

UNDERSTANDING THE EPIGENOME USING SYSTEM GENETICS

SANDER WILLEM TIMMER



This thesis is submitted for the degree of Doctor of Philosophy

Clare Hall College
University of Cambridge

29 October 2013

DECLARATION

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This dissertation does not exceed the specified length limit of 60000 words as defined by the Biology Degree Committee.

This dissertation has been typeset using $\text{\LaTeX}2_{\epsilon}$ in 12 pt Palatino, one and half spaced, according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

Cambridge, 29 October 2013

Sander Willem Timmer

To my family

*It is important in life not to be strong, but to feel strong,
to measure yourself at least once.
If you want something in life, reach out and grab it.*

Christopher McCandless

ACKNOWLEDGMENTS

First and foremost, thanks are due to my supervisor Ewan Birney. His enthusiastic guidance into my scientific journey was tremendously inspiring and crucial to make this thesis possible. In addition, I also want to specially mention the great supervision of Ian Dunham. Both Ewan and Ian were essential to get this thesis done.

This thesis would not have been finished without all the proofreading and comments made by Ewan Birney, Ian Dunham, Nenad Bartonicek, Nils Kölling, and Angela Goncalves. In addition, the post-viva suggestions by Carl Anderson and Mark McCarthy helped improve this work tremendously. Thank you all for the useful comments, discussion, and time you spent sorting me out.

I had the pleasure of sharing my office with a great assortment people over the years. I'd like to thank my fellow office members: Mikhail Spivakov, Albert Vilella, Dace Ruklisa, Markus Fritz, Petra Schwalie, Andre Faure, Nils Kölling, Thomas Rensch, Sandro Morganella, Camille Berthelot, Helena Kilpinen, Valentina Iotchkova, Hannah Meyer, and Leland Taylor for the great times.

The EMBL world is full with delightful people and it was fun to be a part of this European flagship experience. I made many new friends within the Predoc community. The name Gerard will never be the same.

I also would like to thank my housemates at the Wiggle Mansion: Nenad Bartonicek, Christine Seeliger, Nils Kölling, and

Maria Xenophontos. Together with heaps of porridge, bacon, pizza and gallons of gin we kept the Wiggle machine running.

Cambridge is a special place and I own many good memories of it. It is the place of great academic inspiration, the Regal, the Hopbine, and rowing. I am forever in debt with the rowers of Clare Hall Boat Club with whom I spent countless hours in the early-morning paddling up-and-down the River Cam. It was a great experience to defend the honour of our college during Bumps. It has been a great pleasure living in the shire!

To my French with benefits, Marion Libouban; thank you for being so kind and patient while I was grumpily typing away on this thesis.

Finally, a very special thanks go to my parents, sister, and brother. Thank you for your love, support, encouragement, and generosity.

*Cambridge,
29 October 2013*

UNDERSTANDING THE EPIGENOME USING SYSTEM GENETICS

Sander Willem Timmer

Genetics has been successful in associating DNA sequence variants to both dichotomous and continuous traits in a variety of organisms, from plant and farm animal studies to human disease. With the advent of high-throughput genotyping, there has been an almost routine generation of genome-wide association studies (GWAS) between human disease traits and genomic regions. Despite this success, a particular frustration is that the majority of associated loci are in non-coding regions of the genome and thus interpretation is hard.

To improve characterisation of non-coding regions, molecular assays can be used as a phenotype, and subsequently be used to explain how genetics alter molecular mechanisms. In this thesis, the interplay of three molecular assays that are involved in regulating gene expression is studied. On 60 individuals, several assays are performed: FAIRE-chip, CTCF-seq, RNA-seq and DNA-seq.

In the first part, the discovery and characteristics of FAIRE-QTLs is presented. The identified FAIRE-QTLs show strong overlap with other molecular QTLs, histone modifications, and transcription factors.

The second part consists of the integration of genome-wide molecular assays in a human population to reconstruct the human epigenome. Each of the molecular assays is associated with each of the other assays to discover phenotype-to-phenotype correlations. Furthermore, QTL data are used to dissect the causality for these phenotype-to-phenotype correlations in a system genetic manner.

The third part presents a comprehensive view of CTCF binding on the X chromosome, and its implications for X-chromosome inactivation. A novel X chromosome-wide CTCF effect is observed. Using the gender of each of the cell lines, observations are made about which CTCF sites are dosage-compensated, active on both chromosomes, or are only bound in females.

CONTENTS

i	INTRODUCTION	1
1	INTRODUCTION	3
1.1	History of genetics	4
1.1.1	Inheritance	4
1.1.2	Inheritance is by genetics	5
1.1.3	The modern genetic heritability era	6
1.2	The Human epigenome	8
1.3	CCCTC-binding factor (CTCF)	11
1.3.1	Cohesin	12
1.3.2	CTCF loops	12
1.4	The sex chromosomes	13
1.4.1	Discovery of X-chromosome inactivation	14
1.4.2	The X-inactivation center	15
1.4.3	CTCF in X-chromosome inactivation	15
1.5	Cell lines for molecular phenotyping	16
1.5.1	Lymphoblastoid cell lines	17
1.6	Quantification of molecular phenotypes	18
1.6.1	Microarrays and tiling arrays	18
1.6.2	Next generation sequencing for molecular quantification	20
1.7	Chromatin conformation and TF binding affinity	21
1.7.1	DNase I hypersensitive sites	23
1.7.2	Formaldehyde-Assisted Isolation of Regulatory Elements	24
1.7.3	Assay for Transposase-Accessible Chromatin	24
1.7.4	Comparison of chromatin accessibility assays	25
1.8	Discovery of quantitative heritable traits	26

