant mouse and human tissues/cell types were identified. These datasets were then downloaded as both raw sequencing data and

published loops from GEO (Edgar, Domrachev, and Lash 2002; Barrett et al. 2012) or European Bioinformatics Institute (EBI) ArrayExpress (Athar et al. 2019). The selected datasets included: ECHi-C (which also included 299 baited promoters) from human MSC differentiating towards adipose (GEO accession: GSE140782 (Madsen et al. 2020)), PCHi-C from primary human white adipocytes (GEO accession: GSE110619 (Pan et al. 2018)), PCHi-C from human Simpson-Golabi-Behmel syndrome (SGBS) preadipocytes (ArrayExpress accession: E-MTAB-10200 (Sobreira et al. 2021)), and mouse 3T3-L1 preadipocytes differentiating towards adipose (GEO accession: GSE95533 (Siersbæk et al. 2017)).

In the majority of the datasets selected, published loops were detected using CHiCAGO (Table 4.4). Therefore, for these datasets, significant loops were extracted from downloaded published loops using a significance threshold of CHiCAGO score ≥5 (Pan et al. 2018; Madsen et al. 2020; Siersbæk et al. 2017). The published loops reported by Sobreira et al. for SGBS preadipocytes were the only exception, as they were detected using HOMER (findHiCInteractionsByChr.pl) (Heinz et al. 2010). Therefore, significant loops were extracted from this dataset using an FDR-corrected p-value threshold <0.05 (as these loops had no score they were assigned a score of 5 for visualisation). It should be noted that there is an inconsistency between the parameters reported in the methods compared with those reported on ArrayExpress for the Sobreira et al. SGBS dataset (the methods state the use of the parameters: -res 2000 -superRes 5000 in HOMER findHiCInteractionsByChr.pl; however, the file name in ArrayExpress indicates the use of -res 20000 –superRes 10000) (Sobreira et al. 2021). The published loops for mouse 3T3-L1 preadipocytes (differentiating towards adipocytes) (Siersbæk et al. 2017) were converted from mm9 to mm10 using LiftOver (Hinrichs et al. 2006).

In order to recall loops from the raw datasets using a standardised pipeline, reads were aligned to hg19 (excluding haplotypes) or mm10 using HiCUP (v.0.8.1) (Wingett et al. 2015). If raw datasets were provided as multiple runs, they were merged before processing with HiCUP but replicates were processed separately. HiCUP was run using default ditag parameters (longest=700 and shortest=50). The bam files produced by HiCUP were then converted to the CHiCAGO input format using bam2chicago_V02.sh from CHiCAGO tools (Cairns et al. 2016; Freire-Pritchett et al. 2021). Restriction fragment files (.rmap) were generated using the HiCUP digest file (Wingett et al. 2015). Bait maps (bed files containing the baited restriction fragments and annotations (.baitmap)) were produced for each PHi-C/EHi-C dataset as outlined Chapter 4: Supplementary Table 1. In brief, the coordinates of the baits were downloaded and baited fragments were selected if the bait was fully contained within them. CHiCAGO (v.1.14.0) was then run using the default settings for either 6-cutter enzymes or 4-cutter enzymes (Table 4.1) and replicates were merged. As data generated with MboI or other 4 cutter enzymes can suffer from low coverage due to fragment sizes which are smaller than produced by 6-cutter enzymes, datasets produced using MboI were subject to a further analysis step. In this step, datasets were binned at 5 kb (including the baited fragment), synthetically creating a resolution similar to 6-cutter data. As recommended in Freire-Pritchett et al. loops were called on binned data using the default settings for 6-cutter enzymes (Freire-Pritchett et al. 2021). As with the published CHiCAGO loops, a CHiCAGO score of ≥5 was used to establish significance in recalled loops.

Table 4.1: Default Chicago settings for 6-cutter and 4-cutter enzymes.

Enzyme	Min FragLen	Max FragLen	Max LBrownEst	Binsize	Weight Alpha	Weight Beta	Weight Gamma	Weight Delta
6-cutter enzyme (HindIII)	150	40000	1.5e6	20000	34.1157346557331	−2.58688050486759	−17.134784581965	−7.076092455215 41
4-cutter enzyme (MboI)	75	1200	75000	1500	24.5	−2.16	−21.2	−9.2

4.2.7. Published and recalled loop comparison

To investigate the reproducibility of PCHi-C/ECHi-C loops significant under the recommended CHiCAGO settings vs the original published loops, loops were compared between published and recalled datasets. To do this only significant *cis* loops (loops between two regions on the same chromosome) were considered. A loop was deemed to be present in both datasets if there was any non-zero overlap between both of the looping anchor fragments between datasets (i.e. the bait fragment in one dataset overlapped the bait fragment in the other AND the other end fragment in one dataset overlapped the other end fragment in the other, OR the bait fragment in one dataset overlapped the other end fragment in the other at each end of the loop).

4.2.8. Overlapping loops with SNPs and gene TSSs

Loops identified from PCHi-C and ECHi-C were used to link candidate SNPs to potential regulatory target genes. Loops involving SNPs were termed SNP loops and were determined

by overlapping the coordinates of candidate SNPs with either the original published PCHi-C/ECHi-C loops or the recalled PCHi-C/ECHi-C loops which I generated for the same datasets (*cis* loops only). A loop was deemed to be a SNP loop if at least one candidate SNP (defined as 1 bp region) fell within the baited or other end fragment. Hg19 gene TSS in the TBX15-WAR2 locus and surrounding area (chr1:117500000-121050000) were downloaded from BioMart (Ensembl release 105) (Smedley et al. 2009). A loop was deemed to overlap a gene TSS if the TSS fell within the baited or other end fragment. Loops which overlapped both a SNP and a TSS at the same end were not the focus of this work as any potential regulatory impact of these SNPs is less likely to be due to looping.

In order to repeat the above analysis with mouse PCHi-C data, SNP coordinates were converted from hg19 to mm10 using LiftOver (Hinrichs et al. 2006). The coordinates for 251 of 638 SNPs were successfully converted to mm10; the SNPs which could not be converted returned the error message "Deleted in new" indicating that they intersected no chains. A further 7 SNPs were removed as their mm10 locations were not on the same chromosome at the *Tbx15*-*Wars2* locus. Mm10 TSSs were downloaded for the region (chr3:97923872-101065203) from BioMart (Ensembl release 102) (Smedley et al. 2009).

4.2.9. Human vs mouse TAD loop comparison score

To assess the conservation of the regulatory circuitry during human and mouse adipogenesis genome-wide, I assessed the proportion of human chromatin loops during adipogenesis which were also detected in mouse adipogenesis within each TAD. TADs were used to define loci boundaries because regulatory interactions are thought to infrequently cross TAD boundaries (Panigrahi and O'Malley 2021). As previously demonstrated (Chapter 2), TADs

called with Arrowhead or TopDom can vary considerably but there is greater agreement between the two TAD callers in the placement of boundaries. In order to define the boundaries of each locus based on TADs, I considered Arrowhead and TopDom TADs together. Any TADs detected by either TAD caller, which overlapped another TAD by at least one base, were merged. In this way, the boundaries of each locus were defined by the largest TAD that can be formed by the TAD boundaries detected by either TAD caller over the locus. Hg19 gene coordinates were downloaded from (Ensembl release 105) and assigned to TADs based on their TSS.

The datasets selected for comparison were: ECHi-C (also including 299 baited promoters) from human MSC differentiating towards adipose (Madsen et al. 2020) and PCHi-C from mouse 3T3-L1 preadipocytes differentiating towards adipose (Siersbæk et al. 2017). These datasets were selected as they both contain multiple adipogenesis time points. There are several experimental differences between these datasets, including the fact that the human dataset is ECHi-C whereas the mouse dataset is PCHi-C, and that the human data was generated using MboI whereas the mouse dataset was generated with HindIII. I sought to minimise the effect of these differences on the looping data used in the comparison. To do this, I compared *cis* loops which were significant in at least one time point during adipogenesis between the binned human recalled loops and the mouse recalled loops. For the human dataset, loops called on binned data were used as this allowed the detection of loops at a comparable resolution to the mouse data (Freire-Pritchett et al. 2021). I also filtered the human data to remove interactions which did not involve a promoter at one or both ends (assessed based on overlap with a TSS), creating “pseudo” PCHi-C loops. As done in the original study (Siersbæk et al. 2017), any loops involving chr10:107176310-108421650 or chr10:116737743-118739308 were removed from the mouse dataset as this region is

amplified in 3T3-L1 cells. For consistency, any loops overlapping the equivalent regions in hg19 were also removed from the human dataset.

In order to compare significant loops over human adipogenesis with significant loops over mouse adipogenesis, mouse loop coordinates were converted from mm10 to hg19 using liftOver (Hinrichs et al. 2006). If the coordinates of the loops anchors at either end could not be converted, or the converted loops were no longer a *cis* interaction, the loop was excluded. A loop was deemed to be present in both human and mouse if there was any non-zero overlap between both looping anchor fragments. Human loops were then assigned to a TAD if >50% of the length of both loop anchors fell within the TAD. The total number of human loops as well as the proportion and number of loops conserved in mouse were calculated per TAD. When calculating the proportion of conserved loops per TAD, only TADs which had at least one significant human loop were considered.

4.2.10. GWAS and international mouse phenotyping consortium (IMPC) TADs

To test for potential differences in the loop conservation score between loci which did or did not contain a WHRadjBMI GWAS association, significant ($p< 5 \times 10^{-8}$) WHRadjBMI GWAS SNP associations from (Pulit et al. 2019) were downloaded from the GWAS catalog (Buniello et al. 2019). SNP association coordinates were converted to hg19 using the BioMart R package (GRCh37) (Smedley et al. 2009). Any SNPs which could not be converted were excluded. GWAS TADs were defined as the subset of TADs which contained a WHRadjBMI SNP. As a control I also identified GWAS TADs using significant schizophrenia-associated SNPs from (Pardiñas et al. 2018).

I also tested for potential differences in loop conservation between loci which did or did not contain a mouse homolog with a significant adipose phenotype in IMPC. To do this, mouse knockout genes associated with “abnormal adipose tissue morphology” were downloaded from IMPC (release 15.1) (Dickinson et al. 2016). MGI symbols were converted to human homologs using the BioMart R package (Smedley et al. 2009). TADs which contained a human homolog of a mouse knockout gene associated with “abnormal adipose tissue morphology” were defined as IMPC TADs.

The number of human loops, the number of human loops conserved in mouse or the proportion of human loops conserved in mouse were compared between GWAS TADs, non-GWAS TADs, IMPC TADs and non-IMPC TADs using a two-sample permutation test.

4.2.11. Genome visualisation

Genomic regions were visualised using the WASHU Epigenome browser v53.5.2 (D. Li et al. 2019). ChromHMM annotation tracks were added by connecting to “All Chromatin states tracks” in the “Reference human epigenomes from Roadmap Epigenomics Consortium” collection in the “public data hubs” tracks in the browser. Primary chromHMM tracks were selected, which corresponded to the core 15-state model (Ernst and Kellis 2017; Ernst et al. 2011; Ernst and Kellis 2010).

4.3. Results

4.3.1. Human TBX15-WARS2 locus and its chromatin architecture

In order to detect TADs in the TBX15-WARS2 locus in human, I searched for publicly available Hi-C data from relevant tissues. To my knowledge, there is currently no publicly available human adipose Hi-C data. I therefore downloaded and analysed MSC Hi-C data from Madsen et al.. MSC are multipotent stem cells which can differentiate into cell types such as adipocytes, osteocytes and chronocytes (Ullah, Subbarao, and Rho 2015). The Hi-C data was analysed using Juicer (see Methods: 4.2.1) (Durand, Shamim, et al. 2016) and found to contain 221,632,999 Hi-C contacts (out of a total of 487,032,119 sequenced read pairs). The map resolution of the dataset was 7.8 kb and its *cis*/*trans* ratio was 9.61 (~90.6% of contacts were *cis*), indicating that whilst the level of sequencing is low in this dataset, the data has a high signal to noise ratio (experimental noise is expected to affect *cis* and *trans* contacts equally, resulting in a lower *cis*/*trans* ratio (Lajoie, Dekker, and Kaplan 2015)). Using this Hi-C data, I detected TADs with both Arrowhead (Rao et al. 2014; Durand, Shamim, et al. 2016) and TopDom (Shin et al. 2016). Both TAD callers detect a similar TAD in the locus (Figure 4.1). As regulatory interactions are thought to infrequently cross TAD boundaries (Panigrahi and O'Malley 2021), I defined the TBX15-WAR2 locus as a 1.55 Mb region based on the largest human MSC TAD I detected in the locus (chr1:118500000-120050000 detected with Arrowhead) (Figure 4.1). The region contains a total of 27 gene annotations, including 7 protein coding genes: HAO2, HSD3B1, HSD3B2, SPAG17, TBX15, WARS2, and WDR3 (Table 4.2).



Figure 4.1: Human TBX15-WARS2 locus. *WASHU genome browser view displaying the full locus (top panel) and zoomed in on the credible set region (bottom panel) (hg19). Tracks display: gencode gene annotations, mm10 to hg19 blastz (indicating conservation between human and mouse), the location of the credible set region, the positions of the candidate SNPs, compartment annotations (A compartment (red) and B compartment (blue)), TADs called with Arrowhead and TopDom, chromHMM chromatin states in adipose, MSC, bone, and brain-related cell types. Gaps in the human tracks correspond to gaps in the mouse to human genome alignment shown in the blastz track.*

To further investigate the chromatin architecture of the TBX15-WARS2 locus, I next assessed its compartment annotation in the MSC Hi-C data. I found that the majority of the locus is located in the B compartment, indicating that it is likely to be heterochromatic and transcriptionally inactive. However, there are two regions in the A compartment, one small region overlying SPAG17, and another larger region overlying TBX15, WARS2 and the credible set region. This suggests that these regions are likely to be euchromatic and transcriptionally active (Figure 4.1).

To further assess the chromatin state of the TBX15-WARS2 locus I assessed the chromHMM predicted states in tissues and cell types relevant to obesity/fat distribution. These tissues and cell types included: brain, MSC, and cell types which can differentiate from MSC such as adipose, chondrocytes and osteoblast. ChromHMM takes ChIP-seq datasets for 5 chromatin marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) and uses a HMM to segment and label the genome with 15 chromatin states (Figure 4.1) (Ernst and Kellis 2017; Ernst et al. 2011; Ernst and Kellis 2010). ChromHMM states from adipose and MSC related cell types largely agreed with the MSC compartment annotations of this locus. The majority of the locus was annotated as falling in the B compartment and this is supported by predicted chromatin states such as “Heterochromatin” and “Weak repressed polycomb” in these regions. In the smaller A compartment region overlying SPAG17, the predicted chromatin states include “enhancer”, “weak transcription” and some small regions of “flanking active TSS”. At the other end of the locus, in the A compartment region overlying TBX15, WARS2 and the credible set region, there are large areas with the predicted state “weak transcription”, as well as some smaller regions of “enhancer” and “flanking active TSS”. The chromatin state of the credible set region is predicted to be largely “weak transcription” and “enhancer”. However, the chromatin state predictions for brain related tissues largely do not match the A compartment

annotations in this locus as they contain mainly polycomb, quiescent annotations or no annotations in these regions. Together, this suggests that there is activity in this locus in cell types related to MSC differentiation rather than brain (Figure 4.1).

Table 4.2: Genes in the TBX15-WARS2 locus. *Biotype, description and phenotypes from BioMart shown.*

Associated Gene Name	Ensembl Gene ID	Gene Biotype	Description	Phenotype description
WARS2-IT1	ENSG00000224238	antisense	WARS2 intronic transcript 1 (non-protein coding)	
RP11-418J17.1	ENSG00000231365	antisense		
RP4-712E4.1	ENSG00000226172	lincRNA		
RP4-712E4.2	ENSG00000239216	lincRNA		
RP11-418J17.3	ENSG00000227712	lincRNA		
RP5-834N19.1	ENSG00000232650	lincRNA		
WDR3	ENSG00000065183	protein coding	WD repeat domain 3	
SPAG17	ENSG00000155761	protein coding	sperm associated antigen 17	
TBX15	ENSG00000092607	protein coding	T-box 15	COUSIN SYNDROME, Pelviscapular dysplasia
WARS2	ENSG00000116874	protein coding	tryptophanyl tRNA synthetase 2, mitochondrial	
HAO2	ENSG00000116882	protein coding	hydroxyacid oxidase 2 (long chain)	
HSD3B2	ENSG00000203859	protein coding	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2	3-@BETA-HYDROXYSTEROID DEHYDROGENASE TYPE II DEFICIENCY OF, Congenital adrenal hyperplasia due to 3-beta-hydroxysteroid dehydrogenase deficiency, Familial prostate cancer
HSD3B1	ENSG00000203857	protein coding	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1	
PSMC1P12	ENSG00000226126	pseudogene	proteasome (prosome, macropain) 26S subunit, ATPase, 1 pseudogene 12	
RP4-630J13.1	ENSG00000229911	pseudogene		
RBMX2P3	ENSG00000223575	pseudogene	RNA binding motif protein, X-linked 2 pseudogene 3	
RPS3AP12	ENSG00000236804	pseudogene	ribosomal protein S3a pseudogene 12	
RPL6P2	ENSG00000227056	pseudogene	ribosomal protein L6 pseudogene 2	
RP5-871G17.5	ENSG00000232780	pseudogene		
HSD3BP2	ENSG00000203858	pseudogene	hydroxy-delta-5-steroid dehydrogenase, 3 beta, pseudogene 2	
GAPDHP23	ENSG00000235595	pseudogene	glyceraldehyde 3 phosphate dehydrogenase pseudogene 23	
HSD3BP1	ENSG00000187481	pseudogene	hydroxy-delta-5-steroid dehydrogenase, 3 beta, pseudogene 1	
GAPDHP58	ENSG00000229570	pseudogene	glyceraldehyde 3 phosphate dehydrogenase pseudogene 58	
RNA5SP56	ENSG00000222209	rRNA	RNA, 5S ribosomal pseudogene 56	
HAO2-IT1	ENSG00000230921	sense intronic	HAO2 intronic transcript 1 (non-protein coding)	
snoU13	ENSG00000238679	snoRNA	Small nucleolar RNA U13	
RNU1-75P	ENSG00000194297	snRNA	RNA, U1 small nuclear 75, pseudogene	

4.3.2. Characterisation of candidate causal obesity/fat distribution associated SNPs in the TBX15-WARS2 locus

In the TBX15-WARS2 locus there are 76 index SNPs which are significantly associated with at least one of WHR, BMI or WHRadjBMI in the GWAS catalog. The large majority of these SNPs are associated with WHRadjBMI rather than BMI or WHR (Figure 4.2A). This suggests the locus is associated with fat distribution and central obesity rather than overall adiposity (Heid et al. 2010). This indicates that the tissues of action are likely to be related to adipose rather than brain, supporting the chromHMM findings.

I curated a list of candidate SNPs which may functionally underlie the association with obesity/fat distribution. This list contained all the index SNPs in the locus along with any proxy SNPs in high LD with them (SNP with LD ≥ 0.8 with any index SNP), adding a further 562 candidate SNPs (Figure 4.2B). This accounts for the possibility that an index SNP is significantly associated with obesity/fat distribution because it is in high LD with the true causal SNP (X. Zhang et al. 2020). The majority of the candidate SNPs identified overlie the larger A compartment region over TBX15 and WARS2 (Figure 4.1). In order to assess the possible functional implications of these SNPs, I assessed their variant consequence as well as their regulomedb annotation. Most of the candidate SNPs had a variant consequence annotation of: non coding intron variant (n=262), no variant consequence (n=202) or intron variant (n=148) (Figure 4.2C). Regulomedb annotates SNPs based on evidence of regulatory activity. Here, the majority of SNPs were annotated as: transcription factor binding OR DNase peak (n=181); other (n=144); motif hit (n=109); transcription factor binding AND DNase peak (n=66); or no annotation (n=53) (Figure 4.2D) (A. P. Boyle et al. 2012). This indicates that many of the candidate SNPs in this region may overlie regulatory elements, supporting the above

findings that many candidate SNPs also overlie an A compartment region containing “weak transcription”, “enhancer”, and “flanking active TSS” chromHMM annotations (Figure 4.1).

All of the credible set SNPs from Liu et al. (rs984222, rs984225 and rs10923712) were intron variants. rs984222 (the index SNP) and rs984225 were annotated in regulomedb as transcription factor binding and DNase peak, and rs10923712 was annotated as transcription factor binding or DNase peak suggesting they may have regulatory function. In order to investigate this in more detail I looked for adipose eQTLs involving these SNPs in GTEx (Ardlie et al. 2015). All three SNPs were involved in the same eQTLs (Chapter 4: Supplementary Table 2); in subcutaneous adipose the eQTLs involved WARS2, RP4-712E4.1, RP11-418J17.1 and TBX15, whilst in visceral (omentum) adipose the eQTLs involved WARS2, TBX15 and RP11-418J17.1. This suggests that one or all of these variants could be involved in regulating the expression of these genes.

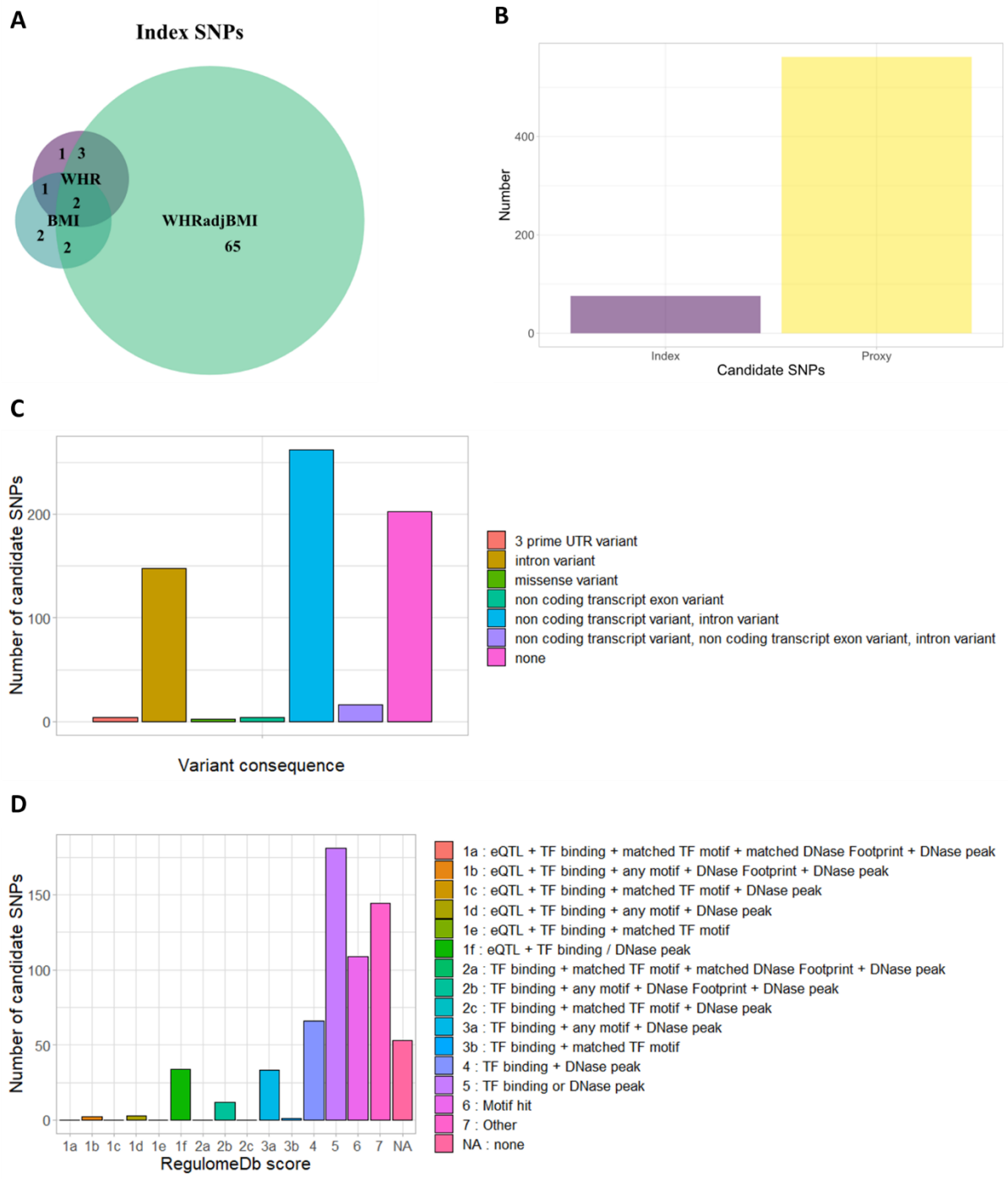


Figure 4.2: Features of candidate SNPs. *A) Venn diagram showing the number of index SNPs associated with each trait (WHR, BMI and WHRadjBMI). B) Bar chart showing the number of index SNPs and proxy SNPs included. C) Distribution of variant consequence annotations observed over all candidate SNPs. D) Distribution of regulomedb annotations observed over all candidate SNPs.*

4.3.3. TADs and compartments in this locus are largely conserved in mouse

The *Tbx15*-*Wars2* locus falls on chromosome 3 in the mouse genome and is largely conserved between human and mouse. However, the credible set region is shorter in the mouse genome (0.839 kb in mm10 compared to 1.591 kb in hg19) (Figure 4.1). Furthermore, the coordinates of only 251 of 638 candidate human SNPs could be converted to the mouse genome using LiftOver (See Methods: 4.2.8). In order to further characterise the similarity of this locus between mouse and human, I assessed the TAD and compartment structure in mouse adipogenesis Hi-C datasets. To do this I analysed Hi-C data from the 3T3-L1 preadipocyte cell line before and after differentiation from He et al. , as well as Hi-C data from 3T3-L1 before differentiation day 0 (D0), after 4 hours (4h) or 2 days (D2) of differentiation from Siersbæk et al. (Table 4.3). As done in human MSCs, TADs were detected with both Arrowhead and TopDom (Durand, Shamim, et al. 2016; Shin et al. 2016).

Table 4.3: Statistics relating to the Mouse adipose Hi-C data. *Including the number of replicates, the number sequenced read pairs, the number successfully aligned/ valid Hi-C contacts, the total number of cis contacts (within chromosome contacts), the ratio of cis (within chromosome contact)/ trans (between chromosome contact) contacts and the map resolution (defined as the minimum bin size at which 80% of bins have at least 1000 contacts and commonly used to select the minimum bin size at which a Hi-C dataset can sensibly be analysed).*

Dataset	**No. replicates**	**Total no. of sequenced read pairs**	**Total no. of Hi-C contacts**	**Total no. of *cis* contacts**	***Cis*/*trans* ratio**	**Map resolution**	**Original study**
3T3-L1 (un-differentiated)	1	680,401,514	406,579,475	283,552,023	2.3	9.15 kb	He et al. 2018
3T3-L1 (differentiated (3T3-Y)	1	690,059,035	374,282,086	295,379,203	3.74	10 kb	He et al. 2018
3T3-L1 D0 differentiation	2	353,344,770	210,580,826	191,124,112	9.82	16.95 kb	Siersbæk et al. 2017
3T3-L1 4h differentiation	2	332,383,308	176,057,158	163,121,602	12.61	18.25 kb	Siersbæk et al. 2017
3T3-L1 D2 differentiation	2	352,572,432	198,996,870	181,805,875	10.58	17.2 kb	Siersbæk et al. 2017

At the *Tbx15*-*Wars2* locus, in the He et al. adipocyte and pre-adipocyte datasets, TADs were detected with both TAD callers. However, in the Siersbæk et al. datasets, TADs could only be detected using both TAD callers at day 2 of differentiation (at day 0 and 4 hours, TADs were only found using TopDom). This is likely to be due to the dominance of false negatives in TADs called by Arrowhead when the dataset has insufficient sequencing depth (see Chapter 2). Most of the TopDom TADs detected in the mouse locus predict a TAD boundary in the middle of the locus (around *Tbx15* and *Wars2*) which was not detected in human MSC TADs. There are several possible explanations for this observation, including: a TAD boundary is present in the middle of the locus in mouse but not human, the TAD boundary forms during adipogenesis

(the human TAD was detected using MSC Hi-C so the boundary may not have been present at this stage) or, since TAD structures are hierarchical, the detected mouse boundary could reflect a different level of the hierarchy than detected in human (e.g. a sub TAD boundary). The fact that in the He et al. data, the TopDom TAD in pre-adipocytes spans the whole locus (with no middle boundary) whereas the TopDom TAD in adipocytes is split into two TADs, supports the possibility that this boundary could be formed later during adipogenesis. On the other hand, the Arrowhead TADs from the same datasets both predict a TAD which span the majority of the locus (with no middle boundary), supporting the possibility that this boundary could be a sub TAD boundary (Figure 4.3).

The compartment annotations generated from the He et al. adipocyte and pre-adipocyte datasets suggests that the entire mouse locus is in the B compartment. This would suggest that this region is likely to be inactive and heterochromatic during adipogenesis. This finding is not supported by the compartment annotations generated from the Siersbæk et al. Hi-C datasets, which all suggest that whilst the majority of the locus falls within the B compartment, the region overlying the *Tbx15* and *Wars2* genes (as well as the lifted over credible set region and the majority of the candidate SNPs) and the region overlying *Wdr3* and *Gdap2* (towards the end of *Spag17*) fall within the A compartment (Figure 4.3). This is very similar to the compartment structure detected in human MSC Hi-C.

These results highlight discrepancies between the two mouse adipogenesis datasets, as well as differences between Arrowhead and TopDom. However, for both TADs and compartments, at least one of the datasets investigated displays similar chromatin architecture to that detected in human MSC. Therefore, there is evidence of conservation between mouse and human higher order chromatin architecture in these cell types.

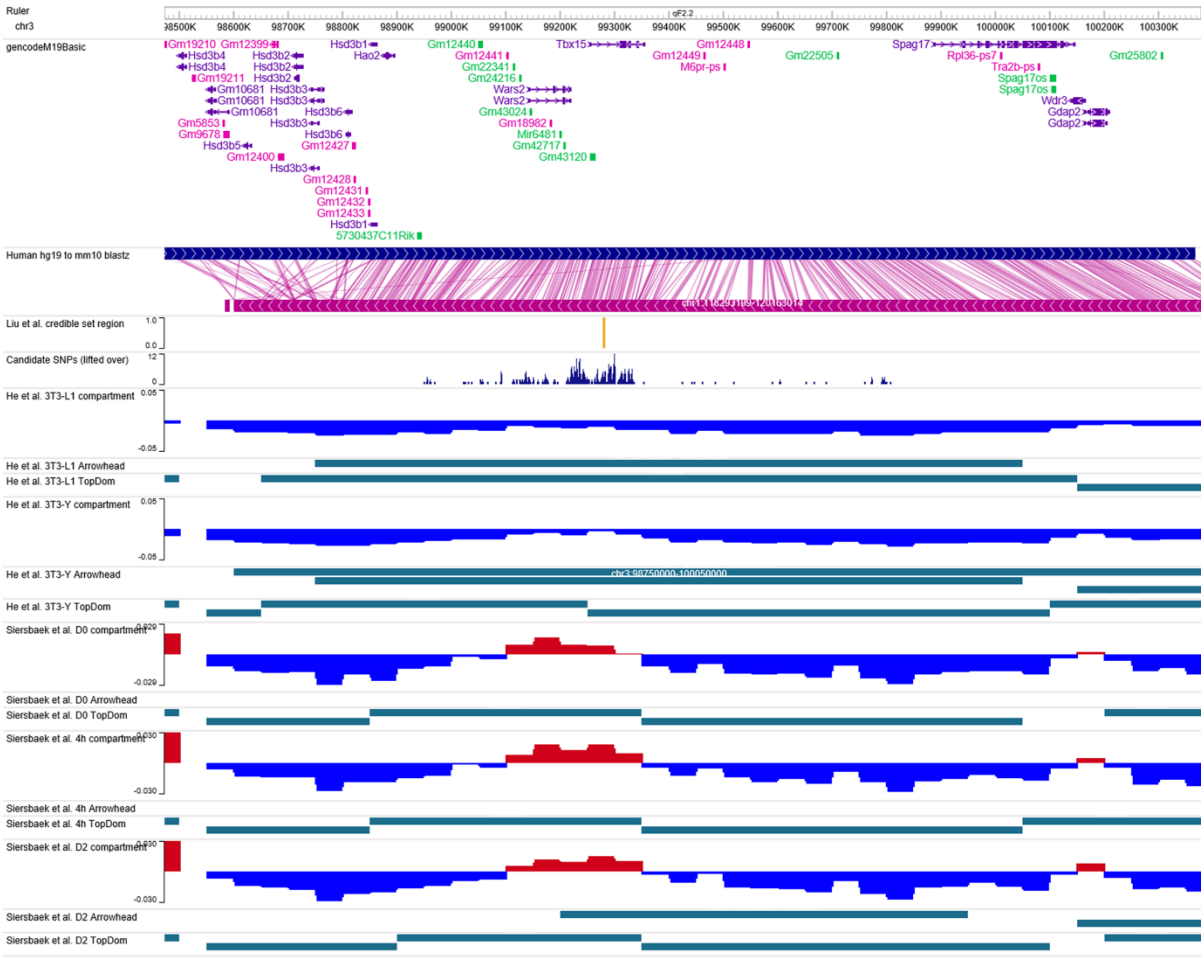


Figure 4.3: Mouse TADs and compartments in Tbx15-Wars2 locus. *WASHU genome browser view displaying the Tbx15-Wars2 locus in mouse. Tracks display: gencode gene annotations, hg19 to mm10 blastz, the equivalent location of the human credible set region after mapping to the mouse genome, the equivalent positions of all human candidate SNPs after mapping to the mouse genome (shown as a density track, any SNPs whose location could not be mapped to the mouse genome have been excluded), compartment annotations (A compartment (red) and B compartment (blue)), TADs called with Arrowhead and TopDom. TADs and compartments detected in pre-adipocyte cell line 3T3-L1 before and after differentiation (3T3-Y) from He et al. and 3T3-L1 cell line after 0 days (D0), 4 hours (4h) or 2 days (D2) of differentiation.*

4.3.4. Chromatin looping overlying candidate SNPs in the TBX15-WARS2 locus

The majority of candidate SNPs in the TBX15-WARS2 locus are in non-coding regions that potentially have regulatory function. Regulatory elements are thought to impact the transcription of their target genes by coming into close physical proximity with them, forming enhancer-promoter loops (Mora et al. 2015; Panigrahi and O'Malley 2021; Matharu and Ahituv 2015; Bonev and Cavalli 2016; Oudelaar and Higgs 2020). These regulatory loops can be profiled via 3C-based methods which include a capture step to enrich for either end of the regulatory interaction e.g. promoters (PCHi-C) (Schoenfelder et al. 2018) or enhancers (ECHi-C) (Madsen et al. 2020).

In order to investigate the potential regulatory loops overlying candidate SNPs in the TBX15-WARS2 locus, I identified looping interactions involving candidate SNPs in mouse and human pre-adipocytes, adipose or adipogenesis. Significant loops were detected using a Chicago score threshold of ≥5 for loops detected with Chicago or an FDR corrected p-value threshold <0.05 for loops detected using Homer (See Methods: 4.2.6). Loops which overlapped a candidate SNP at one or both ends were termed SNP loops. Potential regulatory SNP loops were defined as SNP loops which overlap a SNP but not a TSS at one end. Potential regulatory SNP loops could have regulatory impact via looping and were carried forward for further analysis. Other combinations of loops overlaying SNPs and TSS (other SNP loops) may have little to no regulatory impact associated with the loop and were not considered further (Figure 4.4).

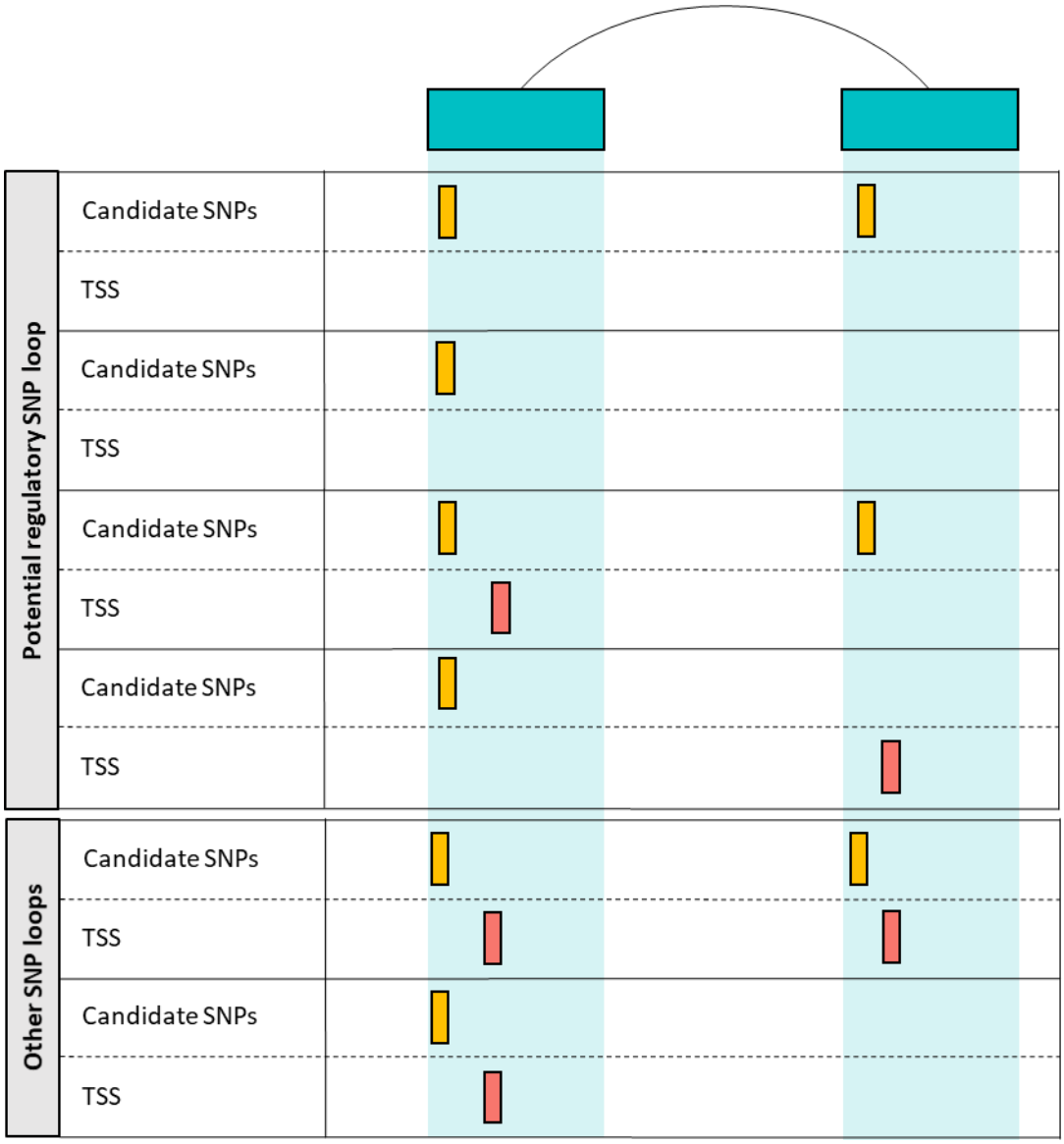


Figure 4.4: Schematic outlining the definition of potential regulatory SNP loops*. SNP loops overlap a candidate SNP at one or both ends of the loop. Potential regulatory SNP loops overlap a SNP but not a TSS (all annotated gene TSSs in the TBX15-WARS2 region downloaded from Ensembl were considered) at one or both ends of the loop. The mirror image of the depicted SNP and TSS overlaps (i.e. swapping the end of the loop at which the overlaps occur) have been omitted for simplicity.*

4.3.4.1. Overview of published looping datasets

To overlap candidate SNPs with chromatin loops, I initially downloaded published PCHi-C and ECHi-C loops from GEO or ArrayExpress. The selected published human looping datasets included data generated in SGBS pre-adipocytes (Sobreira et al. 2021), human white adipose (HWA) (Pan et al. 2018) and MSC differentiating towards adipocytes (three time points) (Madsen et al. 2020). This allowed the assessment of looping interactions at different stages of human adipogenesis. In order to assess the conservation of looping between human and mouse adipogenesis, a dataset generated in mouse preadipocytes differentiating towards adipocytes (three time points) was also selected (Siersbæk et al. 2017). Whilst this dataset does not perfectly match the tissues/developmental stages of the datasets selected in human, to my knowledge it is the only capture Hi-C currently publicly available for mouse adipose/adipogenesis. The loops published from the selected datasets were mainly produced using the HiCUP pipeline and Chicago R package. However, there are some important differences in the generation of the published loops between datasets including: PCHi-C or ECHi-C, the restriction enzyme, the software used, the software parameters and/or the use of binning (Table 4.4).

Table 4.4: Features of selected published human and mouse PCHi-C and ECHi-C data. *[#]Parameters specified according to ArrayExpress. *= minFragLen = 150, maxFragLen = 40000, maxLBrownEst = 1.5e6, binsize=20000, weightAlpha = 34.1157346557331, weightBeta = -2.58688050486759, weightGamma = -17.1347845819659, weightDelta = -7.07609245521541.*

Tissue/ cell type	Time point	Data type	Enzyme	Software	Original study
Human					
Primary human white adipocytes	N/A	PCHi-C	HindIII	HiCUP, CHiCAGO (default settings*)	Pan et al. 2018
Bone marrow–derived MSC	Day 0	ECHi-C (with 299 baited promoters)	MboI	HiCUP, CHiCAGO (with bespoke fragment pooling (median fragment length 1,963 bp) and custom-distance weights)	Madsen et al. 2020
Bone marrow–derived MSC (differentiating towards adipocytes)	Day 1	ECHi-C (with 299 baited promoters)	MboI	HiCUP, CHiCAGO (with bespoke fragment pooling (median fragment length 1,963 bp) and custom-distance weights)	Madsen et al. 2020
Bone marrow–derived MSC (differentiating towards adipocytes)	Day 10	ECHi-C (with 299 baited promoters)	MboI	HiCUP, CHiCAGO (with bespoke fragment pooling (median fragment length 1,963 bp) and custom-distance weights)	Madsen et al. 2020
SGBS pre-adipocytes	N/A	PCHi-C	MboI	Bowtie2, HOMER (findHiCInteractionsByChr.pl with parameters: -res 2000 -superRes 10000 [#])	Sobreira et al. 2021
Mouse					
3T3-L1 preadipocytes (differentiating towards adipocytes)	Day 0	PCHi-C	HindIII	HiCUP, CHiCAGO (with additional filtering and parameters undeclared)	Siersbæk et al. 2017
3T3-L1 preadipocytes (differentiating towards adipocytes)	4 hours	PCHi-C	HindIII	HiCUP, CHiCAGO (with additional filtering and parameters undeclared)	Siersbæk et al. 2017
3T3-L1 preadipocytes (differentiating towards adipocytes)	Day 2	PCHi-C	HindIII	HiCUP, CHiCAGO (with additional filtering and parameters undeclared)	Siersbæk et al. 2017

4.3.4.2. SNP loops present in published looping datasets

4.3.4.2.1. Human SGBS preadipocytes

I first investigated the significant loop interactions for the SGBS preadipocytes (Sobreira et al. 2021) (FDR corrected p-value threshold <0.05). There were two SNP loops in this region (Chapter 4: Supplementary Figure 1). However, neither of the SNP loops were defined as potential regulatory SNP loops.

There were no loops which overlapped with any of the SNPs in the credible set or any of the proxy SNPs in the credible set, suggesting that no candidate SNPs were involved in detectable regulatory interactions in SGBS preadipocytes.

4.3.4.2.2. Primary human white adipocytes

I next investigated the significant loops in HWA (Pan et al. 2018) (Figure 4.5). There were 21 SNP loops in this locus (Chapter 4: Supplementary Figure 2). However, only two were classified as potential regulatory SNP loops. Both of these loops interacted with the TBX15 TSS at one end, suggesting that they could be regulating the transcription of TBX15 (Figure 4.5).

Again, no interactions in this dataset overlapped any of the SNPs in the credible set or any of the proxy SNPs of the credible set, suggesting that this region was not involved in detectable regulatory interactions in HWA.

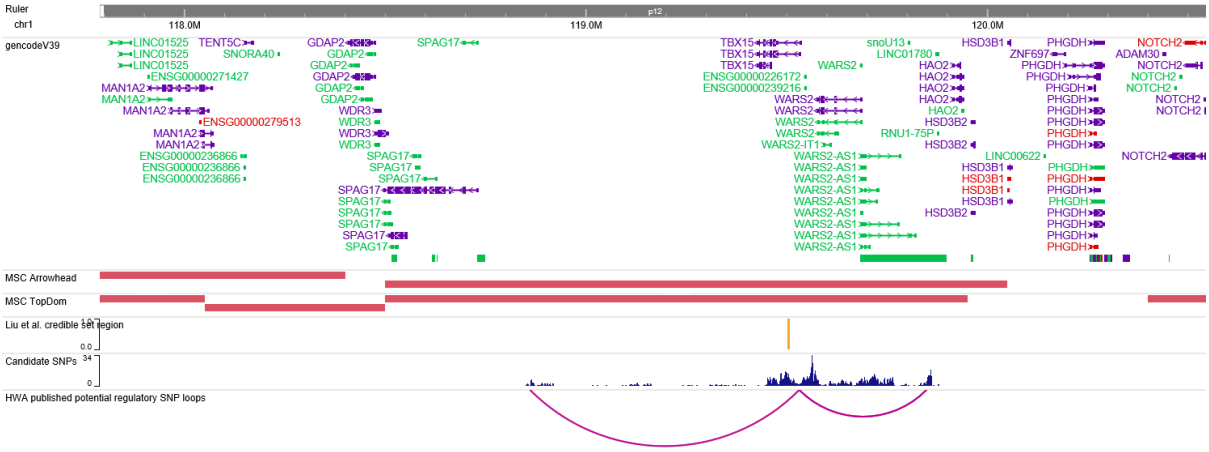


Figure 4.5: Published potential regulatory SNP loops in HWA in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and potential regulatory SNP loops.*

4.3.4.2.3. Human MSC differentiating towards adipocytes

In order to investigate regulatory looping in the locus at different stages of adipogenesis, I next investigated the overlap of candidate SNPs with published ECHi-C loops during human adipogenesis (Madsen et al. 2020). In this dataset, loops were profiled at three separate adipogenesis time points (MSC (day 0), differentiation day 1 and differentiation day 10). This dataset was also very useful because the capture library captured looping from enhancer regions, meaning interactions involving enhancers were enriched in the dataset. In total, there were 46/50, 43/46, and 58/62 potential regulatory SNP loops/SNP loops at day 0, day 1 and day 10 respectively (Figure 4.6 and Chapter 4: Supplementary Figure 3).

The index SNP (rs984222) in the credible set region overlapped an adipogenesis enhancer (along with 5 other candidate SNPs) and there were 9, 10, and 8 looping interactions involving this enhancer in day 0, day 1 and day 10 respectively. These loops included looping interactions with the TBX15 TSS at day 1 and day 10, and with the WARS2 and the RP11-418J17.1 TSS at day 10. This suggests that rs984222 overlies an enhancer involved in regulating TBX15 and WARS2 during adipogenesis.

Many of the potential regulatory SNP loops did not interact with a TSS, making interpretation of their regulatory function difficult. I therefore focused on the potential regulatory SNP loops which were interacting with a gene TSS, suggesting that they may be regulating the expression of the gene. These loops included four interactions between candidate SNPs and the TBX15 TSS, and three interactions between candidate SNPs and the TSSs of WARS2 and RP11-418J17.1 (significant in at least one time point).



Figure 4.6: Published potential regulatory SNP loops in human MSC differentiating towards adipocytes in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and potential regulatory SNP loops at day 0, day 1 and day 10.*

4.3.4.2.4. Mouse pre-adipocytes differentiating towards adipocytes

Using PCHi-C from mouse pre-adipocytes differentiating towards adipocytes at three time points (day 0, 4 hours and day 2), I assessed how well the regulatory looping landscape at this locus is conserved in mice (Siersbæk et al. 2017). There were 23, 13 and 15 SNP loops in the *Tbx15-Wars2* locus at day 0, 4 hours and day 2 respectively (Chapter 4: Supplementary Figure 4). Of these SNP loops, there were 1 and 4 potential regulatory SNP loops at day 0 and 4 hours respectively (there were no loops of this type at day 2) (Figure 4.7). The potential regulatory

SNP loops involved the *Hsd3b6* TSS at day 0; and the *Wdr3*, *Gdap2*, *Hsd3b1*, and *Spag17* TSSs at 4 hours, suggesting they could be regulating the expression of these genes. No significant loop in any time point overlapped any SNP from the credible set or a proxy SNP of the credible set SNPs, suggesting that the SNPs in this region were not involved in detectable regulatory looping during mouse adipogenesis in this dataset.



Figure 4.7: Published potential regulatory SNP loops in mouse adipogenesis in the Tbx15-Wars2 locus. *WASHU genome browser view displaying the Tbx15-Wars2 locus and surrounding area. Tracks display: gencode gene annotations, 3T3-L1 TADs called with Arrowhead and TopDom, the equivalent location of the human credible set region after mapping to the mouse genome, the equivalent positions of all human candidate SNPs after mapping to the mouse genome (shown as a density track, any SNPs whose location could not be mapped to the mouse genome have been excluded) and potential regulatory SNP loops at day 0, 4 hour and day 2.*

4.3.4.3. Limitations of published datasets

As noted above, there are several technical differences between the loops published by the selected datasets which are important to consider. The majority of the datasets are PCHi-C meaning that the Hi-C library has been enriched for interactions involving promoters (Schoenfelder et al. 2018). However, the human MSC differentiating towards adipocyte dataset is ECHi-C, meaning that the Hi-C library has been enriched for interactions involving enhancers identified in the corresponding cell type (Madsen et al. 2020). This allows the investigation of loops captured from the other end of the regulatory interaction. Capturing regulatory interactions occurring during adipogenesis from both ends could facilitate a more detailed dissection of regulatory interactions. There was also a difference in the enzyme used to generate the datasets, i.e. the 6-cutter restriction enzyme HindIII, or the 4-cutter restriction enzyme MboI. This is important because the fragment size produced by digestion with HindIII is much larger than MboI (average fragment size in the human genome ~3,700 bp vs ~430 bp respectively (Freire-Pritchett et al. 2021)). This means that the number of possible restriction fragment pairs is greater for MboI, which can lead to greater data sparsity in experiments using MboI. Typically, this can lead to reduced ability to detect longer range interactions in experiments using MboI and conversely reduced ability to detect shorter range interactions in experiments using HindIII due to decreased sensitivity (Freire-Pritchett et al. 2021). The popular loop calling software CHiCAGO recommends that different settings should be used when analysing an experiment generated using MboI (the default settings are designed for use with HindIII digested data). These MboI-specific settings are designed to avoid the detection of spurious low count interactions as significant (Freire-Pritchett et al. 2021; Cairns et al. 2016). CHiCAGO was originally used as the analysis software in many of the selected studies; however, the parameter choice has not always been reported (Table 4.4). Further to

this, in some cases the published loops have been detected after the deployment of bespoke fragment pooling (Madsen et al. 2020). However, it is not clear whether further CHiCAGO parameter optimisation was undertaken or required to account for this.

4.3.4.4. Overview of the datasets and methodology for loops recalled via a standardised pipeline

Most of the published loop datasets assessed were processed using the same analysis tools (HiCUP and CHiCAGO) (Table 4.4). However, as noted, there was variation in how these analysis tools were applied and reported. In order to assess whether the published significant SNP loops were robust to the parameter choice/ loop calling tool used in the original publication, I recalled the loops in each dataset using HiCUP to align and quality control (QC) reads and then CHiCAGO to detect loops. HiCUP QC indicated that in all human datasets, the majority of the reads were "valid pairs", indicating that the libraries were good quality. However, the mouse adipogenesis datasets had a larger fraction of reads with QC issues than the human datasets, particularly re-ligation (reads representing adjacent restriction fragments rather than 3D contacts) and same dangling ends (reads which overlap the restriction cut site but map to the same restriction fragment) resulting in a greater proportion of reads being filtered (Wingett et al. 2015). However, in the majority of mouse datasets, >50% of sequenced reads were "valid pairs" (Figure 4.8). Loops were then detected using the CHiCAGO settings recommended in Freire-Pritchett et al. (see Methods: 4.2.6). For datasets generated with MboI, the smaller restriction fragment size (compared to HindIII) leads to lower interaction level coverage, which can be increased by binning. Therefore, as

recommended by Freire-Pritchett et al., datasets produced using MboI were analysed in 5 kb bins alongside the un-binned data (Freire-Pritchett et al. 2021).

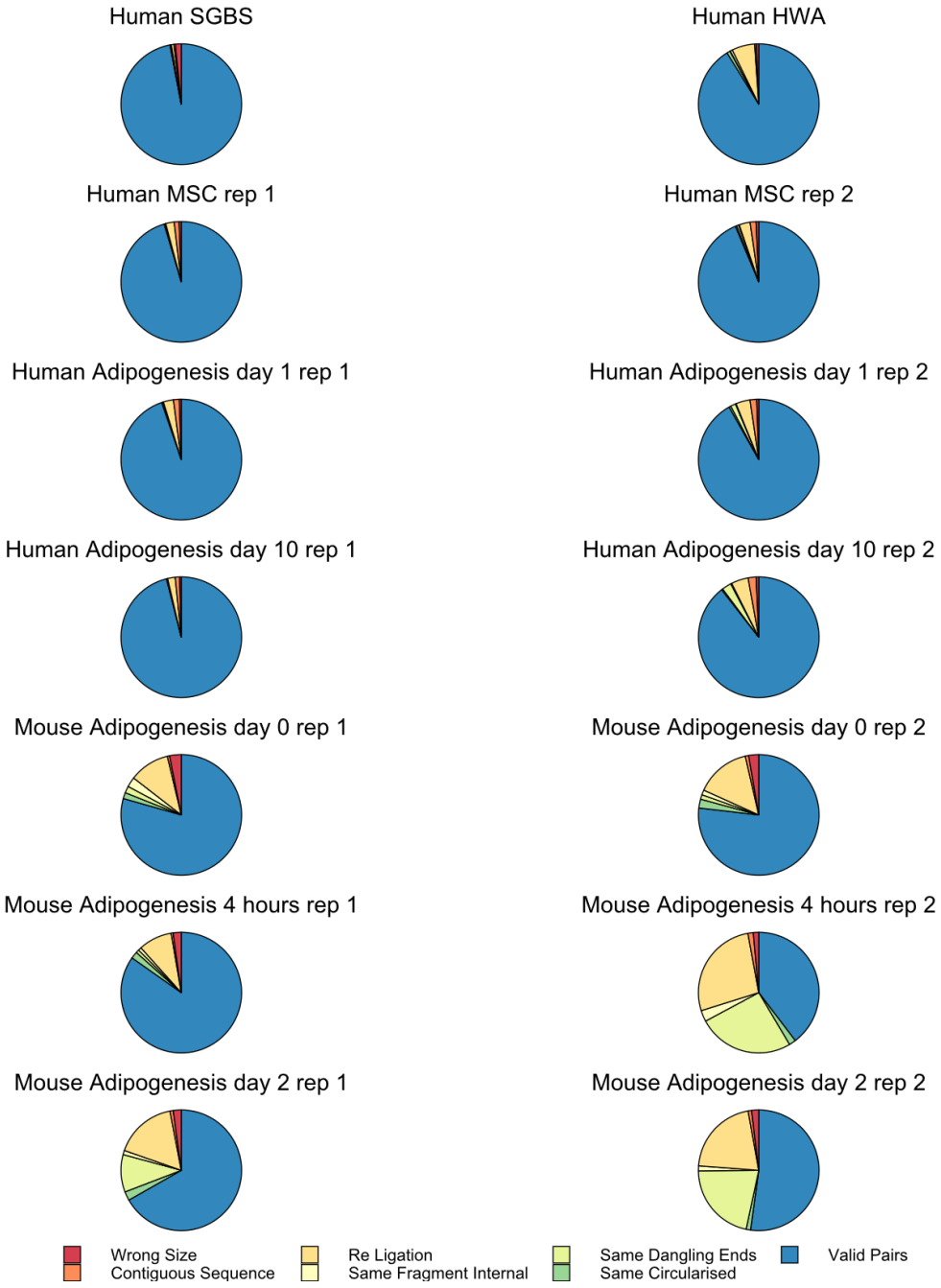


Figure 4.8: Proportion of reads in each HiCUP QC category per dataset. *Valid read pairs are indicated in blue. The proportion of reads filtered due to wrong size, contiguous sequence, re ligation, same fragment internal, same dangling ends or same circularised QC issues are indicated by other colours.*

I compared the number and size of significant *cis* loops in my recalled datasets with the original published datasets (Figure 4.9). For datasets produced using HindIII (HWA and mouse adipogenesis), I found a high percentage of loops were detected in both the published and recalled datasets. Between 84.65- 94.29% of published loops were also detected in the recalled dataset and between 76.91%- 87.6% of recalled loops were also detected in the published dataset. This indicates that loops published in these datasets are highly reproducible with the recommended CHiCAGO settings. In the human adipogenesis datasets, the percentage of published loops which were also detected in the recalled datasets was lower; between 56.78- 58.61% of published loops were detected in the recalled datasets. This is likely due to the use of custom settings and fragment binning in the data published by this study. The highest agreement between recalled and published loops in this dataset was achieved in the binned data; 76.48- 77.15% of recalled loops in binned data were also detected in the published dataset. In the SGBS dataset, the agreement between published and recalled loops was the lowest observed. Only 12.15% of published loops in this dataset were also detected in the recalled loops, indicating very low reproducibility of the published loops with the recommended CHiCAGO settings. This is likely due to the use of a different (Hi-C specific) software in this study (Figure 4.9A). For datasets generated with HindIII (HWA and mouse adipogenesis), the size distribution was very similar between published and recalled loops, whereas for datasets generated with MboI (SGBS and human adipogenesis), as expected, smaller loops were detected using un-binned data compared to binned data (Freire-Pritchett et al. 2021), with the binned data having a more similar size distribution to the published loops (Figure 4.9B).



Figure 4.9: Global comparison of published and recalled loops. *A) The number of significant cis loops detected in the published datasets (top) and the recalled datasets (bottom). The percentage of loops detected in both the published and recalled datasets is indicated. B) Distribution of the length of detected loops in both published and recalled data. Length is calculated as the distance between the midpoints of the loop anchors.*

To assess whether the same loops overlapped SNPs in the published and recalled datasets, I repeated the SNP overlap using the recalled loops and compared the detected SNP loops and potential regulatory SNP loops to those detected using the published loops.

4.3.4.5. SNP loops present in recalled looping datasets

4.3.4.5.1. Human SGBS preadipocytes

In the published SGBS looping data, there were only two SNP loops, but neither was a potential regulatory SNP loop. In the recalled looping data there were also very few SNP loops. In the un-binned SGBS recalled loops I detected no SNP loops, whereas in the binned data there were 2 SNP loops (neither of which were detected in the published loops) (Chapter 4: Supplementary Figure 5). The SNP loops included 1 potential regulatory SNP loop. This loop overlapped the TBX15 TSS, indicating that it could be involved in regulating the expression of TBX15 (Figure 4.10).

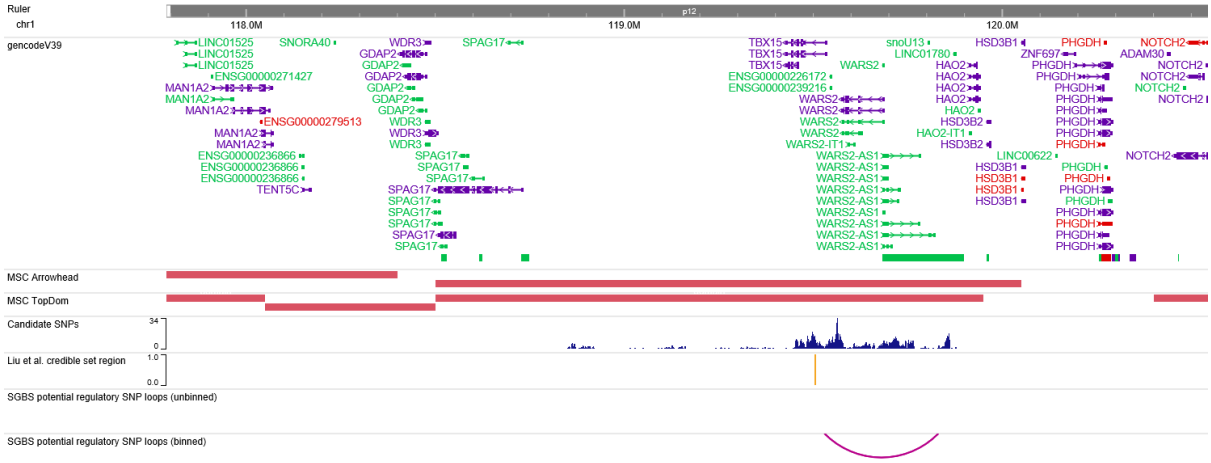


Figure 4.10: Recalled potential regulatory SNP loops in SGBS in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and potential regulatory SNP loops (detected with un-binned and binned data).*

4.3.4.5.2. Primary human white adipocytes

In HWA, the published looping data indicated that there were 21 SNP loops in the TBX15-WARS2 locus. Of these, two SNP loops were potential regulatory interaction SNP loops. Both were involved in interactions with the TBX15 TSS. In the recalled looping data, I detected 23 SNP loops (Chapter 4: Supplementary Figure 6) (of which 16 (69.57% of recalled SNP loops) were also detected in the published loops). Both of the published potential regulatory SNP loops interacting with TBX15 were also detected in the recalled looping data, indicating that they are reproducibly significant with standard CHiCAGO settings. However, there were also two further potential regulatory SNP loops. Both loops overlapped the HAO2 and HAO2-IT1

indicating that they may be involved in regulating the expression of these genes in HWA (Figure 4.11).

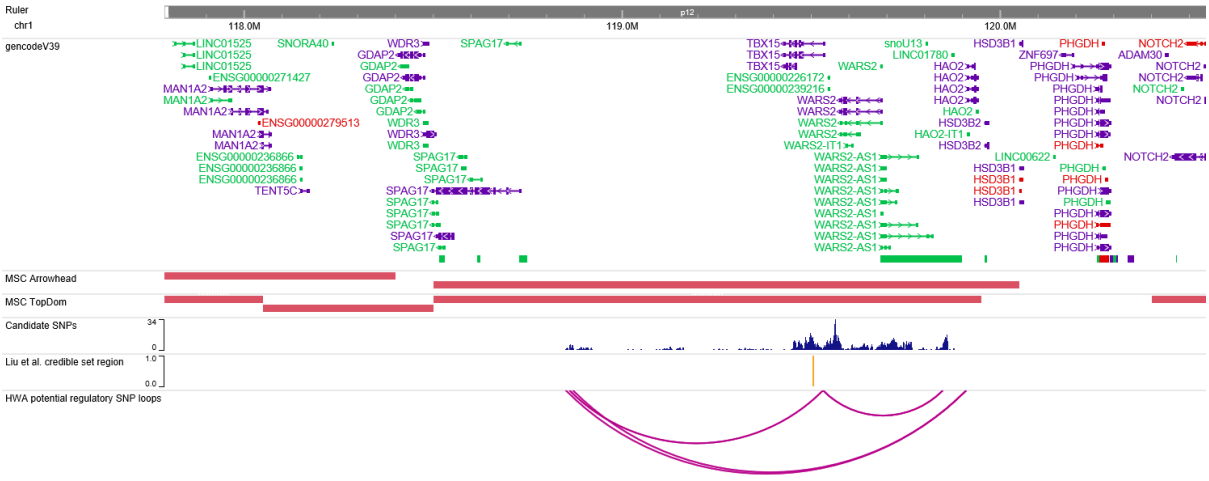


Figure 4.11: Recalled potential regulatory SNP loops in HWA in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and potential regulatory SNP loops.*

4.3.4.5.3. Human MSC differentiating towards adipocytes

In the published looping data, 50, 46, and 62 SNP loops were detected at day 0, day 1 and day 10 respectively. In the recalled looping data I detected 10, 19 and 9 SNP loops in the un-binned data and 64, 49 and 39 SNP loops in the binned data at day 0, day 1 and day 10 respectively (Chapter 4: Supplementary Figure 7). Only 1 (10%), 5 (26.32%) and 3 (33.33%) of SNP loops in the un-binned data and 28 (43.75%), 7 (14.29%) and 11 (28.21%) of SNP loops in binned data

were also detected in the published looping data at day 0, day 1 and day 10 respectively (Chapter 4: Supplementary Figure 9).

Similarly to the SNP loops detected in the published data, many of the SNP loops in the recalled data were potential regulatory interaction SNP loops. There were on average ~12 and ~32 potential regulatory SNP loops per time point in the un-binned and binned data respectively (Figure 4.12).

As seen in the published looping data, SNP loops involving rs984222 (the index SNP in the Liu et al. credible set) were detected in the recalled loops. In the un-binned data, there were two potential regulatory SNP loops involving rs984222 (along with one other candidate SNP) detected at day 1 of adipogenesis. One of these loops was also identified in the published data; however, neither of the loops interacted with a TSS so the regulatory significance of these loops is unclear. In the binned data, there were 11, 9 and 4 potential regulatory SNP loops of which 5, 1 and 1 were also detected in the published looping data at day 0, day 1 and day 10 respectively, overlapping rs984222 (along with 16 other candidate SNPs, including rs984225 which was also in the credible set). These loops included looping interactions with the SPAG17 TSS at day 0 of adipogenesis and RP11-418J17.1 TSS at day 10 of adipogenesis. Interestingly, no significant looping interactions involving rs984222 and the TBX15 or WARS2 TSS were detected in the recalled data, although these were detected in the published looping data. This suggests that these looping interactions are sensitive to the settings used in CHiCAGO or the degree of binning used. As an example, the recalled looping interaction in the binned data between rs984222 and SPAG17 was also detected in the published loops; however, because in the published loops the custom fragment binning produced smaller fragment sizes than in the recalled binned data, it did not overlap the SPAG17 TSS.

Many of the remaining potential regulatory SNP loops did not interact with a gene TSS, making interpretation of their regulatory significance difficult. I therefore only focused on potential regulatory SNP loops which were interacting with a gene TSS in either the binned or un-binned data. At all-time points, potential regulatory SNP loops were detected which interacted with the TSSs of RP11-418J17.1, HAO2 and HAO2-IT1, suggesting that these genes are regulated by candidate SNPs throughout adipogenesis. At day 0 and day 10 (but not day 1), there were potential regulatory SNP loops which interacted with the TSSs of WARS2 and TBX15. Finally, there were potential regulatory SNP loops which interacted with the TSSs of RP4-712E4.1 and RP4-712E4.2 at day 0 only; SPAG17 at day 0 and day 1; and RPL6P2 at day 10 only. These loops indicate that these genes could be regulated by candidate SNPs over human adipogenesis.

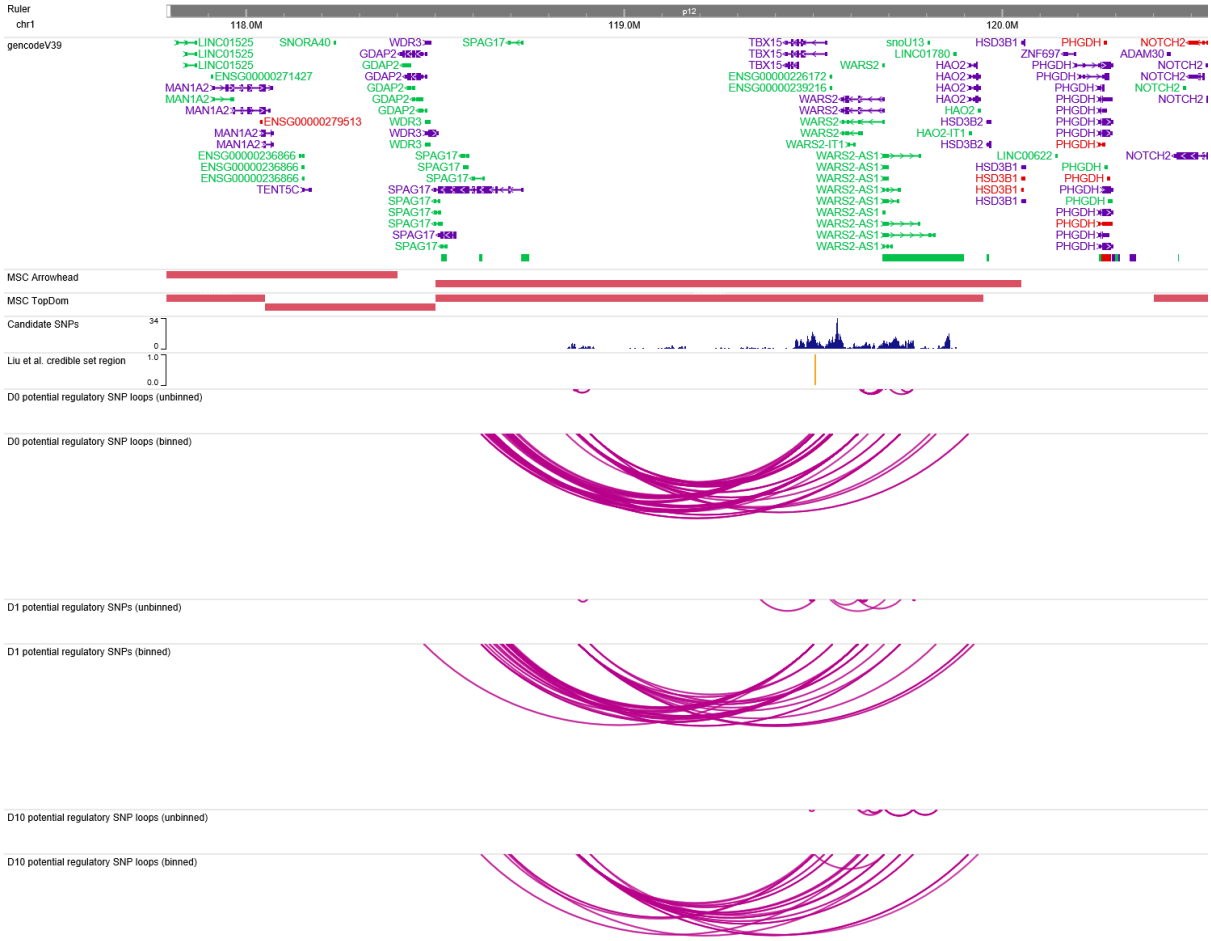


Figure 4.12: Recalled potential regulatory SNP loops in human MSC differentiating towards adipocytes in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and potential regulatory SNP loops at day 0, day 1 and day 10 (detected with un-binned and binned data).*

4.3.4.5.4. Mouse pre-adipocytes differentiating towards adipocytes

In the published looping data, there were 23, 13 and 15 SNP loops in the *Tbx15*-*Wars2* locus at day 0, 4 hours and day 2 of adipogenesis respectively. In the recalled looping data, I detected 28, 19 and 18 SNP loops at day 0, 4 hours and day 2 of adipogenesis respectively (Chapter 4: Supplementary Figure 8). Similarly to the relationship observed globally, many of the recalled SNP loops in this locus were also detected in the published looping datasets: 21 (75%), 12 (63.16%) and 14 (77.78%) of recalled SNP loops at day 0, 4 hours and day 2 of adipogenesis respectively, were also detected in the published data (Chapter 4: Supplementary Figure 9).

The potential regulatory SNP loops in the recalled data were identical to the published data. There was one potential regulatory SNP loop at day 0 of adipogenesis which interacted with the TSS of *Hsd3b6*. At 4 hours of adipogenesis there were 4 potential regulatory SNP loops interacting with the TSSs of *Wdr3* and *Gdap2 (2 loops)*; *Hsd3b1*; or *SPAG17* (Figure 4.13), indicating that these loops are reproducibly significant with standard CHiCAGO settings.

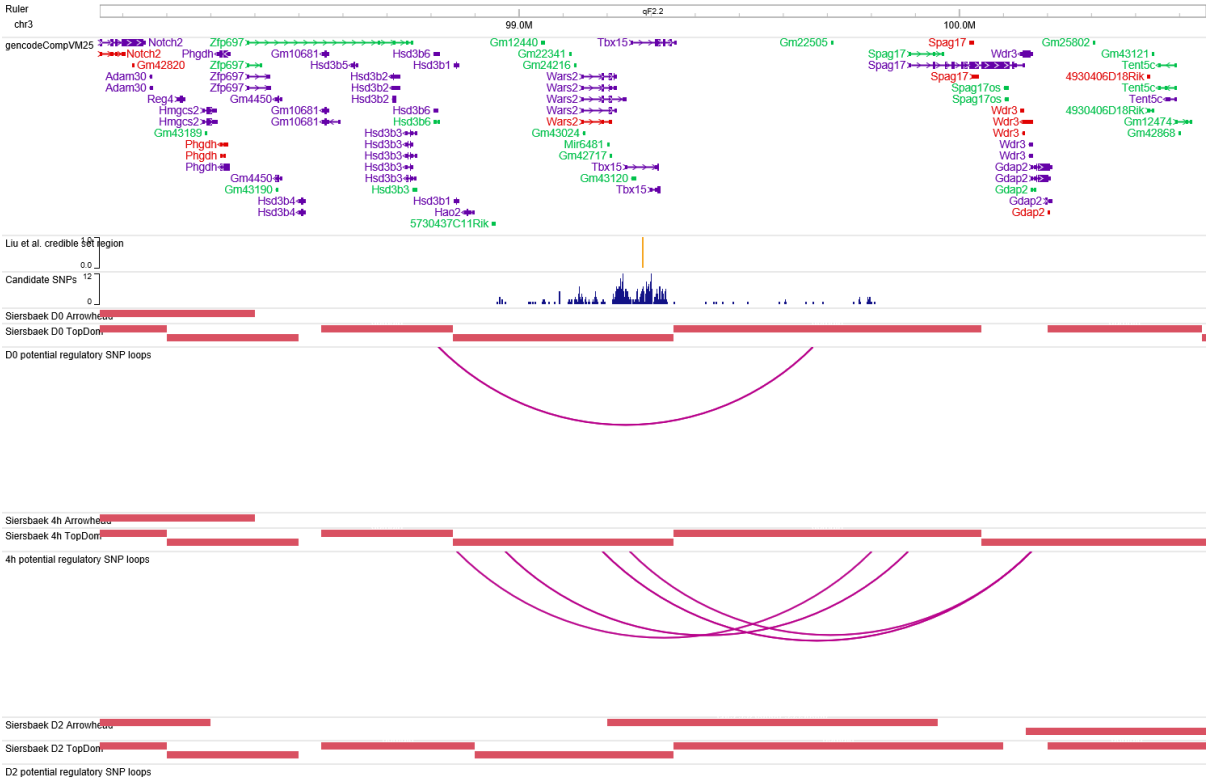


Figure 4.13: Recalled potential regulatory SNP loops in mouse pre-adipocytes differentiating towards adipocytes in the Tbx15-Wars2 locus. *WASHU genome browser view displaying the Tbx15-WarS2 locus and surrounding area. Tracks display: gencode gene annotations, 3T3-L1 TADs called with Arrowhead and TopDom, the equivalent location of the human credible set region after mapping to the mouse genome, the equivalent positions of all human candidate SNPs after mapping to the mouse genome (shown as a density track, any SNPs whose location could not be mapped to the mouse genome have been excluded) and potential regulatory SNP loops at day 0, 4 hours and day 2.*

4.3.4.6. No consensus between chromatin loops in different datasets

I reasoned that SNPs in this region which are involved in regulatory interactions important in adipogenesis are likely to have been detected in multiple of the datasets analysed. In order to investigate this, I prioritised SNP-gene interactions based on the number of datasets in which they had been detected (Table 4.5). I also assessed whether the interaction between the SNP and the gene was detected in both the published and recalled data, as this provides an indication of whether the significance of the interaction is robust to the software and parameters used. I found that the only SNP-gene interactions which were significantly detected in multiple samples were common to different human adipogenesis time points from the same study (Madsen et al. 2020). There were a total of 14 SNPs and 5 genes involved in these interactions, including rs984222 interacting with TBX15. However, the majority of these SNP-gene associations were not detected in both published and recalled datasets, suggesting they are not robust to the software and parameters used. Overall, the lack of consensus SNP-gene interactions from different studies makes it difficult to prioritise particular SNP-gene interactions in this locus. Unfortunately, the lack of common SNP-gene interactions between mouse and human adipogenesis in the TBX15-WARS2 locus could also suggest that the regulatory circuitry in this region is not conserved between human and mouse.

Table 4.5: SNP-gene interactions occurring in multiple analysed looping datasets. *A dash indicates that the SNP-gene interaction was not detected in either published or recalled loops, True (green) indicates that the SNP-gene interaction was detected in both published and recalled loops, and False (red) indicates that the SNP-gene interaction was detected in either published or recalled loops.*

Candidate SNP	Gene	HWA	SGBS	Human adipogenesis day 0	Human adipogenesis day 1	Human adipogenesis day 10	Mouse adipogenesis day 0	Mouse adipogenesis 4 hours	Mouse adipogenesis day 2	N
rs10802074	RP11-418J17.1	-	-	TRUE	FALSE	FALSE	-	-	-	3
rs10802074	TBX15	-	-	FALSE	FALSE	FALSE	-	-	-	3
rs10802074	WARS2	-	-	TRUE	FALSE	FALSE	-	-	-	3
rs2885226	RP11-418J17.1	-	-	TRUE	FALSE	FALSE	-	-	-	3
rs2885226	TBX15	-	-	FALSE	FALSE	FALSE	-	-	-	3
rs2885226	WARS2	-	-	TRUE	FALSE	FALSE	-	-	-	3
rs12752809	RP11-418J17.1	-	-	FALSE	FALSE	FALSE	-	-	-	3
rs10923708	TBX15	-	-	-	FALSE	FALSE	-	-	-	2
rs10923709	TBX15	-	-	-	FALSE	FALSE	-	-	-	2
rs10923710	TBX15	-	-	-	FALSE	FALSE	-	-	-	2
rs10923711	TBX15	-	-	-	FALSE	FALSE	-	-	-	2
rs12045138	RP11-418J17.1	-	-	-	FALSE	TRUE	-	-	-	2
rs12045138	TBX15	-	-	-	FALSE	FALSE	-	-	-	2
rs12045138	WARS2	-	-	-	FALSE	TRUE	-	-	-	2
rs984222	TBX15	-	-	-	FALSE	FALSE	-	-	-	2
rs984223	TBX15	-	-	-	FALSE	FALSE	-	-	-	2
rs12752809	HAO2	-	-	-	FALSE	FALSE	-	-	-	2
rs12752809	HAO2-IT1	-	-	-	FALSE	FALSE	-	-	-	2
rs12752809	TBX15	-	-	FALSE	-	FALSE	-	-	-	2
rs17038218	HAO2	-	-	FALSE	-	FALSE	-	-	-	2
rs35043843	HAO2	-	-	FALSE	-	FALSE	-	-	-	2
rs35501033	HAO2	-	-	FALSE	-	FALSE	-	-	-	2
rs61808932	TBX15	-	-	-	FALSE	FALSE	-	-	-	2

4.3.5. Global human vs mouse regulatory circuitry conservation score

I hypothesised that if the regulatory circuitry of a locus is reflected by chromatin loops, then loci which are important in adipogenesis will contain a larger number of loops during human adipogenesis, compared to loci which are not important in adipogenesis. Furthermore, loci which are important in adipogenesis will have a higher degree of looping conservation between human and mouse, indicating that they have a higher suitability for modelling in mouse. In order to test these hypotheses, I compared genome-wide mouse and human looping during adipogenesis. For this comparison, I used standardised recalled looping data for the datasets: Human MSC differentiating towards adipocytes and Mouse pre-adipocytes differentiating towards adipocytes. These datasets differ in two key respects, the human data was generated with the enzyme MboI, whereas the mouse data was generated with HindIII (meaning the human data has smaller fragment size); and the human data is ECHi-C, whereas the mouse data is PCHi-C. To increase the comparability of the datasets, only human loops detected using binned fragments were considered (as these loops are detected at a similar resolution to data generated with HindIII). I then further filtered human loops so that only loops overlapping a promoter were considered (thereby creating “pseudo PCHi-C”) (see Methods: 4.2.9). The total number of human loops, the number of human loops conserved in mouse and the proportion of conserved loops per TAD were used in further analysis (Figure 4.14A) (using merged Arrowhead and TopDom TADs (see Methods: 4.2.9)).

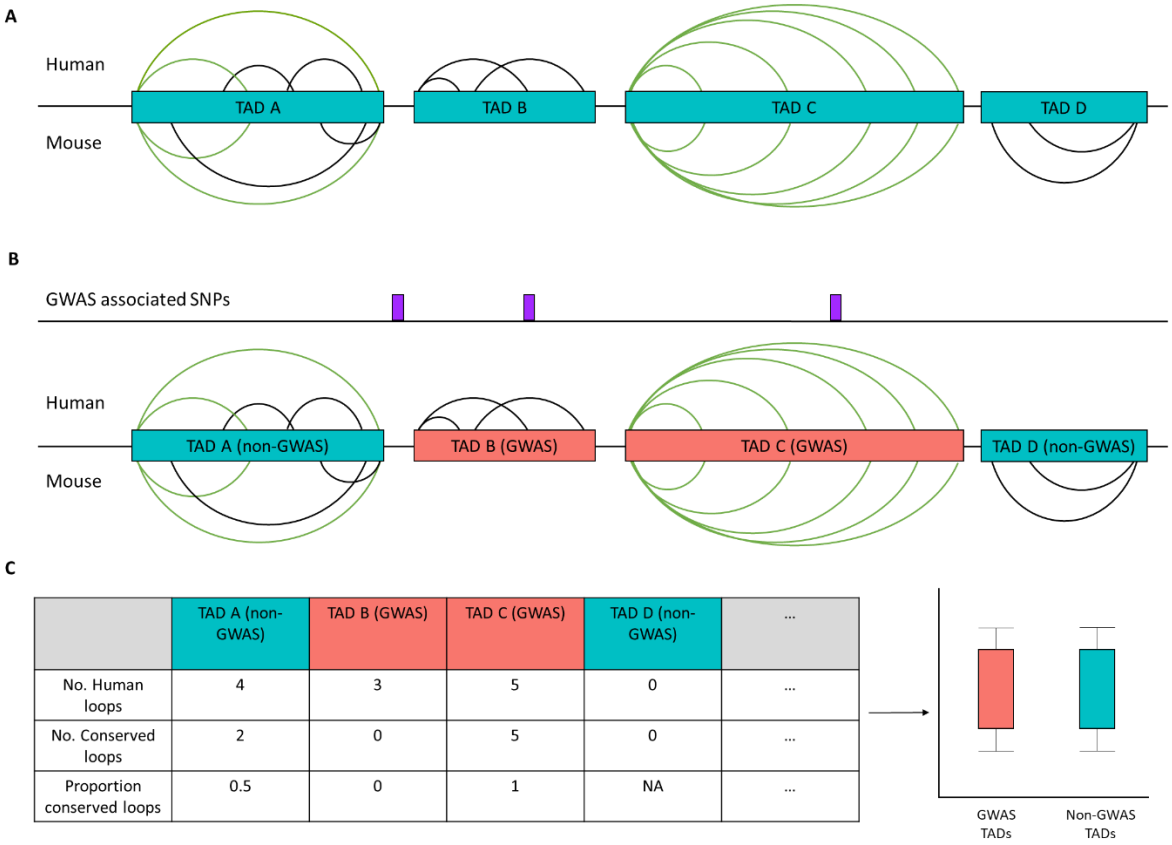


Figure 4.14: Schematic outlining the calculation of the global human vs mouse regulatory circuitry conservation score. *A) Loci were defined by TADs, and human and mouse (lifted over on to hg19) loops within each TAD were compared (conserved loops are shown in green). B) GWAS TADs were defined as TADs containing a significant GWAS-associated SNP. B) The number of humans loops, the number of conserved loops and the proportion of conserved loops were calculated per TAD and distribution of scores was compared between GWAS and non-GWAS TADs.*

TADs were then annotated as GWAS TADs or non-GWAS TADs based on whether they contained a WHRadjBMI associated SNP (from Pulit et al.). The total number of human loops, the number of conserved loops and the proportion of conserved loops were compared between GWAS and non-GWAS TADs (Figure 4.14B-C). I found that GWAS TADs contained significantly more loops during adipogenesis than non-GWAS TADs (p-value<0.001) (mean number of loops in GWAS TADs: ~31 and non-GWAS TADs: ~12). This suggests that loci

associated with WHRadjBMI take part in a higher frequency of chromatin looping during adipogenesis than other loci. I next assessed whether GWAS TADs contained a greater number of loops which are conserved in mouse adipogenesis than non-GWAS TADs. Similarly to the total number of human loops, I found that GWAS TADs contain significantly more loops which are conserved in mouse adipogenesis than non-GWAS TADs (p-value<0.001) (mean number of conserved loops in GWAS TADs: ~4 and non-GWAS TADs: ~1). This suggests that WHRadjBMI associated loci have a greater number of conserved loops between mouse and human adipogenesis than non-GWAS TADs, suggesting that these loci may be good candidates to model the effects of these variants using a mouse model. Finally, I compared the proportion of conserved loops (in TADs which contained >0 human loop) between GWAS TADs and non-GWAS TADs. I found that on average, the proportion of conserved loops was slightly higher in GWAS TADs compared to non-GWAS TADs (mean proportion of conserved loops in GWAS TADs: ~0.15 and non-GWAS TADs: ~0.12); however, this relationship was not statistically significant (Figure 4.15A). This is potentially due to the bias for TADs with fewer loops to have a higher proportion of conserved loops (Chapter 4: Supplementary Figure 10).

I next assessed whether the above relationships were still present if GWAS TADs were defined using a phenotypic trait which was not expected to be linked to adipogenesis. To do this, I used schizophrenia-associated GWAS loci from Pardiñas et al. to define GWAS and non-GWAS TADs (Pardiñas et al. 2018). I found no significant difference between GWAS TADs and non-GWAS for any of the measures investigated, indicating that the relationship between GWAS-associated loci and regulatory loop conservation is dependent on the detection of loops in a trait relevant tissue/cell type, i.e. the regulatory conservation is tissue-specific (Figure 4.15A).

As illustrated above, the TBX15-WARS2 locus contained no looping conservation between human and mouse during adipogenesis. A mouse model of the human association with fat distribution at this loci also showed no significant phenotypes (manuscript in preparation). On the other hand, mouse models of the GWAS association with obesity at the FTO locus have shown significant adipose relevant molecular, cellular and organismal phenotypes (Laber et al. 2021; Sobreira et al. 2021), suggesting that the FTO locus is a good example of a region with a conserved adipogenic phenotype between species. I therefore hypothesised that if loci with higher conservation of regulatory looping between mouse and human are more amenable to modelling in mouse, then the FTO locus should show greater looping conservation compared to the TBX15-WARS2 locus. To assess this, I compared the total number of human loops during adipogenesis, the number of conserved loops, and the proportion of conserved loops in the TAD containing the TBX15-WARS2 locus with the TAD containing the FTO locus. As expected for two loci containing obesity/fat distribution associations, both loci contained a similar total of human loops over adipogenesis (TBX15-WARS2: 49 loops and FTO: 46 loops). However, while ~30% of these loops were conserved in mouse adipogenesis in the FTO locus (14 loops), there were no loops conserved over mouse adipogenesis in the TBX15-WARS2 locus, indicating that compared to the TBX15-WARS2 locus, the conservation of adipogeneic regulatory circuitry between mouse and human is much higher in the FTO locus (Figure 4.15B).

To test whether the relationship between phenotypic and looping conservation seen above is translated to a wider dataset, I downloaded all knockout genes found to be significantly associated with “abnormal adipose tissue morphology” ($p < 0.0001$) from IMPC. I defined IMPC TADs as TADs which contained a human homolog of one of these genes (Dickinson et al. 2016). I found that IMPC GWAS TADs contained the largest number of conserved loops

(mean= ~5), significantly more than IMPC non-GWAS TADs (mean=~2) (p-value<0.01), indicating that TADs with a higher conservation of regulatory circuitry (measured by the number of conserved loops) are more likely to contain genes/variants associated with adipose-related traits in both mouse and human. Interestingly, I found no significant difference between the number of conserved loops in IMPC GWAS TADs and non-IMPC GWAS TADs. This probably reflects the incomplete nature of the IMPC resource (7824 knockout genes have been phenotyped to date); therefore, there are likely to be more genes associated with "abnormal adipose tissue morphology" than have currently been detected (Figure 4.15C).

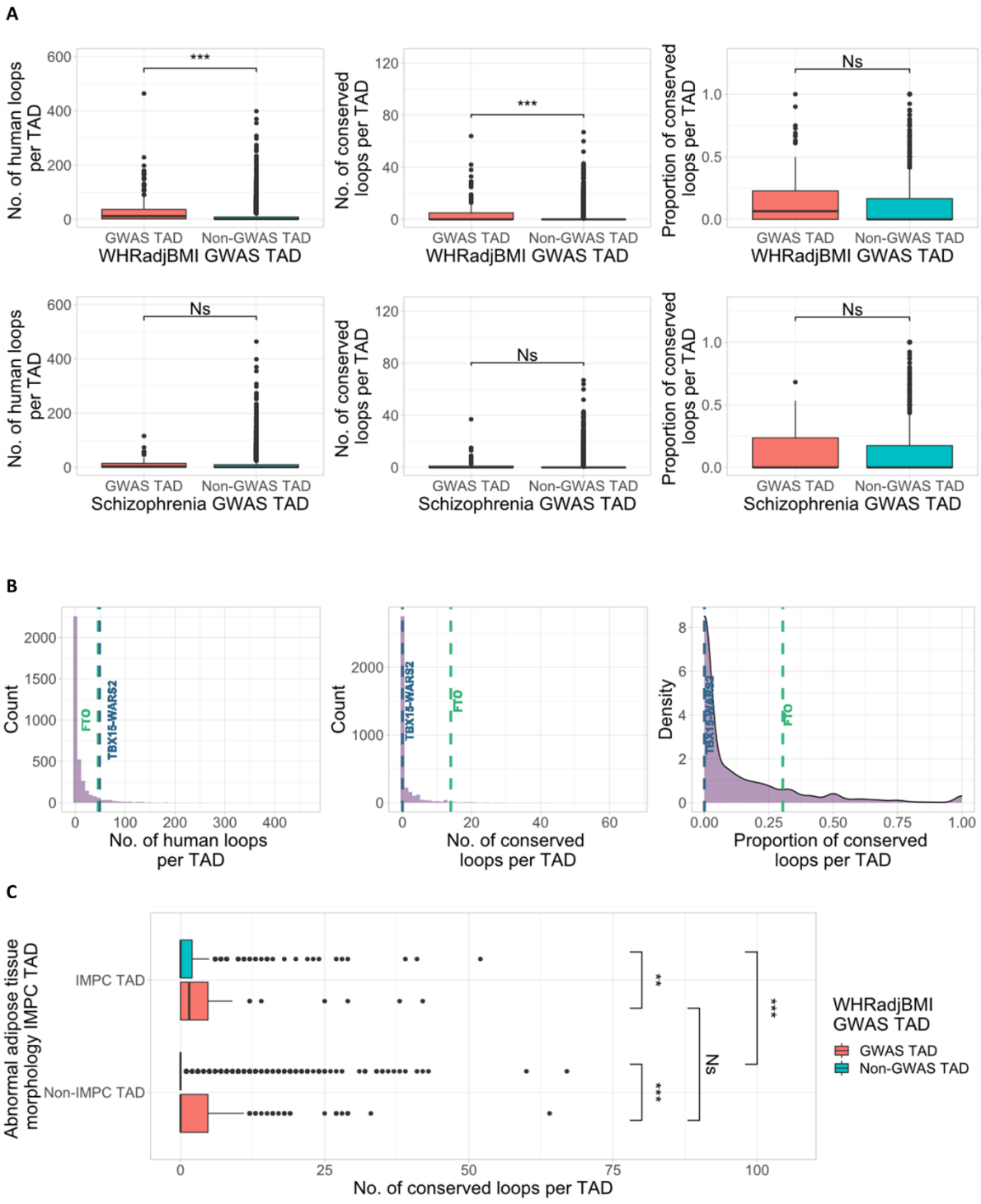


Figure 4.15: Number of human loops, number of human loops conserved in mouse, and the proportion of conserved loops per TAD. *A) GWAS TADs vs non GWAS TADs, GWAS TADs defined by overlap with WHRadjBMI associated SNPs (Pulit et al.) (top) or schizophrenia associated SNPs (Pardinas et al.) (bottom). Significance indicated by permutation testing, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, not significant= NS. FDR correction applied across six p-values. To calculate the proportion of conserved loops per TAD only TADs which contained >0 human loop*

*were considered. B) Distribution of each score, dotted lines indicate the positions of the TADs containing the TBX15-WARS2 locus or FTO. C) Number of conserved loops in IMPC and non-IMPC TADs (defined as TADs containing a human homolog of a mouse knockout gene associated with "Abnormal adipose tissue morphology" in IMPC), split into WHRadjBMI GWAS and non-GWAS TADs. Significance indicated by permutation testing, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, not significant= NS. FDR correction applied across four p-values.*

4.3.5.1. Prioritising fat distribution associated GWAS loci for modelling in mouse

In order to prioritise human WHRadjBMI associated loci for modelling in mouse, I ordered WHRadjBMI GWAS TADs by the number of conserved loops per TAD (Figure 4.14). As ~49% of WHRadjBMI GWAS TADs contained at least one conserved loop, I have shown the top 10 TADs as examples of loci with the highest looping conservation (Table 4.6). These TADs may represent the WHRadjBMI GWAS loci with the greatest conservation of adipogenic regulatory circuitry between human and mouse, and therefore may be most amenable to modelling using mouse. The GWAS TAD with the largest number of conserved loops (64 loops) contained the index SNP rs12138803 which is intronic to DNM3 and PIGC.

Three of the top ranked TADs also contained at least one gene which had a mouse homolog with an "abnormal adipose tissue morphology" IMPC knockdown phenotype. The genes involved included ZHX3, RCAN1, KCNE1, MLX, and ATP6V0A1, which all have a mouse homolog associated with "decreased total body fat amount"; and BECN1, and VPS25 which have mouse homologs associated with "increased total body fat amount". This provides an added layer of information useful in variant to function analysis, as it provides evidence that these genes could be the effector genes driving the GWAS association at these prioritised loci.

Table 4.6: Top ten WHRadjBMI GWAS TADs with highest number of conserved loops between human and mouse. *Genes with a mouse homolog with a significant "abnormal adipose tissue morphology" knockdown phenotype are shown in red. The mapped genes of the associated SNP are highlighted in yellow* (Pulit et al. 2019).