1.8.1	History of QTL mapping	26
1.8.2	Molecular QTLs	27
1.8.3	QTL mapping	28
1.8.4	Multiple testing problem	30
1.9	Scope of this thesis	31
ii	DISCOVERY OF QUANTITATIVE TRAITS	33
2	FINDING QTLs FOR CHROMATIN ACCESSIBILITY	35
2.1	Introduction	35
2.2	Results	36
2.2.1	FAIRE intensity data for 60 CEU individuals	36
2.2.2	Filtering out intra-individual variance	37
2.2.3	Filtering polymorphic probes	43
2.2.4	Properties of the filtered FAIRE probes	44
2.2.5	Finding proximal QTLs	46
2.2.6	Functional analysis of FAIRE-QTLs	49
2.2.7	Comparison of FAIRE associations to other association studies	54
2.2.8	Low coverage sequencing data to improve QTL discovery	59
2.3	Conclusion	64
2.4	Data	67
2.4.1	Genotype data	67
2.4.2	FAIRE-chip data	68
2.4.3	Expression-QTLs	68
2.4.4	GWAS and molecular QTL datasets	68
2.5	Methods	69
2.5.1	Robust Multi-array Average	69
2.5.2	Filtering FAIRE-probes	69
2.5.3	Testing for associations	74
2.5.4	Determining significance of associations	74
2.5.5	Identifying of FAIRE-QTLs previously identified in other phenotypes	75

2.5.6	Identifying known biochemical ENCODE regions using RegulomeDB	75
2.5.7	Regulome DB score	76
2.5.8	Data visualisation	77

iii DISSECTING EPIGENETIC CORRELATIONS USING GENETICS 79

3	CORRELATING MOLECULAR PHENOTYPES	83
3.1	Introduction	83
3.1.1	Molecular interactions	84
3.2	Results	86
3.2.1	Correlating assays	86
3.2.2	CTCF to RNA	87
3.2.3	CTCF to FAIRE	93
3.2.4	CTCF to FAIRE probes	97
3.2.5	FAIRE probes to RNA	105
3.2.6	CTCF to CTCF	110
3.2.7	Validating CTCF-CTCF pairs using ChIA-PET	116
3.2.8	Molecular assays are highly correlating	122
3.3	Conclusions	125
3.3.1	The molecular pathway regulating transcription	125
3.3.2	Lengths of CTCF-CTCF pairs suggest different roles	126
3.4	Data	127
3.4.1	CTCF	127
3.4.2	FAIRE	129
3.4.3	RNA-seq	129
3.4.4	ENCODE datasets on GM12878	130
3.5	Methods	136
3.5.1	Testing for associations	136
3.5.2	Testing associations for significance	136
3.5.3	Comparing CTCF-CTCF pairs with ChIA-PET data	137

3.5.4	Overlapping assays	137
3.5.5	Data visualisation	138
4	THE DRIVER OF EPIGENETIC CORRELATIONS	139
4.1	Introduction	139
4.1.1	Inference of causality	140
4.1.2	Direct versus indirect causality	141
4.2	Results	148
4.2.1	Random direction of causality	148
4.2.2	CTCF to FAIRE	150
4.2.3	CTCF to RNA	155
4.2.4	CTCF to CTCF	160
4.3	Conclusions	166
4.3.1	Common direction of causality	166
4.3.2	CTCF loops	168
4.4	Data	169
4.4.1	Molecular QTLs	169
4.4.2	Karyotypes	169
4.5	Methods	170
4.5.1	Testing for associations	170
4.5.2	Data visualisation	172
iv	X-CHROMOSOME BIOLOGY	173
5	CTCF INVOLVEMENT IN X-CHROMOSOME INACTIVATION	175
5.1	Introduction	175
5.2	Results	176
5.2.1	CTCF binding on chromosome X has a gender-specific profile	176
5.2.2	Overall lack of dosage compensation for CTCF	178
5.2.3	CTCF site classification	180
5.2.4	Biochemical enrichment of different CTCF classes	184
5.2.5	Binding signal and spacing of TF and HM	188
5.2.6	X-chromosome inactivation centre	194

5.2.7	Female specific CTCF sites	198
5.3	Conclusion	200
5.4	Data	202
5.4.1	CTCF	202
5.4.2	DNase I	202
5.4.3	RNA-seq	202
5.5	Methods	203
5.5.1	Classification of CTCF sites	203
5.5.2	Biochemical co-activity with CTCF sites	204
5.5.3	Aggregation plots	204
5.5.4	Data visualisation	205
v	CONCLUSIONS	207
6	FINAL REMARKS AND FUTURE DIRECTIONS	209
6.1	Chromatin as a heritable trait	209
6.2	System genetics	210
6.3	CTCF binding on the X chromosome	213
6.4	Integration of epigenetic data	214
6.4.1	Single molecular level	215
6.5	The epigenome era	217
vi	APPENDIX	219
A	INDIVIDUALS USED IN THIS THESIS	221
A.0.1	Overlap assays	221
B	OVERVIEW ALL FAIRE-QTLS	225
	REFERENCES	241

LIST OF FIGURES

Figure 2.1	FAIRE tiling array design	38
Figure 2.2	FAIRE-probes intra-individual variance	40
Figure 2.3	FAIRE-probes inter-individual variance	41
Figure 2.4	GC content for the FAIRE probes	45
Figure 2.5	GC content for the filtered FAIRE-probes	45
Figure 2.6	FAIRE-QTLs QQ plot	47
Figure 2.7	Example FAIRE-QTL	50
Figure 2.8	Distance FAIRE-QTL to FAIRE-target	51
Figure 2.9	FAIRE-QTLs overlapping GWAS	55
Figure 2.10	FAIRE-QTLs show enrichment for expression-QTLs	57
Figure 2.11	FAIRE-QTLs show enrichment for dsQTL	60
Figure 2.12	1,000 Genomes genotypes reconfirm QTL locus	62
Figure 2.13	1,000 Genomes genotypes determine new QTL	63
Figure 3.1	Complex integration of molecular assays	85
Figure 3.2	CTCF to RNA	90
Figure 3.3	Distance CTCF to RNA	91
Figure 3.4	Violin plot Rho values CTCF to RNA	92
Figure 3.5	QQ-plot CTCF-FAIRE	95
Figure 3.6	Overlap CTCF and FAIRE	96
Figure 3.7	QQ-plot CTCF-FAIRE: difference probes and probesets	98
Figure 3.8	QQ-plot CTCF-FAIRE probes	101
Figure 3.9	Histogram CTCF-FAIREprobes distance	103
Figure 3.10	Violin plot Rho values CTCF-FAIREprobes	104
Figure 3.11	QQ-plot FAIRE-RNA	107
Figure 3.12	Histogram FAIRE-RNA distance	108
Figure 3.13	Violin plot Rho values FAIRE-RNA	109