TAD coordinates	Genes	Total no. human loops	No. con loops	SNP
chr1: 171550000-172700000	SCARNA20, SNORD112, FASLG, DNM3, PRRC2C, RP4-612B18.1, MIR199A2, DNM3OS, RNU6-157P, C1orf105, RNU6-693P, RP4-612B18.3, PFN1P1, DNM3-IT1, RP4-560B9.5, METTL13, RP4-560B9.6, Y_RNA, PIGC, MIR214, RPL4P3, VAMP4, RP4-560B9.4, SUCO, MYOC	88	64	rs12138803
chr20: 39600000-40050000	RP1-1J6.2, RP3-511B24.5, EMILIN3, RN7SL615P, CHD6, RP3-511B24.6, RP4-620E11.5, RNU2-52P, RP3-450M14.1, LPIN3, PLCG1, TOP1, ZHX3, RPL23AP81	63	42	rs1997833
chr21: 34850000-36050000	CMP21-97G8.1, LINC00310, CLIC6, BTF3P6, AP000320.6, GART, AP000569.9, SNORA11, ATP5O, SMIM11, AP000320.7, RCAN1, AP000304.12, CRYZL1, AP000318.1, KCNE1, CMP21-97G8.2, ITSN1, TMEM50B, DNAJC28, SON, RPS5P3, DONSON, C21orf140, AP000322.53, MRPS6, AP000302.58, AP000304.1, KCNE2, RPS5P2, AP000318.2, RN7SL740P, LINC00649, SLC5A3	55	38	rs28451064
chr1: 169850000-170900000	LINC01142, HAUS4P1, MRPS10P1, RP1-79C4.1, AL356475.1, AL354732.1, RP1-79C4.4, PRRX1, SIGLEC30P, KIFAP3, RP11-297H3.3, MIR3119-1, SCYL3, RP11-545I10.2, METTL11B, RP11-332H17.1, ISCUP1, RN7SL269P, RP11-576I22.2, GORAB	198	33	rs10919388 rs10753805
chr17: 27050000-30650000	DPRXP4, SLC6A4, RNU6-1267P, SNORD63, RNU6-990P, RP11-20B24.6, AC010761.14, CRYBA1, SNORD4B, NF1, GIT1, RNU6-1034P, CTD-2349P21.10, AC116407.1, RP11-354P11.4, RP11-338L22.2, RP11-20B24.4, TWF1P1, AC010761.10, RP11-68I3.11, RNY4P13, RP11-805L22.3, ABHD15, KRT17P3, RP11-68I3.2, RP11-68I3.4, RP11-354P11.2, FAM222B, CORO6, RP11-296K13.4, RP11-848P1.2, CRLF3, TBC1D29, ADAP2, AC104984.4, RPL21P123, OMG, RPL23A, RNU6ATAC7P, RP11-354P11.8, SUZ12P, RNF135, RP11-640N20.5, SUZ12, RP11-218M11.1, UTP6, RP11-142O6.1, RP13-753N3.1, SNORA70, RP11-848P1.3, ARGFXP2, AC010761.9, CTD-2370N5.3, RP11-1148O4.2, MIR193A, AC006050.2, RNU6-1134P, RP11-640N20.9, CTD-2349P21.11, RP11-321A17.5, EVI2A, RN7SL316P, RP11-848P1.7, DHRS13, CTD-2349P21.5, RN7SL138P, MIR4733, RPL35AP35, CPD, LRRC37BP1, RP11-848P1.4, TRAF4, AC104996.1, RP11-20B24.5, RP11-271K11.1, TEFM, RHOT1, RP11-805L22.2, RP11-271K11.5, MIR423, TP53I13, CTD-2349P21.6, RP11-68I3.10, CTD-2095E4.1, AC024619.2, SH3GL1P1, MYO18A, RNU6-711P, AK4P1, RP11-805L22.1, TAOK1, LRRC37B, PHF12, AC090616.2, RPL31P58, snoU13, RNU6-920P, FLOT2, ALOX12P1, Y_RNA, RNU6-298P, RP11-640N20.1, AC005562.1, NSRP1, RNU4-34P, AC116407.2, RP11-1148O4.1, MIR365B, RP11-68I3.7, RP11-321A17.3, RP11-82O19.1, NEK8, CTC-542B22.2, MIR451B, RP11-354P11.3, AC003101.1, TLCD1, SH3GL1P2, CTD-2349P21.9, RN7SL45P, MIR3184, RNA5SP437, ANKRD13B, TIAF1, PIPOX, UBL5P2, RP1-41C23.1, RP11-68I3.5, RP11-20B24.7, RP11-321A17.4, MIR4523, ERAL1, ATAD5, EVI2B, RP11-218M11.3, RP11-82O19.2, RP11-640N20.8, BLMH, GOSR1, RAB11FIP4, RPL9P30, CTC-542B22.1, RP11-848P1.9, RP11-640N20.6, NUFIP2, RP11-20B24.3, CTD-2349P21.1, RP11-338L22.3, RP11-848P1.5, SEZ6, AC091177.1, TMIGD1, MIR4724, RHBDL3, SNORD42A, RP11-22N12.2, SSH2, AC010761.6, COPRS, RN7SL79P, RP11-218M11.6, MIR4725, RP11-443G13.2, EFCAB5, SMURF2P1	91	29	rs62070804
chr17: 39850000-41150000	RUNDC1, DHX58, TTC25, RN7SL399P, RNY4P2, CNTD1, LEPREL4, COASY, AOC2, RP11-358B23.1, PLEKHH3, RAMP2-AS1, NT5C3B, CNP, CTD-3193K9.1, KCNH4, PTGES3L-AARSD1, PSME3, RP11-156E6.1, RP11-376M2.2, CTD-2132N18.2, RP11-400F19.8, TUBG1, RNU6-287P, TUBG2, ZNF385C, CTD-3193K9.3, JUP, G6PC, MLX, HAP1, MIR548AT, ATP5LP7, CTD-2132N18.3, PTP4A2P1, HCRT, RNU7-97P, CNTNAP1, AARSD1, STAT3, LINC00671, NAGLU, BECN1, PSMC3IP, WNK4, HSPB9, KLHL11, AC107993.1, HSD17B1, CCR10, MIR5010, EZH1, RP11-194N12.2, RP11-400F19.6, FKBP10, FAM134C, GHDC, STAT5A, AC067852.1, RP11-400F19.18,	70	29	rs9892728

	RP11-358B23.5, DNAJC7, ACLY, ATP6V0A1, RNA5SP442, AC003104.1, STAT5B, CTD-2132N18.4, PTRF, AOC3, RAB5C, NKIRAS2, SNORA40, HMGN2P15, AOC4P, KLHL10, HMGB3P27, COA3, CTD-3193K9.4, PTGES3L, GAST, HSD17B1P1, VPS25, RN7SL871P, KAT2A, RAMP2			
chr7: 25150000-27200000	AC004520.1, AC004079.1, AC004540.4, TSEN15P3, HOXA1, AC004129.9, U3, RP1-170O19.23, CBX3, KIAA0087, HNRNPA2B1, CTD-2227E11.1, C7orf71, AC004129.7, AC004540.5, SNX10, AC004947.2, NPVF, HOXA5, HOXA3, RP1-170O19.21, HOXA-AS3, CTA-242H14.1, AC091705.2, AC010677.5, MIR148A, NFE2L3, HMGB3P20, TPM3P4, HOTAIRM1, CYCS, AC003090.1, NHP2P2, C7orf31, HOXA6, RPL7AP38, RP5-978I12.1, HOXA-AS2, HOXA2, AC091705.1, CTB-119C2.1, AC003985.1, SKAP2, RP1-170O19.22, HOXA7, HOXA4	61	29	rs1534696
chr7: 27200000-28150000	RN7SL365P, RP1-170O19.20, AC073150.6, JAZF1, RP1-170O19.17, HOXA10, TAX1BP1, HOXA11-AS, HOXA9, HOXA11, EVX1, HIBADH, EIF4HP1, AC004009.3, MIR196B, RPL35P4, HOXA10-AS, PSMC1P2, EVX1-AS, HOXA13, HOTTIP, AC004549.6, RP1-170O19.14, HNRNPA1P73	46	28	rs2067087 rs17437657
chr18: 2550000-3700000	RP11-715F3.2, RP11-838N2.8, RN7SL39P, SNORA70, RP11-710M11.1, RP11-737O24.4, RP11-737O24.1, RPL31P59, RP11-838N2.4, MYOM1, RP13-270P17.1, RNU7-25P, METTL4, KATNBL1P3, RP11-737O24.3, LPIN2, CBX3P2, snoU13, RP11-715F3.1, RP13-270P17.2, RP11-703M24.5, RP11-838N2.3, RP13-270P17.3, SMCHD1, NDC80, RPL21P127, MYL12B, RP11-737O24.5, IGLJCOR18, Y_RNA, snoU109, RNU6-340P, RP11-874J12.4, RP11-737O24.2, MYL12A, RP11-838N2.5, EMILIN2, TGIF1, DLGAP1-AS1, DLGAP1-AS2	131	27	rs695115
chr2: 216200000-216800000	AC012668.2, AC012668.3, AC012462.3, FN1, AC093850.2, AC012668.1, AC093850.1, AC122136.1, AC012462.1, LINC00607, AC012462.2, AC122136.2, ATIC	464	27	rs1250259

4.4. Discussion

GWAS has proven to be a powerful tool in linking genetic variants to complex disease. However, due to the fact that the majority of associated variants lie within the non-coding genome and the inheritance pattern of multiple variants can be highly correlated (due to LD structure), functional translation has proven challenging (Cano-Gamez and Trynka 2020). The TBX15-WARS2 locus contains multiple, sex-specific GWAS associations with fat distribution (Pulit et al. 2019; Shungin et al. 2015; Heid et al. 2010); an attempt at fine mapping in this locus narrowed the association to a 1.6 kb credible set region (C. T. Liu et al. 2014). This region was deleted by the Cox group to create a mouse model of this association; however, no molecular or clear organismal phenotype was detected (manuscript in preparation). In order to investigate the regulatory architecture of this locus, including its conservation between mouse and human, I investigated and overlaid the chromatin architecture of the locus (at multiple levels of the chromatin hierarchy) with GWAS-associated variants.

The majority of the index SNPs and proxy SNPs (SNPs in high LD with an index SNP) in this locus are associated with WHRadjBMI rather than BMI or WHR. This suggests that this locus is associated with fat distribution/central adiposity rather than overall adiposity/obesity. At the level of compartment and TAD chromatin organisation, there is reasonable conservation between human and mouse. The credible set region (along with many of the index and proxy SNPs), overlies an A compartment annotation in the human MSC dataset, and in the mouse adipogenesis dataset from Siersbæk et al. (but not from He et al.), suggesting that this region falls within active/euchromatin during adipogenesis in both mouse and human. I further assessed the chromatin state of the A compartment regions in human using chromHMM in adipose, MSC, chondrocytes, osteoblast/bone and brain and found that it contained

annotations for “weak transcription”, “enhancer” and “flanking active TSS” in all cell types except brain. This suggests that this region may have regulatory activity in cell types related to MSC differentiation but not in brain, further supporting that this locus is associated with fat distribution rather than overall adiposity and suggesting that the tissue of action is likely to be related to adipose rather than brain.

Regulatory elements are thought to impact their target genes by making direct physical contact with them (Mora et al. 2015; Panigrahi and O’Malley 2021; Matharu and Ahituv 2015; Bonev and Cavalli 2016; Oudelaar and Higgs 2020). I therefore analysed regulatory looping interactions between SNPs and gene TSSs in multiple human adipose/adipogenesis datasets. I reasoned that regulatory interactions which are important in adipose/adipogenesis would be detected in multiple datasets and therefore this information could be used to prioritise candidate effector SNP-gene interactions. I further reasoned that the presence or absence of equivalent loops to those detected in human during mouse adipogenesis would be a powerful indicator of the degree of conservation of the regulatory circuitry in this locus between mouse and human.

SNPs were initially overlapped with published loops from studies of SGBS (Sobreira et al. 2021), HWA (Pan et al. 2018), human MSC differentiating towards adipocytes (Madsen et al. 2020) and mouse pre-adipocytes differentiating towards adipocytes (Siersbæk et al. 2017). However, it was noted that whilst most of these studies used the same software to detect loops (CHiCAGO), there was variation in the settings used and whether they were reported. In human MSC differentiating to towards adipocytes, loops were detected with custom CHiCAGO settings and pooling methods (where the specific settings were undeclared and the pooled fragment file unpublished). In mouse pre-adipocytes differentiating towards

adipocytes, no pooling was employed but the CHiCAGO settings were undeclared and in SGBS a different software was used (HOMER). A recent protocol paper by the CHiCAGO developers highlighted the importance of using settings appropriate for the fragment size (usually determined by the use of 4-cutter or 6-cutter restriction enzymes) in order to avoid the detection of spurious low count interactions as significant (Freire-Pritchett et al. 2021). It is difficult to assess the impact of this on published results where settings are undeclared. For human MSC differentiating to towards adipocytes, Madsen et al. , employed a custom binning method which created fragments with a median size of 1,963 bp. This is smaller than fragments produced by 6-cutter enzymes and larger than fragments produced by 4-cutter enzymes (e.g. HindIII produces ~3,700 bp and MboI produced ~430 bp fragments in the human genome) (Freire-Pritchett et al. 2021). It was not clear if this study optimised the CHiCAGO settings to account for this or if this would be necessary.

Of further concern is the use of the tool HOMER (findHiCInteractionsByChr.pl) to detect significant loop interactions in SGBS (Sobreira et al. 2021). HOMER is a suite of analysis tools including programs designed specifically to analyse Hi-C data (Heinz et al. 2010). Hi-C data captures interactions genome-wide which are then sequenced in an unbiased fashion (Lajoie, Dekker, and Kaplan 2015). PCHi-C/ECHi-C by design involves an enrichment step biasing the captured interactions towards regions of interest. The HOMER documentation states that "while HOMER can be jury-rigged to glean information from other 3C-sequencing based methods, it has been specifically tailored for Hi-C analysis". It is not clear from Sobreria et al. whether any modification to HOMER was made to account for this.

In order to overcome the inconsistencies of the different tools/parameters used in the published looping datasets, I recalled loops using a standardised loop calling pipeline and

compared the resulting loops to those that were originally published for each dataset. I found high reproducibility in published looping datasets produced with HindIII (HWA and mouse pre-adipocytes differentiating towards adipocytes (Pan et al. 2018; Siersbæk et al. 2017)). In Human MSC differentiating towards adipocytes, which was produced using MboI and published loops detected using custom binning, I found lower reproducibility (Madsen et al. 2020). However, reproducibility was increased by binning fragments before loop calling. The lowest reproducibility was found for the SGBS loops published in Sobreria et al., where it was not clear if the software used was appropriate for the datatype. This result highlights the critical importance of properly benchmarking loop calling software to ensure that correct software/parameters for the input data are chosen, as well as the need to explicitly declare the settings used in publications.

I reasoned that important looping interactions are likely to be detected across multiple datasets. However, I found no SNP-gene interactions in the TBX15-WARS2 locus which were detected in multiple datasets, with the exception of different differentiation time points from the same study. In addition, the majority of interactions detected in multiple differentiation time points were not robustly detected between published and recalled loops. This made it very difficult to prioritise the candidate effector interactions which may be driving the GWAS association in this locus. Furthermore, I detected no common loops between human and mouse adipogenesis at this locus, indicating that the regulatory circuitry may be poorly conserved. This could explain the lack of clear phenotype in the mouse model observed by the Cox group.

In order to assess the association between looping during adipogenesis and GWAS-associated loci genome-wide, I quantified the number of loops per TAD, as well as the

number/proportion of conserved loops between human and mouse adipogenesis. I found that TADs which contained a WHRadjBMI associated GWAS variant contained significantly more loops during human adipogenesis, as well as more conserved loops between human and mouse adipogenesis compared to TADs which did not contain a WHRadjBMI GWAS variant. This suggests that loci involved in fat distribution contain more extensive regulatory circuitry over human adipogenesis, and this regulatory circuitry is enriched for conservation in mouse. I next sought to assess whether conservation of regulatory circuitry between human and mouse is likely to correspond to phenotypic conservation. To do this, I further classified TADs with or without a WHRadjBMI GWAS variant, based on whether they also contained a gene with a mouse homolog significant for "abnormal adipose tissue morphology" in IMPC. This showed that TADs which contained at least one human gene with a mouse homolog significant for "abnormal adipose tissue morphology" and a GWAS variant associated with WHRadjBMI had the highest number of conserved loops between human and mouse. This could suggest that loci with higher conservation of regulatory circuitry are also more likely to show phenotypic conservation. Supporting this, I also showed that whilst the TBX15-WARS2 locus contained many loops during human adipogenesis, none of these loops were conserved during mouse adipogenesis. Compared to the TBX15-WARS2 locus, the FTO locus showed a similar number of loops during human adipogenesis; however, 30% of the loops were conserved during mouse adipogenesis. Unlike the TBX15-WARS2 locus, recent studies have successfully modelled the FTO locus in mouse, detecting molecular, cellular and organismal phenotypes relevant for obesity and adipogenesis (Laber et al. 2021; Sobreira et al. 2021). Together, these results suggest a novel use of looping conservation in the prioritisation of loci for modelling in mouse. The use of looping conservation to pre-select loci with high regulatory conservation has the potential to overcome issues caused by low inter-

species conservation in the non-coding genome when modelling human GWAS variants in mouse (Powell et al. 2022).

It is important to note that the work presented here has several limitations. While we did not find a consensus of interactions between the looping datasets investigated in the TBX15-WARS2 locus, and data from the mouse model indicated that there was no molecular or organismal phenotype, this does not exclude the possibility that we were studying the wrong sex, cell types, developmental stage, environmental cues or variants. If the causal interactions are only present at a specific developmental stage or cell type, our approach involving looking for consensus loops between datasets would not be sufficient. This is because in most cases, multiple looping datasets involving precisely the same developmental stage and cell type were not available. Similarly, given the recently detected sex-specific GWAS associations with fat distribution in this locus, it is possible that the causal interactions are sexually dimorphic (Shungin et al. 2015). However, sex-specific looping datasets are not available. Furthermore, despite attempts to make the datasets used as comparable as possible, the impact of technical differences may remain, such as different capture libraries and different restriction enzymes. The differences in capture library may have particularly affected the global human vs mouse looping comparison (the human dataset mainly captured enhancers, whereas the mouse dataset captured promoters). Since the completion of this work, human SGBS adipogenesis PCHi-C data has been published which has the potential to improve this issue (Joslin et al. 2021). Finally, using the degree of human vs mouse looping conservation during adipogenesis, I have prioritised the top ten candidate WHRadjBMI associated loci for modelling in mouse, three of which contain genes which have homologs with “abnormal adipose tissue morphology” mouse knockout phenotypes. While it is noted that mouse models of the TBX15-WARS and FTO loci support the assertion that loci with higher

conservation may be more amenable to modelling in mouse, more attempts at modelling loci in mouse would be required to prove this.

It is also possible that strategy for defining "potentially regulatory SNP loops" could have missed potentially relevant regulatory loops by excluding loops which had a TSS at both ends. This is something could be addressed in future work and is particularly interesting due to the functional and structural overlap between promoters/TSSs (reviewed in: (Andersson and Sandelin 2019)).

In conclusion, the lack of consensus between looping datasets at the TBX15-WARS2 locus makes the prioritisation of effector interactions at this locus difficult. There was no detected looping conservation between human and mouse adipogenesis, suggesting the regulatory circuitry of this locus may be poorly conserved between human and mouse. I assessed the degree of conservation of regulatory circuitry between human and mouse adipogenesis genome-wide and found that TADs containing WHRadjBMI GWAS variants are enriched for a larger number of conserved loops. This suggests that conservation of regulatory looping could be useful to prioritise GWAS-associated variants for modelling in mouse.

Chapter 5: Discussion

Inside each mammalian nucleus, DNA is compacted into a complex hierarchy of chromatin structures commonly described as chromosome territories, compartments, TADs, loops, and nucleosomes. These chromatin structures create non-random contacts linked to genome function. In this thesis, I have primarily focused on compartments, TADs and loops. Chromatin compartments are highly correlated with euchromatin and heterochromatin (Lieberman-Aiden et al. 2009). TADs have been proposed to constitute "gene regulatory domains", promoting co-regulation of the genes within them and co-localising regulatory elements and their target genes (Dixon, Gorkin, and Ren 2016). Furthermore, looping is thought to be a primary mechanism of enhancer-promoter communication (reviewed in (Popay and Dixon 2022)). Continued research into the understanding of mammalian chromatin architecture is imperative in unravelling the complexity of gene regulatory programs, which gives rise to the known diversity of mammalian cell types. Understanding the level of cell type and species specificity of TADs, as well as the functional relationship between genes that share TADs, is essential to this. The study of chromatin architecture also has great potential in understanding the mechanistic relationship between non-coding GWAS variants and human complex disease. In this thesis, I aimed to investigate the above areas in order to better understand mammalian chromatin architecture and its impact on gene regulation.

5.1. TADs as gene regulatory domains?

TADs are thought to act as regulatory domains, important for correct gene regulation. It is thought that they co-localise genes and their regulatory elements, insulate genes from aberrant regulation, block the spread of activation and transcription in the genome, and are largely tissue and species invariant (Dixon, Gorkin, and Ren 2016). However, despite these strong claims, it is often unclear how widely they are supported by the data.

5.1.1. Tissue and species specificity

It is commonly claimed in the literature that TADs show a large degree of both tissue and species invariance (Dixon et al. 2012; Nora et al. 2012; Dixon et al. 2015; Rao et al. 2014; Schmitt et al. 2016; Vietri Rudan et al. 2015). However, as pointed out in this thesis and by others (Eres and Gilad 2021), these claims are often based upon studies which found that a large number of TADs do in fact vary between both tissues and species. This suggests that there is a possible contradiction in the data and the inferred conclusion. Determining the degree of tissue and species specificity in TADs is crucial to understanding the contribution of TADs to cell type-specific gene expression changes during differentiation and the gene regulatory divergence between species (Eres and Gilad 2021).

To assess tissue or species conservation of TADs, either whole TADs or TAD boundaries can be compared. These strategies offer different but complementary assessments of TAD conservation as it is possible for TAD boundaries to be highly conserved but pair differently to form tissue or species-specific TADs. By comparing TADs and TAD boundaries between tissues in 13 human Hi-C maps and 20 mouse Hi-C maps in Chapter 2, I have demonstrated

that within a species, there is a significantly increased TAD similarity between both TADs and TAD boundaries from the same tissue group compared to TADs and TAD boundaries from different tissue groups respectively. This suggests that both TADs and TAD boundaries show a significant degree of tissue specificity. This is supported by the fact that across all human or mouse datasets, most TADs were only detected in a single tissue group, suggesting that most TADs are tissue-specific. Interestingly, there was also a trend for the increased similarity between TADs and TAD boundaries from the same tissue group but different species for both TAD callers. This could suggest that the tissue specificity of TADs shows a degree of conservation between mouse and human. When comparing between species, TADs and TAD boundaries were significantly more similar within species than between species. Moreover, the effect sizes of these differences was large, indicating a large degree of species specificity in both TADs and TAD boundaries. To further assess the relationship between tissue specificity and species conservation, for every unique TAD detected in any dataset within each species, I quantified the number of tissue groups in which it was present, along with whether it was conserved between species irrespective of the tissue group. This revealed that conserved TADs between species were detected in significantly more tissue groups than TADs which were not conserved between species. This indicates that tissue invariant TADs are more likely to be conserved between species.

5.1.2. Promoting gene co-regulation

In a seminal study by Symmons et al., regulatory sensors were inserted along the mouse genome; the sensors consisted of a lacZ gene and weak promoter that could be regulated by local enhancers. The observed expression patterns of the regulatory sensors revealed that

the genome is split into regulatory domains which closely correlate with TADs (Symmons et al. 2014). Further to this, several studies reported increased expression correlation between genes which share a TAD (Nora et al. 2012; Flavahan et al. 2016; Tarbier et al. 2020), and it was noted that TADs tended to switch between the active and inactive compartments as whole units (Dixon et al. 2015). This led to speculation that TADs promote gene co-regulation and functionally similar genes may fall within the same TAD to facilitate their shared regulatory requirements (suggested in the review (Ruiz-Velasco and Zaugg 2017)).

To assess the functional relationship between genes sharing a TAD, in Chapter 3, I developed a TAD randomisation method which accounted for the linear order of the genome and general features of TADs. In order to focus on algorithmically defined TADs that are likely to have been formed by loop extrusion, I assessed TADs with CTCF binding sites around both boundaries. In these CTCF TADs, I found little evidence for increased similarity in co-expression, shared pathways, PPIs, or molecular function GO terms between pairs of genes sharing a TAD other than that explained by their proximity in the linear order of the genome. When assessing the relationship between paralogous genes and TADs, similarly to a previous study, when all TADs were considered I found that pairs of paralogs were enriched within TADs (Ibn-Salem, Muro, and Andrade-Navarro 2017). However, when split into CTCF and non-CTCF TADs, I discovered that whilst pairs of paralogs were enriched in non-CTCF TADs, they were commonly depleted in CTCF TADs. This may indicate that pairs of paralogs are depleted in TADs formed by loop extrusion and enriched in TADs formed by other mechanisms, e.g. compartment domains. While it is highly likely that there are numerous examples of individual TADs containing co-regulated and functionally similar genes, these results call into question the role of TADs formed by loop extrusion in promoting co-regulation of the genes genome-wide.

5.1.3. Regulatory insulation

TADs are also thought to play a crucial role in insulating genes from aberrant regulatory interactions with enhancers beyond the boundaries of the TAD. In support of this, there are several examples of gene misregulation or disease caused by the loss of a TAD boundary (Lupiáñez et al. 2015; Giorgio et al. 2015; Hanssen et al. 2017). In Chapter 3, I found that genes which singly occupy a TAD have higher evolutionary constraint than genes that co-occupy a TAD and that these genes are enriched for developmental functions. This indicates that they may have high functional importance and privately occupy TADs in order to insulate them from aberrant regulation. Therefore, my results suggest that the insulatory role of TADs in gene regulation is important for at least a subset of genes. I speculate that my results fit with a model in which aberrant regulatory interactions spanning a TAD boundary are prevented by the boundary and aberrant regulatory interactions within TADs are prevented by evolved enhancer-promoter specificity.

5.2. TADs and loops in variant to function analysis

In recent years, there has been increasing interest in the utility of chromatin datasets in mechanistically dissecting the ever-increasing number of published GWAS-associated variants. Due to the insulatory nature of TAD boundaries, TADs are likely to be useful in narrowing down the candidate target genes of GWAS-associated variants. In addition, given that looping is thought to constitute the primary mechanism of enhancer-promoter communication, looping data is also likely to be useful in linking GWAS variants to their target genes.

The Cox group at MRC Harwell are interested in carrying out variant to function analysis of GWAS variants using mouse models. They have generated a mouse model to investigate the human GWAS fat distribution and obesity association in the TBX15-WARS2 locus. By using publicly available Hi-C, PCHi-C and ECHi-C data from disease-relevant tissues and cell types in human and mouse, in Chapter 4 I investigated the mechanistic link between variants in the TBX15-WARS2 locus and fat distribution. I reasoned the regulatory looping interaction underlying this association would be detected in multiple adipogenic datasets. However, no SNP-gene interactions were detected in multiple datasets (except differentiation time points from the same study). This means that it is difficult to prioritise an effector SNP-gene interaction in this locus. It is possible that an effector interaction was not detected because I was looking at the wrong tissues/cell types, developmental stage, sex, environmental stimulus, or variants. Alternatively, it is possible that the mechanism of action of these variants does not involve looping or is a gain of function regulatory interaction which is not present in normal cells. I also found no evidence of looping conservation between human and mouse adipogenesis in the TBX15-WARS2 locus. Furthermore, work by the Cox group to model the GWAS associations in this locus in mouse generated no clear phenotype. Together, these results suggest that the regulatory circuitry of this locus may not be conserved between species.

To prioritise loci which do have a high conservation of regulatory circuitry between human and mouse for modelling in mouse, I compared looping interactions genome-wide between human and mouse adipogenesis. Using this method, I have shown that loci with high looping conservation are also enriched for genes with an adipose-related knockout phenotype in IMPC, suggesting that the loci may also be phenotypically conserved between mouse and human. In addition, I have shown that the FTO locus, which contains an obesity-associated

variant successfully modelled in mouse (Laber et al. 2021; Sobreira et al. 2021), contained greater looping conservation between mouse and human than the TBX15-WARS2 locus. This further supports my hypothesis that conservation in regulatory circuitry will translate into phenotypic conservation. I have suggested the top 10 fat distribution associated loci as prioritised by the loop conservation score for modelling in mouse. The use of looping conservation to assess the conservation of regulatory circuitry between species has the potential to overcome issues caused by low interspecies conservation in modelling non-coding variants in mouse (Powell et al. 2022), and to my knowledge has not been proposed before.

5.3. Future directions

One significant limitation of the work presented in this thesis is the reliance on publicly available data. Published Hi-C, PCHi-C and ECHi-C data is currently limited in terms of the diversity of cell types, tissues and species available, as well as technical differences in data generation. One of the most significant differences between publicly available datasets is the number of contacts captured in the Hi-C map caused by differences in sequencing depth. This is a common difference even between datasets published in the same study and can lead to differences in power to detect TADs (Eres and Gilad 2021). As I have demonstrated in Chapter 2, TAD callers vary in their sensitivity to this, with Arrowhead more sensitive than TopDom. Further to this, as I have previously stated, Hi-C data captures interactions on average over all cell types in the sample. Therefore, datasets generated from tissues that are likely to be made up of a variety of cell types may be less valuable than those generated on a single cell type. To further investigate the relationship between chromatin architecture and gene regulation,

it would be useful to move towards the generation of Hi-C datasets on a variety of cell types in a standardised pipeline. Meanwhile, there are several resources which serve as repositories of 3D chromatin architecture data and annotations across various cell types, e.g. the 3D Genome Browser (Wang et al. 2018) and 3D-Genome Interaction Viewer & database (Yang et al. 2018). These resources are likely to be very useful in locating relevant data as well as providing standardisation of data analysis between datasets.

In part due to the lack of diversity in available data, it is not uncommon in the literature to see analyses which integrate TADs from seemingly irrelevant tissues, e.g. lymphoblastoid TADs to investigate a cardiac enhancer cluster (Man et al. 2019), ESC TADs to investigate bone mineral density associated GWAS variants (Way et al. 2017), and oncogenic cell line TADs to investigate paediatric musculoskeletal GWAS-associated variants (Medina-Gomez et al. 2017). This is often justified by the reported tissue invariance of TADs. However, results outlined in this thesis and in complementary analyses (Sauerwald, Singhal, and Kingsford 2020; McArthur and Capra 2021) suggest that TADs actually display a significant degree of tissue specificity. Therefore, extreme caution should be used when integrating TADs from non-relevant tissues into an analysis and, wherever possible, TAD annotations from appropriate tissues should be selected.

A major criticism of all research on TADs to date relates to the poor definition of TADs themselves (reviewed in (de Wit 2020)). Algorithmically defined TADs, such as those used in the literature and in this thesis, are likely to represent domains formed by a variety of biological mechanisms, e.g. loop extrusion and compartmentalisation (Beagan and Phillips-Cremins 2020). Recent reviews have argued that in future studies of TADs, TADs should be separated based on their underlying mechanism of formation (Beagan and Phillips-Cremins

2020; de Wit 2020). In Chapter 3 of this thesis, I have separated TADs which are most likely to have been formed by loop extrusion involving CTCF (CTCF TADs) from other TADs which may have been formed by other mechanisms (non-CTCF TADs). This demonstrated that paralogs are generally depleted in CTCF TADs, despite their enrichment in non-CTCF TADs, as well as TADs overall. This result illustrates the potential utility of studying the features of TADs in the context of the mechanisms by which they were formed (Beagan and Phillips-Cremins 2020; de Wit 2020). At the very least, it will be essential for future research that the field converges on a definition of TADs as well as a gold standard method for their detection.

In the absence of an accepted definition of TADs and a gold standard detection method, it is hugely important to make sure that conclusions drawn in the study of TADs are robust to the choice of TAD caller. There are several recently published TAD benchmarking papers that are very useful in selecting TAD callers (Zufferey et al. 2018; Dali and Blanchette 2017; Forcato et al. 2017). In addition, as I have demonstrated in Chapter 4, careful benchmarking of loop calling software is also critical. Efforts should be made to ensure that software and settings are appropriate to account for technical variation in datasets. It seems that the interplay between fragment length and software parameters in the popular loop calling software CHiCAGO is a key example of this (Freire-Pritchett et al. 2021; Cairns et al. 2016). To ensure future reproducibility, it is also crucial that the field moves towards publishing all utilised parameters in publications.

As more Hi-C data becomes available and greater standardisation is utilised in its generation, it will be interesting to assess the specific features of the TADs that are tissue or species-specific or variant. In a complementary study, McArthur et al. found that TAD boundaries which are tissue invariant are enriched for complex trait heritability and evolutionary

constraint (McArthur and Capra 2021). As explained above, it would be of particular interest to assess tissue and species invariant or variant TADs in the context of the mechanisms by which they were formed. One hypothesis would be that compartment domains (which are highly correlated with the epigenetic state of the genome) account for the observed tissue specificity of TADs, whereas domains formed by loop extrusion may be more likely to be tissue invariant. In order to assess TAD and TAD boundary variance across cell types and tissues in this thesis, I have used a strict method in which TAD boundaries must occupy the same genomic bin to be classified as invariant between two datasets. In order to extend this work, it would be interesting to separately assess the features of TADs and TAD boundaries which have only shifted slightly compared with those which are entirely tissue-specific. This could identify genes or regulatory elements which change from one TAD to another between cell types and would aid in the investigation of the interplay between TADs and cell type-specific regulatory programs.

I found that the evolutionary constraint of protein coding genes which singly occupy a TAD is significantly higher than the average constraint of genes which share a TAD, suggesting that they may have high functional importance. Furthermore, these genes were enriched for developmental functions. I hypothesise that these genes are developmentally important and may privately occupy a TAD to insulate them from aberrant gene regulation. In order to test this hypothesis, it would be interesting to assess the impact of deleting the boundaries of these TADs, as well as assessing whether there are any examples of described human diseases caused by TAD boundary disruption affecting genes from these TADs (examples in (Lupiáñez, Spielmann, and Mundlos 2016)). It would also be interesting to extend this analysis to also include non-coding genes, as in the current analysis it is possible that TADs which privately contain protein coding genes could also contain non-protein coding genes.

Furthermore the analysis of the functions of genes sharing a TAD could be repeated on non-protein coding genes as well as other non-coding features sharing the same TAD e.g. features/numbers of enhancers and repetitive elements.

By comparing chromatin looping between human and mouse adipogenesis, I have developed a score which I hypothesise can be used to select loci containing GWAS-associated variants with conserved regulatory circuitry, allowing them to be successfully modelled in mouse. Preliminary phenotype data from attempts at modelling GWAS-associated variants at two loci and IMPC knockout phenotype data support my hypothesis that conservation of regulatory circuitry will translate into phenotypic conservation. Further attempts at modelling GWAS-associated loci in mouse are required in order to properly test this. However, this method may help to overcome the inherent difficulties of using mouse models in GWAS variant to function analysis caused by the lower conservation of the non-coding genome (compared to the coding genome) between species. This method could be further extended to include comparisons of looping data across multiple cell types, which would have the potential to aid variant to function analysis across many complex diseases.

5.4. Conclusions

Overall, my thesis has tried but struggled to confirm some of the popular claims surrounding the relationship between TADs and function. In particular, contrary to some beliefs, I found that algorithmically defined TADs show a significant degree of tissue and species specificity. Whether this tissue specificity is driven by domains formed by loop extrusion or compartment domains remains to be established. Similarly, by taking into account the effect of the linear genome, contrary to common assumptions, I show that the role of CTCF TADs in promoting gene expression is limited, but that TADs may be important in the protection of developmentally important genes from aberrant regulation. In order to further investigate the relationship between TADs and function, it will be essential to address the poor definition of TADs and the issue of their detection, which I have discussed throughout this thesis. Finally, my pilot study, comparing loop conservation within TADs between human and mouse, provides a potential strategy to prioritise fat distribution associated GWAS loci for modelling using mouse models. Finding methods to integrate fast emerging but sometimes noisy omics data of different modalities will be imperative for the future understanding of regulatory genomics. A better understanding of gene regulatory mechanisms will facilitate the utility of personalised medicine which will one day become commonplace.

Chapter 6: References

Agnew, Thomas, Michelle Goldsworthy, Carlos Aguilar, Anna Morgan, Michelle Simon, Helen Hilton, Chris Esapa, et al. 2018. "A Wars2 Mutant Mouse Model Displays OXPHOS Deficiencies and Activation of Tissue-Specific Stress Response Pathways." *Cell Reports* 25 (12): 3315-3328.e6. https://doi.org/10.1016/J.CELREP.2018.11.080.

Akdemir, Kadir C., Victoria T. Le, Sahaana Chandran, Yilong Li, Roel G. Verhaak, Rameen Beroukhim, Peter J. Campbell, et al. 2020. "Disruption of Chromatin Folding Domains by Somatic Genomic Rearrangements in Human Cancer." *Nature Genetics 2020 52:3* 52 (3): 294–305. https://doi.org/10.1038/s41588-019-0564-y.

Akiyama, Masato, Yukinori Okada, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, et al. 2017. "Genome-Wide Association Study Identifies 112 New Loci for Body Mass Index in the Japanese Population." *Nature Genetics 2017 49:10* 49 (10): 1458–67. https://doi.org/10.1038/ng.3951.

Andersson, Robin, and Albin Sandelin. 2019. "Determinants of Enhancer and Promoter Activities of Regulatory Elements." *Nature Reviews Genetics 2019 21:2* 21 (2): 71–87. https://doi.org/10.1038/s41576-019-0173-8.

Ardlie, Kristin G., David S. DeLuca, Ayellet V. Segrè, Timothy J. Sullivan, Taylor R. Young, Ellen T. Gelfand, Casandra A. Trowbridge, et al. 2015. "Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." *Science (New York, N.Y.)* 348 (6235): 648–60. https://doi.org/10.1126/SCIENCE.1262110.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics*. NIH Public Access. https://doi.org/10.1038/75556.

Athar, Awais, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, et al. 2019. "ArrayExpress Update - from Bulk to Single-Cell Expression Data." *Nucleic Acids Research* 47 (D1): D711–15. https://doi.org/10.1093/NAR/GKY964.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2012. "NCBI GEO: Archive for Functional Genomics Data Sets—Update." *Nucleic Acids Research* 41 (D1): D991–95. https://doi.org/10.1093/nar/gks1193.

Beagan, Jonathan A., and Jennifer E. Phillips-Cremins. 2020. "On the Existence and Functionality of Topologically Associating Domains." *Nature Genetics*, January. https://doi.org/10.1038/s41588-019-0561-1.

Beagrie, Robert A., Antonio Scialdone, Markus Schueler, Dorothee C.A. Kraemer, Mita Chotalia, Sheila Q. Xie, Mariano Barbieri, et al. 2017. "Complex Multi-Enhancer Contacts Captured by Genome Architecture Mapping (GAM)." *Nature* 543 (7646): 519. https://doi.org/10.1038/NATURE21411.

Belmont, Andrew S. 2002. "Mitotic Chromosome Scaffold Structure: New Approaches to an Old Controversy." *Proceedings of the National Academy of Sciences of the United States*

of America 99 (25): 15855. https://doi.org/10.1073/PNAS.262672799.

Benabdallah, Nezha S., Iain Williamson, Robert S. Illingworth, Lauren Kane, Shelagh Boyle, Dipta Sengupta, Graeme R. Grimes, Pierre Therizols, and Wendy A. Bickmore. 2019. "Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation." *Molecular Cell* 76 (3): 473-484.e7. https://doi.org/10.1016/j.molcel.2019.07.038.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X.

Bertero, Alessandro, Paul A. Fields, Vijay Ramani, Giancarlo Bonora, Galip G. Yardimci, Hans Reinecke, Lil Pabon, William S. Noble, Jay Shendure, and Charles E. Murry. 2019. "Dynamics of Genome Reorganization during Human Cardiogenesis Reveal an RBM20-Dependent Splicing Factory." *Nature Communications* 10 (1): 1538. https://doi.org/10.1038/s41467-019-09483-5.

Berthelot, Camille, Matthieu Muffato, Judith Abecassis, and Hugues Roest Crollius. 2015. "The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions." *Cell Reports* 10 (11): 1913–24. https://doi.org/10.1016/J.CELREP.2015.02.046.

Bianconi, Eva, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, et al. 2013. "An Estimation of the Number of Cells in the Human Body." *Https://Doi.Org/10.3109/03014460.2013.807878* 40 (6): 463–71. https://doi.org/10.3109/03014460.2013.807878.

Bonev, Boyan, and Giacomo Cavalli. 2016. "Organization and Function of the 3D Genome." *Nature Reviews Genetics* 17 (11): 661–78. https://doi.org/10.1038/nrg.2016.112.

Bonev, Boyan, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L Papadopoulos, Yaniv Lubling, Xiaole Xu, et al. 2017. "Multiscale 3D Genome Rewiring during Mouse Neural Development." *Cell* 171 (3): 557-572.e24. https://doi.org/10.1016/j.cell.2017.09.043.

Boyle, Alan P., Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, et al. 2012. "Annotation of Functional Variation in Personal Genomes Using RegulomeDB." *Genome Research* 22 (9): 1790. https://doi.org/10.1101/GR.137323.112.

Boyle, Shelagh, Susan Gilchrist, Joanna M. Bridger, Nicola L. Mahy, Juliet A. Ellis, and Wendy A. Bickmore. 2001. "The Spatial Organization of Human Chromosomes within the Nuclei of Normal and Emerin-Mutant Cells." *Human Molecular Genetics* 10 (3): 211–20. https://doi.org/10.1093/HMG/10.3.211.

Bulger, Michael, and Mark Groudine. 1999. "Looping versus Linking: Toward a Model for Long-Distance Gene Activation." *Genes and Development* 13: 2465–77. http://genesdev.cshlp.org/content/13/19/2465.extract.

Buniello, Annalisa, Jacqueline A.L. Macarthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, 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): D1005–12. https://doi.org/10.1093/NAR/GKY1120.

Cairns, Jonathan, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, et al. 2016. "CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C Data." *Genome Biology* 17 (1): 1–17. https://doi.org/10.1186/S13059-016-0992-2/FIGURES/10.

Calderon, Lesly, Felix D Weiss, Jonathan A Beagan, Marta S Oliveira, Radina Georgieva, Yi-Fang Wang, Thomas S Carroll, et al. 2022. "Cohesin-Dependence of Neuronal Gene Expression Relates to Chromatin Loop Length." *ELife* 11 (April). https://doi.org/10.7554/ELIFE.76539.

Cano-Gamez, Eddie, and Gosia Trynka. 2020. "From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases." *Frontiers in Genetics* 11 (May): 424. https://doi.org/10.3389/FGENE.2020.00424/BIBTEX.

Carbon, Seth, Amelia Ireland, Christopher J Mungall, Shengqiang Shu, Brad Marshall, Suzanna Lewis, Amigo Hub, Web Presence, and Working Group. 2009. "AmiGO: Online Access to Ontology and Annotation Data." *BIOINFORMATICS APPLICATIONS NOTE* 25 (2): 288–89. https://doi.org/10.1093/bioinformatics/btn615.

Carlson, Marc. 2017. "Org.Mm.Eg.Db: Genome Wide Annotation for Mouse. R Package Version 3.5.0." *Bioconductor*. https://doi.org/10.18129/B9.bioc.org.Mm.eg.db.

Chang, Li Hsin, Sourav Ghosh, and Daan Noordermeer. 2020. "TADs and Their Borders: Free Movement or Building a Wall?" *Journal of Molecular Biology*. Academic Press. https://doi.org/10.1016/j.jmb.2019.11.025.

Cho, Won Ki, Jan Hendrik Spille, Micca Hecht, Choongman Lee, Charles Li, Valentin Grube, and Ibrahim I. Cisse. 2018. "Mediator and RNA Polymerase II Clusters Associate in Transcription-Dependent Condensates." *Science* 361 (6400): 412–15. https://doi.org/10.1126/SCIENCE.AAR4199/SUPPL_FILE/AAR4199_S3.MOV.

Ciosk, Rafal, Masaki Shirayama, Anna Shevchenko, Tomoyuki Tanaka, Attila Toth, Andrej Shevchenko, and Kim Nasmyth. 2000. "Cohesin's Binding to Chromosomes Depends on a Separate Complex Consisting of Scc2 and Scc4 Proteins." *Molecular Cell* 5 (2): 243–54. https://doi.org/10.1016/S1097-2765(00)80420-7.

Crane, Emily, Qian Bian, Rachel Patton McCord, Bryan R. Lajoie, Bayly S. Wheeler, Edward J. Ralston, Satoru Uzawa, Job Dekker, and Barbara J. Meyer. 2015. "Condensin-Driven Remodeling of X-Chromosome Topology during Dosage Compensation." *Nature* 523 (7559): 240. https://doi.org/10.1038/NATURE14450.

Cremer, T., C. Cremer, H. Baumann, E. K. Luedtke, K. Sperling, V. Teuber, and C. Zorn. 1982. "Rabl's Model of the Interphase Chromosome Arrangement Tested in Chinise Hamster Cells by Premature Chromosome Condensation and Laser-UV-Microbeam Experiments." *Human Genetics* 60 (1): 46–56. https://doi.org/10.1007/BF00281263/METRICS.

Cremer, T., C. Cremer, T. Schneider, H. Baumann, L. Hens, and M. Kirsch-Volders. 1982. "Analysis of Chromosome Positions in the Interphase Nucleus of Chinese Hamster Cells by Laser-UV-Microirradiation Experiments." *Human Genetics* 62 (3): 201–9.

https://doi.org/10.1007/BF00333519.

Cuartero, Sergi, Felix D. Weiss, Gopuraja Dharmalingam, Ya Guo, Elizabeth Ing-Simmons, Silvia Masella, Irene Robles-Rebollo, et al. 2018. “Control of Inducible Gene Expression Links Cohesin to Hematopoietic Progenitor Self-Renewal and Differentiation.” *Nature Immunology 2018 19:9* 19 (9): 932–41. https://doi.org/10.1038/s41590-018-0184-1.

Dai, Haijiang, Tariq A. Alsalhe, Nasr Chalghaf, Matteo Riccò, Nicola Luigi Bragazzi, and Jianhong Wu. 2020. “The Global Burden of Disease Attributable to High Body Mass Index in 195 Countries and Territories, 1990–2017: An Analysis of the Global Burden of Disease Study.” *PLOS Medicine* 17 (7): e1003198. https://doi.org/10.1371/JOURNAL.PMED.1003198.

Dali, Rola, and Mathieu Blanchette. 2017. “A Critical Assessment of Topologically Associating Domain Prediction Tools.” *Nucleic Acids Research* 45 (6): 2994–3005. https://doi.org/10.1093/nar/gkx145.

Davis, Carrie A., Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, et al. 2018. “The Encyclopedia of DNA Elements (ENCODE): Data Portal Update.” *Nucleic Acids Research* 46 (D1): D794–801. https://doi.org/10.1093/nar/gkx1081.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. “Capturing Chromosome Conformation.” *Science (New York, N.Y.)* 295 (5558): 1306–11. https://doi.org/10.1126/science.1067799.

Deng, Wulan, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, Philip D. Gregory, Ann Dean, and Gerd A. Blobel. 2012. “Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor.” *Cell* 149 (6): 1233–44. https://doi.org/10.1016/J.CELL.2012.03.051.

Deng, Wulan, Jeremy W. Rupon, Ivan Krivega, Laura Breda, Irene Motta, Kristen S. Jahn, Andreas Reik, et al. 2014. “Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping.” *Cell* 158 (4): 849–60. https://doi.org/10.1016/J.CELL.2014.05.050.

Dickinson, Mary E., Ann M. Flenniken, Xiao Ji, Lydia Teboul, Michael D. Wong, Jacqueline K. White, Terrence F. Meehan, et al. 2016. “High-Throughput Discovery of Novel Developmental Phenotypes.” *Nature* 537 (7621): 508–14. https://doi.org/10.1038/nature19356.

Dixon, Jesse R., Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E. Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, et al. 2015. “Chromatin Architecture Reorganization during Stem Cell Differentiation.” *Nature* 518 (7539): 331–36. https://doi.org/10.1038/nature14222.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485 (7398): 376–80. https://doi.org/10.1038/nature11082.

Dixon, Jesse R, David U Gorkin, and Bing Ren. 2016. “Chromatin Domains: The Unit of Chromosome Organization.” *Molecular Cell* 62: 668–80.

https://doi.org/10.1016/j.molcel.2016.05.018.

Dostie, Josée, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, et al. 2006. “Chromosome Conformation Capture Carbon Copy (5C): A Massively Parallel Solution for Mapping Interactions between Genomic Elements.” *Genome Research* 16 (10): 1299–1309. https://doi.org/10.1101/gr.5571506.

Dryden, Nicola H., Laura R. Broome, Frank Dudbridge, Nichola Johnson, Nick Orr, Stefan Schoenfelder, Takashi Nagano, et al. 2014. “Unbiased Analysis of Potential Targets of Breast Cancer Susceptibility Loci by Capture Hi-C.” *Genome Research* 24 (11): 1854–68. https://doi.org/10.1101/GR.175034.114.

Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S.P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. “Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.” *Cell Systems* 3 (1): 95–98. https://doi.org/10.1016/J.CELS.2016.07.002.

Durand, Neva C, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. 2016. “Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom.” *Cell Systems* 3 (1): 99–101. https://doi.org/10.1016/j.cels.2015.07.012.

Durinck, Steffen, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. 2009. “Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package BiomaRt.” *Nature Protocols* 4 (8): 1184–91. https://doi.org/10.1038/NPROT.2009.97.

Dynan, William S., and Robert Tjian. 1985. “Control of Eukaryotic Messenger RNA Synthesis by Sequence-Specific DNA-Binding Proteins.” *Nature 1985 316:6031* 316 (6031): 774–78. https://doi.org/10.1038/316774a0.

Edgar, Ron, Michael Domrachev, and Alex E Lash. 2002. “Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository.” *Nucleic Acids Research* 30 (1): 207–10. http://www.ncbi.nlm.nih.gov/pubmed/11752295.

Eres, Ittai E., and Yoav Gilad. 2021. “A TAD Skeptic: Is 3D Genome Topology Conserved?” *Trends in Genetics*. Elsevier Ltd. https://doi.org/10.1016/j.tig.2020.10.009.

Eres, Ittai E., Kaixuan Luo, Chiaowen Joyce Hsiao, Lauren E. Blake, and Yoav Gilad. 2019. “Reorganization of 3D Genome Structure May Contribute to Gene Regulatory Evolution in Primates.” Edited by Harmit S. Malik. *PLOS Genetics* 15 (7): e1008278. https://doi.org/10.1371/journal.pgen.1008278.

Ernst, Jason, and Manolis Kellis. 2010. “Discovery and Characterization of Chromatin States for Systematic Annotation of the Human Genome.” *Nature Biotechnology 2010 28:8* 28 (8): 817–25. https://doi.org/10.1038/nbt.1662.

———. 2017. “Chromatin-State Discovery and Genome Annotation with ChromHMM.” *Nature Protocols 2017 12:12* 12 (12): 2478–92. https://doi.org/10.1038/nprot.2017.124.