Figure 3.14	QQ-plot CTCF-CTCF pairs	111
Figure 3.15	Histogram number CTCF-CTCF pairs	113
Figure 3.16	Histogram distance CTCF-CTCF pairs	114
Figure 3.17	Violin plot Rho values CTCF-CTCF	115
Figure 3.18	CTCF-CTCF pairs overlapping ChIA-PET	118
Figure 3.19	CTCF-loop length TF and HM profiles	121
Figure 3.20	Example FAIRE-CTCF-RNA correlation	124
Figure 4.1	Causality of a QTL	141
Figure 4.2	Determine causality of a QTL	142
Figure 4.3	Direct versus indirect causality	146
Figure 4.4	Direct versus indirect causality example	147
Figure 4.5	Histogram of causality of random associations	149
Figure 4.6	CTCF-FAIRE probe pairs QTL effect	151
Figure 4.7	CTCF-FAIRE probe - CTCF driven	152
Figure 4.8	CTCF-FAIRE probe - FAIRE driven	154
Figure 4.9	CTCF-RNA pairs QTL effect	156
Figure 4.10	CTCF-RNA - CTCF driven	157
Figure 4.11	CTCF-RNA - RNA driven	159
Figure 4.12	Positive CTCF-CTCF pairs QTL effect	162
Figure 4.13	CTCF-CTCF pairs with shared QTL	163
Figure 4.14	CTCF-CTCF - One causal	165
Figure 5.1	Boxplot chromosomal signals	177
Figure 5.2	Boxplot of MWU autosomes versus X chromosome	179
Figure 5.3	Gender-specific effects on chromosome X	180
Figure 5.4	CTCF X chromosome classification	182
Figure 5.5	Overlap CTCF sites with ENCODE	183
Figure 5.6	CTCF sites with ENCODE	187
Figure 5.7	Aggregate profiles X chromosome	189
Figure 5.8	X chromosome signal	191
Figure 5.9	CTCF classes imposed on DNase I	193
Figure 5.10	Xist intensity	195
Figure 5.11	XIC region	196

Figure 5.12	CTCF sites in XIC	197
Figure 5.13	RNA-FISH results for female-specific ncRNAs	199

LIST OF TABLES

Table 1.1	Overview Histone modifications	10
Table 2.1	Intra-individual variance GM12891	42
Table 2.2	Intra-individual variance GM12892	43
Table 2.3	Replicates inter-individual variance	43
Table 2.4	Overview FAIRE-QTLs	48
Table 2.5	Enrichment of FAIRE-QTLs in ENCODE	53
Table 2.6	GWAS SNPs and QTLs sources	70
Table 2.7	FAIRE probes	71
Table 2.8	Regulome DB score	76
Table 3.1	Overview CTCF to RNA	88
Table 3.2	Overview CTCF to FAIRE	93
Table 3.3	Overview CTCF to FAIRE probes	99
Table 3.4	Overview FAIRE to RNA	105
Table 3.5	Overview CTCF to CTCF	112
Table 3.6	Number of CTCF sites per chromosome	128
Table 3.7	Transcription Factors in ENCODE	131
Table 3.8	Histone modifications in ENCODE	134
Table 3.9	Genome-wide sequencing assays in ENCODE	134
Table 3.10	ENCODE CTCF ChIA-PET dataset	136
Table A.1	Overview individuals	221

Table B.1 Overview FAIRE-QTLs 225

ACRONYMS

ATAC	Assay for Transposase-Accessible Chromatin
bp	Base pairs
Broad	Broad Institute of MIT and Harvard
CEU	Northern and western European ancestry
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP	Chromatin Immunoprecipitation
CTCF	CCCTC-binding factor
DHS	DNase I Hypersensitive Site
DNaseI	DNase I sites
dsQTL	DNaseI-QTL
Duke	Duke University
EBV	Epstein-Barr virus
ENCODE	The Encyclopedia of DNA Elements
eQTL	expression-QTL
F ₁	First generation
F ₂	Second generation
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements
FDR	False Discovery Rate
FISH	Fluorescence <i>In Situ</i> Hybridisation
GWAS	Genome-Wide Association Study
Haib	HudsonAlpha Institute for Biotechnology
HapMap	Haplotype Map

HM	Histone Modification
Indel	Small insertion or deletion
kb	Kilo base pairs
LCL	Lymphoblastoid Cell Lines
LD	Linkage Disequilibrium
Mb	Mega base pairs
mRNA	Messenger RNA
MWU	Mann Whitney U test
ncRNA	Non-coding RNA
NGS	Next Generation Sequencing
OpenChrom	Duke University, The University of North Carolina at Chapel Hill and The University of Texas at Austin
PAR	Pseudo Autosomal Region
PCA	Principle Component Analyses
QQ-plot	Quantile-Quantile-plot
QTL	Quantitative Trait Locus
SEM	Structural Equation Modelling
SNP	Single Nucleotide Polymorphism
SNV	Singe Nucleotide Variance
SV	Structural Variance
Sydh	Stanford University, Yale University, University of Southern California and Harvard University
TF	Transcription Factor
Uw	University of Washington
WTCCC	Wellcome Trust Case Control Consortium
Xa	Active X chromosome
Xi	Inactivated X chromosome
XIC	X-chromosome inactivation center
YRI	Yoruba in Ibadan, Nigeria ancestry

Part I

INTRODUCTION

INTRODUCTION

The last two decades have seen a remarkable increase in quantitative genetics, which involves the association of DNA sequence variants with both dichotomous and continuous traits in a wide range of species. Decreasing costs of high-throughput genotyping and sequencing has led to an almost routine generation of genome-wide association studies (GWAS) between human disease traits and genomic regions ([Hindorff et al., 2009](#); [Johnson and O'Donnell, 2009](#)). Unfortunately, the majority of associated loci are in non-coding regions of the genome, and are therefore difficult to interpret ([Schaub et al., 2012](#)).

To improve our understanding of non-coding regions, it is possible to use molecular assays as a characterisation of phenotype, permitting the exploration of how genetic changes alter molecular mechanisms. By using a system genetics approach, the complex molecular interplay can be studied (see review ([Sieberts and Schadt, 2007](#))). This has been successfully investigated in *Drosophila melanogaster*; for complex traits, ([Ayroles et al., 2009](#)).

In this thesis, I will explore the genetic control and interrelationships of three molecular components: chromatin accessibility measured with Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE), transcription factor (TF) binding for CCCTC-binding factor (CTCF), and gene expression. Serendipitously, I discovered a novel relationship between CTCF and X-chromosome inactivation. Before describing this work, I will first introduce genetics generally, and then review previous work in this area. I will also provide a short introduction into X-chromosome inactivation.

1.1 HISTORY OF GENETICS

1.1.1 *Inheritance*

Even before the discovery of the molecular basis of inheritance, scientists conducted simple experiments on heritability. One of the earliest examples of this is the research performed by Gregor Mendel on pea plants. By measuring seed shape, flower colour, seed coat tint, pod shape, unripe pod colour, flower location, and plant height in both parents and offspring, he was able to establish some of the fundamental laws of heritability (Mendel, 1866). For example, Mendel noted that when a white flower plant is crossed with a purple flower plant, the first generation offspring (F₁) is not a mix of both colours, but consist purely of purple, which is the dominant colour (dominant allele). When plants from the F₁ are crossed in a second generation (F₂), the dominant colour is present in a 3:1 ratio, and thus it is evident that the recessive colour was not lost in the F₁.

Francis Galton aimed to determine whether similar rules applied to humans. Galton began studying variation and heritability in human populations using physical appearances, such as his quantitative study of height (Galton, 1869), to mental characteristics, including genius qualities (Galton, 1907). However, concerningly, he was also the founder of eugenics as he wanted to exploit this field of research to improve mankind (Galton, 1904). In his early research, Galton discovered that abilities are most likely inherited, but that environmental factors could bias his results; thus, he concluded that such studies are more useful when performed with twins (Galton, 1876). To conduct this type of research, Galton innovated several statistical methods, including standard variation and regression towards the mean, (Galton, 1879) that are still widely used today. It was under Galtons supervision that Karl Pearson worked on developing concepts including the p-value (Pearson, 1900), correlation (Pearson, 1920), and principal components analysis (Pearson, 1901).

1.1.2 *Inheritance is by genetics*

Thomas Morgan, a fly geneticist, was a sceptic of Charles Darwin's natural selection theory (Darwin, 1861). Particularly, in his opinion, it seemed highly unlikely that new species could occur solely due to the introduction of slight individual differences (Morgan, 1903). Morgan discovered that some mutant phenotypes in *Drosophila melanogaster* were following Mendel's law. He showed that when a white-eyed male (a male-only mutation) was crossed with a red-eyed female, all offspring (F₁) were red-eyed. When crossbreeding the F₁, the offspring (F₂) again included some white-eyed males. Morgan concluded that some traits were sex-linked and must be carried on one of the sex chromosomes, while other traits are probably carried on other chromosomes (Morgan, 1911). By combining the white-eyed mutants with another sex-linked trait, miniature wing, Morgan made important discoveries about the heritability of these traits. Morgan noted that, although both phenotypes are sex-linked, they are not inherited in such a way that both mutants are always together in F₂. This is, arguably, the first example of association between phenotypes. Morgan called this phenomenon *chromosomal crossover* (Morgan, 1916; Sturtevant et al., 1915). In honour of this discovery, the unit for linkage is called a *morgan*. In the process of these discoveries, Alfred Sturtevant, under the supervision of Morgan, published the first chromosomal map (Sturtevant, 1913). In the process of describing these discoveries, Morgan changed his opinion about Darwin's theory.

When Ronald Fisher published a genetic model showing that variation among a population is consistent with Mendelian inheritance (Fisher, 1919), the new field of population genetics emerged. In addition to his contribution to population genetics, Fisher published a myriad of essential statistical terms and methods including variance (Fisher, 1919), analysis of variance (ANOVA) (Fisher, 1922b), and Fisher's Exact Test (Fisher, 1922a). Fisher also published papers addressing proper use of statist-

ical methods (Fisher, 1925) and recommended practices related to study design (Fisher, 1935). Moreover, in his book about evolution, Fisher described how inheritance is not limited by the extremes of the parents (blending inheritance), but follows the particulate inheritance model, which was first described by Mendel (where particles refer to present-day genes) (Fisher, 1930). It is within this period that Karl Sax first observed that a single gene could control both a qualitative trait (colour) and quantitative trait (weight) of seeds (Sax, 1923).

1.1.3 *The modern genetic heritability era*

In the late 1940s and early 1950s, many great scientific breakthroughs were achieved in the field of molecular biology. This included the discovery that DNA, and not protein, is the carrier of genetic material in 1944 (Avery et al., 1944), followed by the proposition of the structure of DNA in 1953 (Watson and Crick, 1953). The first successful RNA sequencing of the Lac operator in 1973 (Gilbert and Maxam, 1973) and the first synthesis of a gene in 1979 (Khorana, 1979) are among these accomplishments. The next significant step involved inventions that made it possible to read out the actual DNA sequence in a more systematic and straightforward way.

Sanger sequencing, as it is known today, was the first method that made it possible to sequence DNA, by the use of templates and DNA polymerase (Sanger et al., 1977). Another method for sequencing was published at the same time, using chemical fragmentation (Maxam and Gilbert, 1977). However, because Sanger sequencing uses enzymatic synthesis instead of chemical fragmentation, Sanger sequencing is simpler to perform and safer to use. For these reasons, Sanger sequencing became the *de facto* standard for sequencing. Sanger sequencing is performed by radioactively labelling different termination base pairs. Because the termination base pair is attached randomly, reads will be terminated at different lengths. The terminated reads are

then transferred to four independent gel lanes (one lane per base pair), where the order of base pairs can be determined by reading out the lines on each of the four lanes. At later stages, Sanger sequencing was optimised to use dye labelling and capillary electrophoresis, which features made it a highly efficient method.