Ernst, Jason, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, et al. 2011. “Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types.” *Nature 2011 473:7345* 473 (7345): 43–49. https://doi.org/10.1038/nature09906.

Feric, Marina, Nilesh Vaidya, Tyler S. Harmon, Diana M. Mitrea, Lian Zhu, Tiffany M. Richardson, Richard W. Kriwacki, Rohit V. Pappu, and Clifford P. Brangwynne. 2016. "Coexisting Liquid Phases Underlie Nucleolar Subcompartments." *Cell* 165 (7): 1686–97. https://doi.org/10.1016/J.CELL.2016.04.047.

Finucane, Hilary K., Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, et al. 2018. "Heritability Enrichment of Specifically Expressed Genes Identifies Disease-Relevant Tissues and Cell Types." *Nature Genetics 2018 50:4* 50 (4): 621–29. https://doi.org/10.1038/s41588-018-0081-4.

Fiorillo, Luca, Francesco Musella, Mattia Conte, Rieke Kempfer, Andrea M. Chiariello, Simona Bianco, Alexander Kukalev, et al. 2021. "Comparison of the Hi-C, GAM and SPRITE Methods Using Polymer Models of Chromatin." *Nature Methods* 18 (5): 482. https://doi.org/10.1038/S41592-021-01135-1.

Flavahan, William A., Yotam Drier, Brian B. Liau, Shawn M. Gillespie, Andrew S. Venteicher, Anat O. Stemmer-Rachamimov, Mario L. Suvà, and Bradley E. Bernstein. 2016. "Insulator Dysfunction and Oncogene Activation in IDH Mutant Gliomas." *Nature* 529 (7584): 110–14. https://doi.org/10.1038/nature16490.

Flyamer, Ilya M., Johanna Gassler, Maxim Imakaev, Hugo B. Brandão, Sergey V. Ulianov, Nezar Abdennur, Sergey V. Razin, Leonid A. Mirny, and Kikuë Tachibana-Konwalski. 2017. "Single-Nucleus Hi-C Reveals Unique Chromatin Reorganization at Oocyte-to-Zygote Transition." *Nature* 544 (7648): 110–14. https://doi.org/10.1038/nature21711.

Forcato, Mattia, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. 2017. "Comparison of Computational Methods for Hi-C Data Analysis." *Nature Methods* 14 (7). https://doi.org/10.1038/nmeth.4325.

Franke, Martin, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, et al. 2016. "Formation of New Chromatin Domains Determines Pathogenicity of Genomic Duplications." *Nature* 538 (7624): 265–69. https://doi.org/10.1038/nature19800.

Fraser, J., C. Ferrai, A. M. Chiariello, M. Schueler, T. Rito, G. Laudanno, M. Barbieri, et al. 2015. "Hierarchical Folding and Reorganization of Chromosomes Are Linked to Transcriptional Changes in Cellular Differentiation." *Molecular Systems Biology* 11 (12): 852–852. https://doi.org/10.15252/msb.20156492.

Freire-Pritchett, Paula, Helen Ray-Jones, Monica Della Rosa, Chris Q. Eijsbouts, William R. Orchard, Steven W. Wingett, Chris Wallace, Jonathan Cairns, Mikhail Spivakov, and Valeriya Malysheva. 2021. "Detecting Chromosomal Interactions in Capture Hi-C Data with CHiCAGO and Companion Tools." *Nature Protocols 2021 16:9* 16 (9): 4144–76. https://doi.org/10.1038/s41596-021-00567-5.

Freire-Pritchett, Paula, Stefan Schoenfelder, Csilla Várnai, Steven W Wingett, Jonathan Cairns, Amanda J Collier, Raquel García-Vílchez, et al. 2017. "Global Reorganisation of Cis-Regulatory Units upon Lineage Commitment of Human Embryonic Stem Cells." *ELife* 6 (March). https://doi.org/10.7554/eLife.21926.

Fudenberg, Geoffrey, Nezar Abdennur, Maxim Imakaev, Anton Goloborodko, and Leonid A. Mirny. 2017. "Emerging Evidence of Chromosome Folding by Loop Extrusion." *Cold*

Spring Harbor Symposia on Quantitative Biology 82 (January): 45–55. https://doi.org/10.1101/SQB.2017.82.034710/-/DC1.

Fudenberg, Geoffrey, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A. Mirny. 2016. "Formation of Chromosomal Domains by Loop Extrusion." *Cell Reports* 15 (9): 2038–49. https://doi.org/10.1016/J.CELREP.2016.04.085.

Furlong, Eileen E.M., and Michael Levine. 2018. "Developmental Enhancers and Chromosome Topology." *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.aau0320.

Gandhi, Rita, Peter J. Gillespie, and Tatsuya Hirano. 2006. "Human Wapl Is a Cohesin-Binding Protein That Promotes Sister-Chromatid Resolution in Mitotic Prophase." *Current Biology : CB* 16 (24): 2406–17. https://doi.org/10.1016/J.CUB.2006.10.061.

Gassler, Johanna, Hugo B Brandão, Maxim Imakaev, Ilya M Flyamer, Sabrina Ladstätter, Wendy A Bickmore, Jan-Michael Peters, Leonid A Mirny, and Kikuë Tachibana. 2017. "A Mechanism of Cohesin-Dependent Loop Extrusion Organizes Zygotic Genome Architecture." *The EMBO Journal* 36 (24): 3600–3618. https://doi.org/10.15252/EMBJ.201798083.

Ghavi-Helm, Yad, Aleksander Jankowski, Sascha Meiers, Rebecca R. Viales, Jan O. Korbel, and Eileen E. M. Furlong. 2019. "Highly Rearranged Chromosomes Reveal Uncoupling between Genome Topology and Gene Expression." *Nature Genetics* 51 (8): 1272–82. https://doi.org/10.1038/s41588-019-0462-3.

Ghavi-Helm, Yad, Felix A. Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen E.M. Furlong. 2014. "Enhancer Loops Appear Stable during Development and Are Associated with Paused Polymerase." *Nature* 512 (7512): 96–100. https://doi.org/10.1038/NATURE13417.

Gilbertson, Sarah E., Hannah C. Walter, Katherine Gardner, Spencer N. Wren, Golnaz Vahedi, and Amy S. Weinmann. 2022. "Topologically Associating Domains Are Disrupted by Evolutionary Genome Rearrangements Forming Species-Specific Enhancer Connections in Mice and Humans." *Cell Reports* 39 (5). https://doi.org/10.1016/J.CELREP.2022.110769.

Giorgetti, Luca, Bryan R. Lajoie, Ava C. Carter, Mikael Attia, Ye Zhan, Jin Xu, Chong Jian Chen, et al. 2016. "Structural Organization of the Inactive X Chromosome in the Mouse." *Nature* 535 (7613): 575–79. https://doi.org/10.1038/nature18589.

Giorgio, Elisa, Daniel Robyr, Malte Spielmann, Enza Ferrero, Eleonora Di Gregorio, Daniele Imperiale, Giovanna Vaula, et al. 2015. "A Large Genomic Deletion Leads to Enhancer Adoption by the Lamin B1 Gene: A Second Path to Autosomal Dominant Adult-Onset Demyelinating Leukodystrophy (ADLD)." *Human Molecular Genetics* 24 (11): 3143–54. https://doi.org/10.1093/HMG/DDV065.

Haarhuis, Judith H.I., Robin H. van der Weide, Vincent A. Blomen, J. Omar Yáñez-Cuna, Mario Amendola, Marjon S. van Ruiten, Peter H.L. Krijger, et al. 2017. "The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension." *Cell* 169 (4): 693-707.e14. https://doi.org/10.1016/J.CELL.2017.04.013/ATTACHMENT/A7AB8846-8FD4-40D7-8DF0-E55B34A296EC/MMC1.XLSX.

Haberle, Vanja, and Alexander Stark. 2018. “Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation.” *Nature Reviews Molecular Cell Biology 2018 19:10* 19 (10): 621–37. https://doi.org/10.1038/s41580-018-0028-8.

Haddad, Noelle, Cédric Vaillant, and Daniel Jost. 2017. “IC-Finder: Inferring Robustly the Hierarchical Organization of Chromatin Folding.” *Nucleic Acids Research* 45 (10): e81–e81. https://doi.org/10.1093/NAR/GKX036.

Hansen, Anders S. 2020. “CTCF as a Boundary Factor for Cohesin-Mediated Loop Extrusion: Evidence for a Multi-Step Mechanism.” *Nucleus* 11 (1): 132. https://doi.org/10.1080/19491034.2020.1782024.

Hansen, Anders S., Iryna Pustova, Claudia Cattoglio, Robert Tjian, and Xavier Darzacq. 2017. “CTCF and Cohesin Regulate Chromatin Loop Stability with Distinct Dynamics.” *ELife* 6 (May). https://doi.org/10.7554/ELIFE.25776.001.

Hanssen, Lars L.P., Mira T. Kassouf, A. Marieke Oudelaar, Daniel Biggs, Chris Preece, Damien J. Downes, Matthew Gosden, et al. 2017. “Tissue-Specific CTCF–Cohesin-Mediated Chromatin Architecture Delimits Enhancer Interactions and Function in Vivo.” *Nature Cell Biology 2017 19:8* 19 (8): 952–61. https://doi.org/10.1038/ncb3573.

He, Mengnan, Yan Li, Qianzi Tang, Diyan Li, Long Jin, Shilin Tian, Tiandong Che, et al. 2018. “Genome-Wide Chromatin Structure Changes During Adipogenesis and Myogenesis.” *Int. J. Biol. Sci* 14 (11): 1571–85. https://doi.org/10.7150/ijbs.25328.

He, Peng, Brian A. Williams, Diane Trout, Georgi K. Marinov, Henry Amrhein, Libera Berghella, Say Tar Goh, et al. 2020. “The Changing Mouse Embryo Transcriptome at Whole Tissue and Single-Cell Resolution.” *Nature* 583 (7818): 760–67. https://doi.org/10.1038/s41586-020-2536-x.

Heger, Peter, Birger Marin, and Einhard Schierenberg. 2009. “Loss of the Insulator Protein CTCF during Nematode Evolution.” *BMC Molecular Biology* 10 (August): 84. https://doi.org/10.1186/1471-2199-10-84.

Heid, Iris M., Anne U. Jackson, Joshua C. Randall, Tthomas W. Winkler, Llu Qi, Valgerdur Ssteinthorsdottir, Ggudmar Tthorleifsson, et al. 2010. “Meta-Analysis Identifies 13 New Loci Associated with Waist-Hip Ratio and Reveals Sexual Dimorphism in the Genetic Basis of Fat Distribution.” *Nature Genetics 2010 42:11* 42 (11): 949–60. https://doi.org/10.1038/ng.685.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. “Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities.” *Molecular Cell* 38 (4): 576–89. https://doi.org/10.1016/J.MOLCEL.2010.05.004.

Hildebrand, Erica M., and Job Dekker. 2020. “Mechanisms and Functions of Chromosome Compartmentalization.” *Trends in Biochemical Sciences*. Elsevier Ltd. https://doi.org/10.1016/j.tibs.2020.01.002.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, et al. 2006. “The UCSC Genome Browser Database: Update 2006.” *Nucleic Acids Research*

34 (Database issue). https://doi.org/10.1093/NAR/GKJ144.

Hiratani, Ichiro, Tyrone Ryba, Mari Itoh, Tomoki Yokochi, Michaela Schwaiger, Chia Wei Chang, Yung Lyou, Tim M. Townes, Dirk Schübeler, and David M. Gilbert. 2008. "Global Reorganization of Replication Domains During Embryonic Stem Cell Differentiation." *PLOS Biology* 6 (10): e245. https://doi.org/10.1371/JOURNAL.PBIO.0060245.

Hnisz, Denes, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp. 2017. "A Phase Separation Model for Transcriptional Control." *Cell* 169 (1): 13–23. https://doi.org/10.1016/J.CELL.2017.02.007.

Hoffmann, Thomas J., Hélène Choquet, Jie Yin, Yambazi Banda, Mark N. Kvale, Maria Glymour, Catherine Schaefer, Neil Risch, and Eric Jorgenson. 2018. "A Large Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci." *Genetics* 210 (2): 499–515. https://doi.org/10.1534/GENETICS.118.301479.

Hruby, Adela, and Frank B. Hu. 2015. "The Epidemiology of Obesity: A Big Picture." *PharmacoEconomics* 33 (7): 673. https://doi.org/10.1007/S40273-014-0243-X.

Hurst, Laurence D., Csaba Pál, and Martin J. Lercher. 2004. "The Evolutionary Dynamics of Eukaryotic Gene Order." *Nature Reviews Genetics* 5 (4): 299–310. https://doi.org/10.1038/nrg1319.

Ibn-Salem, Jonas, Enrique M. Muro, and Miguel A. Andrade-Navarro. 2017. "Co-Regulation of Paralog Genes in the Three-Dimensional Chromatin Architecture." *Nucleic Acids Research* 45 (1): 81–91. https://doi.org/10.1093/nar/gkw813.

Ing-Simmons, Elizabeth, Vlad C. Seitan, Andre J. Faure, Paul Flicek, Thomas Carroll, Job Dekker, Amanda G. Fisher, Boris Lenhard, and Matthias Merkenschlager. 2015. "Spatial Enhancer Clustering and Regulation of Enhancer-Proximal Genes by Cohesin." *Genome Research* 25 (4): 504. https://doi.org/10.1101/GR.184986.114.

Jackson, Dean A., and Ana Pombo. 1998. "Replicon Clusters Are Stable Units of Chromosome Structure: Evidence That Nuclear Organization Contributes to the Efficient Activation and Propagation of S Phase in Human Cells." *The Journal of Cell Biology* 140 (6): 1285–95. https://doi.org/10.1083/JCB.140.6.1285.

Javierre, Biola M., Sven Sewitz, Jonathan Cairns, Steven W. Wingett, Csilla Várnai, Michiel J. Thiecke, Paula Freire-Pritchett, et al. 2016. "Lineage-Specific Genome Architecture Links Enhancers and Non-Coding Disease Variants to Target Gene Promoters." *Cell* 167 (5): 1369-1384.e19. https://doi.org/10.1016/J.CELL.2016.09.037/ATTACHMENT/5BC79F6F-1B69-4192-8CB8-4247CC2E0F39/MMC4.ZIP.

Joslin, Amelia C., Débora R. Sobreira, Grace T. Hansen, Noboru J. Sakabe, Ivy Aneas, Lindsey E. Montefiori, Kathryn M. Farris, et al. 2021. "A Functional Genomics Pipeline Identifies Pleiotropy and Cross-Tissue Effects within Obesity-Associated GWAS Loci." *Nature Communications 2021 12:1* 12 (1): 1–15. https://doi.org/10.1038/s41467-021-25614-3.

Kadauke, Stephan, and Gerd A. Blobel. 2009. "Chromatin Loops in Gene Regulation." *Biochimica et Biophysica Acta* 1789 (1): 17. https://doi.org/10.1016/J.BBAGRM.2008.07.002.

Karolchik, Donna, Angela S. Hinricks, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet,

David Haussler, and W. James Kent. 2004. “The UCSC Table Browser Data Retrieval Tool.” *Nucleic Acids Research* 32 (DATABASE ISS.). https://doi.org/10.1093/nar/gkh103.

Keane, Thomas M., Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, et al. 2011. “Mouse Genomic Variation and Its Effect on Phenotypes and Gene Regulation.” *Nature* 477 (7364): 289–94. https://doi.org/10.1038/nature10413.

Kichaev, Gleb, Gaurav Bhatia, Po Ru Loh, Steven Gazal, Kathryn Burch, Malika K. Freund, Armin Schoech, Bogdan Pasaniuc, and Alkes L. Price. 2019. “Leveraging Polygenic Functional Enrichment to Improve GWAS Power.” *American Journal of Human Genetics* 104 (1): 65–75. https://doi.org/10.1016/J.AJHG.2018.11.008/ATTACHMENT/66985F37-6CC0-4A42-8AFB-6CAE1352D095/MMC2.XLSX.

Kim, Somi, Nam Kyung Yu, and Bong Kiun Kaang. 2015. “CTCF as a Multifunctional Protein in Genome Regulation and Gene Expression.” *Experimental & Molecular Medicine 2015 47:6* 47 (6): e166–e166. https://doi.org/10.1038/emm.2015.33.

Kim, Yong Hoon, Sajid A Marhon, Yuxiang Zhang, David J Steger, Kyoung-Jae Won, and Mitchell A Lazar. 2018. “Rev-Erba Dynamically Modulates Chromatin Looping to Control Circadian Gene Transcription.” *Science*. Vol. 359. http://science.sciencemag.org/.

Krefting, Jan, Miguel A. Andrade-Navarro, and Jonas Ibn-Salem. 2018. “Evolutionary Stability of Topologically Associating Domains Is Associated with Conserved Gene Regulation.” *BMC Biology* 16 (1): 87. https://doi.org/10.1186/s12915-018-0556-x.

Kueng, Stephanie, Björn Hegemann, Beate H. Peters, Jesse J. Lipp, Alexander Schleiffer, Karl Mechtler, and Jan Michael Peters. 2006. “Wapl Controls the Dynamic Association of Cohesin with Chromatin.” *Cell* 127 (5): 955–67. https://doi.org/10.1016/J.CELL.2006.09.040.

Laber, Samantha, Sara Forcisi, Liz Bentley, Julia Petzold, Franco Moritz, Kirill S. Smirnov, Loubna Al Sadat, et al. 2021. “Linking the FTO Obesity Rs1421085 Variant Circuitry to Cellular, Metabolic, and Organismal Phenotypes in Vivo.” *Science Advances* 7 (30). https://doi.org/10.1126/SCIADV.ABG0108/SUPPL_FILE/SCIADV.ABG0108_SM.PDF.

Lajoie, Bryan R, Job Dekker, and Noam Kaplan. 2015. “The Hitchhiker’s Guide to Hi-C Analysis: Practical Guidelines.” *Methods (San Diego, Calif.)* 72 (January): 65–75. https://doi.org/10.1016/j.ymeth.2014.10.031.

Langer-Safer, P. R., M. Levine, and D. C. Ward. 1982. “Immunological Method for Mapping Genes on Drosophila Polytene Chromosomes.” *Proceedings of the National Academy of Sciences of the United States of America* 79 (14): 4381. https://doi.org/10.1073/PNAS.79.14.4381.

Lappalainen, Tuuli, and Daniel G. MacArthur. 2021. “From Variant to Function in Human Disease Genetics.” *Science* 373 (6562): 1464–68. https://doi.org/10.1126/SCIENCE.ABI8207/ASSET/A27CC20A-79DE-43E2-9E21-6131BB9A0054/ASSETS/IMAGES/LARGE/SCIENCE.ABI8207-T1.JPG.

Lazar, Nathan H., Kimberly A. Nevonen, Brendan O’Connell, Christine McCann, Rachel J. O’Neill, Richard E. Green, Thomas J. Meyer, Mariam Okhovat, and Lucia Carbone. 2018. “Epigenetic Maintenance of Topological Domains in the Highly Rearranged Gibbon

Genome." *Genome Research* 28 (7): 983–97. https://doi.org/10.1101/gr.233874.117.

Leung, Danny, Inkyung Jung, Nisha Rajagopal, Anthony Schmitt, Siddarth Selvaraj, Ah Young Lee, Chia-An Yen, et al. 2015. "Integrative Analysis of Haplotype-Resolved Epigenomes across Human Tissues." *Nature* 518 (7539): 350–54. https://doi.org/10.1038/nature14217.

Lévy-Leduc, Celine, M. Delattre, T. Mary-Huard, and S. Robin. 2014. "Two-Dimensional Segmentation for Analyzing Hi-C Data." *Bioinformatics (Oxford, England)* 30 (17). https://doi.org/10.1093/BIOINFORMATICS/BTU443.

Li, Daofeng, Silas Hsu, Deepak Purushotham, Renee L. Sears, and Ting Wang. 2019. "WashU Epigenome Browser Update 2019." *Nucleic Acids Research* 47 (W1): W158–65. https://doi.org/10.1093/NAR/GKZ348.

Li, Gang, Gail Sudlow, and Andrew S. Belmont. 1998. "Interphase Cell Cycle Dynamics of a Late-Replicating, Heterochromatic Homogeneously Staining Region: Precise Choreography of Condensation/Decondensation and Nuclear Positioning." *The Journal of Cell Biology* 140 (5): 975. https://doi.org/10.1083/JCB.140.5.975.

Lichter, P., T. Cremer, J. Borden, L. Manuelidis, and D. C. Ward. 1988. "Delineation of Individual Human Chromosomes in Metaphase and Interphase Cells by in Situ Suppression Hybridization Using Recombinant DNA Libraries." *Human Genetics* 80 (3): 224–34. https://doi.org/10.1007/BF01790090.

Lieberman-Aiden, Erez, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science (New York, N.Y.)* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Liu, Ching Ti, Martin L. Buchkovich, Thomas W. Winkler, Iris M. Heid, Ingrid B. Borecki, Caroline S. Fox, Karen L. Mohlke, Kari E. North, and L. Adrienne Cupples. 2014. "Multi-Ethnic Fine-Mapping of 14 Central Adiposity Loci." *Human Molecular Genetics* 23 (17): 4738–44. https://doi.org/10.1093/HMG/DDU183.

Liu, Ning Qing, Michela Maresca, Teun van den Brand, Luca Braccioli, Marijne M.G.A. Schijns, Hans Teunissen, Benoit G. Bruneau, Elphège èP Nora, and Elzo de Wit. 2020. "WAPL Maintains a Cohesin Loading Cycle to Preserve Cell-Type-Specific Distal Gene Regulation." *Nature Genetics 2020 53:1* 53 (1): 100–109. https://doi.org/10.1038/s41588-020-00744-4.

Long, Helen S., Simon Greenaway, George Powell, Ann-Marie Mallon, Cecilia M. Lindgren, and Michelle M. Simon. 2022. "Making Sense of the Linear Genome, Gene Function and TADs." *Epigenetics & Chromatin 2022 15:1* 15 (1): 1–19. https://doi.org/10.1186/S13072-022-00436-9.

Loos, Ruth J.F., and Giles S.H. Yeo. 2021. "The Genetics of Obesity: From Discovery to Biology." *Nature Reviews Genetics 2021 23:2* 23 (2): 120–33. https://doi.org/10.1038/s41576-021-00414-z.

Lu, Leina, Xiaoxiao Liu, Wei Kai Huang, Paola Giusti-Rodríguez, Jian Cui, Shanshan Zhang, Wanying Xu, et al. 2020. "Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal

the Function of Non-Coding Genome in Neural Development and Diseases." *Molecular Cell* 79 (3): 521-534.e15. https://doi.org/10.1016/J.MOLCEL.2020.06.007.

Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, et al. 2015. "Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions." *Cell* 161 (5): 1012–25. https://doi.org/10.1016/J.CELL.2015.04.004.

Lupiáñez, Darío G, Malte Spielmann, and Stefan Mundlos. 2016. "Breaking TADs: How Alterations of Chromatin Domains Result in Disease." *Trends in Genetics : TIG* 32 (4): 225–37. https://doi.org/10.1016/j.tig.2016.01.003.

Ma, Hong, Jagath Samarabandu, Rekandu S. Devdhar, Raj Acharya, Ping Chin Cheng, Chunling Meng, and Ronald Berezney. 1998. "Spatial and Temporal Dynamics of DNA Replication Sites in Mammalian Cells." *The Journal of Cell Biology* 143 (6): 1415–25. https://doi.org/10.1083/JCB.143.6.1415.

Madsen, Jesper G.S., Maria S. Madsen, Alexander Rauch, Sofie Traynor, Elvira L. Van Hauwaert, Anders K. Haakonsson, Biola M. Javierre, Mette Hyldahl, Peter Fraser, and Susanne Mandrup. 2020. "Highly Interconnected Enhancer Communities Control Lineage-Determining Genes in Human Mesenchymal Stem Cells." *Nature Genetics* 52 (11): 1227–38. https://doi.org/10.1038/s41588-020-0709-z.

Man, Joyce C.K., Rajiv A. Mohan, Malou van den Boogaard, Catharina R.E. Hilvering, Catherine Jenkins, Vincent Wakker, Valerio Bianchi, et al. 2019. "An Enhancer Cluster Controls Gene Activity and Topology of the SCN5A-SCN10A Locus in Vivo." *Nature Communications 2019 10:1* 10 (1): 1–15. https://doi.org/10.1038/s41467-019-12856-5.

Matharu, Navneet, and Nadav Ahituv. 2015. "Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease." Edited by Elizabeth M. C. Fisher. *PLOS Genetics* 11 (12): e1005640. https://doi.org/10.1371/journal.pgen.1005640.

McArthur, Evonne, and John A. Capra. 2021. "Topologically Associating Domain Boundaries That Are Stable across Diverse Cell Types Are Evolutionarily Constrained and Enriched for Heritability." *American Journal of Human Genetics* 108 (2): 269–83. https://doi.org/10.1016/J.AJHG.2021.01.001/ATTACHMENT/96E0D966-A76B-4B0B-80A9-501B98C37288/MMC1.PDF.

McGinty, Robert K., and Song Tan. 2015. "Nucleosome Structure and Function." *Chemical Reviews* 115 (6): 2255–73. https://doi.org/10.1021/CR500373H/ASSET/IMAGES/LARGE/CR-2014-00373H_0014.JPEG.

Medina-Gomez, Carolina, John P. Kemp, Niki L. Dimou, Eskil Kreiner, Alessandra Chesi, Babette S. Zemel, Klaus Bønnelykke, et al. 2017. "Bivariate Genome-Wide Association Meta-Analysis of Pediatric Musculoskeletal Traits Reveals Pleiotropic Effects at the SREBF1/TOM1L2 Locus." *Nature Communications 2017 8:1* 8 (1): 1–11. https://doi.org/10.1038/s41467-017-00108-3.

Meyenn, Ferdinand von, Rebecca V. Berrens, Simon Andrews, Fátima Santos, Amanda J. Collier, Felix Krueger, Rodrigo Osorno, Wendy Dean, Peter J. Rugg-Gunn, and Wolf Reik.

2016. “Comparative Principles of DNA Methylation Reprogramming during Human and Mouse In Vitro Primordial Germ Cell Specification.” *Developmental Cell* 39 (1): 104–15. https://doi.org/10.1016/j.devcel.2016.09.015.

Mifsud, Borbala, Filipe Tavares-Cadete, Alice N. Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W. Wingett, et al. 2015. “Mapping Long-Range Promoter Contacts in Human Cells with High-Resolution Capture Hi-C.” *Nature Genetics 2015 47:6* 47 (6): 598–606. https://doi.org/10.1038/ng.3286.

Mirny, Leonid A, Maxim Imakaev, and Nezar Abdennur. 2019. “Two Major Mechanisms of Chromosome Organization.” *Current Opinion in Cell Biology* 58 (June): 142–52. https://doi.org/10.1016/j.ceb.2019.05.001.

Misteli, Tom. 2020. “The Self-Organizing Genome: Principles of Genome Architecture and Function.” *Cell* 183 (1): 28–45. https://doi.org/10.1016/J.CELL.2020.09.014.

Miura, Hisashi, Rawin Poonperm, Saori Takahashi, and Ichiro Hiratani. 2018. “Practical Analysis of Hi-C Data: Generating A/B Compartment Profiles.” *Methods in Molecular Biology* 1861: 221–45. https://doi.org/10.1007/978-1-4939-8766-5_16.

Montefiori, Lindsey E., Debora R. Sobreira, Noboru J. Sakabe, Ivy Aneas, Amelia C. Joslin, Grace T. Hansen, Grazyna Bozek, Ivan P. Moskowitz, Elizabeth M. McNally, and Marcelo A. Nóbrega. 2018. “A Promoter Interaction Map for Cardiovascular Disease Genetics.” *ELife* 7 (July). https://doi.org/10.7554/ELIFE.35788.

Mora, Antonio, Geir Kjetil Sandve, Odd Stokke Gabrielsen, and Ragnhild Eskeland. 2015. “In the Loop: Promoter–Enhancer Interactions and Bioinformatics.” *Briefings in Bioinformatics* 17 (6): bbv097. https://doi.org/10.1093/bib/bbv097.

Moreau, P., R. Hen, B. Wasylyk, R. Everett, M. P. Gaub, and P. Chambon. 1981. “The SV40 72 Base Repair Repeat Has a Striking Effect on Gene Expression Both in SV40 and Other Chimeric Recombinants.” *Nucleic Acids Research* 9 (22): 6047–68. https://doi.org/10.1093/NAR/9.22.6047.

Morrison, Olivia, and Jitendra Thakur. 2021. “Molecular Complexes at Euchromatin, Heterochromatin and Centromeric Chromatin.” *International Journal of Molecular Sciences* 22 (13). https://doi.org/10.3390/IJMS22136922.

Murakami, Yota. 2013. “Heterochromatin and Euchromatin.” *Encyclopedia of Systems Biology*, 881–84. https://doi.org/10.1007/978-1-4419-9863-7_1413.

Muro, Enrique M., Jonas Ibn-Salem, and Miguel A. Andrade-Navarro. 2019. “The Distributions of Protein Coding Genes within Chromatin Domains in Relation to Human Disease.” *Epigenetics and Chromatin* 12 (1): 72. https://doi.org/10.1186/s13072-019-0317-2.

Myers, Timothy A., Stephen J. Chanock, and Mitchell J. Machiela. 2020. “LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations.” *Frontiers in Genetics* 11 (February). https://doi.org/10.3389/FGENE.2020.00157.

Nagano, Takashi, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. 2013. “Single-Cell Hi-C Reveals Cell-to-Cell Variability in Chromosome Structure.” *Nature* 502 (7469): 59–64. https://doi.org/10.1038/nature12593.

Nagano, Takashi, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. 2017. "Cell-Cycle Dynamics of Chromosomal Organization at Single-Cell Resolution." *Nature 2017 547:7661* 547 (7661): 61–67. https://doi.org/10.1038/nature23001.

Ndiaye, Fatou K., Marlène Huyvaert, Ana Ortalli, Mickaël Canouil, Cécile Lecoeur, Marie Verbanck, Stéphane Lobbens, et al. 2019. "The Expression of Genes in Top Obesity-Associated Loci Is Enriched in Insula and Substantia Nigra Brain Regions Involved in Addiction and Reward." *International Journal of Obesity (2005)* 44 (2): 539. https://doi.org/10.1038/S41366-019-0428-7.

Neems, Daniel S., Arturo G. Garza-Gongora, Erica D. Smith, and Steven T. Kosak. 2016. "Topologically Associated Domains Enriched for Lineage-Specific Genes Reveal Expression-Dependent Nuclear Topologies during Myogenesis." *Proceedings of the National Academy of Sciences* 113 (12): E1691–1700. https://doi.org/10.1073/PNAS.1521826113.

Ngan, Chew Yee, Chee Hong Wong, Harianto Tjong, Wenbo Wang, Rachel L. Goldfeder, Cindy Choi, Hao He, et al. 2020. "Chromatin Interaction Analyses Elucidate the Roles of PRC2-Bound Silencers in Mouse Development." *Nature Genetics* 52 (3): 264–72. https://doi.org/10.1038/S41588-020-0581-X.

Nichols, Michael H., and Victor G. Corces. 2021. "Principles of 3D Compartmentalization of the Human Genome." *Cell Reports* 35 (13): 109330. https://doi.org/10.1016/J.CELREP.2021.109330.

Niimura, Yoshihito. 2009. "Evolutionary Dynamics of Olfactory Receptor Genes in Chordates: Interaction between Environments and Genomic Contents." *Human Genomics*. BioMed Central. https://doi.org/10.1186/1479-7364-4-2-107.

Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. "Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre." *Nature* 485 (7398): 381–85. https://doi.org/10.1038/nature11049.

Norton, Heidi K., Daniel J. Emerson, Harvey Huang, Jesi Kim, Katelyn R. Titus, Shi Gu, Danielle S. Bassett, and Jennifer E. Phillips-Cremins. 2018. "Detecting Hierarchical Genome Folding with Network Modularity." *Nature Methods 2018 15:2* 15 (2): 119–22. https://doi.org/10.1038/nmeth.4560.

Nuebler, Johannes, Geoffrey Fudenberg, Maxim Imakaev, Nezar Abdennur, and Leonid A. Mirny. 2018. "Chromatin Organization by an Interplay of Loop Extrusion and Compartmental Segregation." *Proceedings of the National Academy of Sciences of the United States of America* 115 (29): E6697–6706. https://doi.org/10.1073/pnas.1717730115.

O'Keefe, R. T., S. C. Henderson, and D. L. Spector. 1992. "Dynamic Organization of DNA Replication in Mammalian Cell Nuclei: Spatially and Temporally Defined Replication of Chromosome-Specific Alpha-Satellite DNA Sequences." *The Journal of Cell Biology* 116 (5): 1095. https://doi.org/10.1083/JCB.116.5.1095.

O'Sullivan, Justin M., Michael D. Hendy, Tatyana Pichugina, Graeme C. Wake, and Jörg

Langowski. 2013. “The Statistical-Mechanics of Chromosome Conformation Capture.” *Nucleus (United States)* 4 (5): 390–98. https://doi.org/10.4161/NUCL.26513/SUPPL_FILE/KNCL_A_10926513_SM0001.ZIP.

Oldach, Phoebe, and Conrad A. Nieduszynski. 2019. “Cohesin-Mediated Genome Architecture Does Not Define DNA Replication Timing Domains.” *Genes* 10 (3). https://doi.org/10.3390/GENES10030196.

Oluwadare, Oluwatosin, and Jianlin Cheng. 2017. “ClusterTAD: An Unsupervised Machine Learning Approach to Detecting Topologically Associated Domains of Chromosomes from Hi-C Data.” *BMC Bioinformatics* 18 (1): 1–14. https://doi.org/10.1186/S12859-017-1931-2/FIGURES/8.

Oudelaar, A. Marieke, James O.J. Davies, Lars L.P. Hanssen, Jelena M. Telenius, Ron Schwessinger, Yu Liu, Jill M. Brown, et al. 2018. “Single-Allele Chromatin Interactions Identify Regulatory Hubs in Dynamic Compartmentalized Domains.” *Nature Genetics 2018 50:12* 50 (12): 1744–51. https://doi.org/10.1038/s41588-018-0253-2.

Oudelaar, A Marieke, and Douglas R Higgs. 2020. “The Relationship between Genome Structure and Function.” *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-00303-x.

Paliou, Christina, Philine Guckelberger, Robert Schöpflin, Verena Heinrich, Andrea Esposito, Andrea M. Chiariello, Simona Bianco, et al. 2019. “Preformed Chromatin Topology Assists Transcriptional Robustness of Shh during Limb Development.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (25): 12390–99. https://doi.org/10.1073/PNAS.1900672116/SUPPL_FILE/PNAS.1900672116.SM02.MP4.

Pan, David Z., Kristina M. Garske, Marcus Alvarez, Yash V. Bhagat, James Boocock, Elina Nikkola, Zong Miao, et al. 2018. “Integration of Human Adipocyte Chromosomal Interactions with Adipose Gene Expression Prioritizes Obesity-Related Genes from GWAS.” *Nature Communications* 9 (1): 1512. https://doi.org/10.1038/s41467-018-03554-9.

Pang, Baoxu, and Michael P. Snyder. 2020. “Systematic Identification of Silencers in Human Cells.” *Nature Genetics 2020 52:3* 52 (3): 254–63. https://doi.org/10.1038/s41588-020-0578-5.

Panigrahi, Anil, and Bert W. O’Malley. 2021. “Mechanisms of Enhancer Action: The Known and the Unknown.” *Genome Biology 2021 22:1* 22 (1): 1–30. https://doi.org/10.1186/S13059-021-02322-1.

Pardiñas, Antonio F., Peter Holmans, Andrew J. Pocklington, Valentina Escott-Price, Stephan Ripke, Noa Carrera, Sophie E. Legge, et al. 2018. “Common Schizophrenia Alleles Are Enriched in Mutation-Intolerant Genes and in Regions under Strong Background Selection.” *Nature Genetics 2018 50:3* 50 (3): 381–89. https://doi.org/10.1038/s41588-018-0059-2.

Phillips-Cremins, Jennifer E., Michael E.G. Sauria, Amartya Sanyal, Tatiana I. Gerasimova, Bryan R. Lajoie, Joshua S.K. Bell, Chin-Tong Ong, et al. 2013. “Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment.” *Cell* 153 (6): 1281–95. https://doi.org/10.1016/j.cell.2013.04.053.

Pinkel, D, J Landegentt, C Collins, J Fuscoet, R Segraves, J Lucas, and J Gray. 1988. "Fluorescence in Situ Hybridization with Human Chromosome-Specific Libraries: Detection of Trisomy 21 and Translocations of Chromosome 4 (Tumor Cytogenetics/Prenatal Diagnosis/Aneuploidy/Interphase)." *Proc. Nati. Acad. Sci. USA* 85: 9138–42.

Piovesan, Allison, Maria Chiara Pelleri, Francesca Antonaros, Pierluigi Strippoli, Maria Caracausi, and Lorenza Vitale. 2019. "On the Length, Weight and GC Content of the Human Genome." *BMC Research Notes* 12 (1). https://doi.org/10.1186/S13104-019-4137-Z.

Popay, Tessa M., and Jesse R. Dixon. 2022. "Coming Full Circle: On the Origin and Evolution of the Looping Model for Enhancer–Promoter Communication." *Journal of Biological Chemistry* 298 (8): 102117. https://doi.org/10.1016/J.JBC.2022.102117.

Pope, Benjamin D., Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L. Vera, et al. 2014. "Topologically Associating Domains Are Stable Units of Replication-Timing Regulation." *Nature* 515 (7527): 402–5. https://doi.org/10.1038/nature13986.

Powell, George, Helen Long, Louisa Zolkiewski, Rebecca Dumbell, Ann Marie Mallon, Cecilia M. Lindgren, and Michelle M. Simon. 2022. "Modelling the Genetic Aetiology of Complex Disease: Human-Mouse Conservation of Noncoding Features and Disease-Associated Loci." *Biology Letters* 18 (3). https://doi.org/10.1098/RSBL.2021.0630.

Powell, George, Michelle Simon, Sara Pulit, Ann-Marie Mallon, and Cecilia Lindgren. 2019. "Tolerance of Nonsynonymous Variation Is Closely Correlated between Human and Mouse Orthologues." *BioRxiv*, June, 657981. https://doi.org/10.1101/657981.

Pulit, Sara L., Charli Stoneman, Andrew P. Morris, Andrew R. Wood, Craig A. Glastonbury, Jessica Tyrrell, Loïc Yengo, et al. 2019. "Meta-Analysis of Genome-Wide Association Studies for Body Fat Distribution in 694 649 Individuals of European Ancestry." *Human Molecular Genetics* 28 (1): 166–74. https://doi.org/10.1093/hmg/ddy327.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Quinodoz, Sofia A., Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M. Lai, et al. 2018. "Higher-Order Inter-Chromosomal Hubs Shape 3D Genome Organization in the Nucleus." *Cell* 174 (3): 744-757.e24. https://doi.org/10.1016/J.CELL.2018.05.024.

Ramani, Vijay, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. 2017. "Massively Multiplex Single-Cell Hi-C." *Nature Methods* 14 (3): 263–66. https://doi.org/10.1038/nmeth.4155.

Rao, Suhas S.P., Su Chen Huang, Brian Glenn St Hilaire, Jesse M. Engreitz, Elizabeth M. Perez, Kyong Rim Kieffer-Kwon, Adrian L. Sanborn, et al. 2017. "Cohesin Loss Eliminates All Loop Domains." *Cell* 171 (2): 305-320.e24. https://doi.org/10.1016/J.CELL.2017.09.026/ATTACHMENT/9126C43F-E65D-4836-AD17-B34516888EC2/MMC2.PDF.

Rao, Suhas S P, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80. https://doi.org/10.1016/j.cell.2014.11.021.

Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. 2019. “G:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update).” *Nucleic Acids Research* 47 (W1): W191–98. https://doi.org/10.1093/nar/gkz369.

Reimand, Jüri, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. 2007. “G:Profiler—a Web-Based Toolset for Functional Profiling of Gene Lists from Large-Scale Experiments.” *Nucleic Acids Research* 35 (Web Server issue): W193. https://doi.org/10.1093/NAR/GKM226.

Rivera-Mulia, Juan Carlos, Quinton Buckley, Takayo Sasaki, Jared Zimmerman, Ruth A Didier, Kristopher Nazor, Jeanne F Loring, et al. 2015. “Dynamic Changes in Replication Timing and Gene Expression during Lineage Specification of Human Pluripotent Stem Cells.” https://doi.org/10.1101/gr.187989.114.

Rosa-Garrido, Manuel, Douglas J. Chapski, Anthony D. Schmitt, Todd H. Kimball, Elaheh Karbassi, Emma Monte, Enrique Balderas, et al. 2017. “High-Resolution Mapping of Chromatin Conformation in Cardiac Myocytes Reveals Structural Remodeling of the Epigenome in Heart Failure.” *Circulation* 136 (17): 1613–25. https://doi.org/10.1161/CIRCULATIONAHA.117.029430.

Rowley, M. Jordan, and Victor G. Corces. 2016. “The Three-Dimensional Genome: Principles and Roles of Long-Distance Interactions.” *Current Opinion in Cell Biology* 40 (June): 8–14. https://doi.org/10.1016/J.CEB.2016.01.009.

Rowley, M. Jordan, Michael H. Nichols, Xiaowen Lyu, Masami Ando-Kuri, I. Sarahi M. Rivera, Karen Hermetz, Ping Wang, Yijun Ruan, and Victor G. Corces. 2017. “Evolutionarily Conserved Principles Predict 3D Chromatin Organization.” *Molecular Cell* 67 (5): 837. https://doi.org/10.1016/J.MOLCEL.2017.07.022.

Ruiz-Velasco, Mariana, and Judith B. Zaugg. 2017. “Structure Meets Function: How Chromatin Organisation Conveys Functionality.” *Current Opinion in Systems Biology* 1 (February): 129–36. https://doi.org/10.1016/J.COISB.2017.01.003.

Ryba, Tyrone, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C. Schulz, Allan J. Robins, Stephen Dalton, and David M. Gilbert. 2010. “Evolutionarily Conserved Replication Timing Profiles Predict Long-Range Chromatin Interactions and Distinguish Closely Related Cell Types.” *Genome Research* 20 (6): 761–70. https://doi.org/10.1101/GR.099655.109.

Sabari, Benjamin R., Alessandra Dall’Agnese, Ann Boija, Isaac A. Klein, Eliot L. Coffey, Krishna Shrinivas, Brian J. Abraham, et al. 2018. “Coactivator Condensation at Super-Enhancers Links Phase Separation and Gene Control.” *Science* 361 (6400). https://doi.org/10.1126/SCIENCE.AAR3958/SUPPL_FILE/AAR3958_SABARI_SM_TABLE_S3.XLSX.

Sakaue, Saori, Masahiro Kanai, Yosuke Tanigawa, Juha Karjalainen, Mitja Kurki, Seizo Koshiba,

Akira Narita, et al. 2021. “A Cross-Population Atlas of Genetic Associations for 220 Human Phenotypes.” *Nature Genetics 2021 53:10* 53 (10): 1415–24. https://doi.org/10.1038/s41588-021-00931-x.

Saksouk, Nehmé, Elisabeth Simboeck, and Jérôme Déjardin. 2015. “Constitutive Heterochromatin Formation and Transcription in Mammals.” *Epigenetics and Chromatin* 8 (1): 1–17. https://doi.org/10.1186/1756-8935-8-3/FIGURES/4.

Sanborn, Adrian L, Suhas S P Rao, Su-Chen Huang, Neva C Durand, Miriam H Huntley, Andrew I Jewett, Ivan D Bochkov, et al. 2015. “Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (47): E6456-65. https://doi.org/10.1073/pnas.1518552112.

Sarnataro, Sergio, Andrea Riba, and Nacho Molina. 2020. “Regulation of Transcription Reactivation Dynamics Exiting Mitosis.” *BioRxiv*, April, 2020.04.15.042853. https://doi.org/10.1101/2020.04.15.042853.

Sauerwald, Natalie, Akshat Singhal, and Carl Kingsford. 2020. “Analysis of the Structural Variability of Topologically Associated Domains as Revealed by Hi-C.” *NAR Genomics and Bioinformatics* 2 (1). https://doi.org/10.1093/nargab/lqz008.

Sayers, Eric W., Evan E. Bolton, J. Rodney Brister, Kathi Canese, Jessica Chan, Donald C. Comeau, Ryan Connor, et al. 2022. “Database Resources of the National Center for Biotechnology Information.” *Nucleic Acids Research* 50 (D1): D20–26. https://doi.org/10.1093/NAR/GKAB1112.

Schmitt, Anthony D., Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L. Tan, Yun Li, et al. 2016. “A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome.” *Cell Reports* 17 (8): 2042–59. https://doi.org/10.1016/J.CELREP.2016.10.061.

Schoenfelder, Stefan, Mayra Furlan-Magaril, Borbala Mifsud, Filipe Tavares-Cadete, Robert Sugar, Biola Maria Javierre, Takashi Nagano, et al. 2015. “The Pluripotent Regulatory Circuitry Connecting Promoters to Their Long-Range Interacting Elements.” *Genome Research* 25 (4): 582–97. https://doi.org/10.1101/GR.185272.114.

Schoenfelder, Stefan, Biola Maria Javierre, Mayra Furlan-Magaril, Steven W. Wingett, and Peter Fraser. 2018. “Promoter Capture Hi-C: High-Resolution, Genome-Wide Profiling of Promoter Interactions.” *Journal of Visualized Experiments : JoVE* 2018 (136): 57320. https://doi.org/10.3791/57320.

Schwarzer, Wibke, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A. Fonseca, et al. 2017. “Two Independent Modes of Chromatin Organization Revealed by Cohesin Removal.” *Nature 2017 551:7678* 551 (7678): 51–56. https://doi.org/10.1038/nature24281.

Serra, François, Davide Baù, Mike Goodstadt, David Castillo, Guillaume Filion, and Marc A. Marti-Renom. 2017. “Automatic Analysis and 3D-Modelling of Hi-C Data Using TADbit Reveals Structural Features of the Fly Chromatin Colors.” *PLOS Computational Biology* 13 (7): e1005665. https://doi.org/10.1371/JOURNAL.PCBI.1005665.

Sexton, Tom, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. 2012. "Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome." *Cell* 148 (3): 458–72. https://doi.org/10.1016/J.CELL.2012.01.010.

Shen, Yin, Feng Yue, David F. McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, et al. 2012. "A Map of the Cis-Regulatory Sequences in the Mouse Genome." *Nature* 488 (7409): 116–20. https://doi.org/10.1038/nature11243.

Shin, Hanjun, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xianghong Jasmine Zhou. 2016. "TopDom: An Efficient and Deterministic Method for Identifying Topological Domains in Genomes." *Nucleic Acids Research* 44 (7): e70–e70. https://doi.org/10.1093/nar/gkv1505.

Shrinivas, Krishna, Benjamin R. Sabari, Eliot L. Coffey, Isaac A. Klein, Ann Boija, Alicia V. Zamudio, Jurian Schuijers, et al. 2019. "Enhancer Features That Drive Formation of Transcriptional Condensates." *Molecular Cell* 75 (3): 549-561.e7. https://doi.org/10.1016/J.MOLCEL.2019.07.009.

Shungin, D., T. Winkler, D. C. Croteau-Chonka, T. Ferreira, A. E. Locke, R. Mägi, R. J. Strawbridge, et al. 2015. "New Genetic Loci Link Adipose and Insulin Biology to Body Fat Distribution." *Nature* 518 (7538): 187. https://doi.org/10.1038/NATURE14132.

Siersbæk, Rasmus, Jesper Grud Skat Madsen, Biola Maria Javierre, Ronni Nielsen, Emilie Kristine Bagge, Jonathan Cairns, Steven William Wingett, et al. 2017. "Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation." *Molecular Cell* 66 (3): 420-435.e5. https://doi.org/10.1016/j.molcel.2017.04.010.

Sima, Jiao, Abhijit Chakraborty, Vishnu Dileep, Marco Michalski, Kyle N. Klein, Nicolas P. Holcomb, Jesse L. Turner, et al. 2019. "Identifying Cis Elements for Spatiotemporal Control of Mammalian DNA Replication." *Cell* 176 (4): 816-830.e18. https://doi.org/10.1016/J.CELL.2018.11.036.

Simonis, Marieke, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. 2006. "Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture–on-Chip (4C)." *Nature Genetics* 38 (11): 1348–54. https://doi.org/10.1038/ng1896.

Smedley, Damian, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. 2009. "BioMart - Biological Queries Made Easy." *BMC Genomics* 10 (1): 22. https://doi.org/10.1186/1471-2164-10-22.