All of these breakthroughs in techniques to read out sequences of DNA on a large scale led to the establishment of several large sequencing projects. Noteworthy are *Haemophilus influenzae* as the first living organism that was fully sequenced (Fleischmann et al., 1995), *Caenorhabditis elegans* as the first animal to be fully sequenced (C. elegans Sequencing Consortium, 1998), and the first human chromosome (22) that was fully sequenced (Dunham et al., 1999). These efforts led to two major consortia, one public (Lander et al., 2001) and one industrial (Venter et al., 2001), to sequence the human genome and deliver the first complete reference human genome. The promise of attaining a working human genome sequence resulted in a new wave in large-scale genetics efforts. A vital point demonstrated by these projects was that most of the human genome is not contained within genes and is not transcribed (i.e. non-coding), which is sometimes dubbed junk DNA. Several large-scale interdisciplinary projects were founded to understand what, if any, functionality was present within these non-coding parts of the genome. The ENCODE project discovered that nearly the entire genome shows biochemical activity (ENCODE Project Consortium, 2012; ENCODE Project Consortium et al., 2007), indicating large amounts of regulatory functionality in the non-coding portions of the genome. Other projects focussed on the discovery of common genetic variations worldwide (1000 Genomes Project Consortium et al., 2010, 2012) and the determination of the normal variation in gene-expression (Lappalainen et al., 2013). Another consortium focussed on discovering the predisposition to common diseases (Lohmueller et al., 2003; Wellcome Trust Case Control Consortium et al., 2007).

1.2 THE HUMAN EPIGENOME

The history and definition of epigenetics are interpreted differently depending on the research field addressing this topic. In the eyes of Conrad Waddington, the epigenetic state and fate of a cell is like a ball rolling down a hill, finding the local minimum (Waddington, 1957). Once a state is set, it will stay in that state or change into a future state, but it can never be reversed back into its original pluripotent state. Waddington stated that genetics cannot explain the difference between cell types (because the genetic background will not change); he demonstrated this using *Drosophila* embryos, and concluded that the observed phenomena must be due to epigenetics (Waddington, 1952). In Waddington's time, epigenetics was most often used to explain differences that genetics could not explain. In the 1970s, a different definition for epigenetics emerged. With the discovery of DNA methylation, made simultaneously by Riggs (Riggs, 1975) and Holliday (Holliday and Pugh, 1975), epigenetics became known to additionally include chemical modifications. Bird added to the field of epigenetics the knowledge that CpG islands are regions rich with CpG sites, on which DNA methylation can occur at the cytosine of the CpG dinucleotides (Bird et al., 1985; Bird, 1986).

The structural composition of DNA in cells is largely determined by the wrapping of DNA inside nucleosomes around histones, necessary to fit a genome into the nucleus (Woodcock and Ghosh, 2010). Each nucleosome is composed of four basic histones: H3, H4, H2A and H2B. The histones H3 and H4 are bound in one dimer, and the histones H2A and H2B are bound in another dimer. The final nucleosome consists of two of each of the different types of dimers that are bound together in a histone octamer complex (Luger et al., 1997).

The wrapping of the DNA around this histone octamer complex is coordinated by biochemical events such as methylation, acetylation, phosphorylation and ubiquitination of different

residues of the histone. Most of these events occur at the unstructured tails of the histone and happen post-translation ([Allfrey et al., 1964](#)). The current status of research indicates that there are more than 100 different posttranslational modifications, each of them affecting the histones and the nucleosome. Many of these modifications are still poorly understood in terms of what causes them, which effects they might have, and whether these are directed effects or random biochemical events ([Jensen et al., 2002](#)) (see review ([Bannister and Kouzarides, 2011](#); [Strahl and Allis, 2000](#))).

The chemical modifications of the DNA and histone proteins exist in a complex interplay with genetics. This interplay of genetics and epigenetics, that determines the chromatin structure, is known as the epigenome ([Bernstein et al., 2007](#)). Changes to the epigenome, regulated by the genome ([Birney, 2011](#)), can be cell-state specific ([Mikkelsen et al., 2007](#)) or may correlate with diseases like cancer ([Jones and Laird, 1999](#)) (see review ([Jones and Baylin, 2007](#); [Portela and Esteller, 2010](#))), ageing ([Horvath, 2013](#)) (see review ([Gentilini et al., 2013](#))), and development (see review ([Feng et al., 2010](#); [Smith and Meissner, 2013](#))). The difference between the epigenome and the genome is that the epigenome is constantly altering, while the genome is not.

The functional annotation of all histone modifications is still far from complete. However, some of the more common modifications have been well-studied. Acetylation and methylation of lysine is one of the most well-understood histone modifications. The acetylation of lysine occurs, in almost all cases, in regions of chromatin accessibility and in transcribed regions. Methylation of lysine, in contrast, has a different effect depending on which residue is methylated. Table 1.1 ([ENCODE Project Consortium, 2012](#)) summarises some of the most well-studied histone modifications (see review ([Bannister and Kouzarides, 2011](#); [Strahl and Allis, 2000](#))).

Table 1.1: Overview Histone modifications and their putative functional annotation ([Bannister and Kouzarides, 2011](#); [ENCODE Project Consortium, 2012](#); [Strahl and Allis, 2000](#)).

Name	Putative function
H2A.Z	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Mark of active regulatory elements with preference for promoters
H3K9me1	Preference for the 5' end of genes
H3K9me3	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Mark of active regulatory elements may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Elongation mark associated with transcribed portions of genes, with preference for 3' regions after intron 1
H3K79me2	Transcriptionassociated mark, with preference for 5' end of genes
H4K20me1	Preference for 5' end of genes

1.3 CCCTC-BINDING FACTOR (CTCF)

There are currently over 3,200 transcription factors (TFs) known, and these proteins are involved in all known biological pathways (Yusuf et al., 2012). One central TF that is known for its extremely diverse functionality is the Zinc finger CCCTC-binding factor (CTCF). Since its discovery in chickens, as a C-myc repressor in 1990 (Lobanenkov et al., 1990), it has been successfully associated as being involved in chromatin organisation, gene regulation, enhancer-blocking, insulating, DNA-looping, imprinting, and X-chromosome inactivation. CTCF is considered to be one of the key chromatin regulatory TFs, and is known to be essential for establishing and maintaining the 3-dimensional genome structure (see review (Phillips and Corces, 2009)).

The 11 zinc fingers of CTCF bind the DNA with a 50-60 base pair (bp) footprint within the middle a DNase I hypersensitive site (Boyle et al., 2011). From the 11 zinc fingers, only 4 are essential for CTCF binding, and these bind mainly to the highly conserved 12bp motif (Ohlsson et al., 2001) that is conserved in both vertebrates (Kim et al., 2007; Schmidt et al., 2012) and fruit flies (Holohan et al., 2007). CTCF binding also shows tissue-specific binding activity (Schmidt et al., 2010). Furthermore, differences in the CTCF motif between individuals can result in different CTCF binding (Maurano et al., 2012).

In regions adjacent to CTCF, there is often an increase in methylation (Stadler et al., 2011). Interestingly, when Stadler introduced genetic mutations inside the CTCF binding motif, both CTCF binding and the adjacent methylation were lowered (Stadler et al., 2011), indicating that CTCF is driving the methylation.

CTCF is known to be an insulator binding protein (Wallace and Felsenfeld, 2007). There is also high enrichment of CTCF around chromatin barriers known for active and repressive histone modifications (Cuddapah et al., 2009). All of this previous work emphasises CTCF's role in the regulation of gene expres-

sion by actively controlling regulatory histone modifications (Phillips and Corces, 2009).

1.3.1 *Cohesin*

CTCF is also implicated in chromosome structure. The cohesin protein complex is a well-known complex of proteins that, together with CTCF, acts as a structural binding unit for chromosomes. The full cohesin complex consists of five conserved core proteins: SMC1, SMC3, RAD21, SA1 and SA2. Both SMC1 and SMC3 are part of the highly conserved family of Structural Maintenance of Chromosomes (SMC) proteins that act through ATPase (Merkenschlager and Odom, 2013). Cohesin, when bound with CTCF, is thought to trap DNA molecules using a so-called "embrace model" (Haering et al., 2008). Additionally, cohesin can form a complex with a mediator to bind two cohesin loops, creating DNA loops (Kagey et al., 2010).

1.3.2 *CTCF loops*

In previous studies using ChiA-PET (Section 1.7.0.2) in humans (Merkenschlager and Odom, 2013; Sanyal et al., 2012), as well as model organisms like mouse (Handoko et al., 2011), there is evidence that CTCF insulating loops can be local (cis), remote (trans) and interchromosomal. Handoko illustrated that, for the mouse, depending on the length of the CTCF loop, the CTCF sites are correlated with different histone patterns. Loops with length <200kb overlap mainly with active histone modifications, such as H3K4 methylation. Loops longer than 200kb correlate with known repressive histone modifications (Handoko et al., 2011).

1.4 THE SEX CHROMOSOMES

In humans, there are 22 pairs of autosomes and one pair of allosomes, or sex chromosomes. Females have two copies of the X chromosome (XX), and males have one copy of the X chromosome and one of the Y chromosome (XY). The actual phenotypic determination of sex is found in genes located on the Y chromosome; the majority of genes on the X chromosome are not related to sex-specific functionality (Ross et al., 2005). Because of the different copy numbers of X chromosomes, males and females show differential disease susceptibility (see reviews (Libert et al., 2010; Ober et al., 2008)). For females, for example, monosomy of the X-chromosome can lead to the Turner syndrome.

It is commonly accepted that the current set of sex chromosomes is the results of the, still on-going, evolutionary transformation from a normal pair of autosomes (Lahn and Page, 1999). Lahn showed that 19 ancestral genes, with homologs on both the X and Y-chromosomes, have similar nucleotide divergence. The ancestral allele for male determination is now solely located at the Y chromosome (Graves, 2006). As the Y chromosome evolved over time, genes beneficial to males crossed over towards the Y chromosome (Burgoyne, 1982), adjacent to the sex-determination locus (Skaletsky et al., 2003). Both the X and Y-chromosomes contain two pseudoautosomal regions (PAR), PAR₁ and PAR₂, which recombine during meiosis (Graves et al., 1998; Hinch et al., 2014). In mammals, to the best of established knowledge, only humans have 2 PAR regions; other closely-related species, including chimpanzee and rhesus macaque, have only a single PAR region (Bachtrog, 2013; Carninci et al., 2006). This observation provides strong evidence that both PAR regions evolved independently.

1.4.1 *Discovery of X-chromosome inactivation*

If both copies of the X chromosome in females were to be transcribed, there would be an imbalance between males and females in sex-chromosome gene transcription, which could impair normal functionality. For this reason, mammals have evolved a genetically-driven system called *dosage compensation*, in which one of the female's X-chromosome expressions is silenced during an early developmental stage. By using this system, there is one active X chromosome (Xa) and one inactive X (Xi) in females. The obvious benefit of this system is that that X chromosome genes have the same transcription levels in males as in females.

The study of the mechanisms driving X-chromosome inactivation, and thus dosage compensation, has been ongoing since Susumu Ohno showed in 1959 that each X chromosome of the X-chromosomes pair in females appears different under the microscope (Ohno et al., 1959). One of the X chromosomes looks like a functional autosome, while the other is condensed and heterochromatic, also called a Barr body (Barr and Bertram, 1949). Following these discoveries in 1961, Mary Lyon proposed that X-chromosome inactivation is a random process (Lyon, 1961). This system of random inactivation was proposed after the observation that in female mice, the coat colour has patches of normal and mutant colour and is not uniform as would be expected if either the paternal or the maternal X chromosome were consistently inactivated. Coat colour, including mottled patterns (Lyon, 1961), had been linked to mutations in several X-linked genes. As the fur coat is not uniform, this could only be due to random inactivation of the paternal or maternal X chromosome. This would result in gene expression that originated sometimes from the mutated gene and sometimes from the original gene.

1.4.2 *The X-inactivation center*

X-chromosome inactivation in mammals stands as one of the major illustrations of epigenetic gene regulation (see review (Pollex and Heard, 2012)). A major driving force for achieving silencing in one of the X chromosomes is the X-inactivation center (XIC) (Penny et al., 1996), remarkably, XIC is not essential for maintenance of the inactivation (Brown and Willard, 1994). At the XIC, there are 4 noncoding RNA genes, *Xist*, *Tsix*, *Jpx* and *Ftx*. *Xist* is the main ncRNA that silences the inactive X chromosome (Brown et al., 1991; Herzing et al., 1997). *Tsix* is the antisense gene of *Xist* and is involved in the early stages of X-chromosome inactivation, though not the silencing step itself (Lee et al., 1999). Interestingly, moving the XIC to an autosome will lead to the inactivation of one of the chromosome pair, as shown in a study in which *Xist* was used to silence one chromosome in trisomy (Jiang et al., 2013). During inactivation, the *Xist* ncRNA binds and then coats the entire X chromosome (Clemson et al., 1996), except for the regions that host escape genes (Carrel and Willard, 2005), and alters the 3D structure of the X chromosome (Engreitz et al., 2013) into a Barr body.