Sobreira, Débora R., Amelia C. Joslin, Qi Zhang, Iain Williamson, Grace T. Hansen, Kathryn M. Farris, Noboru J. Sakabe, et al. 2021. "Extensive Pleiotropism and Allelic Heterogeneity Mediate Metabolic Effects of *IRX3* and *IRX5*." *Science* 372 (6546): 1085–91. https://doi.org/10.1126/science.abf1008.

Soler-Oliva, María E., José A. Guerrero-Martínez, Valentina Bachetti, and José C. Reyes. 2017. "Analysis of the Relationship between Coexpression Domains and Chromatin 3D Organization." *PLoS Computational Biology* 13 (9). https://doi.org/10.1371/journal.pcbi.1005708.

Solovei, Irina, Katharina Thanisch, and Yana Feodorova. 2016. “How to Rule the Nucleus: Divide et Impera.” *Current Opinion in Cell Biology* 40: 47–59. https://doi.org/10.1016/j.ceb.2016.02.014.

Stadhouders, Ralph, Enrique Vidal, François Serra, Bruno Di Stefano, François Le Dily, Javier Quilez, Antonio Gomez, et al. 2018. “Transcription Factors Orchestrate Dynamic Interplay between Genome Topology and Gene Regulation during Cell Reprogramming.” *Nature Genetics* 50 (2): 238–49. https://doi.org/10.1038/s41588-017-0030-7.

Stamboulian, Moses, Rafael F. Guerrero, Matthew W. Hahn, and Predrag Radivojac. 2020. “The Ortholog Conjecture Revisited: The Value of Orthologs and Paralogs in Function Prediction.” *Bioinformatics* 36 (Suppl 1): I219–26. https://doi.org/10.1093/BIOINFORMATICS/BTAA468.

Strom, Amy R., Alexander V. Emelyanov, Mustafa Mir, Dmitry V. Fyodorov, Xavier Darzacq, and Gary H. Karpen. 2017. “Phase Separation Drives Heterochromatin Domain Formation.” *Nature 2017 547:7662* 547 (7662): 241–45. https://doi.org/10.1038/nature22989.

Sun, Quan, Cheynna A. Crowley, Le Huang, Jia Wen, Jiawen Chen, Erik L. Bao, Paul L. Auer, et al. 2022. “From GWAS Variant to Function: A Study of ∼148,000 Variants for Blood Cell Traits.” *Human Genetics and Genomics Advances* 3 (1): 100063. https://doi.org/10.1016/J.XHGG.2021.100063.

Symmons, Orsolya, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nassari, Wibke Schwarzer, Laurence Ettwiller, and François Spitz. 2014. “Functional and Topological Characteristics of Mammalian Regulatory Domains.” *Genome Research* 24 (3): 390–400. https://doi.org/10.1101/gr.163519.113.

Szklarczyk, Damian, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, et al. 2019. “STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets.” *Nucleic Acids Research* 47 (D1): D607–13. https://doi.org/10.1093/nar/gky1131.

Tanabe, Hideyuki, Felix A. Habermann, Irina Solovei, Marion Cremer, and Thomas Cremer. 2002. “Non-Random Radial Arrangements of Interphase Chromosome Territories: Evolutionary Considerations and Functional Implications.” *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 504 (1–2): 37–45. https://doi.org/10.1016/S0027-5107(02)00077-5.

Tarbier, Marcel, Sebastian D. Mackowiak, João Frade, Silvina Catuara-Solarz, Inna Biryukova, Eleni Gelali, Diego Bárcena Menéndez, et al. 2020. “Nuclear Gene Proximity and Protein Interactions Shape Transcript Covariations in Mammalian Single Cells.” *Nature Communications 2020 11:1* 11 (1): 1–12. https://doi.org/10.1038/s41467-020-19011-5.

Tena, Juan J., and José M. Santos-Pereira. 2021. “Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease.” *Frontiers in Cell and Developmental Biology* 0 (July): 1817. https://doi.org/10.3389/FCELL.2021.702787.

The Gene Ontology Consortium. 2018. “The Gene Ontology Resource: 20 Years and Still GOing Strong.” *Nucleic Acids Research* 47 (D1): D330–38.

https://doi.org/10.1093/nar/gky1055.

Thévenin, Annelyse, Liat Ein-Dor, Michal Ozery-Flato, and Ron Shamir. 2014. “Functional Gene Groups Are Concentrated within Chromosomes, among Chromosomes and in the Nuclear Space of the Human Genome.” *Nucleic Acids Research* 42 (15): 9854–61. https://doi.org/10.1093/nar/gku667.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. “Genome-Wide Association Studies.” *Nature Reviews Methods Primers 2021 1:1* 1 (1): 1–21. https://doi.org/10.1038/s43586-021-00056-9.

Ulianov, Sergey V., Ekaterina E. Khrameeva, Alexey A. Gavrilov, Ilya M. Flyamer, Pavel Kos, Elena A. Mikhaleva, Aleksey A. Penin, et al. 2016. “Active Chromatin and Transcription Play a Key Role in Chromosome Partitioning into Topologically Associating Domains.” *Genome Research* 26 (1): 70–84. https://doi.org/10.1101/gr.196006.115.

Ullah, Imran, Raghavendra Baregundi Subbarao, and Gyu Jin Rho. 2015. “Human Mesenchymal Stem Cells - Current Trends and Future Prospective.” *Bioscience Reports* 35 (2): 191. https://doi.org/10.1042/BSR20150025.

Vermunt, Marit W, Di Zhang, and Gerd A Blobel. 2018. “The Interdependence of Gene-Regulatory Elements and the 3D Genome.” *J. Cell Biol* 218 (1): 12–26. https://doi.org/10.1083/jcb.201809040.

Vian, Laura, Aleksandra Pękowska, Suhas S.P. Rao, Kyong Rim Kieffer-Kwon, Seolkyoung Jung, Laura Baranello, Su Chen Huang, et al. 2018. “The Energetics and Physiological Impact of Cohesin Extrusion.” *Cell* 173 (5): 1165. https://doi.org/10.1016/J.CELL.2018.03.072.

Vickaryous, Matthew K., and Brian K. Hall. 2006. “Human Cell Type Diversity, Evolution, Development, and Classification with Special Reference to Cells Derived from the Neural Crest.” *Biological Reviews* 81 (3): 425–55. https://doi.org/10.1017/S1464793106007068.

Vietri Rudan, Matteo, Christopher Barrington, Stephen Henderson, Christina Ernst, Duncan T. Odom, Amos Tanay, and Suzana Hadjur. 2015. “Comparative Hi-C Reveals That CTCF Underlies Evolution of Chromosomal Domain Architecture.” *Cell Reports* 10 (8): 1297–1309. https://doi.org/10.1016/j.celrep.2015.02.004.

Wang, Yanli, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, et al. 2018. “The 3D Genome Browser: A Web-Based Browser for Visualizing 3D Genome Organization and Long-Range Chromatin Interactions.” *Genome Biology* 19 (1): 1–12. https://doi.org/10.1186/S13059-018-1519-9/FIGURES/5.

Way, Gregory P, Daniel W Youngstrom, Kurt D Hankenson, Casey S Greene, and Struan FA Grant. 2017. “Implicating Candidate Genes at GWAS Signals by Leveraging Topologically Associating Domains.” *European Journal of Human Genetics* 25 (11): 1286–89. https://doi.org/10.1038/ejhg.2017.108.

Wendt, Kerstin S., Keisuke Yoshida, Takehiko Itoh, Masashige Bando, Birgit Koch, Erika Schirghuber, Shuichi Tsutsumi, et al. 2008. “Cohesin Mediates Transcriptional Insulation by CCCTC-Binding Factor.” *Nature 2008 451:7180* 451 (7180): 796–801. https://doi.org/10.1038/nature06634.

Whalen, Sean, and Katherine S. Pollard. 2019. “Most Chromatin Interactions Are Not in Linkage Disequilibrium.” *Genome Research* 29 (3): 334–43. https://doi.org/10.1101/GR.238022.118/-/DC1.

Wingett, Steven, Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews. 2015. “HiCUP: Pipeline for Mapping and Processing Hi-C Data.” *F1000Research* 4: 1310. https://doi.org/10.12688/f1000research.7334.1.

Wit, Elzo de. 2020. “TADs as the Caller Calls Them.” *Journal of Molecular Biology* 432 (3): 638–42. https://doi.org/10.1016/J.JMB.2019.09.026.

Wit, Elzo de, and Wouter de Laat. 2012. “A Decade of 3C Technologies: Insights into Nuclear Organization.” *Genes & Development* 26 (1): 11. https://doi.org/10.1101/GAD.179804.111.

Wittkopp, Patricia J., and Gizem Kalay. 2011. “Cis-Regulatory Elements: Molecular Mechanisms and Evolutionary Processes Underlying Divergence.” *Nature Reviews Genetics 2011 13:1* 13 (1): 59–69. https://doi.org/10.1038/nrg3095.

Woodcock, C. L., and R. A. Horowitz. 1997. “Electron Microscopy of Chromatin.” *Methods* 12 (1): 84–95. https://doi.org/10.1006/METH.1997.0450.

Wu, Hua-Jun, Alexandro Landshammer, Elena K. Stamenova, Adriano Bolondi, Helene Kretzmer, Alexander Meissner, and Franziska Michor. 2021. “Topological Isolation of Developmental Regulators in Mammalian Genomes.” *Nature Communications 2021 12:1* 12 (1): 1–19. https://doi.org/10.1038/s41467-021-24951-7.

Wutz, Gordana, Csilla Várnai, Kota Nagasaka, David A Cisneros, Roman R Stocsits, Wen Tang, Stefan Schoenfelder, et al. 2017. “Topologically Associating Domains and Chromatin Loops Depend on Cohesin and Are Regulated by CTCF, WAPL, and PDS5 Proteins.” *The EMBO Journal* 36 (24): 3573–99. https://doi.org/10.15252/EMBJ.201798004.

Xi, Wang, and Michael A. Beer. 2021. “Loop Competition and Extrusion Model Predicts CTCF Interaction Specificity.” *Nature Communications 2021 12:1* 12 (1): 1–15. https://doi.org/10.1038/s41467-021-21368-0.

Yaffe, Eitan, Shlomit Farkash-Amar, Andreas Polten, Zohar Yakhini, Amos Tanay, and Itamar Simon. 2010. “Comparative Analysis of DNA Replication Timing Reveals Conserved Large-Scale Chromosomal Architecture.” *PLOS Genetics* 6 (7): e1001011. https://doi.org/10.1371/JOURNAL.PGEN.1001011.

Yaffe, Eitan, and Amos Tanay. 2011. “Probabilistic Modeling of Hi-C Contact Maps Eliminates Systematic Biases to Characterize Global Chromosomal Architecture.” *Nature Genetics* 43 (11): 1059–65. https://doi.org/10.1038/ng.947.

Yan, Jian, Shi-An A Chen, Andrea Local, Tristin Liu, Yunjiang Qiu, Kristel M Dorighi, Sebastian Preissl, et al. 2018. “Histone H3 Lysine 4 Monomethylation Modulates Long-Range Chromatin Interactions at Enhancers.” *Cell Research* 28 (2): 204–20. https://doi.org/10.1038/cr.2018.1.

Yan, Koon Kiu, Shaoke Lou, and Mark Gerstein. 2017. “MrTADFinder: A Network Modularity Based Approach to Identify Topologically Associating Domains in Multiple Resolutions.” *PLOS Computational Biology* 13 (7): e1005647.

https://doi.org/10.1371/JOURNAL.PCBI.1005647.

Yang, Dongchan, Insu Jang, Jinhyuk Choi, Min Seo Kim, Andrew J. Lee, Hyunwoong Kim, Junghyun Eom, Dongsup Kim, Inkyung Jung, and Byungwook Lee. 2018. "3DIV: A 3D-Genome Interaction Viewer and Database." *Nucleic Acids Research* 46 (D1): D52–57. https://doi.org/10.1093/NAR/GKX1017.

Yu, G., F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang. 2010. "GOSemSim: An R Package for Measuring Semantic Similarity among GO Terms and Gene Products." *Bioinformatics* 26 (7): 976–78. https://doi.org/10.1093/bioinformatics/btq064.

Yu, Miao, and Bing Ren. 2017. "The Three-Dimensional Organization of Mammalian Genomes." *Annual Review of Cell and Developmental Biology* 33 (1): 265–89. https://doi.org/10.1146/annurev-cellbio-100616-060531.

Zhang, Xiao, Kenneth C. Ehrlich, Fangtang Yu, Xiaojun Hu, Xiang He Meng, Hong Wen Deng, Hui Shen, and Melanie Ehrlich. 2020. "Osteoporosis- and Obesity-Risk Interrelationships: An Epigenetic Analysis of GWAS-Derived SNPs at the Developmental Gene TBX15." *Epigenetics* 15 (6–7): 728–49. https://doi.org/10.1080/15592294.2020.1716491.

Zhang, Yanxiao, Ting Li, Sebastian Preissl, Maria Luisa Amaral, Jonathan D. Grinstein, Elie N. Farah, Eugin Destici, et al. 2019. "Transcriptionally Active HERV-H Retrotransposons Demarcate Topologically Associating Domains in Human Pluripotent Stem Cells." *Nature Genetics* 51 (9): 1380–88. https://doi.org/10.1038/S41588-019-0479-7.

Zhu, Zhaozhong, Yanjun Guo, Huwenbo Shi, Cong Lin Liu, Ronald Allan Panganiban, Wonil Chung, Luke J. O'Connor, et al. 2020. "Shared Genetic and Experimental Links between Obesity-Related Traits and Asthma Subtypes in UK Biobank." *Journal of Allergy and Clinical Immunology* 145 (2): 537–49. https://doi.org/10.1016/J.JACI.2019.09.035/ATTACHMENT/747D2831-2D78-4023-ABD6-A93A085394CD/MMC22.PDF.

Zufferey, Marie, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. 2018. "Comparison of Computational Methods for the Identification of Topologically Associating Domains." *Genome Biology* 19 (1): 217. https://doi.org/10.1186/s13059-018-1596-9.

Agnew, Thomas, Michelle Goldsworthy, Carlos Aguilar, Anna Morgan, Michelle Simon, Helen Hilton, Chris Esapa, et al. 2018. "A Wars2 Mutant Mouse Model Displays OXPHOS Deficiencies and Activation of Tissue-Specific Stress Response Pathways." *Cell Reports* 25 (12): 3315-3328.e6. https://doi.org/10.1016/J.CELREP.2018.11.080.

Akdemir, Kadir C., Victoria T. Le, Sahaana Chandran, Yilong Li, Roel G. Verhaak, Rameen Beroukhim, Peter J. Campbell, et al. 2020. "Disruption of Chromatin Folding Domains by Somatic Genomic Rearrangements in Human Cancer." *Nature Genetics 2020 52:3* 52 (3): 294–305. https://doi.org/10.1038/s41588-019-0564-y.

Akiyama, Masato, Yukinori Okada, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, et al. 2017. "Genome-Wide Association Study Identifies 112 New Loci for Body Mass Index in the Japanese Population." *Nature Genetics 2017 49:10* 49 (10): 1458–67. https://doi.org/10.1038/ng.3951.

Andersson, Robin, and Albin Sandelin. 2019. "Determinants of Enhancer and Promoter

Activities of Regulatory Elements." *Nature Reviews Genetics 2019 21:2* 21 (2): 71–87. https://doi.org/10.1038/s41576-019-0173-8.

Ardlie, Kristin G., David S. DeLuca, Ayellet V. Segrè, Timothy J. Sullivan, Taylor R. Young, Ellen T. Gelfand, Casandra A. Trowbridge, et al. 2015. "Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." *Science (New York, N.Y.)* 348 (6235): 648–60. https://doi.org/10.1126/SCIENCE.1262110.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics*. NIH Public Access. https://doi.org/10.1038/75556.

Athar, Awais, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, et al. 2019. "ArrayExpress Update - from Bulk to Single-Cell Expression Data." *Nucleic Acids Research* 47 (D1): D711–15. https://doi.org/10.1093/NAR/GKY964.

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2012. "NCBI GEO: Archive for Functional Genomics Data Sets—Update." *Nucleic Acids Research* 41 (D1): D991–95. https://doi.org/10.1093/nar/gks1193.

Beagan, Jonathan A., and Jennifer E. Phillips-Cremins. 2020. "On the Existence and Functionality of Topologically Associating Domains." *Nature Genetics*, January. https://doi.org/10.1038/s41588-019-0561-1.

Beagrie, Robert A., Antonio Scialdone, Markus Schueler, Dorothee C.A. Kraemer, Mita Chotalia, Sheila Q. Xie, Mariano Barbieri, et al. 2017. "Complex Multi-Enhancer Contacts Captured by Genome Architecture Mapping (GAM)." *Nature* 543 (7646): 519. https://doi.org/10.1038/NATURE21411.

Belmont, Andrew S. 2002. "Mitotic Chromosome Scaffold Structure: New Approaches to an Old Controversy." *Proceedings of the National Academy of Sciences of the United States of America* 99 (25): 15855. https://doi.org/10.1073/PNAS.262672799.

Benabdallah, Nezha S., Iain Williamson, Robert S. Illingworth, Lauren Kane, Shelagh Boyle, Dipta Sengupta, Graeme R. Grimes, Pierre Therizols, and Wendy A. Bickmore. 2019. "Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation." *Molecular Cell* 76 (3): 473-484.e7. https://doi.org/10.1016/j.molcel.2019.07.038.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X.

Bertero, Alessandro, Paul A. Fields, Vijay Ramani, Giancarlo Bonora, Galip G. Yardimci, Hans Reinecke, Lil Pabon, William S. Noble, Jay Shendure, and Charles E. Murry. 2019. "Dynamics of Genome Reorganization during Human Cardiogenesis Reveal an RBM20-Dependent Splicing Factory." *Nature Communications* 10 (1): 1538. https://doi.org/10.1038/s41467-019-09483-5.

Berthelot, Camille, Matthieu Muffato, Judith Abecassis, and Hugues Roest Crollius. 2015. "The 3D Organization of Chromatin Explains Evolutionary Fragile Genomic Regions." *Cell*

Reports 10 (11): 1913–24. https://doi.org/10.1016/J.CELREP.2015.02.046.

Bianconi, Eva, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, et al. 2013. "An Estimation of the Number of Cells in the Human Body." *Https://Doi.Org/10.3109/03014460.2013.807878* 40 (6): 463–71. https://doi.org/10.3109/03014460.2013.807878.

Bonev, Boyan, and Giacomo Cavalli. 2016. "Organization and Function of the 3D Genome." *Nature Reviews Genetics* 17 (11): 661–78. https://doi.org/10.1038/nrg.2016.112.

Bonev, Boyan, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L Papadopoulos, Yaniv Lubling, Xiaole Xu, et al. 2017. "Multiscale 3D Genome Rewiring during Mouse Neural Development." *Cell* 171 (3): 557-572.e24. https://doi.org/10.1016/j.cell.2017.09.043.

Boyle, Alan P., Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, et al. 2012. "Annotation of Functional Variation in Personal Genomes Using RegulomeDB." *Genome Research* 22 (9): 1790. https://doi.org/10.1101/GR.137323.112.

Boyle, Shelagh, Susan Gilchrist, Joanna M. Bridger, Nicola L. Mahy, Juliet A. Ellis, and Wendy A. Bickmore. 2001. "The Spatial Organization of Human Chromosomes within the Nuclei of Normal and Emerin-Mutant Cells." *Human Molecular Genetics* 10 (3): 211–20. https://doi.org/10.1093/HMG/10.3.211.

Bulger, Michael, and Mark Groudine. 1999. "Looping versus Linking: Toward a Model for Long-Distance Gene Activation." *Genes and Development* 13: 2465–77. http://genesdev.cshlp.org/content/13/19/2465.extract.

Buniello, Annalisa, Jacqueline A.L. Macarthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, 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): D1005–12. https://doi.org/10.1093/NAR/GKY1120.

Cairns, Jonathan, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, et al. 2016. "CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C Data." *Genome Biology* 17 (1): 1–17. https://doi.org/10.1186/S13059-016-0992-2/FIGURES/10.

Calderon, Lesly, Felix D Weiss, Jonathan A Beagan, Marta S Oliveira, Radina Georgieva, Yi-Fang Wang, Thomas S Carroll, et al. 2022. "Cohesin-Dependence of Neuronal Gene Expression Relates to Chromatin Loop Length." *ELife* 11 (April). https://doi.org/10.7554/ELIFE.76539.

Cano-Gamez, Eddie, and Gosia Trynka. 2020. "From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases." *Frontiers in Genetics* 11 (May): 424. https://doi.org/10.3389/FGENE.2020.00424/BIBTEX.

Carbon, Seth, Amelia Ireland, Christopher J Mungall, Shengqiang Shu, Brad Marshall, Suzanna Lewis, Amigo Hub, Web Presence, and Working Group. 2009. "AmiGO: Online Access to Ontology and Annotation Data." *BIOINFORMATICS APPLICATIONS NOTE* 25 (2): 288–89.

https://doi.org/10.1093/bioinformatics/btn615.

Carlson, Marc. 2017. “Org.Mm.Eg.Db: Genome Wide Annotation for Mouse. R Package Version 3.5.0.” *Bioconductor*. https://doi.org/10.18129/B9.bioc.org.Mm.eg.db.

Chang, Li Hsin, Sourav Ghosh, and Daan Noordermeer. 2020. “TADs and Their Borders: Free Movement or Building a Wall?” *Journal of Molecular Biology*. Academic Press. https://doi.org/10.1016/j.jmb.2019.11.025.

Cho, Won Ki, Jan Hendrik Spille, Micca Hecht, Choongman Lee, Charles Li, Valentin Grube, and Ibrahim I. Cisse. 2018. “Mediator and RNA Polymerase II Clusters Associate in Transcription-Dependent Condensates.” *Science* 361 (6400): 412–15. https://doi.org/10.1126/SCIENCE.AAR4199/SUPPL_FILE/AAR4199_S3.MOV.

Ciosk, Rafal, Masaki Shirayama, Anna Shevchenko, Tomoyuki Tanaka, Attila Toth, Andrej Shevchenko, and Kim Nasmyth. 2000. “Cohesin’s Binding to Chromosomes Depends on a Separate Complex Consisting of Scc2 and Scc4 Proteins.” *Molecular Cell* 5 (2): 243–54. https://doi.org/10.1016/S1097-2765(00)80420-7.

Crane, Emily, Qian Bian, Rachel Patton McCord, Bryan R. Lajoie, Bayly S. Wheeler, Edward J. Ralston, Satoru Uzawa, Job Dekker, and Barbara J. Meyer. 2015. “Condensin-Driven Remodeling of X-Chromosome Topology during Dosage Compensation.” *Nature* 523 (7559): 240. https://doi.org/10.1038/NATURE14450.

Cremer, T., C. Cremer, H. Baumann, E. K. Luedtke, K. Sperling, V. Teuber, and C. Zorn. 1982. “Rabl’s Model of the Interphase Chromosome Arrangement Tested in Chinise Hamster Cells by Premature Chromosome Condensation and Laser-UV-Microbeam Experiments.” *Human Genetics* 60 (1): 46–56. https://doi.org/10.1007/BF00281263/METRICS.

Cremer, T., C. Cremer, T. Schneider, H. Baumann, L. Hens, and M. Kirsch-Volders. 1982. “Analysis of Chromosome Positions in the Interphase Nucleus of Chinese Hamster Cells by Laser-UV-Microirradiation Experiments.” *Human Genetics* 62 (3): 201–9. https://doi.org/10.1007/BF00333519.

Cuartero, Sergi, Felix D. Weiss, Gopuraja Dharmalingam, Ya Guo, Elizabeth Ing-Simmons, Silvia Masella, Irene Robles-Rebollo, et al. 2018. “Control of Inducible Gene Expression Links Cohesin to Hematopoietic Progenitor Self-Renewal and Differentiation.” *Nature Immunology 2018 19:9* 19 (9): 932–41. https://doi.org/10.1038/s41590-018-0184-1.

Dai, Haijiang, Tariq A. Alsalhe, Nasr Chalghaf, Matteo Riccò, Nicola Luigi Bragazzi, and Jianhong Wu. 2020. “The Global Burden of Disease Attributable to High Body Mass Index in 195 Countries and Territories, 1990–2017: An Analysis of the Global Burden of Disease Study.” *PLOS Medicine* 17 (7): e1003198. https://doi.org/10.1371/JOURNAL.PMED.1003198.

Dali, Rola, and Mathieu Blanchette. 2017. “A Critical Assessment of Topologically Associating Domain Prediction Tools.” *Nucleic Acids Research* 45 (6): 2994–3005. https://doi.org/10.1093/nar/gkx145.

Davis, Carrie A., Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, et al. 2018. “The Encyclopedia of DNA Elements (ENCODE): Data Portal Update.” *Nucleic Acids Research* 46 (D1): D794–801.

https://doi.org/10.1093/nar/gkx1081.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. “Capturing Chromosome Conformation.” *Science (New York, N.Y.)* 295 (5558): 1306–11. https://doi.org/10.1126/science.1067799.

Deng, Wulan, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, Philip D. Gregory, Ann Dean, and Gerd A. Blobel. 2012. “Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor.” *Cell* 149 (6): 1233–44. https://doi.org/10.1016/J.CELL.2012.03.051.

Deng, Wulan, Jeremy W. Rupon, Ivan Krivega, Laura Breda, Irene Motta, Kristen S. Jahn, Andreas Reik, et al. 2014. “Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping.” *Cell* 158 (4): 849–60. https://doi.org/10.1016/J.CELL.2014.05.050.

Dickinson, Mary E., Ann M. Flenniken, Xiao Ji, Lydia Teboul, Michael D. Wong, Jacqueline K. White, Terrence F. Meehan, et al. 2016. “High-Throughput Discovery of Novel Developmental Phenotypes.” *Nature* 537 (7621): 508–14. https://doi.org/10.1038/nature19356.

Dixon, Jesse R., Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E. Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, et al. 2015. “Chromatin Architecture Reorganization during Stem Cell Differentiation.” *Nature* 518 (7539): 331–36. https://doi.org/10.1038/nature14222.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485 (7398): 376–80. https://doi.org/10.1038/nature11082.

Dixon, Jesse R, David U Gorkin, and Bing Ren. 2016. “Chromatin Domains: The Unit of Chromosome Organization.” *Molecular Cell* 62: 668–80. https://doi.org/10.1016/j.molcel.2016.05.018.

Dostie, Josée, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, et al. 2006. “Chromosome Conformation Capture Carbon Copy (5C): A Massively Parallel Solution for Mapping Interactions between Genomic Elements.” *Genome Research* 16 (10): 1299–1309. https://doi.org/10.1101/gr.5571506.

Dryden, Nicola H., Laura R. Broome, Frank Dudbridge, Nichola Johnson, Nick Orr, Stefan Schoenfelder, Takashi Nagano, et al. 2014. “Unbiased Analysis of Potential Targets of Breast Cancer Susceptibility Loci by Capture Hi-C.” *Genome Research* 24 (11): 1854–68. https://doi.org/10.1101/GR.175034.114.

Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S.P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. “Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments.” *Cell Systems* 3 (1): 95–98. https://doi.org/10.1016/J.CELS.2016.07.002.

Durand, Neva C, James T Robinson, Muhammad S Shamim, Ido Machol, Jill P Mesirov, Eric S Lander, and Erez Lieberman Aiden. 2016. “Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom.” *Cell Systems* 3 (1): 99–101.

https://doi.org/10.1016/j.cels.2015.07.012.

Durinck, Steffen, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. 2009. “Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package BiomaRt.” *Nature Protocols* 4 (8): 1184–91. https://doi.org/10.1038/NPROT.2009.97.

Dynan, William S., and Robert Tjian. 1985. “Control of Eukaryotic Messenger RNA Synthesis by Sequence-Specific DNA-Binding Proteins.” *Nature 1985 316:6031* 316 (6031): 774–78. https://doi.org/10.1038/316774a0.

Edgar, Ron, Michael Domrachev, and Alex E Lash. 2002. “Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository.” *Nucleic Acids Research* 30 (1): 207–10. http://www.ncbi.nlm.nih.gov/pubmed/11752295.

Eres, Ittai E., and Yoav Gilad. 2021. “A TAD Skeptic: Is 3D Genome Topology Conserved?” *Trends in Genetics*. Elsevier Ltd. https://doi.org/10.1016/j.tig.2020.10.009.

Eres, Ittai E., Kaixuan Luo, Chiaowen Joyce Hsiao, Lauren E. Blake, and Yoav Gilad. 2019. “Reorganization of 3D Genome Structure May Contribute to Gene Regulatory Evolution in Primates.” Edited by Harmit S. Malik. *PLOS Genetics* 15 (7): e1008278. https://doi.org/10.1371/journal.pgen.1008278.

Ernst, Jason, and Manolis Kellis. 2010. “Discovery and Characterization of Chromatin States for Systematic Annotation of the Human Genome.” *Nature Biotechnology 2010 28:8* 28 (8): 817–25. https://doi.org/10.1038/nbt.1662.

———. 2017. “Chromatin-State Discovery and Genome Annotation with ChromHMM.” *Nature Protocols 2017 12:12* 12 (12): 2478–92. https://doi.org/10.1038/nprot.2017.124.

Ernst, Jason, Pouya Kheradpour, Tarjei S. Mikkelsen, Noam Shoresh, Lucas D. Ward, Charles B. Epstein, Xiaolan Zhang, et al. 2011. “Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types.” *Nature 2011 473:7345* 473 (7345): 43–49. https://doi.org/10.1038/nature09906.

Feric, Marina, Nilesh Vaidya, Tyler S. Harmon, Diana M. Mitrea, Lian Zhu, Tiffany M. Richardson, Richard W. Kriwacki, Rohit V. Pappu, and Clifford P. Brangwynne. 2016. “Coexisting Liquid Phases Underlie Nucleolar Subcompartments.” *Cell* 165 (7): 1686–97. https://doi.org/10.1016/J.CELL.2016.04.047.

Finucane, Hilary K., Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, et al. 2018. “Heritability Enrichment of Specifically Expressed Genes Identifies Disease-Relevant Tissues and Cell Types.” *Nature Genetics 2018 50:4* 50 (4): 621–29. https://doi.org/10.1038/s41588-018-0081-4.

Fiorillo, Luca, Francesco Musella, Mattia Conte, Rieke Kempfer, Andrea M. Chiariello, Simona Bianco, Alexander Kukalev, et al. 2021. “Comparison of the Hi-C, GAM and SPRITE Methods Using Polymer Models of Chromatin.” *Nature Methods* 18 (5): 482. https://doi.org/10.1038/S41592-021-01135-1.

Flavahan, William A., Yotam Drier, Brian B. Liau, Shawn M. Gillespie, Andrew S. Venteicher, Anat O. Stemmer-Rachamimov, Mario L. Suvà, and Bradley E. Bernstein. 2016. “Insulator Dysfunction and Oncogene Activation in IDH Mutant Gliomas.” *Nature* 529 (7584): 110–14. https://doi.org/10.1038/nature16490.

Flyamer, Ilya M., Johanna Gassler, Maxim Imakaev, Hugo B. Brandão, Sergey V. Ulianov, Nezar Abdennur, Sergey V. Razin, Leonid A. Mirny, and Kikuë Tachibana-Konwalski. 2017. "Single-Nucleus Hi-C Reveals Unique Chromatin Reorganization at Oocyte-to-Zygote Transition." *Nature* 544 (7648): 110–14. https://doi.org/10.1038/nature21711.

Forcato, Mattia, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. 2017. "Comparison of Computational Methods for Hi-C Data Analysis." *Nature Methods* 14 (7). https://doi.org/10.1038/nmeth.4325.

Franke, Martin, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, et al. 2016. "Formation of New Chromatin Domains Determines Pathogenicity of Genomic Duplications." *Nature* 538 (7624): 265–69. https://doi.org/10.1038/nature19800.

Fraser, J., C. Ferrai, A. M. Chiariello, M. Schueler, T. Rito, G. Laudanno, M. Barbieri, et al. 2015. "Hierarchical Folding and Reorganization of Chromosomes Are Linked to Transcriptional Changes in Cellular Differentiation." *Molecular Systems Biology* 11 (12): 852–852. https://doi.org/10.15252/msb.20156492.

Freire-Pritchett, Paula, Helen Ray-Jones, Monica Della Rosa, Chris Q. Eijsbouts, William R. Orchard, Steven W. Wingett, Chris Wallace, Jonathan Cairns, Mikhail Spivakov, and Valeriya Malysheva. 2021. "Detecting Chromosomal Interactions in Capture Hi-C Data with CHiCAGO and Companion Tools." *Nature Protocols 2021 16:9* 16 (9): 4144–76. https://doi.org/10.1038/s41596-021-00567-5.

Freire-Pritchett, Paula, Stefan Schoenfelder, Csilla Várnai, Steven W Wingett, Jonathan Cairns, Amanda J Collier, Raquel García-Vílchez, et al. 2017. "Global Reorganisation of Cis-Regulatory Units upon Lineage Commitment of Human Embryonic Stem Cells." *ELife* 6 (March). https://doi.org/10.7554/eLife.21926.

Fudenberg, Geoffrey, Nezar Abdennur, Maxim Imakaev, Anton Goloborodko, and Leonid A. Mirny. 2017. "Emerging Evidence of Chromosome Folding by Loop Extrusion." *Cold Spring Harbor Symposia on Quantitative Biology* 82 (January): 45–55. https://doi.org/10.1101/SQB.2017.82.034710/-/DC1.

Fudenberg, Geoffrey, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A. Mirny. 2016. "Formation of Chromosomal Domains by Loop Extrusion." *Cell Reports* 15 (9): 2038–49. https://doi.org/10.1016/J.CELREP.2016.04.085.

Furlong, Eileen E.M., and Michael Levine. 2018. "Developmental Enhancers and Chromosome Topology." *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.aau0320.

Gandhi, Rita, Peter J. Gillespie, and Tatsuya Hirano. 2006. "Human Wapl Is a Cohesin-Binding Protein That Promotes Sister-Chromatid Resolution in Mitotic Prophase." *Current Biology : CB* 16 (24): 2406–17. https://doi.org/10.1016/J.CUB.2006.10.061.

Gassler, Johanna, Hugo B Brandão, Maxim Imakaev, Ilya M Flyamer, Sabrina Ladstätter, Wendy A Bickmore, Jan-Michael Peters, Leonid A Mirny, and Kikuë Tachibana. 2017. "A Mechanism of Cohesin-Dependent Loop Extrusion Organizes Zygotic Genome Architecture." *The EMBO Journal* 36 (24): 3600–3618. https://doi.org/10.15252/EMBJ.201798083.

Ghavi-Helm, Yad, Aleksander Jankowski, Sascha Meiers, Rebecca R. Viales, Jan O. Korbel, and Eileen E. M. Furlong. 2019. “Highly Rearranged Chromosomes Reveal Uncoupling between Genome Topology and Gene Expression.” *Nature Genetics* 51 (8): 1272–82. https://doi.org/10.1038/s41588-019-0462-3.

Ghavi-Helm, Yad, Felix A. Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen E.M. Furlong. 2014. “Enhancer Loops Appear Stable during Development and Are Associated with Paused Polymerase.” *Nature* 512 (7512): 96–100. https://doi.org/10.1038/NATURE13417.

Gilbertson, Sarah E., Hannah C. Walter, Katherine Gardner, Spencer N. Wren, Golnaz Vahedi, and Amy S. Weinmann. 2022. “Topologically Associating Domains Are Disrupted by Evolutionary Genome Rearrangements Forming Species-Specific Enhancer Connections in Mice and Humans.” *Cell Reports* 39 (5). https://doi.org/10.1016/J.CELREP.2022.110769.

Giorgetti, Luca, Bryan R. Lajoie, Ava C. Carter, Mikael Attia, Ye Zhan, Jin Xu, Chong Jian Chen, et al. 2016. “Structural Organization of the Inactive X Chromosome in the Mouse.” *Nature* 535 (7613): 575–79. https://doi.org/10.1038/nature18589.

Giorgio, Elisa, Daniel Robyr, Malte Spielmann, Enza Ferrero, Eleonora Di Gregorio, Daniele Imperiale, Giovanna Vaula, et al. 2015. “A Large Genomic Deletion Leads to Enhancer Adoption by the Lamin B1 Gene: A Second Path to Autosomal Dominant Adult-Onset Demyelinating Leukodystrophy (ADLD).” *Human Molecular Genetics* 24 (11): 3143–54. https://doi.org/10.1093/HMG/DDV065.

Haarhuis, Judith H.I., Robin H. van der Weide, Vincent A. Blomen, J. Omar Yáñez-Cuna, Mario Amendola, Marjon S. van Ruiten, Peter H.L. Krijger, et al. 2017. “The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension.” *Cell* 169 (4): 693-707.e14. https://doi.org/10.1016/J.CELL.2017.04.013/ATTACHMENT/A7AB8846-8FD4-40D7-8DF0-E55B34A296EC/MMC1.XLSX.

Haberle, Vanja, and Alexander Stark. 2018. “Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation.” *Nature Reviews Molecular Cell Biology 2018 19:10* 19 (10): 621–37. https://doi.org/10.1038/s41580-018-0028-8.

Haddad, Noelle, Cédric Vaillant, and Daniel Jost. 2017. “IC-Finder: Inferring Robustly the Hierarchical Organization of Chromatin Folding.” *Nucleic Acids Research* 45 (10): e81–e81. https://doi.org/10.1093/NAR/GKX036.

Hansen, Anders S. 2020. “CTCF as a Boundary Factor for Cohesin-Mediated Loop Extrusion: Evidence for a Multi-Step Mechanism.” *Nucleus* 11 (1): 132. https://doi.org/10.1080/19491034.2020.1782024.

Hansen, Anders S., Iryna Pustova, Claudia Cattoglio, Robert Tjian, and Xavier Darzacq. 2017. “CTCF and Cohesin Regulate Chromatin Loop Stability with Distinct Dynamics.” *ELife* 6 (May). https://doi.org/10.7554/ELIFE.25776.001.

Hanssen, Lars L.P., Mira T. Kassouf, A. Marieke Oudelaar, Daniel Biggs, Chris Preece, Damien J. Downes, Matthew Gosden, et al. 2017. “Tissue-Specific CTCF–Cohesin-Mediated Chromatin Architecture Delimits Enhancer Interactions and Function in Vivo.” *Nature Cell Biology 2017 19:8* 19 (8): 952–61. https://doi.org/10.1038/ncb3573.

He, Mengnan, Yan Li, Qianzi Tang, Diyan Li, Long Jin, Shilin Tian, Tiandong Che, et al. 2018. "Genome-Wide Chromatin Structure Changes During Adipogenesis and Myogenesis." *Int. J. Biol. Sci* 14 (11): 1571–85. https://doi.org/10.7150/ijbs.25328.

He, Peng, Brian A. Williams, Diane Trout, Georgi K. Marinov, Henry Amrhein, Libera Berghella, Say Tar Goh, et al. 2020. "The Changing Mouse Embryo Transcriptome at Whole Tissue and Single-Cell Resolution." *Nature* 583 (7818): 760–67. https://doi.org/10.1038/s41586-020-2536-x.

Heger, Peter, Birger Marin, and Einhard Schierenberg. 2009. "Loss of the Insulator Protein CTCF during Nematode Evolution." *BMC Molecular Biology* 10 (August): 84. https://doi.org/10.1186/1471-2199-10-84.

Heid, Iris M., Anne U. Jackson, Joshua C. Randall, Tthomas W. Winkler, Llu Qi, Valgerdur Ssteinthorsdottir, Ggudmar Tthorleifsson, et al. 2010. "Meta-Analysis Identifies 13 New Loci Associated with Waist-Hip Ratio and Reveals Sexual Dimorphism in the Genetic Basis of Fat Distribution." *Nature Genetics 2010 42:11* 42 (11): 949–60. https://doi.org/10.1038/ng.685.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4): 576–89. https://doi.org/10.1016/J.MOLCEL.2010.05.004.

Hildebrand, Erica M., and Job Dekker. 2020. "Mechanisms and Functions of Chromosome Compartmentalization." *Trends in Biochemical Sciences*. Elsevier Ltd. https://doi.org/10.1016/j.tibs.2020.01.002.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, et al. 2006. "The UCSC Genome Browser Database: Update 2006." *Nucleic Acids Research* 34 (Database issue). https://doi.org/10.1093/NAR/GKJ144.

Hiratani, Ichiro, Tyrone Ryba, Mari Itoh, Tomoki Yokochi, Michaela Schwaiger, Chia Wei Chang, Yung Lyou, Tim M. Townes, Dirk Schübeler, and David M. Gilbert. 2008. "Global Reorganization of Replication Domains During Embryonic Stem Cell Differentiation." *PLOS Biology* 6 (10): e245. https://doi.org/10.1371/JOURNAL.PBIO.0060245.

Hnisz, Denes, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp. 2017. "A Phase Separation Model for Transcriptional Control." *Cell* 169 (1): 13–23. https://doi.org/10.1016/J.CELL.2017.02.007.

Hoffmann, Thomas J., Hélène Choquet, Jie Yin, Yambazi Banda, Mark N. Kvale, Maria Glymour, Catherine Schaefer, Neil Risch, and Eric Jorgenson. 2018. "A Large Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci." *Genetics* 210 (2): 499–515. https://doi.org/10.1534/GENETICS.118.301479.

Hruby, Adela, and Frank B. Hu. 2015. "The Epidemiology of Obesity: A Big Picture." *PharmacoEconomics* 33 (7): 673. https://doi.org/10.1007/S40273-014-0243-X.

Hurst, Laurence D., Csaba Pál, and Martin J. Lercher. 2004. "The Evolutionary Dynamics of Eukaryotic Gene Order." *Nature Reviews Genetics* 5 (4): 299–310.

https://doi.org/10.1038/nrg1319.

Ibn-Salem, Jonas, Enrique M. Muro, and Miguel A. Andrade-Navarro. 2017. "Co-Regulation of Paralog Genes in the Three-Dimensional Chromatin Architecture." *Nucleic Acids Research* 45 (1): 81–91. https://doi.org/10.1093/nar/gkw813.

Ing-Simmons, Elizabeth, Vlad C. Seitan, Andre J. Faure, Paul Flicek, Thomas Carroll, Job Dekker, Amanda G. Fisher, Boris Lenhard, and Matthias Merkenschlager. 2015. "Spatial Enhancer Clustering and Regulation of Enhancer-Proximal Genes by Cohesin." *Genome Research* 25 (4): 504. https://doi.org/10.1101/GR.184986.114.

Jackson, Dean A., and Ana Pombo. 1998. "Replicon Clusters Are Stable Units of Chromosome Structure: Evidence That Nuclear Organization Contributes to the Efficient Activation and Propagation of S Phase in Human Cells." *The Journal of Cell Biology* 140 (6): 1285–95. https://doi.org/10.1083/JCB.140.6.1285.

Javierre, Biola M., Sven Sewitz, Jonathan Cairns, Steven W. Wingett, Csilla Várnai, Michiel J. Thiecke, Paula Freire-Pritchett, et al. 2016. "Lineage-Specific Genome Architecture Links Enhancers and Non-Coding Disease Variants to Target Gene Promoters." *Cell* 167 (5): 1369-1384.e19. https://doi.org/10.1016/J.CELL.2016.09.037/ATTACHMENT/5BC79F6F-1B69-4192-8CB8-4247CC2E0F39/MMC4.ZIP.

Joslin, Amelia C., Débora R. Sobreira, Grace T. Hansen, Noboru J. Sakabe, Ivy Aneas, Lindsey E. Montefiori, Kathryn M. Farris, et al. 2021. "A Functional Genomics Pipeline Identifies Pleiotropy and Cross-Tissue Effects within Obesity-Associated GWAS Loci." *Nature Communications 2021 12:1* 12 (1): 1–15. https://doi.org/10.1038/s41467-021-25614-3.

Kadauke, Stephan, and Gerd A. Blobel. 2009. "Chromatin Loops in Gene Regulation." *Biochimica et Biophysica Acta* 1789 (1): 17. https://doi.org/10.1016/J.BBAGRM.2008.07.002.

Karolchik, Donna, Angela S. Hinricks, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. 2004. "The UCSC Table Browser Data Retrieval Tool." *Nucleic Acids Research* 32 (DATABASE ISS.). https://doi.org/10.1093/nar/gkh103.

Keane, Thomas M., Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, et al. 2011. "Mouse Genomic Variation and Its Effect on Phenotypes and Gene Regulation." *Nature* 477 (7364): 289–94. https://doi.org/10.1038/nature10413.

Kichaev, Gleb, Gaurav Bhatia, Po Ru Loh, Steven Gazal, Kathryn Burch, Malika K. Freund, Armin Schoech, Bogdan Pasaniuc, and Alkes L. Price. 2019. "Leveraging Polygenic Functional Enrichment to Improve GWAS Power." *American Journal of Human Genetics* 104 (1): 65–75. https://doi.org/10.1016/J.AJHG.2018.11.008/ATTACHMENT/66985F37-6CC0-4A42-8AFB-6CAE1352D095/MMC2.XLSX.

Kim, Somi, Nam Kyung Yu, and Bong Kiun Kaang. 2015. "CTCF as a Multifunctional Protein in Genome Regulation and Gene Expression." *Experimental & Molecular Medicine 2015 47:6* 47 (6): e166–e166. https://doi.org/10.1038/emm.2015.33.

Kim, Yong Hoon, Sajid A Marhon, Yuxiang Zhang, David J Steger, Kyoung-Jae Won, and Mitchell A Lazar. 2018. "Rev-Erba Dynamically Modulates Chromatin Looping to Control Circadian Gene Transcription." *Science*. Vol. 359. http://science.sciencemag.org/.

Krefting, Jan, Miguel A. Andrade-Navarro, and Jonas Ibn-Salem. 2018. “Evolutionary Stability of Topologically Associating Domains Is Associated with Conserved Gene Regulation.” *BMC Biology* 16 (1): 87. https://doi.org/10.1186/s12915-018-0556-x.

Kueng, Stephanie, Björn Hegemann, Beate H. Peters, Jesse J. Lipp, Alexander Schleiffer, Karl Mechtler, and Jan Michael Peters. 2006. “Wapl Controls the Dynamic Association of Cohesin with Chromatin.” *Cell* 127 (5): 955–67. https://doi.org/10.1016/J.CELL.2006.09.040.

Laber, Samantha, Sara Forcisi, Liz Bentley, Julia Petzold, Franco Moritz, Kirill S. Smirnov, Loubna Al Sadat, et al. 2021. “Linking the FTO Obesity Rs1421085 Variant Circuitry to Cellular, Metabolic, and Organismal Phenotypes in Vivo.” *Science Advances* 7 (30). https://doi.org/10.1126/SCIADV.ABG0108/SUPPL_FILE/SCIADV.ABG0108_SM.PDF.

Lajoie, Bryan R, Job Dekker, and Noam Kaplan. 2015. “The Hitchhiker’s Guide to Hi-C Analysis: Practical Guidelines.” *Methods (San Diego, Calif.)* 72 (January): 65–75. https://doi.org/10.1016/j.ymeth.2014.10.031.

Langer-Safer, P. R., M. Levine, and D. C. Ward. 1982. “Immunological Method for Mapping Genes on Drosophila Polytene Chromosomes.” *Proceedings of the National Academy of Sciences of the United States of America* 79 (14): 4381. https://doi.org/10.1073/PNAS.79.14.4381.

Lappalainen, Tuuli, and Daniel G. MacArthur. 2021. “From Variant to Function in Human Disease Genetics.” *Science* 373 (6562): 1464–68. https://doi.org/10.1126/SCIENCE.ABI8207/ASSET/A27CC20A-79DE-43E2-9E21-6131BB9A0054/ASSETS/IMAGES/LARGE/SCIENCE.ABI8207-T1.JPG.

Lazar, Nathan H., Kimberly A. Nevonen, Brendan O’Connell, Christine McCann, Rachel J. O’Neill, Richard E. Green, Thomas J. Meyer, Mariam Okhovat, and Lucia Carbone. 2018. “Epigenetic Maintenance of Topological Domains in the Highly Rearranged Gibbon Genome.” *Genome Research* 28 (7): 983–97. https://doi.org/10.1101/gr.233874.117.

Leung, Danny, Inkyung Jung, Nisha Rajagopal, Anthony Schmitt, Siddarth Selvaraj, Ah Young Lee, Chia-An Yen, et al. 2015. “Integrative Analysis of Haplotype-Resolved Epigenomes across Human Tissues.” *Nature* 518 (7539): 350–54. https://doi.org/10.1038/nature14217.

Lévy-Leduc, Celine, M. Delattre, T. Mary-Huard, and S. Robin. 2014. “Two-Dimensional Segmentation for Analyzing Hi-C Data.” *Bioinformatics (Oxford, England)* 30 (17). https://doi.org/10.1093/BIOINFORMATICS/BTU443.

Li, Daofeng, Silas Hsu, Deepak Purushotham, Renee L. Sears, and Ting Wang. 2019. “WashU Epigenome Browser Update 2019.” *Nucleic Acids Research* 47 (W1): W158–65. https://doi.org/10.1093/NAR/GKZ348.

Li, Gang, Gail Sudlow, and Andrew S. Belmont. 1998. “Interphase Cell Cycle Dynamics of a Late-Replicating, Heterochromatic Homogeneously Staining Region: Precise Choreography of Condensation/Decondensation and Nuclear Positioning.” *The Journal of Cell Biology* 140 (5): 975. https://doi.org/10.1083/JCB.140.5.975.

Lichter, P., T. Cremer, J. Borden, L. Manuelidis, and D. C. Ward. 1988. “Delineation of

Individual Human Chromosomes in Metaphase and Interphase Cells by in Situ Suppression Hybridization Using Recombinant DNA Libraries." *Human Genetics* 80 (3): 224–34. https://doi.org/10.1007/BF01790090.

Lieberman-Aiden, Erez, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science (New York, N.Y.)* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Liu, Ching Ti, Martin L. Buchkovich, Thomas W. Winkler, Iris M. Heid, Ingrid B. Borecki, Caroline S. Fox, Karen L. Mohlke, Kari E. North, and L. Adrienne Cupples. 2014. "Multi-Ethnic Fine-Mapping of 14 Central Adiposity Loci." *Human Molecular Genetics* 23 (17): 4738–44. https://doi.org/10.1093/HMG/DDU183.

Liu, Ning Qing, Michela Maresca, Teun van den Brand, Luca Braccioli, Marijne M.G.A. Schijns, Hans Teunissen, Benoit G. Bruneau, Elphège èP Nora, and Elzo de Wit. 2020. "WAPL Maintains a Cohesin Loading Cycle to Preserve Cell-Type-Specific Distal Gene Regulation." *Nature Genetics 2020 53:1* 53 (1): 100–109. https://doi.org/10.1038/s41588-020-00744-4.

Long, Helen S., Simon Greenaway, George Powell, Ann-Marie Mallon, Cecilia M. Lindgren, and Michelle M. Simon. 2022. "Making Sense of the Linear Genome, Gene Function and TADs." *Epigenetics & Chromatin 2022 15:1* 15 (1): 1–19. https://doi.org/10.1186/S13072-022-00436-9.

Loos, Ruth J.F., and Giles S.H. Yeo. 2021. "The Genetics of Obesity: From Discovery to Biology." *Nature Reviews Genetics 2021 23:2* 23 (2): 120–33. https://doi.org/10.1038/s41576-021-00414-z.

Lu, Leina, Xiaoxiao Liu, Wei Kai Huang, Paola Giusti-Rodríguez, Jian Cui, Shanshan Zhang, Wanying Xu, et al. 2020. "Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-Coding Genome in Neural Development and Diseases." *Molecular Cell* 79 (3): 521-534.e15. https://doi.org/10.1016/J.MOLCEL.2020.06.007.

Lupiáñez, Darío G., Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, et al. 2015. "Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions." *Cell* 161 (5): 1012–25. https://doi.org/10.1016/J.CELL.2015.04.004.

Lupiáñez, Darío G, Malte Spielmann, and Stefan Mundlos. 2016. "Breaking TADs: How Alterations of Chromatin Domains Result in Disease." *Trends in Genetics : TIG* 32 (4): 225–37. https://doi.org/10.1016/j.tig.2016.01.003.

Ma, Hong, Jagath Samarabandu, Rekandu S. Devdhar, Raj Acharya, Ping Chin Cheng, Chunling Meng, and Ronald Berezney. 1998. "Spatial and Temporal Dynamics of DNA Replication Sites in Mammalian Cells." *The Journal of Cell Biology* 143 (6): 1415–25. https://doi.org/10.1083/JCB.143.6.1415.

Madsen, Jesper G.S., Maria S. Madsen, Alexander Rauch, Sofie Traynor, Elvira L. Van Hauwaert, Anders K. Haakonsson, Biola M. Javierre, Mette Hyldahl, Peter Fraser, and Susanne Mandrup. 2020. "Highly Interconnected Enhancer Communities Control Lineage-Determining Genes in Human Mesenchymal Stem Cells." *Nature Genetics* 52

(11): 1227–38. https://doi.org/10.1038/s41588-020-0709-z.

Man, Joyce C.K., Rajiv A. Mohan, Malou van den Boogaard, Catharina R.E. Hilvering, Catherine Jenkins, Vincent Wakker, Valerio Bianchi, et al. 2019. "An Enhancer Cluster Controls Gene Activity and Topology of the SCN5A-SCN10A Locus in Vivo." *Nature Communications 2019 10:1* 10 (1): 1–15. https://doi.org/10.1038/s41467-019-12856-5.

Matharu, Navneet, and Nadav Ahituv. 2015. "Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease." Edited by Elizabeth M. C. Fisher. *PLOS Genetics* 11 (12): e1005640. https://doi.org/10.1371/journal.pgen.1005640.

McArthur, Evonne, and John A. Capra. 2021. "Topologically Associating Domain Boundaries That Are Stable across Diverse Cell Types Are Evolutionarily Constrained and Enriched for Heritability." *American Journal of Human Genetics* 108 (2): 269–83. https://doi.org/10.1016/J.AJHG.2021.01.001/ATTACHMENT/96E0D966-A76B-4B0B-80A9-501B98C37288/MMC1.PDF.

McGinty, Robert K., and Song Tan. 2015. "Nucleosome Structure and Function." *Chemical Reviews* 115 (6): 2255–73. https://doi.org/10.1021/CR500373H/ASSET/IMAGES/LARGE/CR-2014-00373H_0014.JPEG.

Medina-Gomez, Carolina, John P. Kemp, Niki L. Dimou, Eskil Kreiner, Alessandra Chesi, Babette S. Zemel, Klaus Bønnelykke, et al. 2017. "Bivariate Genome-Wide Association Meta-Analysis of Pediatric Musculoskeletal Traits Reveals Pleiotropic Effects at the SREBF1/TOM1L2 Locus." *Nature Communications 2017 8:1* 8 (1): 1–11. https://doi.org/10.1038/s41467-017-00108-3.

Meyenn, Ferdinand von, Rebecca V. Berrens, Simon Andrews, Fátima Santos, Amanda J. Collier, Felix Krueger, Rodrigo Osorno, Wendy Dean, Peter J. Rugg-Gunn, and Wolf Reik. 2016. "Comparative Principles of DNA Methylation Reprogramming during Human and Mouse In Vitro Primordial Germ Cell Specification." *Developmental Cell* 39 (1): 104–15. https://doi.org/10.1016/j.devcel.2016.09.015.

Mifsud, Borbala, Filipe Tavares-Cadete, Alice N. Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W. Wingett, et al. 2015. "Mapping Long-Range Promoter Contacts in Human Cells with High-Resolution Capture Hi-C." *Nature Genetics 2015 47:6* 47 (6): 598–606. https://doi.org/10.1038/ng.3286.

Mirny, Leonid A, Maxim Imakaev, and Nezar Abdennur. 2019. "Two Major Mechanisms of Chromosome Organization." *Current Opinion in Cell Biology* 58 (June): 142–52. https://doi.org/10.1016/j.ceb.2019.05.001.

Misteli, Tom. 2020. "The Self-Organizing Genome: Principles of Genome Architecture and Function." *Cell* 183 (1): 28–45. https://doi.org/10.1016/J.CELL.2020.09.014.

Miura, Hisashi, Rawin Poonperm, Saori Takahashi, and Ichiro Hiratani. 2018. "Practical Analysis of Hi-C Data: Generating A/B Compartment Profiles." *Methods in Molecular Biology* 1861: 221–45. https://doi.org/10.1007/978-1-4939-8766-5_16.

Montefiori, Lindsey E., Debora R. Sobreira, Noboru J. Sakabe, Ivy Aneas, Amelia C. Joslin,

Grace T. Hansen, Grazyna Bozek, Ivan P. Moskowitz, Elizabeth M. McNally, and Marcelo A. Nóbrega. 2018. “A Promoter Interaction Map for Cardiovascular Disease Genetics.” *ELife* 7 (July). https://doi.org/10.7554/ELIFE.35788.

Mora, Antonio, Geir Kjetil Sandve, Odd Stokke Gabrielsen, and Ragnhild Eskeland. 2015. “In the Loop: Promoter–Enhancer Interactions and Bioinformatics.” *Briefings in Bioinformatics* 17 (6): bbv097. https://doi.org/10.1093/bib/bbv097.

Moreau, P., R. Hen, B. Wasylyk, R. Everett, M. P. Gaub, and P. Chambon. 1981. “The SV40 72 Base Repair Repeat Has a Striking Effect on Gene Expression Both in SV40 and Other Chimeric Recombinants.” *Nucleic Acids Research* 9 (22): 6047–68. https://doi.org/10.1093/NAR/9.22.6047.

Morrison, Olivia, and Jitendra Thakur. 2021. “Molecular Complexes at Euchromatin, Heterochromatin and Centromeric Chromatin.” *International Journal of Molecular Sciences* 22 (13). https://doi.org/10.3390/IJMS22136922.

Murakami, Yota. 2013. “Heterochromatin and Euchromatin.” *Encyclopedia of Systems Biology*, 881–84. https://doi.org/10.1007/978-1-4419-9863-7_1413.

Muro, Enrique M., Jonas Ibn-Salem, and Miguel A. Andrade-Navarro. 2019. “The Distributions of Protein Coding Genes within Chromatin Domains in Relation to Human Disease.” *Epigenetics and Chromatin* 12 (1): 72. https://doi.org/10.1186/s13072-019-0317-2.

Myers, Timothy A., Stephen J. Chanock, and Mitchell J. Machiela. 2020. “LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations.” *Frontiers in Genetics* 11 (February). https://doi.org/10.3389/FGENE.2020.00157.

Nagano, Takashi, Yaniv Lubling, Tim J. Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D. Laue, Amos Tanay, and Peter Fraser. 2013. “Single-Cell Hi-C Reveals Cell-to-Cell Variability in Chromosome Structure.” *Nature* 502 (7469): 59–64. https://doi.org/10.1038/nature12593.

Nagano, Takashi, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. 2017. “Cell-Cycle Dynamics of Chromosomal Organization at Single-Cell Resolution.” *Nature 2017 547:7661* 547 (7661): 61–67. https://doi.org/10.1038/nature23001.

Ndiaye, Fatou K., Marlène Huyvaert, Ana Ortalli, Mickaël Canouil, Cécile Lecoeur, Marie Verbanck, Stéphane Lobbens, et al. 2019. “The Expression of Genes in Top Obesity-Associated Loci Is Enriched in Insula and Substantia Nigra Brain Regions Involved in Addiction and Reward.” *International Journal of Obesity (2005)* 44 (2): 539. https://doi.org/10.1038/S41366-019-0428-7.

Neems, Daniel S., Arturo G. Garza-Gongora, Erica D. Smith, and Steven T. Kosak. 2016. “Topologically Associated Domains Enriched for Lineage-Specific Genes Reveal Expression-Dependent Nuclear Topologies during Myogenesis.” *Proceedings of the National Academy of Sciences* 113 (12): E1691–1700. https://doi.org/10.1073/PNAS.1521826113.

Ngan, Chew Yee, Chee Hong Wong, Harianto Tjong, Wenbo Wang, Rachel L. Goldfeder, Cindy Choi, Hao He, et al. 2020. “Chromatin Interaction Analyses Elucidate the Roles of PRC2-

Bound Silencers in Mouse Development." *Nature Genetics* 52 (3): 264–72. https://doi.org/10.1038/S41588-020-0581-X.

Nichols, Michael H., and Victor G. Corces. 2021. "Principles of 3D Compartmentalization of the Human Genome." *Cell Reports* 35 (13): 109330. https://doi.org/10.1016/J.CELREP.2021.109330.

Niimura, Yoshihito. 2009. "Evolutionary Dynamics of Olfactory Receptor Genes in Chordates: Interaction between Environments and Genomic Contents." *Human Genomics*. BioMed Central. https://doi.org/10.1186/1479-7364-4-2-107.

Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. "Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre." *Nature* 485 (7398): 381–85. https://doi.org/10.1038/nature11049.

Norton, Heidi K., Daniel J. Emerson, Harvey Huang, Jesi Kim, Katelyn R. Titus, Shi Gu, Danielle S. Bassett, and Jennifer E. Phillips-Cremins. 2018. "Detecting Hierarchical Genome Folding with Network Modularity." *Nature Methods 2018 15:2* 15 (2): 119–22. https://doi.org/10.1038/nmeth.4560.

Nuebler, Johannes, Geoffrey Fudenberg, Maxim Imakaev, Nezar Abdennur, and Leonid A. Mirny. 2018. "Chromatin Organization by an Interplay of Loop Extrusion and Compartmental Segregation." *Proceedings of the National Academy of Sciences of the United States of America* 115 (29): E6697–6706. https://doi.org/10.1073/pnas.1717730115.

O'Keefe, R. T., S. C. Henderson, and D. L. Spector. 1992. "Dynamic Organization of DNA Replication in Mammalian Cell Nuclei: Spatially and Temporally Defined Replication of Chromosome-Specific Alpha-Satellite DNA Sequences." *The Journal of Cell Biology* 116 (5): 1095. https://doi.org/10.1083/JCB.116.5.1095.

O'Sullivan, Justin M., Michael D. Hendy, Tatyana Pichugina, Graeme C. Wake, and Jörg Langowski. 2013. "The Statistical-Mechanics of Chromosome Conformation Capture." *Nucleus (United States)* 4 (5): 390–98. https://doi.org/10.4161/NUCL.26513/SUPPL_FILE/KNCL_A_10926513_SM0001.ZIP.

Oldach, Phoebe, and Conrad A. Nieduszynski. 2019. "Cohesin-Mediated Genome Architecture Does Not Define DNA Replication Timing Domains." *Genes* 10 (3). https://doi.org/10.3390/GENES10030196.

Oluwadare, Oluwatosin, and Jianlin Cheng. 2017. "ClusterTAD: An Unsupervised Machine Learning Approach to Detecting Topologically Associated Domains of Chromosomes from Hi-C Data." *BMC Bioinformatics* 18 (1): 1–14. https://doi.org/10.1186/S12859-017-1931-2/FIGURES/8.

Oudelaar, A. Marieke, James O.J. Davies, Lars L.P. Hanssen, Jelena M. Telenius, Ron Schwessinger, Yu Liu, Jill M. Brown, et al. 2018. "Single-Allele Chromatin Interactions Identify Regulatory Hubs in Dynamic Compartmentalized Domains." *Nature Genetics 2018 50:12* 50 (12): 1744–51. https://doi.org/10.1038/s41588-018-0253-2.

Oudelaar, A Marieke, and Douglas R Higgs. 2020. "The Relationship between Genome

Structure and Function." *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-00303-x.

Paliou, Christina, Philine Guckelberger, Robert Schöpflin, Verena Heinrich, Andrea Esposito, Andrea M. Chiariello, Simona Bianco, et al. 2019. "Preformed Chromatin Topology Assists Transcriptional Robustness of Shh during Limb Development." *Proceedings of the National Academy of Sciences of the United States of America* 116 (25): 12390–99. https://doi.org/10.1073/PNAS.1900672116/SUPPL_FILE/PNAS.1900672116.SM02.MP4.

Pan, David Z., Kristina M. Garske, Marcus Alvarez, Yash V. Bhagat, James Boocock, Elina Nikkola, Zong Miao, et al. 2018. "Integration of Human Adipocyte Chromosomal Interactions with Adipose Gene Expression Prioritizes Obesity-Related Genes from GWAS." *Nature Communications* 9 (1): 1512. https://doi.org/10.1038/s41467-018-03554-9.

Pang, Baoxu, and Michael P. Snyder. 2020. "Systematic Identification of Silencers in Human Cells." *Nature Genetics 2020 52:3* 52 (3): 254–63. https://doi.org/10.1038/s41588-020-0578-5.

Panigrahi, Anil, and Bert W. O'Malley. 2021. "Mechanisms of Enhancer Action: The Known and the Unknown." *Genome Biology 2021 22:1* 22 (1): 1–30. https://doi.org/10.1186/S13059-021-02322-1.

Pardiñas, Antonio F., Peter Holmans, Andrew J. Pocklington, Valentina Escott-Price, Stephan Ripke, Noa Carrera, Sophie E. Legge, et al. 2018. "Common Schizophrenia Alleles Are Enriched in Mutation-Intolerant Genes and in Regions under Strong Background Selection." *Nature Genetics 2018 50:3* 50 (3): 381–89. https://doi.org/10.1038/s41588-018-0059-2.

Phillips-Cremins, Jennifer E., Michael E.G. Sauria, Amartya Sanyal, Tatiana I. Gerasimova, Bryan R. Lajoie, Joshua S.K. Bell, Chin-Tong Ong, et al. 2013. "Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment." *Cell* 153 (6): 1281–95. https://doi.org/10.1016/j.cell.2013.04.053.

Pinkel, D, J Landegentt, C Collins, J Fuscoet, R Segraves, J Lucas, and J Gray. 1988. "Fluorescence in Situ Hybridization with Human Chromosome-Specific Libraries: Detection of Trisomy 21 and Translocations of Chromosome 4 (Tumor Cytogenetics/Prenatal Diagnosis/Aneuploidy/Interphase)." *Proc. Nati. Acad. Sci. USA* 85: 9138–42.

Piovesan, Allison, Maria Chiara Pelleri, Francesca Antonaros, Pierluigi Strippoli, Maria Caracausi, and Lorenza Vitale. 2019. "On the Length, Weight and GC Content of the Human Genome." *BMC Research Notes* 12 (1). https://doi.org/10.1186/S13104-019-4137-Z.

Popay, Tessa M., and Jesse R. Dixon. 2022. "Coming Full Circle: On the Origin and Evolution of the Looping Model for Enhancer–Promoter Communication." *Journal of Biological Chemistry* 298 (8): 102117. https://doi.org/10.1016/J.JBC.2022.102117.

Pope, Benjamin D., Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L. Vera, et al. 2014. "Topologically Associating Domains Are Stable Units of Replication-Timing Regulation." *Nature* 515 (7527): 402–5. https://doi.org/10.1038/nature13986.

Powell, George, Helen Long, Louisa Zolkiewski, Rebecca Dumbell, Ann Marie Mallon, Cecilia M. Lindgren, and Michelle M. Simon. 2022. “Modelling the Genetic Aetiology of Complex Disease: Human-Mouse Conservation of Noncoding Features and Disease-Associated Loci.” *Biology Letters* 18 (3). https://doi.org/10.1098/RSBL.2021.0630.

Powell, George, Michelle Simon, Sara Pulit, Ann-Marie Mallon, and Cecilia Lindgren. 2019. “Tolerance of Nonsynonymous Variation Is Closely Correlated between Human and Mouse Orthologues.” *BioRxiv*, June, 657981. https://doi.org/10.1101/657981.

Pulit, Sara L., Charli Stoneman, Andrew P. Morris, Andrew R. Wood, Craig A. Glastonbury, Jessica Tyrrell, Loïc Yengo, et al. 2019. “Meta-Analysis of Genome-Wide Association Studies for Body Fat Distribution in 694 649 Individuals of European Ancestry.” *Human Molecular Genetics* 28 (1): 166–74. https://doi.org/10.1093/hmg/ddy327.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Quinodoz, Sofia A., Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M. Lai, et al. 2018. “Higher-Order Inter-Chromosomal Hubs Shape 3D Genome Organization in the Nucleus.” *Cell* 174 (3): 744-757.e24. https://doi.org/10.1016/J.CELL.2018.05.024.

Ramani, Vijay, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. 2017. “Massively Multiplex Single-Cell Hi-C.” *Nature Methods* 14 (3): 263–66. https://doi.org/10.1038/nmeth.4155.

Rao, Suhas S.P., Su Chen Huang, Brian Glenn St Hilaire, Jesse M. Engreitz, Elizabeth M. Perez, Kyong Rim Kieffer-Kwon, Adrian L. Sanborn, et al. 2017. “Cohesin Loss Eliminates All Loop Domains.” *Cell* 171 (2): 305-320.e24. https://doi.org/10.1016/J.CELL.2017.09.026/ATTACHMENT/9126C43F-E65D-4836-AD17-B34516888EC2/MMC2.PDF.

Rao, Suhas S P, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80. https://doi.org/10.1016/j.cell.2014.11.021.

Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. 2019. “G:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update).” *Nucleic Acids Research* 47 (W1): W191–98. https://doi.org/10.1093/nar/gkz369.

Reimand, Jüri, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. 2007. “G:Profiler—a Web-Based Toolset for Functional Profiling of Gene Lists from Large-Scale Experiments.” *Nucleic Acids Research* 35 (Web Server issue): W193. https://doi.org/10.1093/NAR/GKM226.

Rivera-Mulia, Juan Carlos, Quinton Buckley, Takayo Sasaki, Jared Zimmerman, Ruth A Didier, Kristopher Nazor, Jeanne F Loring, et al. 2015. “Dynamic Changes in Replication Timing and Gene Expression during Lineage Specification of Human Pluripotent Stem Cells.” https://doi.org/10.1101/gr.187989.114.

Rosa-Garrido, Manuel, Douglas J. Chapski, Anthony D. Schmitt, Todd H. Kimball, Elaheh Karbassi, Emma Monte, Enrique Balderas, et al. 2017. "High-Resolution Mapping of Chromatin Conformation in Cardiac Myocytes Reveals Structural Remodeling of the Epigenome in Heart Failure." *Circulation* 136 (17): 1613–25. https://doi.org/10.1161/CIRCULATIONAHA.117.029430.

Rowley, M. Jordan, and Victor G. Corces. 2016. "The Three-Dimensional Genome: Principles and Roles of Long-Distance Interactions." *Current Opinion in Cell Biology* 40 (June): 8–14. https://doi.org/10.1016/J.CEB.2016.01.009.

Rowley, M. Jordan, Michael H. Nichols, Xiaowen Lyu, Masami Ando-Kuri, I. Sarahi M. Rivera, Karen Hermetz, Ping Wang, Yijun Ruan, and Victor G. Corces. 2017. "Evolutionarily Conserved Principles Predict 3D Chromatin Organization." *Molecular Cell* 67 (5): 837. https://doi.org/10.1016/J.MOLCEL.2017.07.022.

Ruiz-Velasco, Mariana, and Judith B. Zaugg. 2017. "Structure Meets Function: How Chromatin Organisation Conveys Functionality." *Current Opinion in Systems Biology* 1 (February): 129–36. https://doi.org/10.1016/J.COISB.2017.01.003.

Ryba, Tyrone, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C. Schulz, Allan J. Robins, Stephen Dalton, and David M. Gilbert. 2010. "Evolutionarily Conserved Replication Timing Profiles Predict Long-Range Chromatin Interactions and Distinguish Closely Related Cell Types." *Genome Research* 20 (6): 761–70. https://doi.org/10.1101/GR.099655.109.

Sabari, Benjamin R., Alessandra Dall'Agnese, Ann Boija, Isaac A. Klein, Eliot L. Coffey, Krishna Shrinivas, Brian J. Abraham, et al. 2018. "Coactivator Condensation at Super-Enhancers Links Phase Separation and Gene Control." *Science* 361 (6400). https://doi.org/10.1126/SCIENCE.AAR3958/SUPPL_FILE/AAR3958_SABARI_SM_TABLE_S3.XLSX.

Sakaue, Saori, Masahiro Kanai, Yosuke Tanigawa, Juha Karjalainen, Mitja Kurki, Seizo Koshiba, Akira Narita, et al. 2021. "A Cross-Population Atlas of Genetic Associations for 220 Human Phenotypes." *Nature Genetics 2021 53:10* 53 (10): 1415–24. https://doi.org/10.1038/s41588-021-00931-x.

Saksouk, Nehmé, Elisabeth Simboeck, and Jérôme Déjardin. 2015. "Constitutive Heterochromatin Formation and Transcription in Mammals." *Epigenetics and Chromatin* 8 (1): 1–17. https://doi.org/10.1186/1756-8935-8-3/FIGURES/4.

Sanborn, Adrian L, Suhas S P Rao, Su-Chen Huang, Neva C Durand, Miriam H Huntley, Andrew I Jewett, Ivan D Bochkov, et al. 2015. "Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-Type and Engineered Genomes." *Proceedings of the National Academy of Sciences of the United States of America* 112 (47): E6456-65. https://doi.org/10.1073/pnas.1518552112.

Sarnataro, Sergio, Andrea Riba, and Nacho Molina. 2020. "Regulation of Transcription Reactivation Dynamics Exiting Mitosis." *BioRxiv*, April, 2020.04.15.042853. https://doi.org/10.1101/2020.04.15.042853.

Sauerwald, Natalie, Akshat Singhal, and Carl Kingsford. 2020. "Analysis of the Structural Variability of Topologically Associated Domains as Revealed by Hi-C." *NAR Genomics and*

Bioinformatics 2 (1). https://doi.org/10.1093/nargab/lqz008.

Sayers, Eric W., Evan E. Bolton, J. Rodney Brister, Kathi Canese, Jessica Chan, Donald C. Comeau, Ryan Connor, et al. 2022. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 50 (D1): D20–26. https://doi.org/10.1093/NAR/GKAB1112.

Schmitt, Anthony D., Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L. Tan, Yun Li, et al. 2016. "A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome." *Cell Reports* 17 (8): 2042–59. https://doi.org/10.1016/J.CELREP.2016.10.061.

Schoenfelder, Stefan, Mayra Furlan-Magaril, Borbala Mifsud, Filipe Tavares-Cadete, Robert Sugar, Biola Maria Javierre, Takashi Nagano, et al. 2015. "The Pluripotent Regulatory Circuitry Connecting Promoters to Their Long-Range Interacting Elements." *Genome Research* 25 (4): 582–97. https://doi.org/10.1101/GR.185272.114.

Schoenfelder, Stefan, Biola Maria Javierre, Mayra Furlan-Magaril, Steven W. Wingett, and Peter Fraser. 2018. "Promoter Capture Hi-C: High-Resolution, Genome-Wide Profiling of Promoter Interactions." *Journal of Visualized Experiments : JoVE* 2018 (136): 57320. https://doi.org/10.3791/57320.

Schwarzer, Wibke, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A. Fonseca, et al. 2017. "Two Independent Modes of Chromatin Organization Revealed by Cohesin Removal." *Nature 2017 551:7678* 551 (7678): 51–56. https://doi.org/10.1038/nature24281.

Serra, François, Davide Baù, Mike Goodstadt, David Castillo, Guillaume Filion, and Marc A. Marti-Renom. 2017. "Automatic Analysis and 3D-Modelling of Hi-C Data Using TADbit Reveals Structural Features of the Fly Chromatin Colors." *PLOS Computational Biology* 13 (7): e1005665. https://doi.org/10.1371/JOURNAL.PCBI.1005665.

Sexton, Tom, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. 2012. "Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome." *Cell* 148 (3): 458–72. https://doi.org/10.1016/J.CELL.2012.01.010.

Shen, Yin, Feng Yue, David F. McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, et al. 2012. "A Map of the Cis-Regulatory Sequences in the Mouse Genome." *Nature* 488 (7409): 116–20. https://doi.org/10.1038/nature11243.

Shin, Hanjun, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xianghong Jasmine Zhou. 2016. "TopDom: An Efficient and Deterministic Method for Identifying Topological Domains in Genomes." *Nucleic Acids Research* 44 (7): e70–e70. https://doi.org/10.1093/nar/gkv1505.

Shrinivas, Krishna, Benjamin R. Sabari, Eliot L. Coffey, Isaac A. Klein, Ann Boija, Alicia V. Zamudio, Jurian Schuijers, et al. 2019. "Enhancer Features That Drive Formation of Transcriptional Condensates." *Molecular Cell* 75 (3): 549-561.e7. https://doi.org/10.1016/J.MOLCEL.2019.07.009.

Shungin, D., T. Winkler, D. C. Croteau-Chonka, T. Ferreira, A. E. Locke, R. Mägi, R. J.

Strawbridge, et al. 2015. “New Genetic Loci Link Adipose and Insulin Biology to Body Fat Distribution.” *Nature* 518 (7538): 187. https://doi.org/10.1038/NATURE14132.

Siersbæk, Rasmus, Jesper Grud Skat Madsen, Biola Maria Javierre, Ronni Nielsen, Emilie Kristine Bagge, Jonathan Cairns, Steven William Wingett, et al. 2017. “Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation.” *Molecular Cell* 66 (3): 420-435.e5. https://doi.org/10.1016/j.molcel.2017.04.010.

Sima, Jiao, Abhijit Chakraborty, Vishnu Dileep, Marco Michalski, Kyle N. Klein, Nicolas P. Holcomb, Jesse L. Turner, et al. 2019. “Identifying Cis Elements for Spatiotemporal Control of Mammalian DNA Replication.” *Cell* 176 (4): 816-830.e18. https://doi.org/10.1016/J.CELL.2018.11.036.

Simonis, Marieke, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. 2006. “Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture–on-Chip (4C).” *Nature Genetics* 38 (11): 1348–54. https://doi.org/10.1038/ng1896.

Smedley, Damian, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. 2009. “BioMart - Biological Queries Made Easy.” *BMC Genomics* 10 (1): 22. https://doi.org/10.1186/1471-2164-10-22.

Sobreira, Débora R., Amelia C. Joslin, Qi Zhang, Iain Williamson, Grace T. Hansen, Kathryn M. Farris, Noboru J. Sakabe, et al. 2021. “Extensive Pleiotropism and Allelic Heterogeneity Mediate Metabolic Effects of *IRX3* and *IRX5*.” *Science* 372 (6546): 1085–91. https://doi.org/10.1126/science.abf1008.

Soler-Oliva, María E., José A. Guerrero-Martínez, Valentina Bachetti, and José C. Reyes. 2017. “Analysis of the Relationship between Coexpression Domains and Chromatin 3D Organization.” *PLoS Computational Biology* 13 (9). https://doi.org/10.1371/journal.pcbi.1005708.

Solovei, Irina, Katharina Thanisch, and Yana Feodorova. 2016. “How to Rule the Nucleus: Divide et Impera.” *Current Opinion in Cell Biology* 40: 47–59. https://doi.org/10.1016/j.ceb.2016.02.014.

Stadhouders, Ralph, Enrique Vidal, François Serra, Bruno Di Stefano, François Le Dily, Javier Quilez, Antonio Gomez, et al. 2018. “Transcription Factors Orchestrate Dynamic Interplay between Genome Topology and Gene Regulation during Cell Reprogramming.” *Nature Genetics* 50 (2): 238–49. https://doi.org/10.1038/s41588-017-0030-7.

Stamboulian, Moses, Rafael F. Guerrero, Matthew W. Hahn, and Predrag Radivojac. 2020. “The Ortholog Conjecture Revisited: The Value of Orthologs and Paralogs in Function Prediction.” *Bioinformatics* 36 (Suppl 1): I219–26. https://doi.org/10.1093/BIOINFORMATICS/BTAA468.

Strom, Amy R., Alexander V. Emelyanov, Mustafa Mir, Dmitry V. Fyodorov, Xavier Darzacq, and Gary H. Karpen. 2017. “Phase Separation Drives Heterochromatin Domain Formation.” *Nature 2017 547:7662* 547 (7662): 241–45. https://doi.org/10.1038/nature22989.

Sun, Quan, Cheynna A. Crowley, Le Huang, Jia Wen, Jiawen Chen, Erik L. Bao, Paul L. Auer, et

al. 2022. “From GWAS Variant to Function: A Study of ∼148,000 Variants for Blood Cell Traits.” *Human Genetics and Genomics Advances* 3 (1): 100063. https://doi.org/10.1016/J.XHGG.2021.100063.

Symmons, Orsolya, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nassari, Wibke Schwarzer, Laurence Ettwiller, and François Spitz. 2014. “Functional and Topological Characteristics of Mammalian Regulatory Domains.” *Genome Research* 24 (3): 390–400. https://doi.org/10.1101/gr.163519.113.

Szklarczyk, Damian, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, et al. 2019. “STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets.” *Nucleic Acids Research* 47 (D1): D607–13. https://doi.org/10.1093/nar/gky1131.

Tanabe, Hideyuki, Felix A. Habermann, Irina Solovei, Marion Cremer, and Thomas Cremer. 2002. “Non-Random Radial Arrangements of Interphase Chromosome Territories: Evolutionary Considerations and Functional Implications.” *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 504 (1–2): 37–45. https://doi.org/10.1016/S0027-5107(02)00077-5.

Tarbier, Marcel, Sebastian D. Mackowiak, João Frade, Silvina Catuara-Solarz, Inna Biryukova, Eleni Gelali, Diego Bárcena Menéndez, et al. 2020. “Nuclear Gene Proximity and Protein Interactions Shape Transcript Covariations in Mammalian Single Cells.” *Nature Communications 2020 11:1* 11 (1): 1–12. https://doi.org/10.1038/s41467-020-19011-5.

Tena, Juan J., and José M. Santos-Pereira. 2021. “Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease.” *Frontiers in Cell and Developmental Biology* 0 (July): 1817. https://doi.org/10.3389/FCELL.2021.702787.

The Gene Ontology Consortium. 2018. “The Gene Ontology Resource: 20 Years and Still GOing Strong.” *Nucleic Acids Research* 47 (D1): D330–38. https://doi.org/10.1093/nar/gky1055.

Thévenin, Annelyse, Liat Ein-Dor, Michal Ozery-Flato, and Ron Shamir. 2014. “Functional Gene Groups Are Concentrated within Chromosomes, among Chromosomes and in the Nuclear Space of the Human Genome.” *Nucleic Acids Research* 42 (15): 9854–61. https://doi.org/10.1093/nar/gku667.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. “Genome-Wide Association Studies.” *Nature Reviews Methods Primers 2021 1:1* 1 (1): 1–21. https://doi.org/10.1038/s43586-021-00056-9.

Ulianov, Sergey V., Ekaterina E. Khrameeva, Alexey A. Gavrilov, Ilya M. Flyamer, Pavel Kos, Elena A. Mikhaleva, Aleksey A. Penin, et al. 2016. “Active Chromatin and Transcription Play a Key Role in Chromosome Partitioning into Topologically Associating Domains.” *Genome Research* 26 (1): 70–84. https://doi.org/10.1101/gr.196006.115.

Ullah, Imran, Raghavendra Baregundi Subbarao, and Gyu Jin Rho. 2015. “Human Mesenchymal Stem Cells - Current Trends and Future Prospective.” *Bioscience Reports* 35 (2): 191. https://doi.org/10.1042/BSR20150025.

Vermunt, Marit W, Di Zhang, and Gerd A Blobel. 2018. “The Interdependence of Gene-Regulatory Elements and the 3D Genome.” *J. Cell Biol* 218 (1): 12–26. https://doi.org/10.1083/jcb.201809040.

Vian, Laura, Aleksandra Pękowska, Suhas S.P. Rao, Kyong Rim Kieffer-Kwon, Seolkyoung Jung, Laura Baranello, Su Chen Huang, et al. 2018. “The Energetics and Physiological Impact of Cohesin Extrusion.” *Cell* 173 (5): 1165. https://doi.org/10.1016/J.CELL.2018.03.072.

Vickaryous, Matthew K., and Brian K. Hall. 2006. “Human Cell Type Diversity, Evolution, Development, and Classification with Special Reference to Cells Derived from the Neural Crest.” *Biological Reviews* 81 (3): 425–55. https://doi.org/10.1017/S1464793106007068.

Vietri Rudan, Matteo, Christopher Barrington, Stephen Henderson, Christina Ernst, Duncan T. Odom, Amos Tanay, and Suzana Hadjur. 2015. “Comparative Hi-C Reveals That CTCF Underlies Evolution of Chromosomal Domain Architecture.” *Cell Reports* 10 (8): 1297–1309. https://doi.org/10.1016/j.celrep.2015.02.004.

Wang, Yanli, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, et al. 2018. “The 3D Genome Browser: A Web-Based Browser for Visualizing 3D Genome Organization and Long-Range Chromatin Interactions.” *Genome Biology* 19 (1): 1–12. https://doi.org/10.1186/S13059-018-1519-9/FIGURES/5.

Way, Gregory P, Daniel W Youngstrom, Kurt D Hankenson, Casey S Greene, and Struan FA Grant. 2017. “Implicating Candidate Genes at GWAS Signals by Leveraging Topologically Associating Domains.” *European Journal of Human Genetics* 25 (11): 1286–89. https://doi.org/10.1038/ejhg.2017.108.

Wendt, Kerstin S., Keisuke Yoshida, Takehiko Itoh, Masashige Bando, Birgit Koch, Erika Schirghuber, Shuichi Tsutsumi, et al. 2008. “Cohesin Mediates Transcriptional Insulation by CCCTC-Binding Factor.” *Nature 2008 451:7180* 451 (7180): 796–801. https://doi.org/10.1038/nature06634.

Whalen, Sean, and Katherine S. Pollard. 2019. “Most Chromatin Interactions Are Not in Linkage Disequilibrium.” *Genome Research* 29 (3): 334–43. https://doi.org/10.1101/GR.238022.118/-/DC1.

Wingett, Steven, Philip Ewels, Mayra Furlan-Magaril, Takashi Nagano, Stefan Schoenfelder, Peter Fraser, and Simon Andrews. 2015. “HiCUP: Pipeline for Mapping and Processing Hi-C Data.” *F1000Research* 4: 1310. https://doi.org/10.12688/f1000research.7334.1.

Wit, Elzo de. 2020. “TADs as the Caller Calls Them.” *Journal of Molecular Biology* 432 (3): 638–42. https://doi.org/10.1016/J.JMB.2019.09.026.

Wit, Elzo de, and Wouter de Laat. 2012. “A Decade of 3C Technologies: Insights into Nuclear Organization.” *Genes & Development* 26 (1): 11. https://doi.org/10.1101/GAD.179804.111.

Wittkopp, Patricia J., and Gizem Kalay. 2011. “Cis-Regulatory Elements: Molecular Mechanisms and Evolutionary Processes Underlying Divergence.” *Nature Reviews Genetics 2011 13:1* 13 (1): 59–69. https://doi.org/10.1038/nrg3095.

Woodcock, C. L., and R. A. Horowitz. 1997. “Electron Microscopy of Chromatin.” *Methods* 12 (1): 84–95. https://doi.org/10.1006/METH.1997.0450.

Wu, Hua-Jun, Alexandro Landshammer, Elena K. Stamenova, Adriano Bolondi, Helene Kretzmer, Alexander Meissner, and Franziska Michor. 2021. “Topological Isolation of Developmental Regulators in Mammalian Genomes.” *Nature Communications 2021 12:1* 12 (1): 1–19. https://doi.org/10.1038/s41467-021-24951-7.

Wutz, Gordana, Csilla Várnai, Kota Nagasaka, David A Cisneros, Roman R Stocsits, Wen Tang, Stefan Schoenfelder, et al. 2017. “Topologically Associating Domains and Chromatin Loops Depend on Cohesin and Are Regulated by CTCF, WAPL, and PDS5 Proteins.” *The EMBO Journal* 36 (24): 3573–99. https://doi.org/10.15252/EMBJ.201798004.

Xi, Wang, and Michael A. Beer. 2021. “Loop Competition and Extrusion Model Predicts CTCF Interaction Specificity.” *Nature Communications 2021 12:1* 12 (1): 1–15. https://doi.org/10.1038/s41467-021-21368-0.

Yaffe, Eitan, Shlomit Farkash-Amar, Andreas Polten, Zohar Yakhini, Amos Tanay, and Itamar Simon. 2010. “Comparative Analysis of DNA Replication Timing Reveals Conserved Large-Scale Chromosomal Architecture.” *PLOS Genetics* 6 (7): e1001011. https://doi.org/10.1371/JOURNAL.PGEN.1001011.

Yaffe, Eitan, and Amos Tanay. 2011. “Probabilistic Modeling of Hi-C Contact Maps Eliminates Systematic Biases to Characterize Global Chromosomal Architecture.” *Nature Genetics* 43 (11): 1059–65. https://doi.org/10.1038/ng.947.

Yan, Jian, Shi-An A Chen, Andrea Local, Tristin Liu, Yunjiang Qiu, Kristel M Dorighi, Sebastian Preissl, et al. 2018. “Histone H3 Lysine 4 Monomethylation Modulates Long-Range Chromatin Interactions at Enhancers.” *Cell Research* 28 (2): 204–20. https://doi.org/10.1038/cr.2018.1.

Yan, Koon Kiu, Shaoke Lou, and Mark Gerstein. 2017. “MrTADFinder: A Network Modularity Based Approach to Identify Topologically Associating Domains in Multiple Resolutions.” *PLOS Computational Biology* 13 (7): e1005647. https://doi.org/10.1371/JOURNAL.PCBI.1005647.

Yang, Dongchan, Insu Jang, Jinhyuk Choi, Min Seo Kim, Andrew J. Lee, Hyunwoong Kim, Junghyun Eom, Dongsup Kim, Inkyung Jung, and Byungwook Lee. 2018. “3DIV: A 3D-Genome Interaction Viewer and Database.” *Nucleic Acids Research* 46 (D1): D52–57. https://doi.org/10.1093/NAR/GKX1017.

Yu, G., F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang. 2010. “GOSemSim: An R Package for Measuring Semantic Similarity among GO Terms and Gene Products.” *Bioinformatics* 26 (7): 976–78. https://doi.org/10.1093/bioinformatics/btq064.

Yu, Miao, and Bing Ren. 2017. “The Three-Dimensional Organization of Mammalian Genomes.” *Annual Review of Cell and Developmental Biology* 33 (1): 265–89. https://doi.org/10.1146/annurev-cellbio-100616-060531.

Zhang, Xiao, Kenneth C. Ehrlich, Fangtang Yu, Xiaojun Hu, Xiang He Meng, Hong Wen Deng, Hui Shen, and Melanie Ehrlich. 2020. “Osteoporosis- and Obesity-Risk Interrelationships: An Epigenetic Analysis of GWAS-Derived SNPs at the Developmental Gene TBX15.” *Epigenetics* 15 (6–7): 728–49. https://doi.org/10.1080/15592294.2020.1716491.

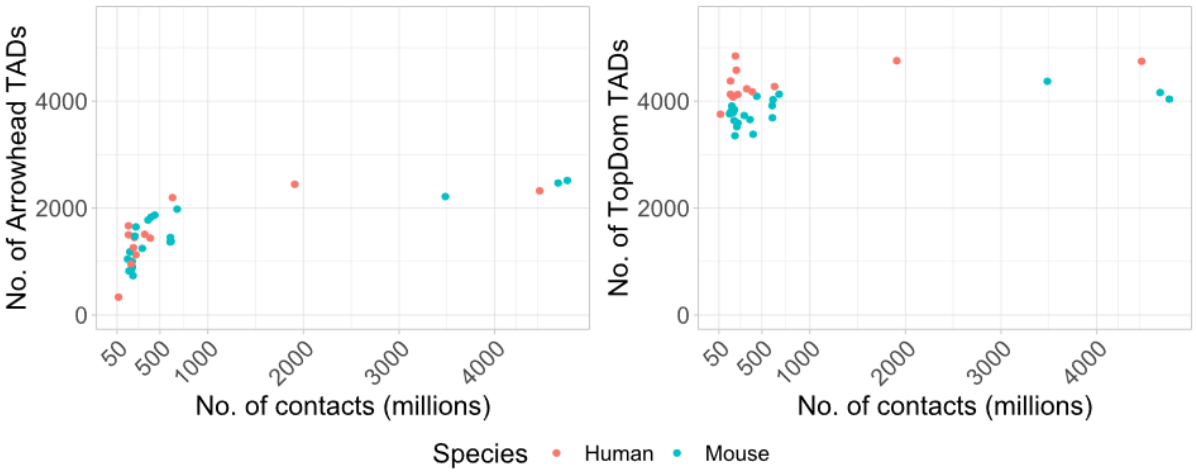
Zhang, Yanxiao, Ting Li, Sebastian Preissl, Maria Luisa Amaral, Jonathan D. Grinstein, Elie N.

Farah, Eugin Destici, et al. 2019. “Transcriptionally Active HERV-H Retrotransposons Demarcate Topologically Associating Domains in Human Pluripotent Stem Cells.” *Nature Genetics* 51 (9): 1380–88. https://doi.org/10.1038/S41588-019-0479-7.

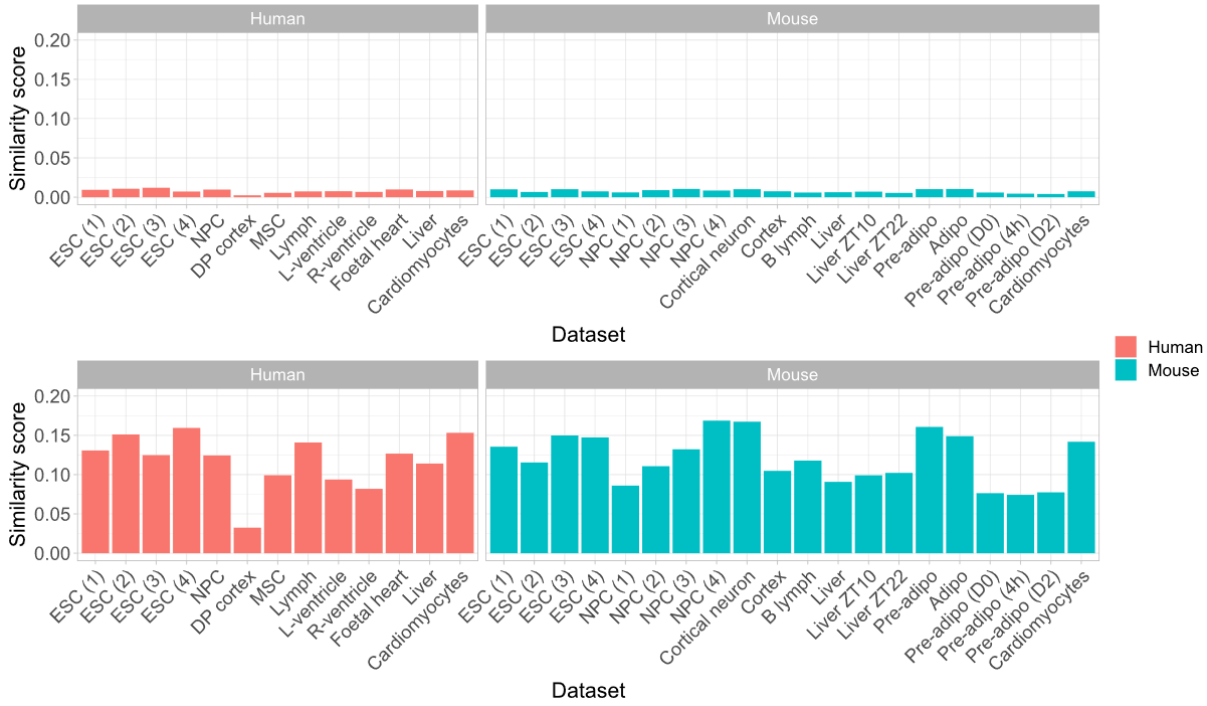
Zhu, Zhaozhong, Yanjun Guo, Huwenbo Shi, Cong Lin Liu, Ronald Allan Panganiban, Wonil Chung, Luke J. O’Connor, et al. 2020. “Shared Genetic and Experimental Links between Obesity-Related Traits and Asthma Subtypes in UK Biobank.” *Journal of Allergy and Clinical Immunology* 145 (2): 537–49. https://doi.org/10.1016/J.JACI.2019.09.035/ATTACHMENT/747D2831-2D78-4023-ABD6-A93A085394CD/MMC22.PDF.

Zufferey, Marie, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. 2018. “Comparison of Computational Methods for the Identification of Topologically Associating Domains.” *Genome Biology* 19 (1): 217. https://doi.org/10.1186/s13059-018-1596-9.