1.4.3 *CTCF in X-chromosome inactivation*

CTCF has been identified as a factor involved in regulating X-chromosome inactivation (Chao et al., 2002). Between *Tsix* and *Xist* is a CTCF site that shows strong binding and is most likely involved with *Xist* induction (Spencer et al., 2011). In terms of X-chromosome inactivation, CTCF binds to the XIC (Chao et al., 2002) and is most likely involved in inducing *Xist* in early X-chromosome inactivation (Spencer et al., 2011). CTCF is highly clustered with the transcription factor YY1, and is a key factor in controlling *Xist* expression (Donohoe et al., 2007). During inactivation, YY1 anchors the *Xist* RNA to the DNA to silence the specific X chromosome (Thorvaldsen et al., 2011). Mutations

in some specific CTCF sites can lead to skewed X-chromosome inactivation ([Pugacheva et al., 2005](#)).

1.5 CELL LINES FOR MOLECULAR PHENOTYPING

When studying the basic processes of biology, working with primary tissue is challenging. Many biochemical assays require at least 10^6 cells. It is difficult to access one particular cell type from a living donor at this scale. For this reason, many studies are performed on immortalised cell lines, which grow well in the laboratory. This technique was established with the HeLa cell line, the first immortal cell line obtained from a single cancer patient in the 1950s. The HeLa cell line ([Gey et al., 1952](#)) was essential for the discovery of the polio vaccine ([Scherer et al., 1953](#)). HeLa cell lines were the beginning of a long tradition of research performed on cell lines (see review ([Masters, 2002](#))). The recent publication of the full genome sequence of HeLa indicates how essential this immortal cell line has been, even for present-day research ([Landry et al., 2013](#)).

Following HeLa, many other cell lines were established that were derived from cancer tumours. For example, K562 is derived from leukaemia ([Lozzio and Lozzio, 1975](#)) and MCF-7 from breast cancer ([Soule et al., 1973](#)). Although all of these cell lines are a useful resource for replenishable cells, they provide a poor platform for comparing differences between healthy individuals. As each cell line originates from a different tissue and has its own "broken" cancer genome ([Greenman et al., 2007](#)), these cell lines cannot be used within a population study, as direct comparison of these cell lines would indicate cancer- or tissue-specific differences, rather than individual specific differences.

1.5.1 *Lymphoblastoid cell lines*

A breakthrough in creating cell lines from healthy individuals involved the ability to transform B lymphocytes by Epstein-Barr virus (EBV) into lymphoblastoid cell lines (LCLs) (see review (Miller, 1971)). These cell lines are, like HeLa, immortalised and thus capable of indefinite growth, making them ideal to study biological effects on a population level, while being able to compare to previously-performed experiments on cells from the same line. Because LCLs originate from B lymphocytes, there are some genomic regions, such as the IG locus, that are continuously reorganised (Alt et al., 1992). For this reason, in studies using LCLs, these regions are often excluded.

Because LCLs can provide almost unlimited quantities of cells, they are used heavily in many large research consortia as well as in smaller research studies (see review (Sie et al., 2009)). Beyond this feature, there is the additional benefit that a large number of LCLs, from different human populations, have been genotyped and phased within the HapMap consortium (International HapMap Consortium, 2003, 2005). LCLs are also being sequenced (most on low-coverage) during Phase 1 of the 1,000 Genomes Project (1000 Genomes Project Consortium et al., 2010, 2012), providing access to high-quality genotypes. For most of the LCLs that are part of the 1,000 Genomes Project, the transcriptome has been quantified, using RNA sequencing, in the gEUVADIS consortium (Lappalainen et al., 2013).

Several molecular assays have been successfully performed on a subset of LCLs. In these studies, inter- and intra-individual differences were discovered for gene expression (Montgomery et al., 2010; Reddy et al., 2012), chromatin state (Ernst et al., 2011; McDaniell et al., 2010), proteome (Joubert-Caron et al., 2000), and the discovery of molecular Quantitative Trait Loci (QTLs) other than expression-QTLs, like DNase I sensitivity QTLs (dsQTLs) (Degner et al., 2012). In ENCODE, a large-scale consortium to map all functional elements of the genome (ENCODE Project

[Consortium, 2012; ENCODE Project Consortium et al., 2007](#)), one LCL, GM12878, has been used as one of the core cell lines, meaning that many molecular assays have been applied to it.

Due to the extensive research that has been performed on these cell lines, there is a huge potential to build a multidimensional view of molecular phenotypes that can be used for systematic genetic dissection. This will, in the long term, increase our understanding of molecular pathways and contribute to enhancing our understanding of the molecular role of many previously identified noncoding GWAS results.

1.6 QUANTIFICATION OF MOLECULAR PHENOTYPES

Several molecular techniques have been developed over time to quantify a large number of genomic sites simultaneously. By using microarrays (Section [1.6.1](#)), and later next generation sequencing (NGS) (Section [1.6.2](#)), it has become feasible to record large scale measurements of molecular phenotypes. Although there are differences between microarrays and NGS, both methods have in common that the first step requires isolating the DNA of interest, and the second step involves quantifying the isolated DNA.

1.6.1 *Microarrays and tiling arrays*

With the development of the first microarrays, it became possible to measure the abundance of a wide range of specific sequences simultaneously. A classic microarray is a matrix of points with specific probes. In essence, the first microarrays were analogous to many northern/southern blots performed in parallel ([Augenlicht and Kobrin, 1982; Maskos and Southern, 1992](#)) that could, for example, quantify gene expression ([Schena et al., 1995](#)) for an entire transcriptome at once ([Lashkari et al., 1997](#)).

The basis of a microarray involves having a range of probes that hybridise to complementary sequences that are fluorescently labelled. During measurement, complementary sequences with the probe will bind, and non-hybridised material is washed away. By measuring the fluorescent signal for each probe, a probe-specific quantitative measure is obtained.