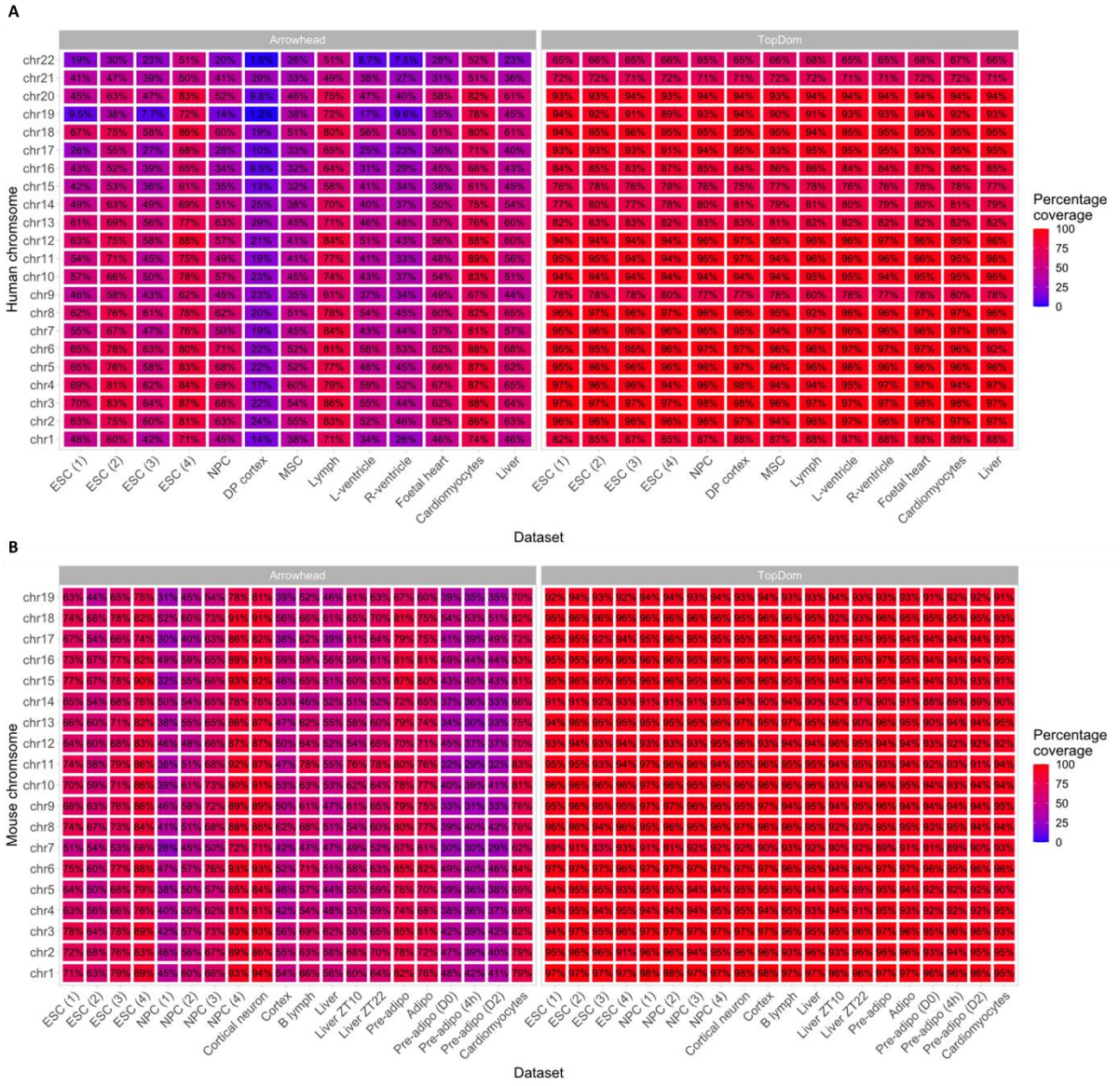
Chapter 7: Supplementary



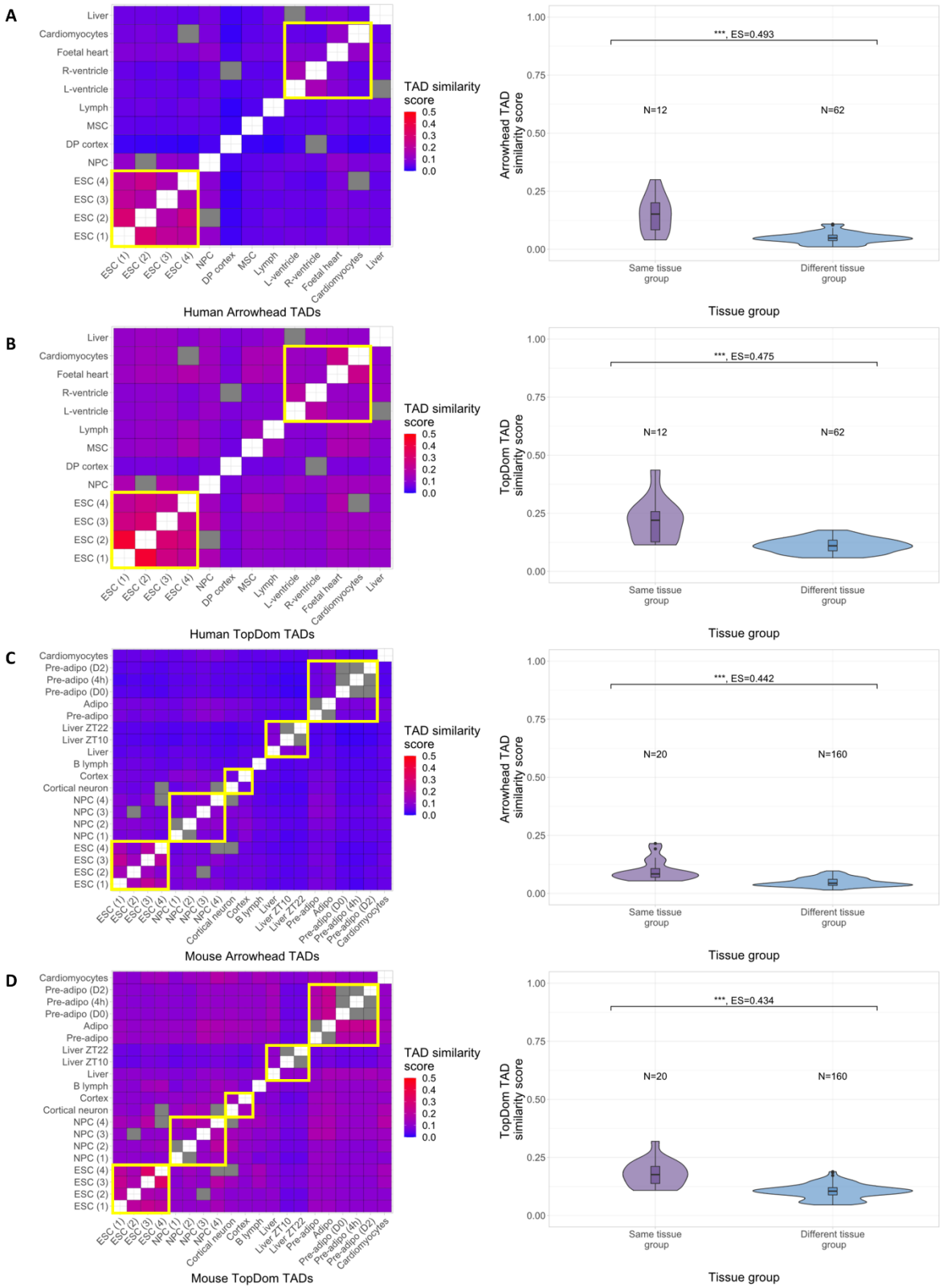
Chapter 2: Supplementary Figure 1: Number of detected autosomal TADs vs number of Hi-C contacts in all datasets. *TADs detected with Arrowhead (left) and TopDom (right) at 50 kb bin size resolution with default parameters (m=2000 and w=5 respectively).*



Chapter 2: Supplementary Figure 2: Similarity score between Arrowhead and TopDom TADs (top) and boundaries (bottom) detected in each dataset. *A higher similarity score indicates higher similarity.*

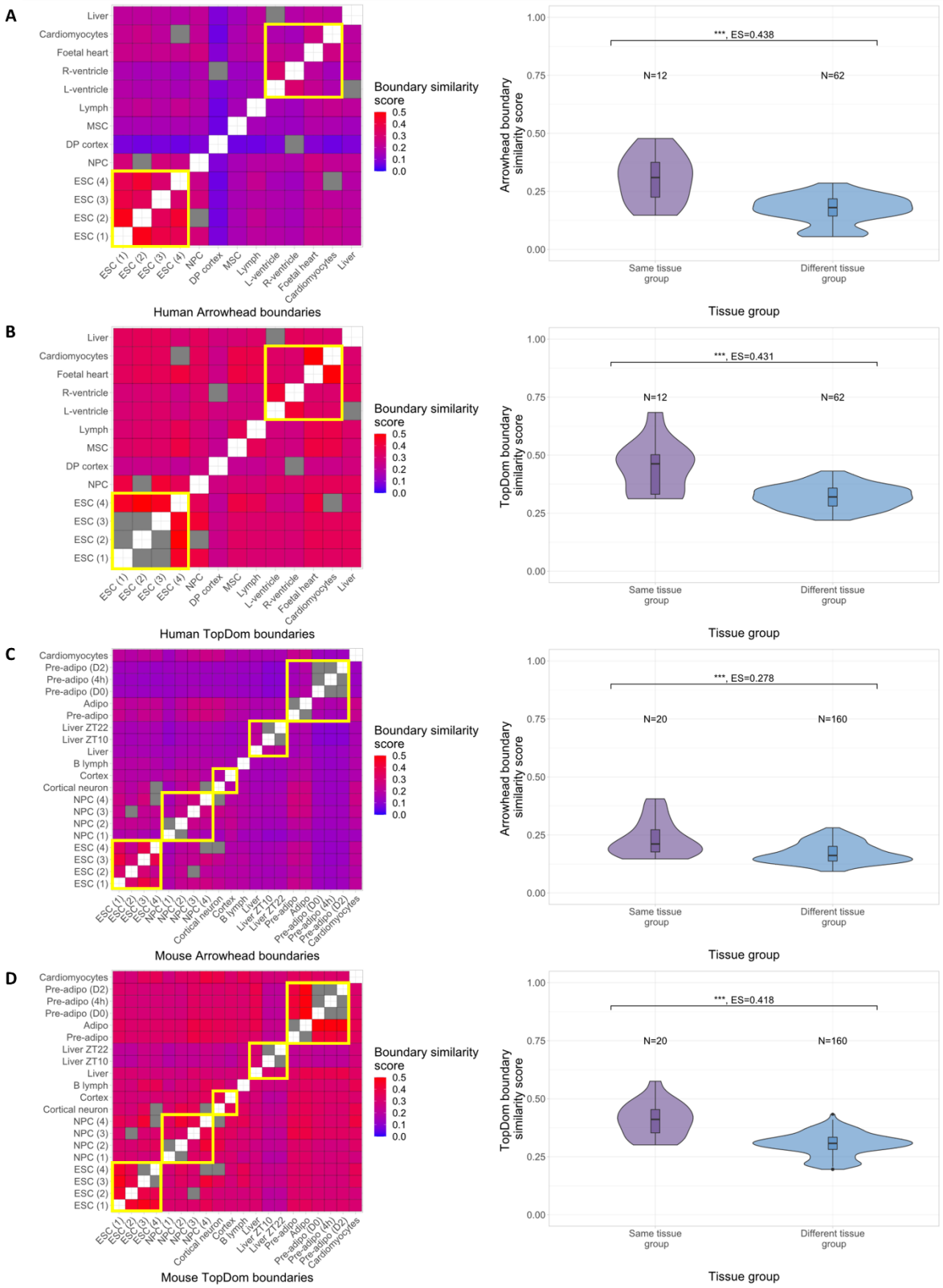


Chapter 2: Supplementary Figure 3: Percentage chromosome coverage of Arrowhead or TopDom autosomal TADs. *TADs detected at 50 kb bin size resolution with default parameters. A) Human datasets. B) Mouse datasets.*



Chapter 2: Supplementary Figure 4: Autosomal TAD comparison between all pairs of datasets, excluding those from the same data repository. *Left: heat maps indicating pairwise similarity scores between datasets. A higher similarity score indicates higher similarity. Pairs of datasets in the same*

*tissue/cell type group are indicated by yellow squares. Excluded scores are shown in grey. Right: distribution of scores between pairs of datasets in the same tissue group vs different tissue groups. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across four p-values. A) Human Arrowhead TADs. B) Human TopDom TADs. C) Mouse Arrowhead TADs. D) Mouse TopDom TADs.*



Chapter 2: Supplementary Figure 5: Autosomal TAD boundary comparison between all pairs of datasets, excluding those from the same data repository. *Left: heat maps indicating pairwise similarity scores between datasets. A higher similarity score indicates higher similarity. Pairs of*

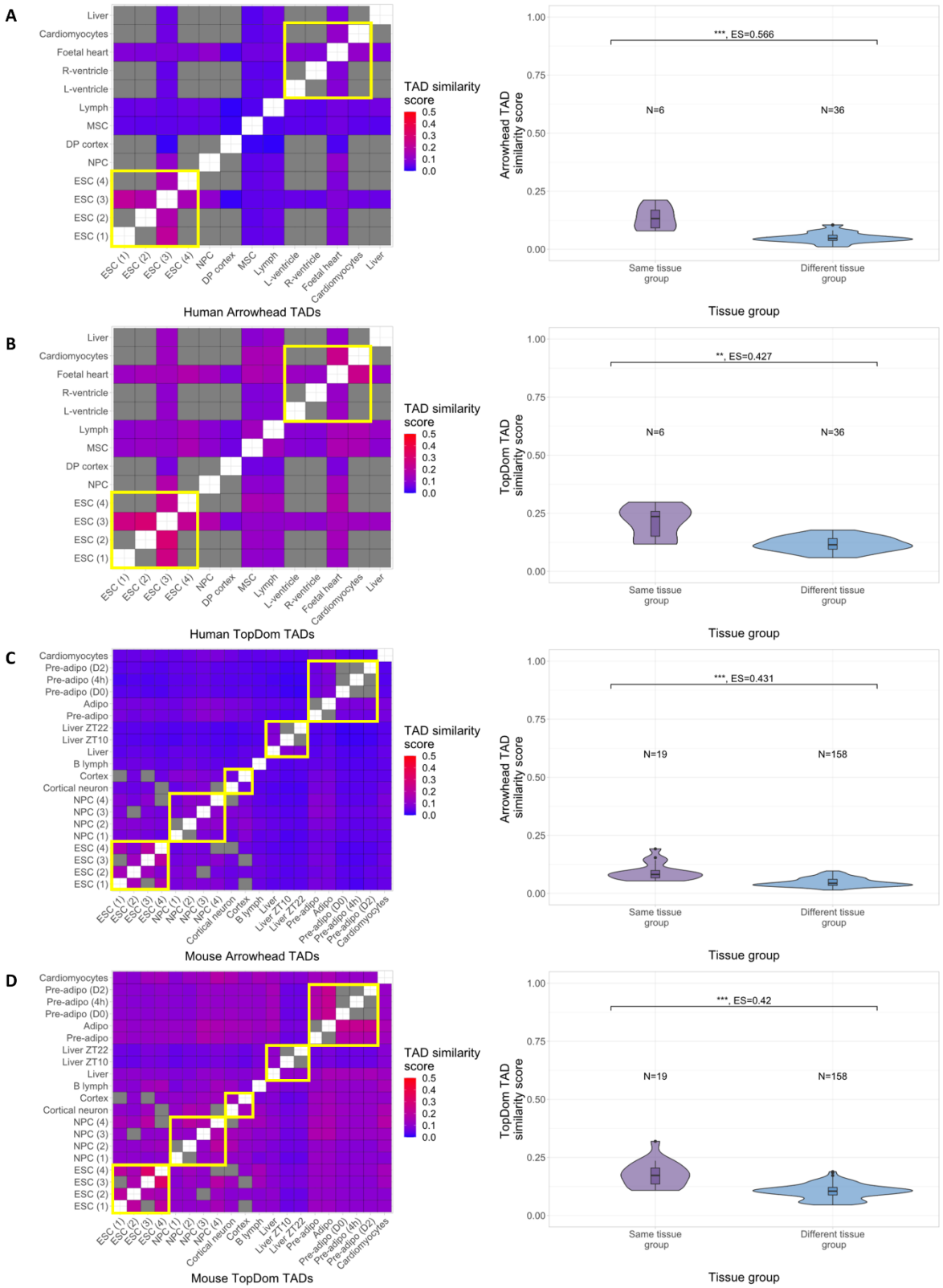
*datasets in the same tissue/cell type group are indicated by yellow squares. Excluded scores are shown in grey. Right: distribution of scores between pairs of datasets in the same tissue group vs different tissue groups. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across four p-values. A) Human Arrowhead TAD boundaries. B) Human TopDom TAD boundaries. C) Mouse Arrowhead TAD boundaries. D) Mouse TopDom TAD boundaries.*

Chapter 2: Supplementary Table 1: First and last authors of publications associated with selected datasets.

Human datasets				
Dataset	GEO or NCBI ID	Reference	First authors	Last authors
ESC (1)	GEO:GSE35156	Dixon et al. 2012	Jesse R. Dixon	Bing Ren
ESC (2)	GEO:GSE52457	Dixon et al. 2015	Jesse R. Dixon, Inkyung Jung, Siddarth Selvaraj	Bing Ren
NPC	GEO:GSE52457	Dixon et al. 2015	Jesse R. Dixon, Inkyung Jung, Siddarth Selvaraj	Bing Ren
Left ventricle	GEO:GSE58752	Leung et al. 2015	Danny Leung, Inkyung Jung, Nisha Rajagopal	Bing Ren
Liver	GEO:GSE58752	Leung et al. 2015	Danny Leung, Inkyung Jung, Nisha Rajagopal	Bing Ren
Lymphoblastoid cells	GEO:GSE63525	Rao et al. 2014	Suhas S.P. Rao, Miriam H. Huntley, Adrian L. Sanborn	Eric S. Lander, Erez Lieberman Aiden
ESC (3)	GEO:GSE86821	Freire-Pritchett et al. 2017	Paula Freire-Pritchett, Stefan Schoenfelder	Peter J Rugg-Gunn, Mikhail Spivakov
Dorsolateral prefrontal cortex	GEO:GSE87112	Schmitt et al. 2016	Anthony D. Schmitt, Ming Hu	Bing Ren
Right ventricle	GEO:GSE87112	Schmitt et al. 2016	Anthony D. Schmitt, Ming Hu	Bing Ren
Foetal heart	GEO:GSE106687	Bertero et al. 2019	Alessandro Bertero, Paul A. Fields	Charles E. Murry

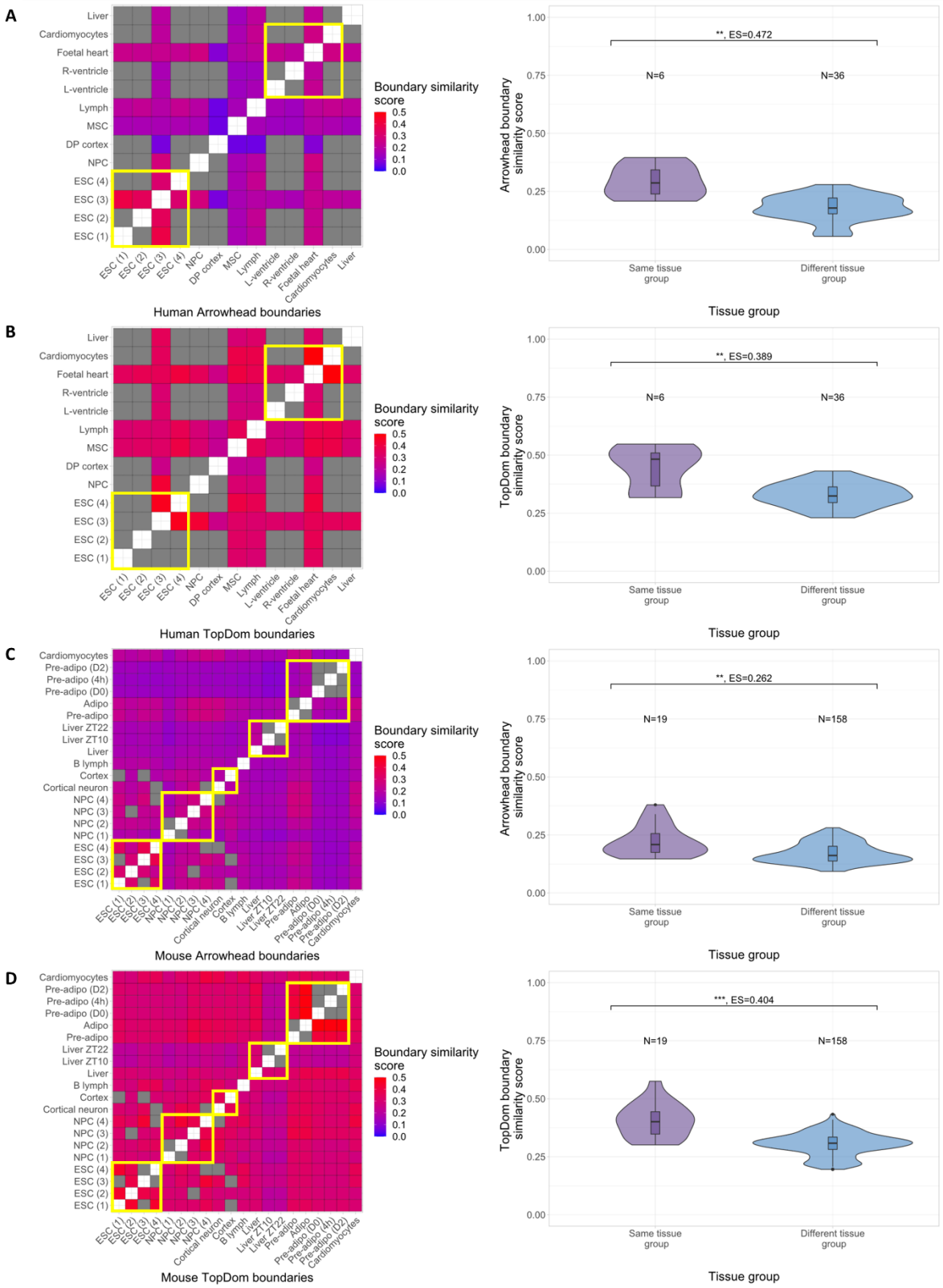
ES derived ventricular cardiomyocyte	GEO:GSE116862	Zhang et al. 2019	Yanxiao Zhang, Ting Li, Sebastian Preissl, Marios Arvanitis	Neil C. Chi , Bing Ren, Alexis Battle
ESC (4)	GEO:GSE116862	Zhang et al. 2019	Yanxiao Zhang, Ting Li, Sebastian Preissl, Marios Arvanitis	Neil C. Chi , Bing Ren, Alexis Battle
MSC	GEO:GSE140782	Madsen et al. 2020	Jesper G. S. Madsen	Susanne Mandrup
Mouse datasets				
Dataset	GEO or NCBI ID	Reference	First authors	Last authors
Cortex	GEO:GSE34587	Shen et al. 2012	Yin Shen, Feng Yue	Bing Ren
ESC (1)	GEO:GSE35156	Dixon et al. 2012	Jesse R. Dixon	Bing Ren
NPC (1)	GEO:GSE59027	Fraser et al. 2015	James Fraser, Carmelo Ferrai, Andrea M Chiariello, Markus Schueler, Tiago Rito,Giovanni Laudanno	Colin ASemple, Josée Dostie, Ana Pombo, Mario Nicodemi
NPC (2)	GEO:GSE59027	Fraser et al. 2015	James Fraser, Carmelo Ferrai, Andrea M Chiariello, Markus Schueler, Tiago Rito,Giovanni Laudanno	Colin ASemple, Josée Dostie, Ana Pombo, Mario Nicodemi
B-lymphoblasts	GEO:GSE63525	Rao et al. 2014	Suhas S.P. Rao, Miriam H. Huntley, Adrian L. Sanborn	Eric S. Lander, Erez Lieberman Aiden
Liver	GEO:GSE65126	Vietri Rudan et al. 2015	Matteo Vietri Rudan	Suzana Hadjur
ESC (2)	GEO:GSE72697	Giorgetti et al. 2016	Luca Giorgetti, Bryan R. Lajoie, Ava C. Carter, Mikael Attia	Edith Heard, Job Dekker
NPC (3)	GEO:GSE72697	Giorgetti et al. 2016	Luca Giorgetti, Bryan R. Lajoie, Ava C. Carter, Mikael Attia	Edith Heard, Job Dekker
ESC (3)	GEO:GSE74055	Yan et al. 2018	Jian Yan, Shi-An A Chen	Bing Ren
3T3-L1 preadipocytes	NCBI:SRA:SRP082116	He et al. 2018	Mengnan He, Yan Li, Qianzi Tang	Xuewei Li, Mingzhou Li

3T3-Y adipocytes	NCBI:SRA:SRP082116	He et al. 2018	Mengnan He, Yan Li, Qianzi Tang	Xuewei Li, Mingzhou Li
Pre-adipocytes differentiating towards adipocytes, day 0	GEO:GSE95533	Siersbæk et al. 2017	Rasmus Siersbæk, Jesper Grud Skat Madsen, Biola Maria Javierre, Ronni Nielsen	Peter Fraser, Susanne Mandrup
Pre-adipocytes differentiating towards adipocytes, 4 hours	GEO:GSE95533	Siersbæk et al. 2017	Rasmus Siersbæk, Jesper Grud Skat Madsen, Biola Maria Javierre, Ronni Nielsen	Peter Fraser, Susanne Mandrup
Pre-adipocytes differentiating towards adipocytes, day 2	GEO:GSE95533	Siersbæk et al. 2017	Rasmus Siersbæk, Jesper Grud Skat Madsen, Biola Maria Javierre, Ronni Nielsen	Peter Fraser, Susanne Mandrup
ESC (4)	GEO:GSE96107	Bonev et al. 2017	Boyan Bonev	Giacomo Cavalli
NPC (4)	GEO:GSE96107	Bonev et al. 2017	Boyan Bonev	Giacomo Cavalli
Cortical neuron	GEO:GSE96107	Bonev et al. 2017	Boyan Bonev	Giacomo Cavalli
Cardiac myocytes	GEO:GSE96692	Rosa-Garrido et al. 2017	Manuel Rosa-Garrido	Thomas M. Vondriska
Liver ZT10	GEO:GSE104129	Y. H. Kim et al. 2018	Yong Hoon Kim, Sajid A. Marhon	Kyoung-Jae Won, Mitchell A. Lazar
Liver ZT22	GEO:GSE104129	Y. H. Kim et al. 2018	Yong Hoon Kim, Sajid A. Marhon	Kyoung-Jae Won, Mitchell A. Lazar



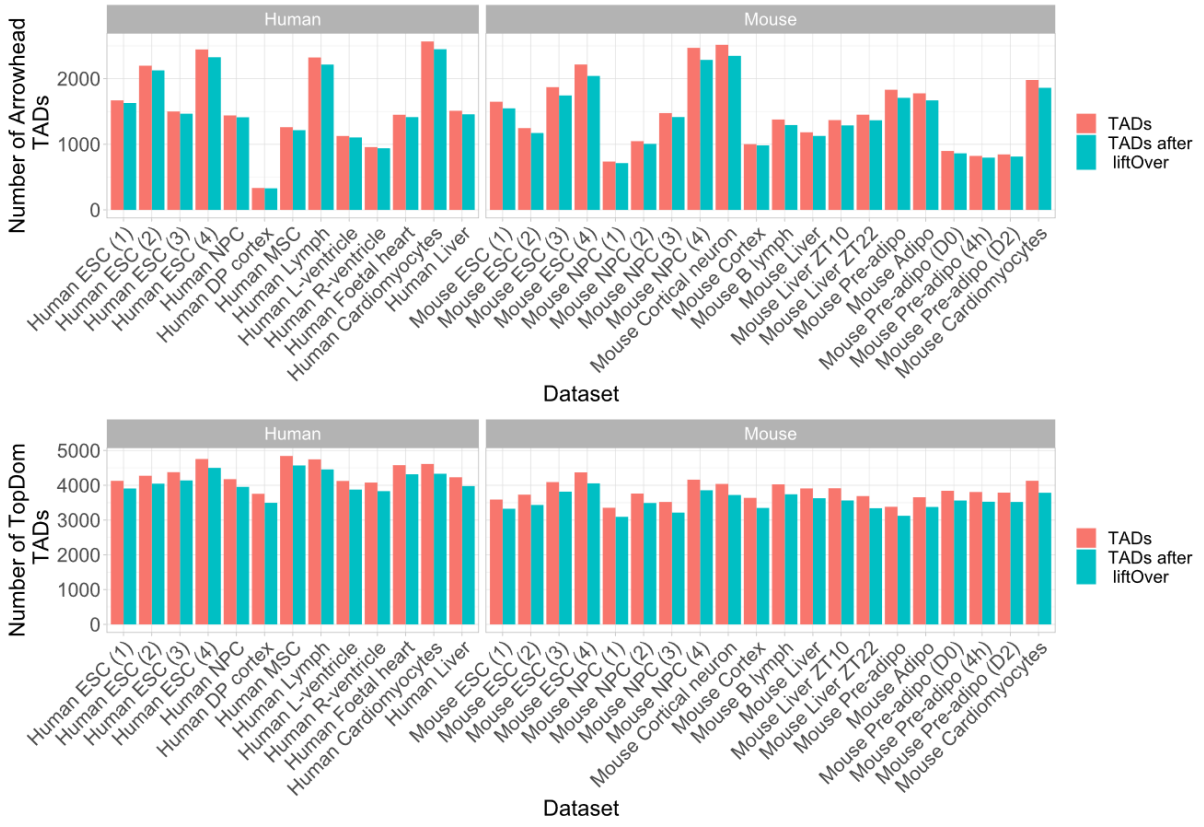
Chapter 2: Supplementary Figure 6: Autosomal TAD comparison between all pairs of datasets, excluding those with a common first or last author. *Left: heat maps indicating pairwise similarity scores between datasets. A higher similarity score indicates higher similarity. Pairs of datasets in the*

*same tissue/cell type group are indicated by yellow squares. Excluded scores are shown in grey. Right: distribution of scores between pairs of datasets in the same tissue group vs different tissue groups. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across four p-values. A) Human Arrowhead TADs. B) Human TopDom TADs. C) Mouse Arrowhead TADs. D) Mouse TopDom TADs.*

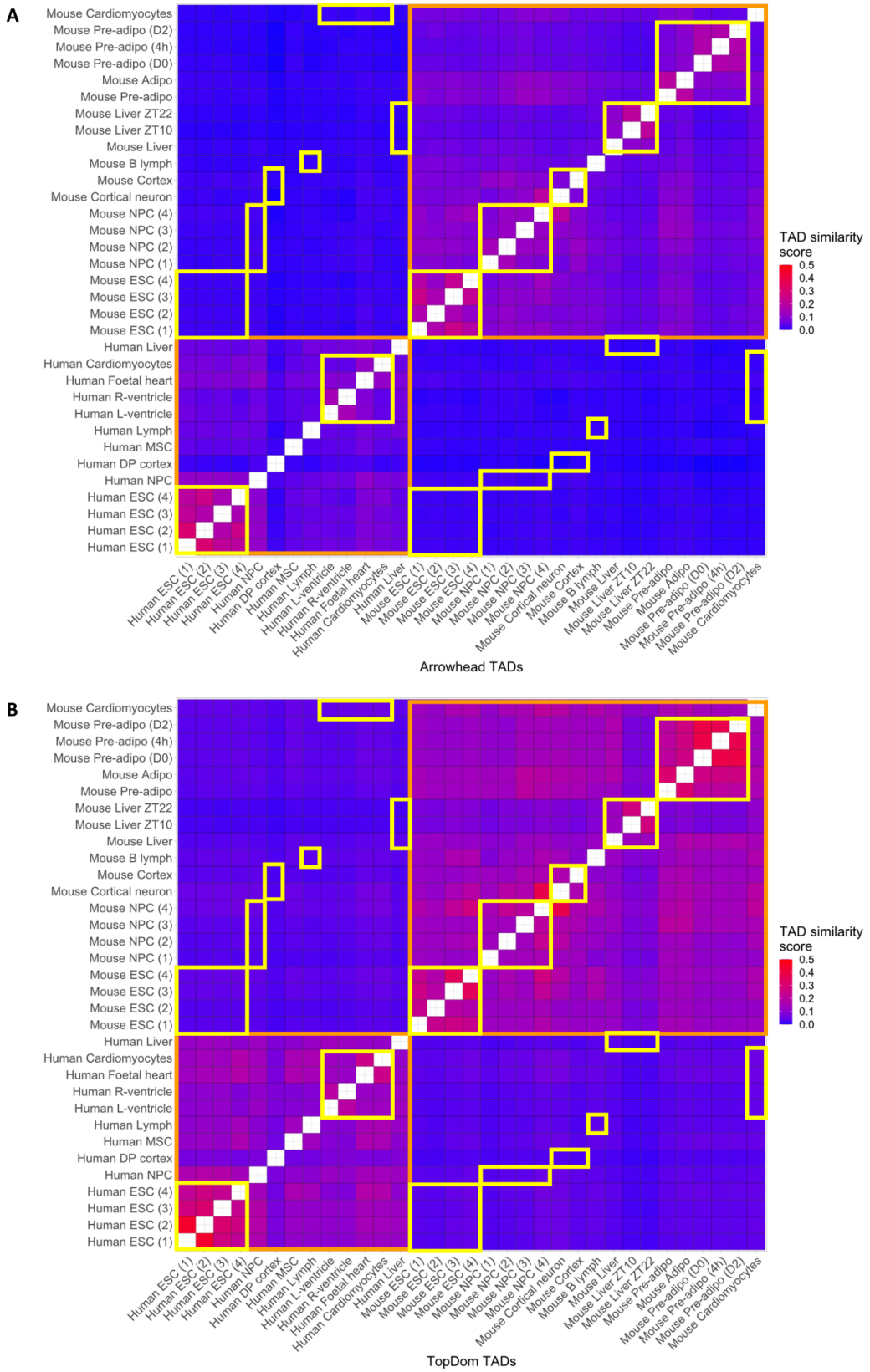


Chapter 2: Supplementary Figure 7: Autosomal TAD boundary comparison between all pairs of datasets, excluding those with a common first or last author. *Left: heat maps indicating pairwise similarity scores between datasets. A higher similarity score indicates higher similarity. Pairs of*

*datasets in the same tissue/cell type group are indicated by yellow squares. Excluded scores are shown in grey. Right: distribution of scores between pairs of datasets in the same tissue group vs different tissue groups. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across four p-values. A) Human Arrowhead TAD boundaries. B) Human TopDom TAD boundaries. C) Mouse Arrowhead TAD boundaries. D) Mouse TopDom TAD boundaries.*

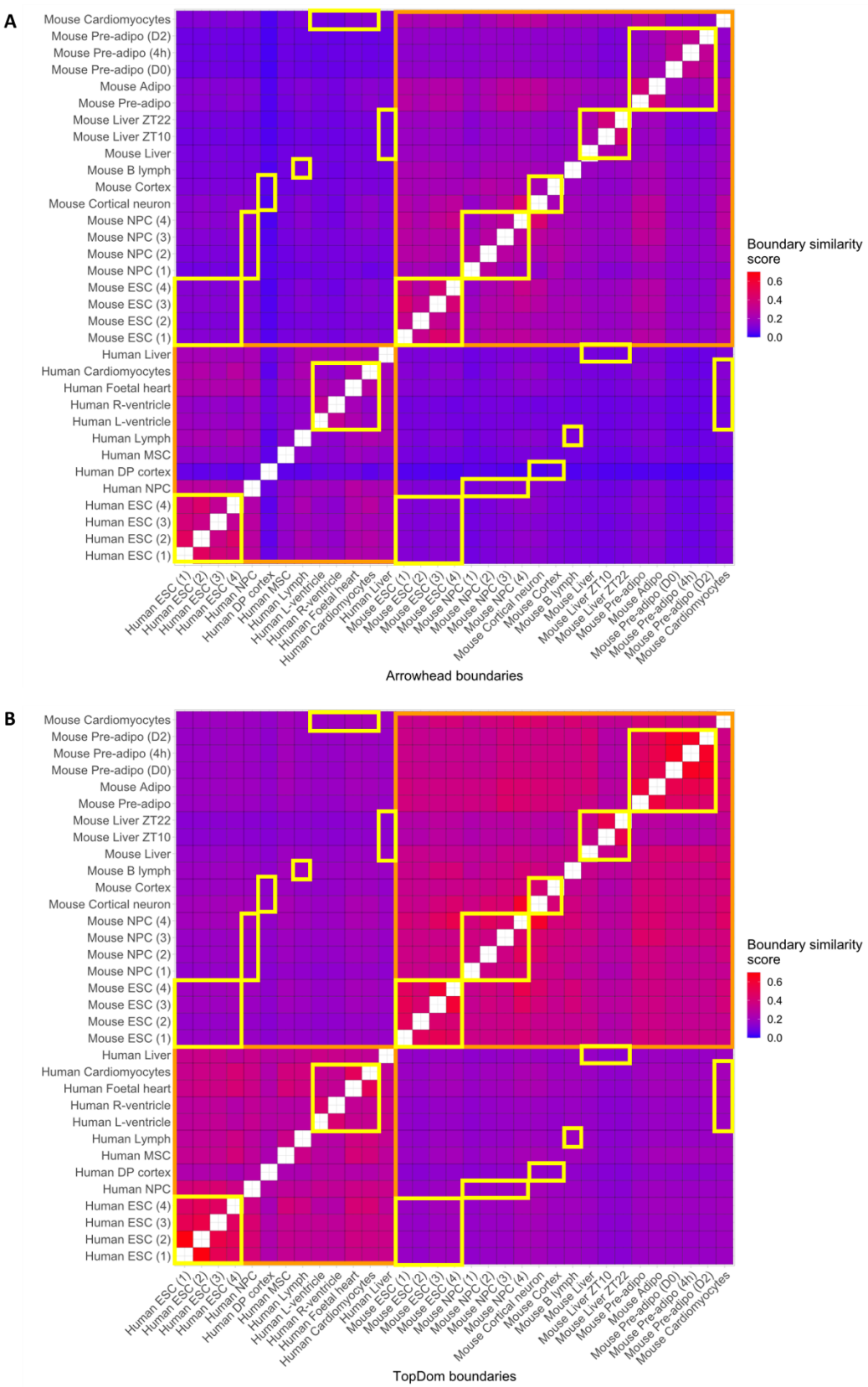


Chapter 2: Supplementary Figure 8: Number of autosomal TADs before and after liftOver between hg19 and mm10. *Arrowhead TADs (top), TopDom TADs (bottom).*



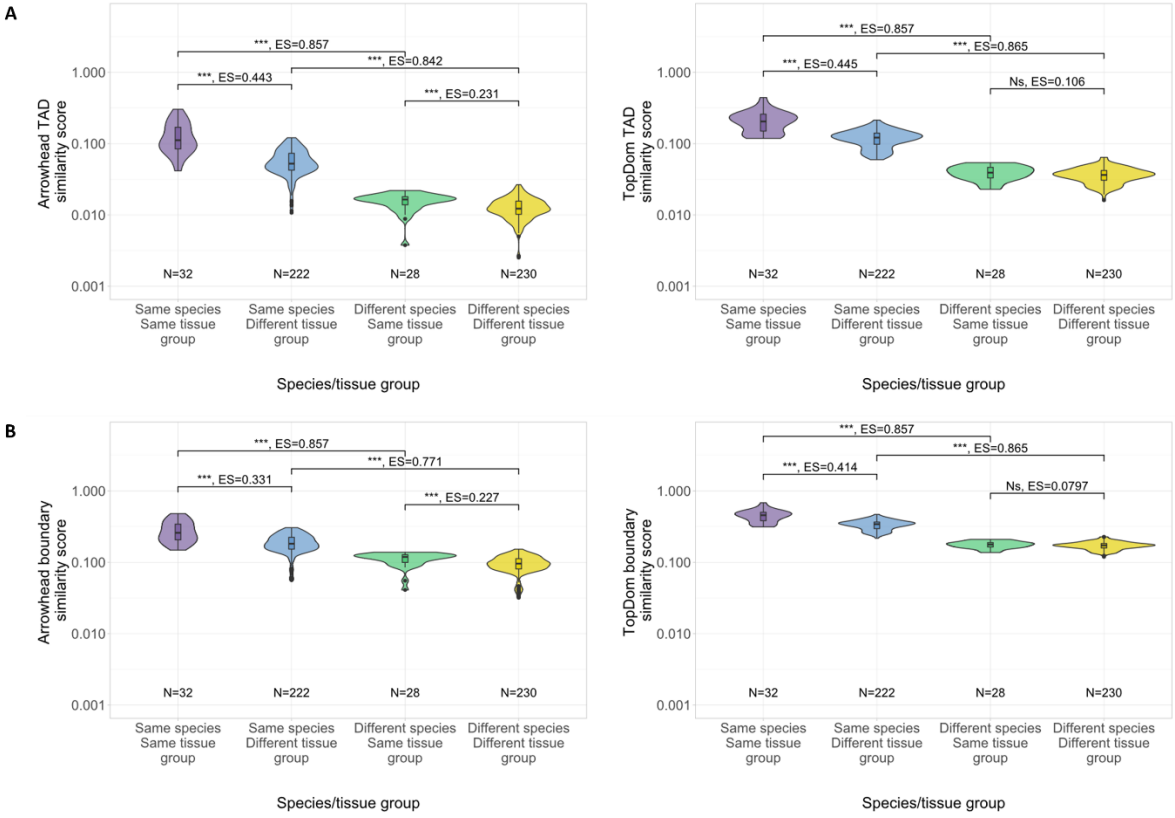
Chapter 2: Supplementary Figure 9: Heat maps indicating pairwise autosomal TAD similarity scores between datasets. *A higher similarity score indicates higher similarity. Pairs of datasets from the same*

species are indicated by orange squares and pairs of datasets in the same tissue/cell type group are indicated by yellow squares.

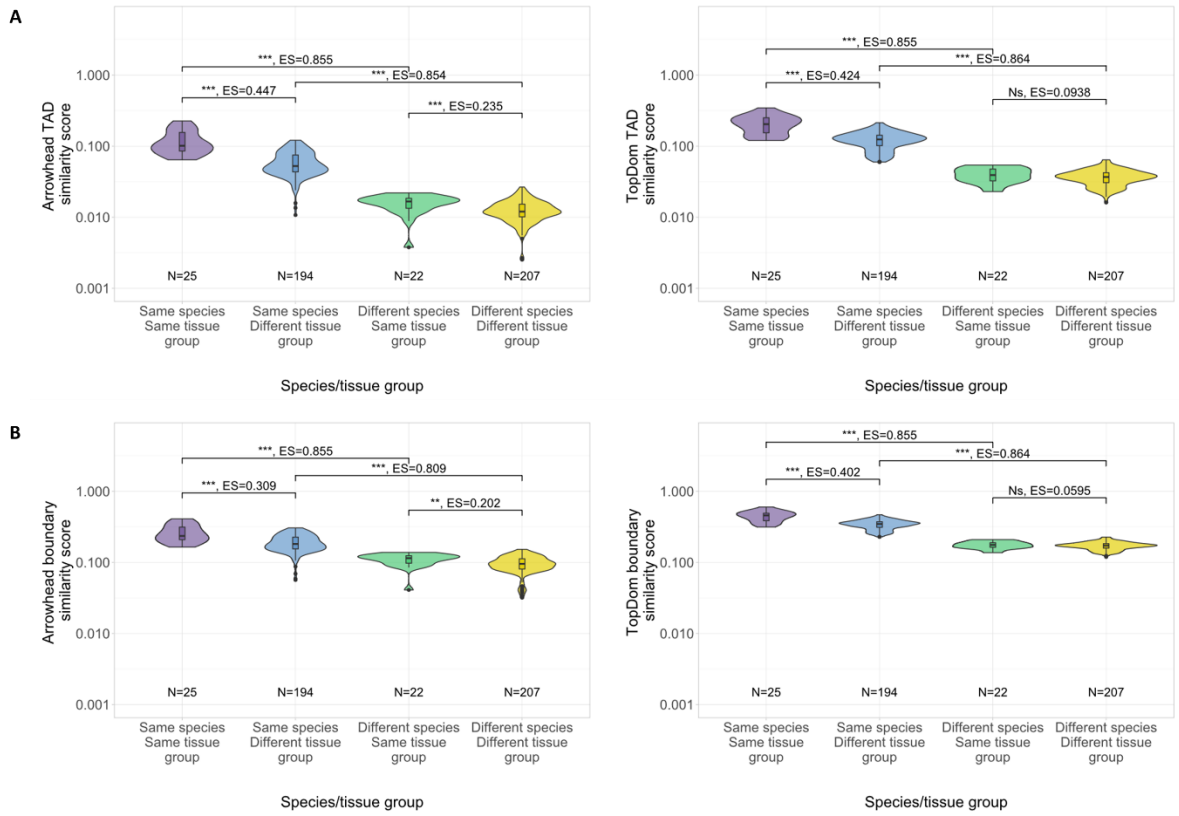


Chapter 2: Supplementary Figure 10: Heat maps indicating pairwise TAD autosomal boundary similarity scores between datasets. *A higher similarity score indicates higher similarity. Pairs of*

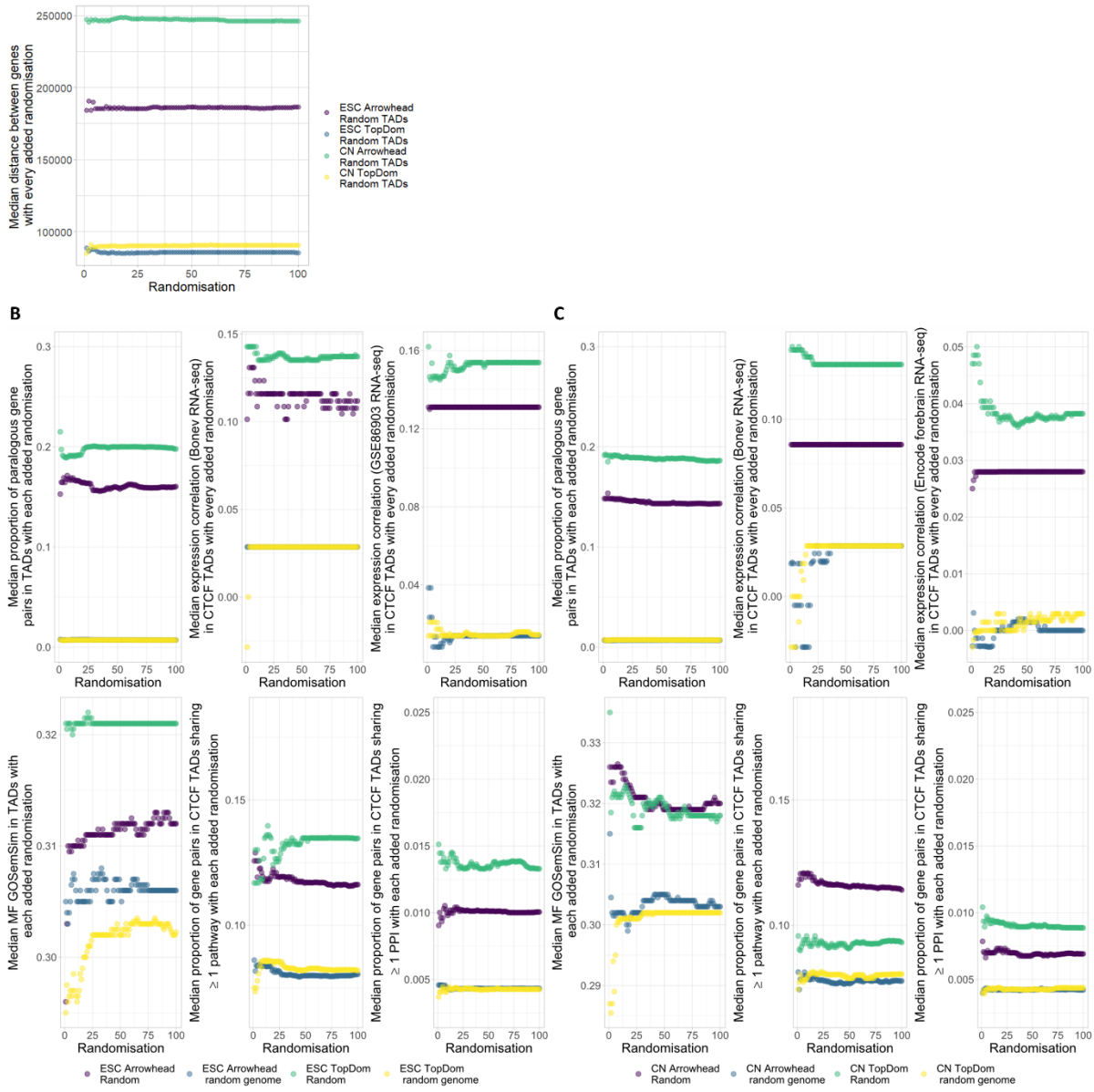
datasets from the same species are indicated by orange squares and pairs of datasets in the same tissue/cell type group are indicated by yellow squares.



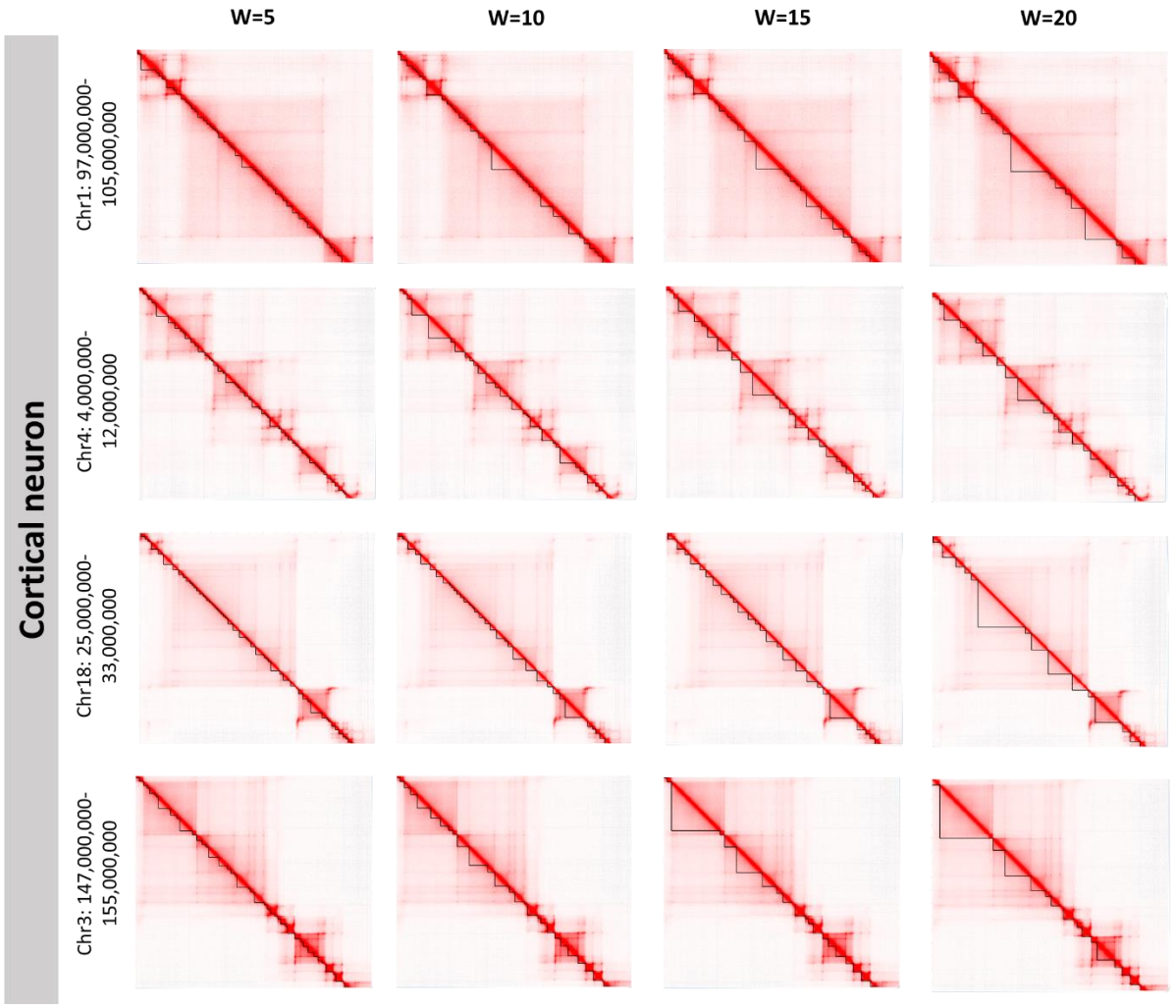
Chapter 2: Supplementary Figure 11: Distribution of autosomal TADs and TAD boundaries similarity scores between pairs of datasets in the same tissue group vs different tissue groups, excluding those from the same data repository. *Left: Arrowhead TADs/boundaries, right: TopDom TADs/boundaries. A higher similarity score indicates higher similarity. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, NS= not significant, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across eight p-values (i.e. within A and B separately). A) TADs. B) TADs boundaries.*



Chapter 2: Supplementary Figure 12: Distribution of autosomal TADs and TAD boundaries similarity scores between pairs of datasets in the same tissue group vs different tissue groups, excluding those with a common first or last author. *Left: Arrowhead TADs/boundaries, right: TopDom TADs/boundaries. A higher similarity score indicates higher similarity. Significance indicated by Wilcoxon test, FDR corrected p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *, NS= not significant, ES= Effect size calculated using r for Wilcoxon. FDR correction applied across eight p-values (i.e. within A and B separately). A) TADs. B) TADs boundaries.*



Chapter 3: Supplementary Figure 1: Median value of measures with each added set of random autosomal TADs. *Measures: A) Median distance between genes. B) For ESC random TADs/random genome TADs: Median proportion of paralogous gene pairs, median expression correlation between pairs of genes calculated using RNA-seq from Bonev et al. or GSE86903 et al., median MF GO semantic similarity between gene pairs, median proportion of pairs of genes sharing ≥1 pathway and median proportion of pairs of genes sharing ≥1 PPI. C) For cortical neuron random TADs/random genome TADs: Median proportion of paralogous gene pairs, median expression correlation between gene pairs calculated using RNA-seq from Bonev et at. or encode forebrain, median MF GO semantic similarity between gene pairs, median proportion of pairs of genes sharing ≥1 pathway and median proportion of pairs of genes sharing ≥1 PPI.*



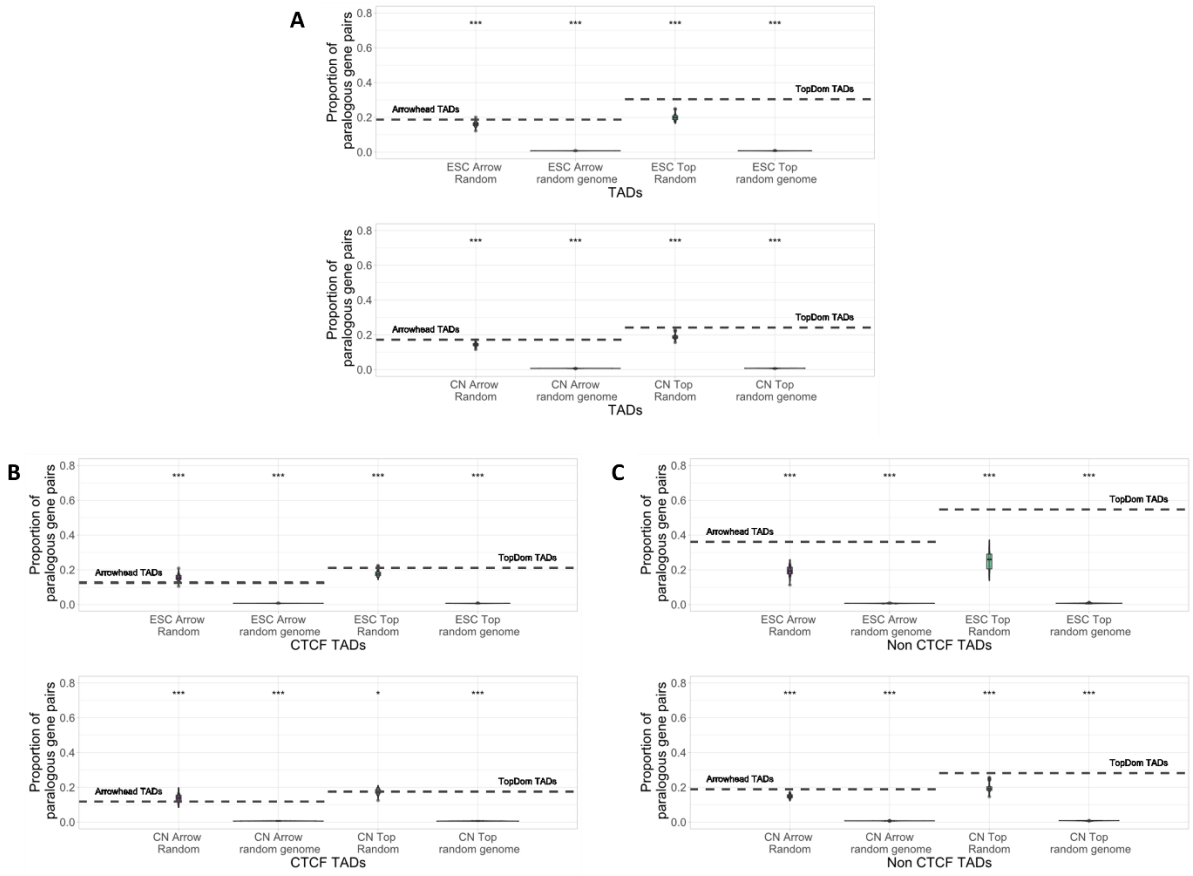
Chapter 3: Supplementary Figure 2: Cortical neuron TADs detected with TopDom and different values of the "w" parameter. *Hi-C maps shown for four example loci. The intensity of red indicates the degree of interaction in the Hi-C map. TADs are annotated as black triangles in lower half of the Hi-C map.*



Chapter 3: Supplementary Figure 3: Comparison of the overlap structure of random cortical neuron TADs produced by my randomisation method and previously published methods (autosomal TADs). *Overlap structure of cortical neuron TADs with an example set of random cortical neuron TADs produced by my method, by the method outlined in Rao et al. or by the method outlined in Nora et al. Non-overlapping TADs were excluded. Remaining TADs were classified based on whether they took part in any "nested" (one TAD falls completely within another) or "non-nested" (one TAD partially overlaps another) overlaps.*

Chapter 3: Supplementary Table 1: Comparison of autosomal CTCF and non CTCF TADs. *Number (top) or percentage (bottom) of CTCF/non CTCF TADs in the focal dataset which have an equivalent CTCF/non CTCF TADs (within ±2bins of both boundaries) in the comparison dataset. For ESC (left) and cortical neuron (right). CTCF TADs are defined as TADs with a CTCF ChIP-seq binding peak within ±10 kb of both TAD boundaries. Non-CTCF TADs are defined as TADs with a CTCF ChIP-seq binding peak within ±10 kb of only one or neither boundary. Compartment annotations are derived from the eigenvector of the first principle component of the Pearson correlation matrix of the Hi-C map.*

ESC				
Focal dataset	**Comparison dataset**			
	ESC Arrowhead CTCF	ESC Arrowhead non CTCF	ESC TopDom CTCF	ESC TopDom non CTCF
ESC Arrowhead CTCF	5206	0	1601	246
ESC Arrowhead non CTCF	0	3165	381	479
ESC TopDom CTCF	1601	381	3481	0
ESC TopDom non CTCF	246	479	0	2469
Focal dataset	**Comparison dataset**			
	ESC Arrowhead CTCF	ESC Arrowhead non CTCF	ESC TopDom CTCF	ESC TopDom non CTCF
ESC Arrowhead CTCF	100.00%	0.00%	30.75%	4.73%
ESC Arrowhead non CTCF	0.00%	100.00%	12.04%	15.13%
ESC TopDom CTCF	45.99%	10.95%	100.00%	0.00%
ESC TopDom non CTCF	9.96%	19.40%	0.00%	100.00%
Cortical neuron				
Focal dataset	**Comparison dataset**			
	CN Arrowhead CTCF	CN Arrowhead non CTCF	CN TopDom CTCF	CN TopDom non CTCF
CN Arrowhead CTCF	2203	0	523	333
CN Arrowhead non CTCF	0	5798	528	1020
CN TopDom CTCF	523	528	1879	0
CN TopDom non CTCF	333	1020	0	3551
Focal dataset	**Comparison dataset**			
	CN Arrowhead CTCF	CN Arrowhead non CTCF	CN TopDom CTCF	CN TopDom non CTCF
CN Arrowhead CTCF	100.00%	0.00%	23.74%	15.12%
CN Arrowhead non CTCF	0.00%	100.00%	9.11%	17.59%
CN TopDom CTCF	27.83%	28.10%	100.00%	0.00%
CN TopDom non CTCF	9.38%	28.72%	0.00%	100.00%



Chapter 3: Supplementary Figure 4: Proportion of paralogous gene pairs in TADs vs proportion in random TADs (autosomal). Equivalent figure to Figure 3.9 but showing the full distribution of proportions for random TADs and random genome TADs. *Significance indicated by median p-value of Fisher's exact test between real TADs and each of 100 sets of random TADs. FDR correction applied across eight median p-values (i.e. within A, B and C separately). FDR corrected median p-value: p<0.001 = ***, p<0.01 = **, p<0.05 = *. A) Proportion in all autosomal TADs (dotted line) and distributions of proportions in 100 sets each of random TADs and random genome TADs. B) Proportion in CTCF autosomal TADs (dotted line) and distributions of proportions in 100 sets each of random CTCF TADs and random genome CTCF TADs. C) Proportion in non-CTCF autosomal TADs (dotted line) and distributions of proportions in 100 sets each of random non-CTCF TADs and random genome non-CTCF TADs.*

Chapter 3: Supplementary Table 2: Effect sizes corresponding to the p-values displayed in Figure 12: Average constraint of genes sharing an autosomal TAD.

No. of genes in TAD	No. of genes in TAD	ESC Arrowhead TADs Effect size	CN Arrowhead TADs Effect size	ESC TopDom TADs Effect size	CN TopDom TADs Effect size
1	**2**	0.07	0.0919	0.0634	0.0699
1	**3**	0.0879	0.0988	0.105	0.0577
1	**4**	0.102	0.123	0.122	0.123
1	**5**	0.129	0.108	0.15	0.0803
1	**6+**	0.138	0.129	0.205	0.148
2	**3**	0.0224	0.00766	0.0424	-0.0144
2	**4**	0.0438	0.0394	0.0681	0.062
2	**5**	0.0838	0.0356	0.11	0.0169
2	**6+**	0.0582	0.0188	0.15	0.0748
3	**4**	0.0252	0.0363	0.0374	0.0861
3	**5**	0.0722	0.0318	0.0895	0.0357
3	**6+**	0.0325	0.0122	0.113	0.0911
4	**5**	0.0508	-0.00178	0.0649	-0.0502
4	**6+**	0.0026	-0.027	0.0763	0.00547
5	**6+**	-0.0419	-0.0197	0.0145	0.0462

Chapter 3: Supplementary Table 3: Effect sizes corresponding to the p-values displayed for autosomal CTCF TADs in Figure 14: Average constraint of genes sharing an autosomal CTCF TADs or non-CTCF TADs.

No. of genes in TAD	No. of genes in TAD	ESC Arrowhead TADs Effect size	CN Arrowhead TADs Effect size	ESC TopDom TADs Effect size	CN TopDom TADs Effect size
1	**2**	0.0756	0.0721	0.0535	0.103
1	**3**	0.126	0.0859	0.131	0.0855
1	**4**	0.137	0.186	0.129	0.117
1	**5**	0.173	0.122	0.162	0.0913
1	**6+**	0.132	0.142	0.187	0.146
2	**3**	0.0566	0.0126	0.076	-0.0167
2	**4**	0.0761	0.145	0.0748	0.0237
2	**5**	0.128	0.0767	0.123	-0.00956
2	**6+**	0.0443	0.0559	0.131	0.0305
3	**4**	0.0236	0.143	0.00941	0.0461
3	**5**	0.0859	0.0699	0.0727	0.0133
3	**6+**	-0.0151	0.0451	0.0593	0.0531
4	**5**	0.0675	-0.0587	0.0774	-0.0376
4	**6+**	-0.0466	-0.107	0.0482	-0.00265
5	**6+**	-0.103	-0.0305	-0.0251	0.0338

Chapter 3: Supplementary Table 4: Effect sizes corresponding to the p-values displayed for autosomal non-CTCF TADs in Figure 14: Average constraint of genes sharing an autosomal CTCF TADs or non-CTCF TADs.

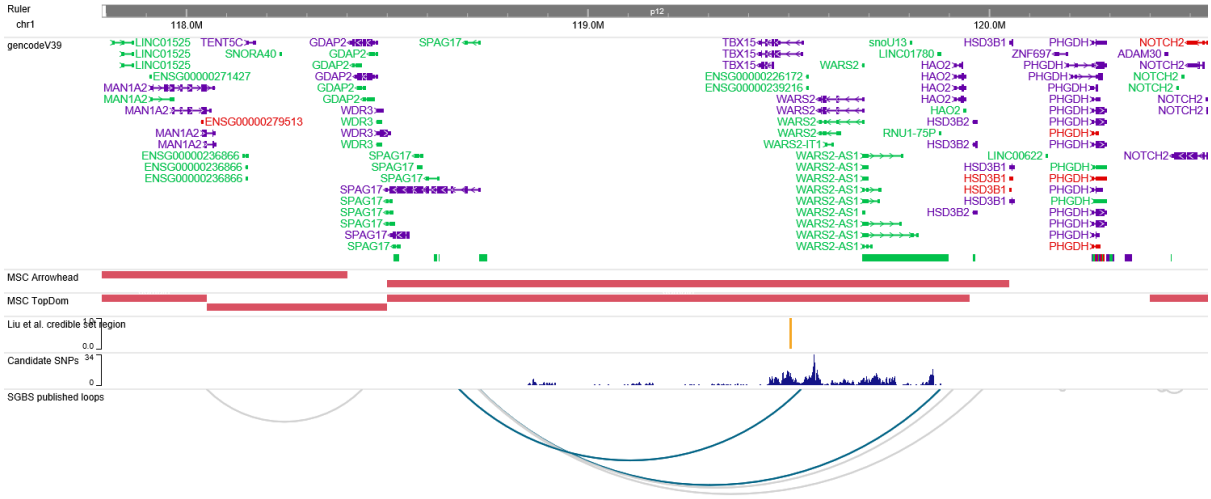
No. of genes in TAD	No. of genes in TAD	ESC Arrowhead TADs Effect size	CN Arrowhead TADs Effect size	ESC TopDom TADs Effect size	CN TopDom TADs Effect size
1	**2**	0.0734	0.0997	0.122	0.0523
1	**3**	0.0232	0.104	0.0923	0.0415
1	**4**	0.0494	0.0945	0.176	0.123
1	**5**	0.047	0.0995	0.159	0.0787
1	**6+**	0.219	0.12	0.34	0.151
2	**3**	-0.0532	0.00423	-0.0111	-0.015
2	**4**	-0.0205	-0.0055	0.107	0.0811
2	**5**	-0.0193	0.0138	0.0926	0.0357
2	**6+**	0.164	0.000185	0.285	0.101
3	**4**	0.0328	-0.0101	0.13	0.106
3	**5**	0.0282	0.0112	0.112	0.0553
3	**6+**	0.213	-0.00378	0.284	0.116
4	**5**	0.00313	0.0229	-0.00926	-0.0451
4	**6+**	0.168	0.00275	0.178	0.0169
5	**6+**	0.156	-0.0155	0.178	0.0485

Chapter 4: Supplementary Table 1: Method used to generate the bait map required to run ChiCAGO for each dataset used.

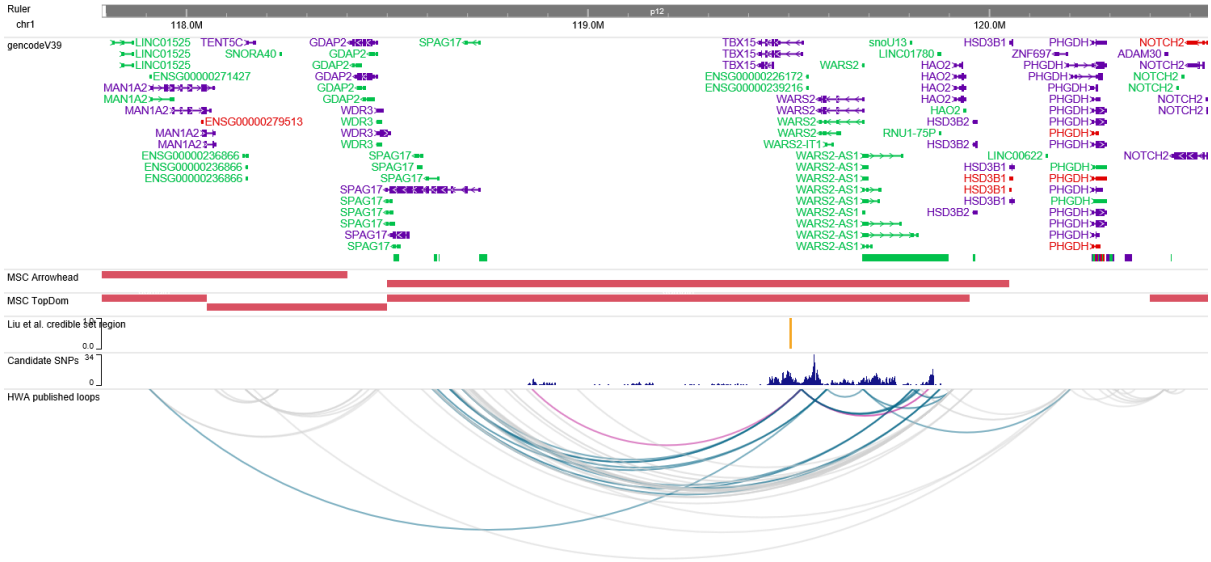
Datasets	Bait maps
Pan et al. 2018: Human white adipose	In Pan et al. 2018 it states: "The RNA baits were designed in Mifsud et al. ", therefore the bait coordinates were downloaded and extracted from the supplementary of Mifsud et al. 2015 and the bait file was generated as the subset of hg19 HindIII restriction fragments (generated with HiCUP) which fully contained a bait.
Madsen et al. 2020: Human MSC differentiating towards adipocytes	The bait coordinates were downloaded from the supplementary of Madsen et al. 2020 and the bait file was generated as the subset of hg19 MboI restriction fragments (generated with HiCUP) which fully contained a bait.
Sobreira et al. 2021: Human SGBS preadipocytes	In Sobreira et al. 2021 it states: “The in situ PCHi-C libraries were performed as in Montefiore et al.”, therefore the bait coordinates were downloaded and extracted from the supplementary of Montefiori et al. 2018 and the bait file was generated as the subset of hg19 MboI restriction fragments (generated with HiCUP) which fully contained a bait.
Siersbæk et al. 2017: Mouse 3T3-L1 differentiating towards adipocytes	In Siersbæk et al. 2017 it states: “39,021 biotinylated 120-mer RNA baits were designed to the ends of HindIII restriction fragments that overlap 25,747 Ensembl-annotated promoters of protein-coding, noncoding, antisense, snRNA, miRNA and snoRNA transcripts (Schoenfelder et al. 2015).”, therefore bait coordinates (mm9) were downloaded and extracted from the supplementary of Schoenfelder et al. 2015. The bait coordinates were then converted to mm10 using liftover (Hinrichs et al. 2006). The bait file was generated as the subset of mm10 HindIII restriction fragments (generated with HiCUP) which fully contained a bait.

Chapter 4: Supplementary Table 2: GTEx adipose eQTLS involving SNPs from the credible set defined in Liu et al. *All three SNPs are involved in the same eQTLs in both tissues.*

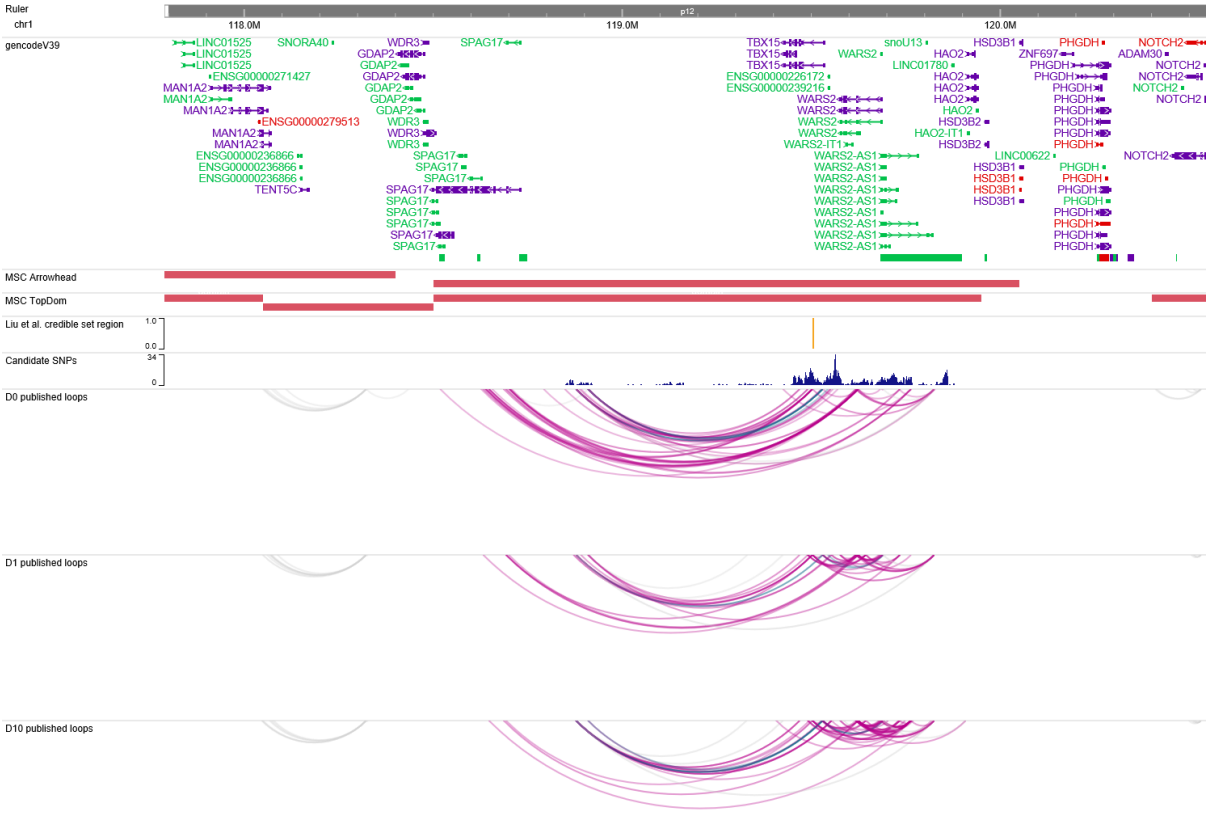
Gencode Id	Gene Symbol	SNP Id	P-Value	NES	Tissue
ENSG00000116874.11	WARS2	rs10923712	6.80E-38	-0.49	Adipose - Subcutaneous
ENSG00000226172.2	RP4-712E4.1	rs10923712	1.10E-19	-0.44	Adipose - Subcutaneous
ENSG00000231365.5	RP11-418J17.1	rs10923712	2.70E-12	-0.3	Adipose - Subcutaneous
ENSG00000092607.13	TBX15	rs10923712	5.00E-08	0.14	Adipose - Subcutaneous
ENSG00000116874.11	WARS2	rs10923712	1.60E-21	-0.39	Adipose - Visceral (Omentum)
ENSG00000092607.13	TBX15	rs10923712	8.40E-11	0.24	Adipose - Visceral (Omentum)
ENSG00000231365.5	RP11-418J17.1	rs10923712	1.80E-10	-0.26	Adipose - Visceral (Omentum)
ENSG00000116874.11	WARS2	rs984222	6.80E-38	-0.49	Adipose - Subcutaneous
ENSG00000226172.2	RP4-712E4.1	rs984222	1.10E-19	-0.44	Adipose - Subcutaneous
ENSG00000231365.5	RP11-418J17.1	rs984222	2.70E-12	-0.3	Adipose - Subcutaneous
ENSG00000092607.13	TBX15	rs984222	5.00E-08	0.14	Adipose - Subcutaneous
ENSG00000116874.11	WARS2	rs984222	1.60E-21	-0.39	Adipose - Visceral (Omentum)
ENSG00000092607.13	TBX15	rs984222	8.40E-11	0.24	Adipose - Visceral (Omentum)
ENSG00000231365.5	RP11-418J17.1	rs984222	1.80E-10	-0.26	Adipose - Visceral (Omentum)
ENSG00000116874.11	WARS2	rs984225	6.80E-38	-0.49	Adipose - Subcutaneous
ENSG00000226172.2	RP4-712E4.1	rs984225	1.10E-19	-0.44	Adipose - Subcutaneous
ENSG00000231365.5	RP11-418J17.1	rs984225	2.70E-12	-0.3	Adipose - Subcutaneous
ENSG00000092607.13	TBX15	rs984225	5.00E-08	0.14	Adipose - Subcutaneous
ENSG00000116874.11	WARS2	rs984225	1.60E-21	-0.39	Adipose - Visceral (Omentum)
ENSG00000092607.13	TBX15	rs984225	8.40E-11	0.24	Adipose - Visceral (Omentum)
ENSG00000231365.5	RP11-418J17.1	rs984225	1.80E-10	-0.26	Adipose - Visceral (Omentum)



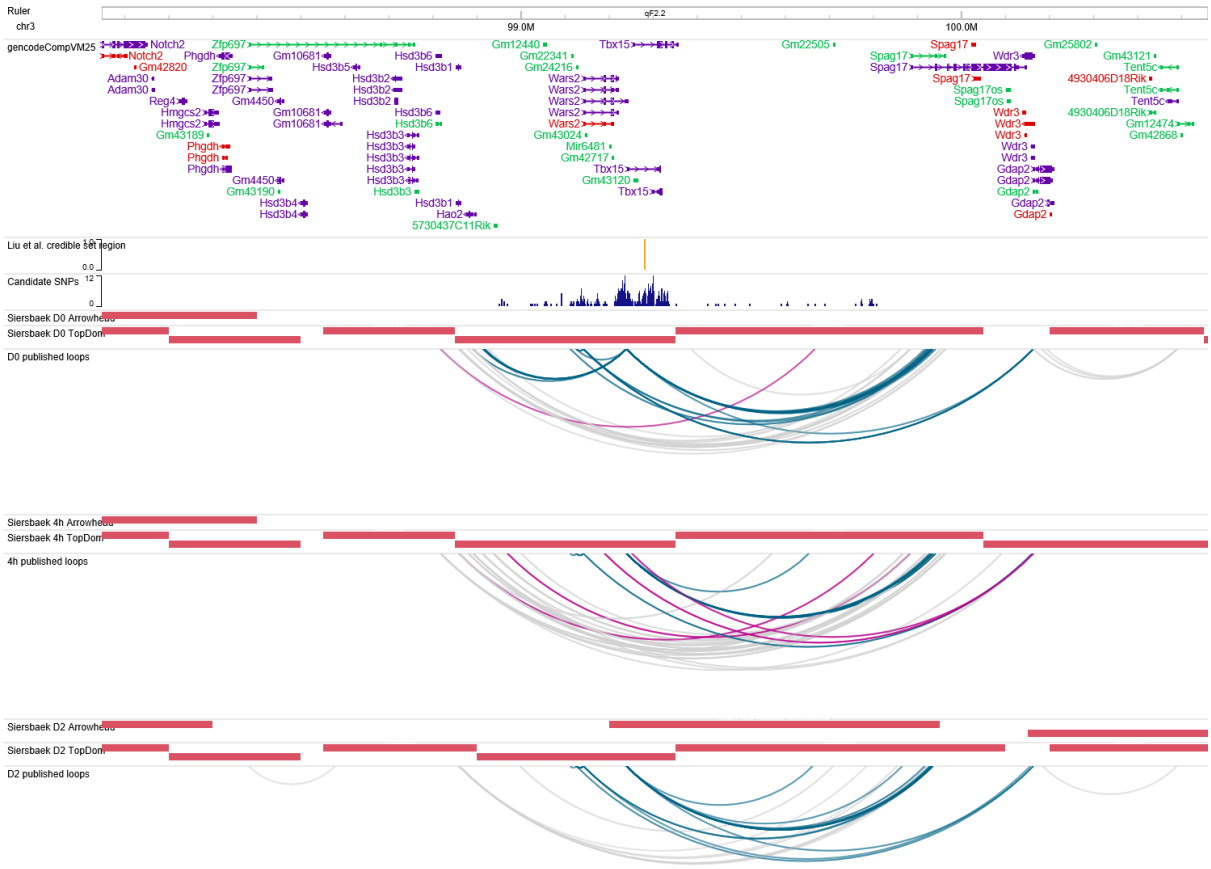
Chapter 4: Supplementary Figure 1: Published loops in SGBS in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and chromatin loops (other SNP loops: blue and non SNP loops: grey).*



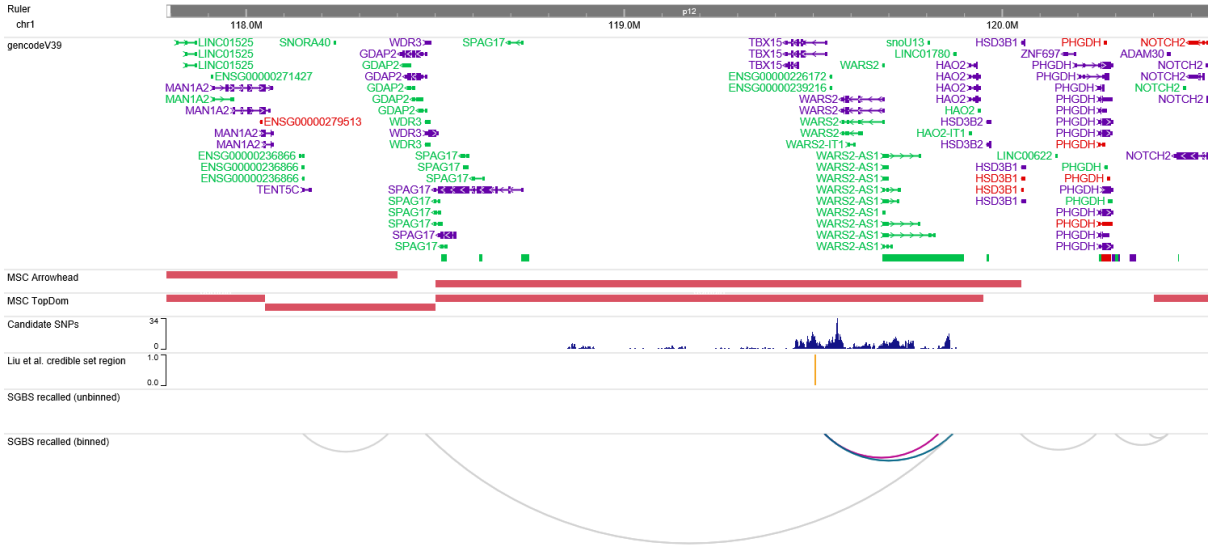
Chapter 4: Supplementary Figure 2: Published loops in HWA in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and chromatin loops (potential regulatory SNP loops: pink, other SNP loops: blue and non SNP loops: grey).*



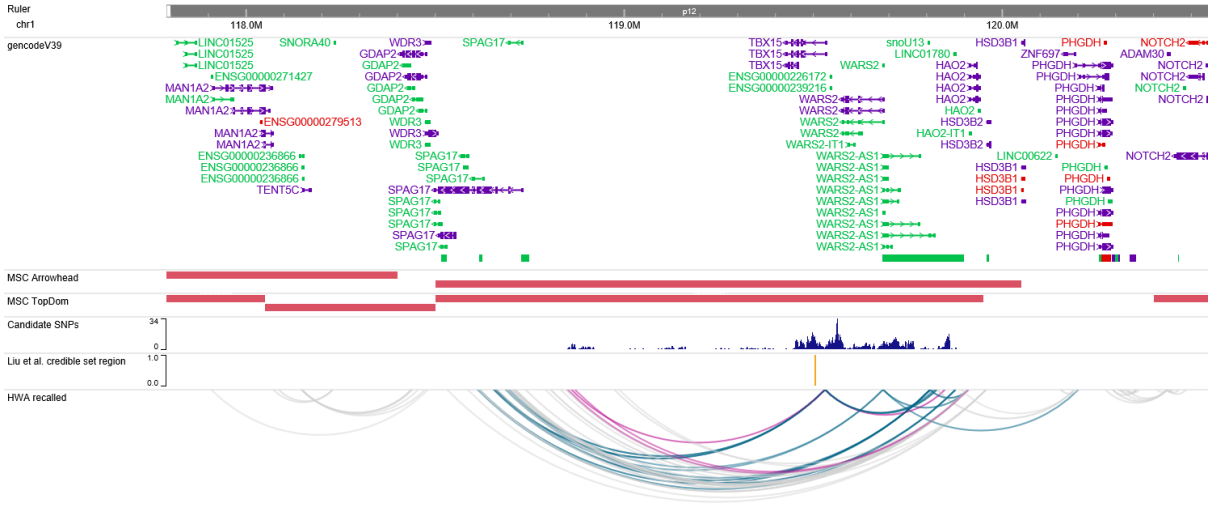
Chapter 4: Supplementary Figure 3: Published loops in human MSC differentiating towards adipocytes in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and chromatin loops at day 0, day 1 and day 10 (potential regulatory SNP loops: pink, other SNP loops: blue and non SNP loops: grey).*



Chapter 4: Supplementary Figure 4: Published loops in mouse pre-adipocytes differentiating towards adipocytes in the Tbx15-Wars2 locus. *WASHU genome browser view displaying the Tbx15-Wars2 locus and surrounding area. Tracks display: gencode gene annotations, TADs called with Arrowhead and TopDom, the equivalent location of the human credible set region after mapping to the mouse genome, the equivalent positions of all human candidate SNPs after mapping to the mouse genome (shown as a density track, any SNPs whose location could not be mapped to the mouse genome have been excluded) and chromatin loops at day 0, 4 hours and day 2 (potential regulatory SNP loops: pink, other SNP loops: blue and non SNP loops: grey).*



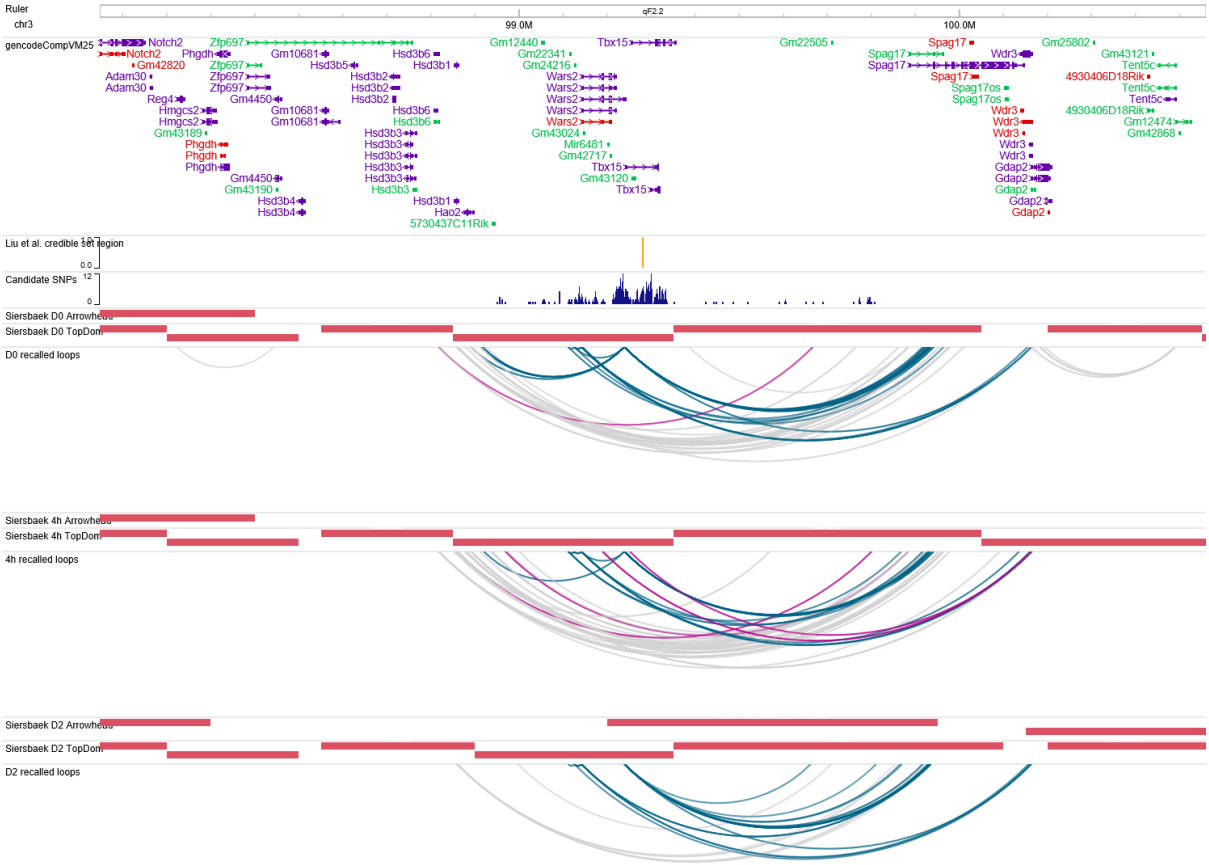
Chapter 4: Supplementary Figure 5: Recalled loops in SGBS in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and chromatin loops detected with un-binned and binned data (potential regulatory SNP loops: pink, other SNP loops: blue and non SNP loops: grey).*



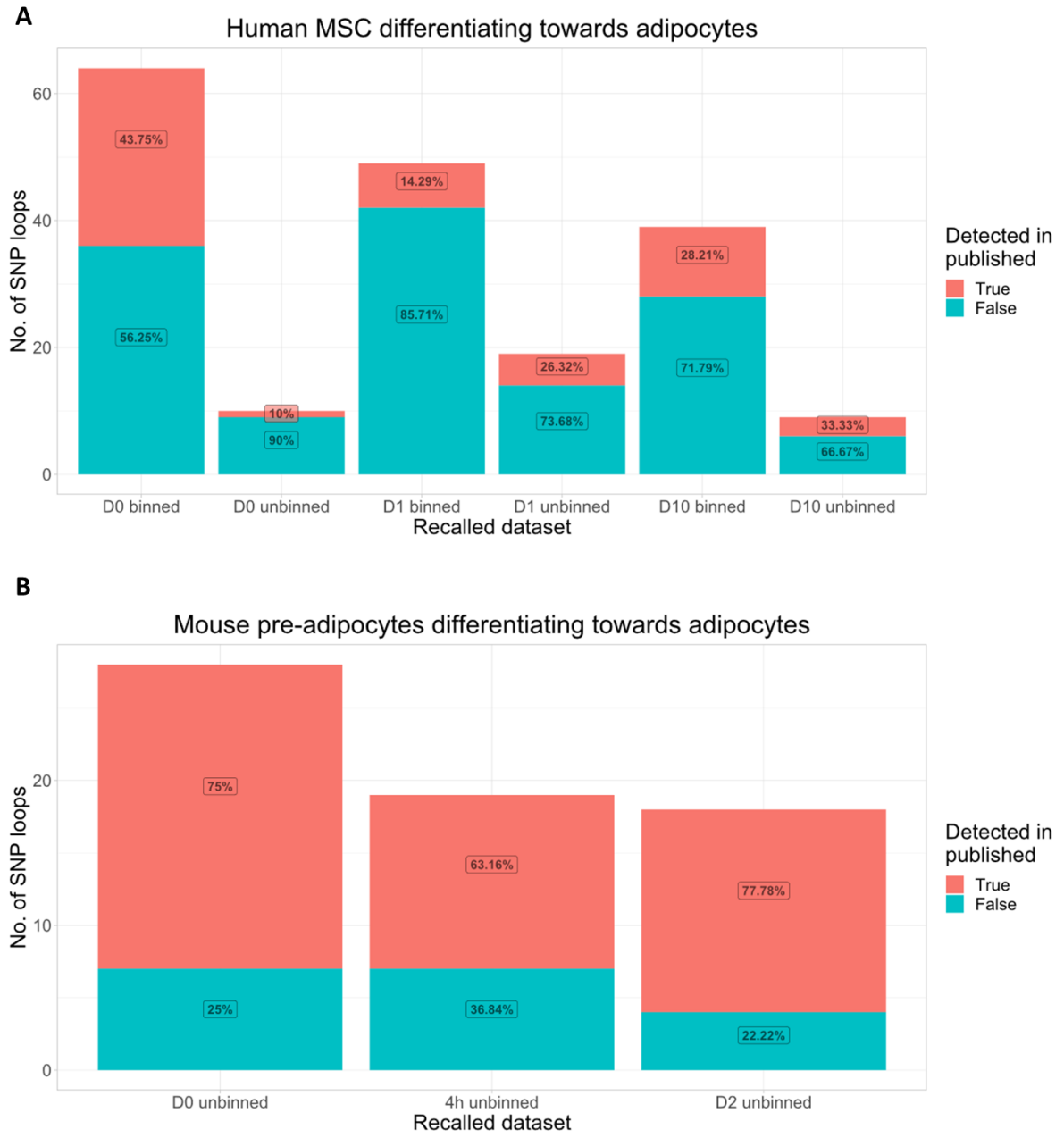
***Chapter 4: Supplementary Figure 6**: **Recalled loops in HWA in the TBX15-WARS2 locus.** WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and chromatin loops detected with un-binned and binned data (potential regulatory SNP loops: pink, other SNP loops: blue and non SNP loops: grey).*



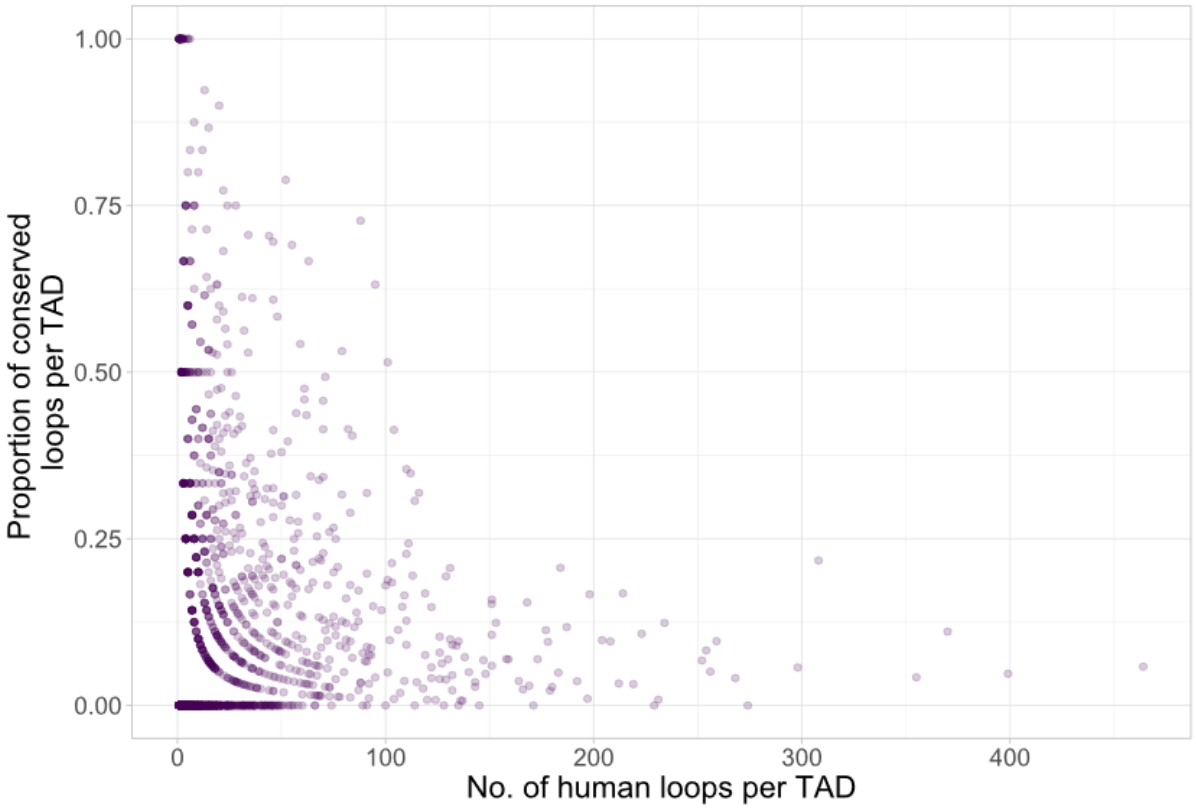
Chapter 4: Supplementary Figure 7: Recalled loops in Human MSC differentiating towards adipocytes in the TBX15-WARS2 locus. *WASHU genome browser view displaying the TBX15-WARS2 locus and surrounding area. Tracks display: gencode gene annotations, MSC TADs called with Arrowhead and TopDom, the location of the credible set region, the positions of all candidate SNPs (shown as a density track) and chromatin loops at day 0, day 1 and day 10 called detected with un-binned and binned data (potential regulatory SNP loops: pink, other SNP loops: blue and non SNP loops: grey).*



Chapter 4: Supplementary Figure 8: Recalled loops in mouse pre-adipocytes differentiating towards adipocytes in the Tbx15-Wars2 locus. *WASHU genome browser view displaying the Tbx15-Wars2 locus and surrounding area. Tracks display: gencode gene annotations, TADs called with Arrowhead and TopDom, the equivalent location of the human credible set region after mapping to the mouse genome, the equivalent positions of all human candidate SNPs after mapping to the mouse genome (shown as a density track, any SNPs whose location could not be mapped to the mouse genome have been excluded) and chromatin loops at day 0, 4 hours and day 2 (potential regulatory SNP loops: pink, other SNP loops: blue and non SNP loops: grey).*



Chapter 4: Supplementary Figure 9: Number of loops in recalled datasets and percentage also detected in published loops. *A) Human MSC differentiating towards adipocytes. B) Mouse pre-adipocytes differentiating towards adipocytes.*



Chapter 4: Supplementary Figure 10: Correlation between the total number of human loops and the proportion of conserved loops per TAD.