Tiling arrays use the same biochemical methodology as microarrays, but differ in their probe design. Whereas microarrays will target precisely known sequences of interest (such as a gene or transcript), in tiling arrays, the genomic sequence is constantly sampled with evenly distributed probes. Because of this continuous sampling, tiling arrays will provide an almost continuous readout for the targeted genomic regions.

1.6.1.1 *Examples of use of microarrays*

Tiling arrays have been used extensively for measuring molecular signals in non-coding regions. This technique has been used to measure signals for transcription factor binding, histone modifications, and splicing variants ([Malone and Oliver, 2011](#)). Also, genome-wide assays such as FAIRE, DNase I ([Crawford et al., 2006](#)), and methylation ([ENCODE Project Consortium et al., 2007](#)) have been studied extensively using tiling arrays.

Another widely-used application of microarrays is to genotype specific single-nucleotide polymorphisms (SNPs) ([Fan et al., 2000](#)), small insertion or deletion (indels), or copy number variations ([Pinkel et al., 1998](#)) (see review ([Gresham et al., 2008](#))). On these arrays, probes are designed to match the genetic variations of interest. In this way, it is possible to acquire a read-out of the allele with and without SNP, and by comparing these two signals, to determine what the genotype of the SNP is.

1.6.2 *Next generation sequencing for molecular quantification*

Microarrays have been proven to be highly successful in a wide variety of settings, but involve some serious drawbacks. The first is that microarrays rely on probe design, and thus will only measure whatever the probes are designed to measure. The second, and most significant, is that the hybridisation of probes with complementary sequences is not always equally strong. This could mean, for example, that although the expression of two genes is identical, microarray quantifications can show differences.

For these reasons, microarrays are being replaced by next-generation sequencing (NGS) quantifications. NGS is being used for transcriptomics (see review ([Wang et al., 2009](#))), for ChIP experiments (see review ([Furey, 2012](#))), and for chromatin confirmation studies (see review ([Schones and Zhao, 2008](#))). This swift change has been driven by the rapid drop in price of NGS, while offering more extensive genome-wide coverage and higher sensitivity, and the additional benefit of providing a continual quantitative measurement of the genome rather than only for predefined regions. Similar to microarrays, the measurements themselves are independent of the isolation of genetic material, making NGS suitable for a wide range of research topics.

There are currently numerous NGS platforms available (see review ([Metzker, 2010](#))). The most notable and widely used examples include an array-based pyrosequencing approach known as 454 sequencing ([Margulies et al., 2005](#)), a sequencing-by-synthesis method called Illumina sequencing ([Bentley, 2006](#)), and sequencing-by-ligation method named SOLiD sequencing ([Valouev et al., 2008](#)).

New sequencing technologies are being developed continuously. Third-generation sequencing methods (e.g. Nanopore sequencing) promise longer reads and resolution of a single molecule ([Oxford Nanopore Technologies, 2012](#)). At the same time, second-generation sequencing methods are still improving

by becoming faster, enhancing scalability, and reducing overall costs.

I will now outline a number of molecular phenotyping methods that can be used in combination with microarrays or sequencing.

1.7 CHROMATIN CONFORMATION AND TRANSCRIPTION FACTOR BINDING AFFINITY

Several techniques, for example FAIRE (Giresi and Lieb, 2009; Giresi et al., 2007), DNase I (Crawford et al., 2006), and ATAC (Buenrostro et al., 2013), have been developed in the field of molecular biology to assay the conformation of chromatin. In addition, several Chromatin Immunoprecipitation (ChIP) assays are able to measure the binding of transcription factors or histone modifications (Furey, 2012; Wang et al., 2012). Each of these methods can be read out with either microarrays or sequencing.

1.7.0.1 *Chromatin Immunoprecipitation*

ChIP is an immunoprecipitation technique that permits the study of the interaction between proteins and DNA (Kuo and Allis, 1999). With ChIP, it is possible to determine whether a specific protein is bound to a specific genomic region. ChIP can be used to determine the binding of TF or the presence of histone modification (HM). Initially, ChIP was used to investigate specific cases (e.g. TF x binds genomic region y). In 2000, the combination of ChIP with microarrays (ChIP-chip) made genome-wide quantifications possible (Iyer et al., 2001; Ren et al., 2000). In 2007, ChIP was quantified using sequencing (ChIP-seq) for the first time (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007).

A ChIP experiment consists of the following steps:

1. crosslinking of DNA
2. sonication of DNA

3. pull down of DNA fragments using immunoprecipitation
4. quantification using microarrays or sequencing

1.7.0.2 *Chromatin Interaction Analysis by Paired-End Tag Sequencing*

The genome is highly complex and its spatial conformation is essential to its function (Dekker et al., 2002; Lieberman-Aiden et al., 2009). For some events to occur, different parts of the same, or sometimes even different, chromosomes are required to interact with each other. Some factors, like CTCF (Handoko et al., 2011), are known to influence the structure of the genome in long-range interactions that enable chromatin interactions (Sanyal et al., 2012). A method of studying these chromosomal interactions is by using chromatin interaction analysis by paired-end tag (ChIA-PET) sequencing (Fullwood et al., 2009). The ChIA-PET technique effectively combines paired-end sequencing with ChIP. Using ChIA-PET, it is possible to measure binding between different regions of the genome through transcription factors. Because both paired-end tags (PET) are sequenced, the distinction can be made between non-chimeric CTCF loops (both PETs are identical) and chimeric CTCF loops (both PETs are different) (Li et al., 2010).

A ChIA-PET ChIP experiment consists of the following steps:

1. crosslinking of DNA
2. sonication of DNA
3. linkers are introduced to bind the free DNA ends and ligate together
4. ligated linkers are digested at short distance from the linkers
5. pull DNA fragments down using immunoprecipitation
6. quantification of DNA fragments using paired-end sequencing

1.7.1 *DNase I hypersensitive sites*

A key element in epigenetic, or epigenome, studies is the understanding and identification of regulatory elements. These regulatory elements can directly and indirectly regulate the binding of epigenetic marks and gene expression. Areas of the DNA that are known to be regulatory for transcription show a loss of nucleosomes. This lack of nucleosomes can be measured by the ability of nucleases to act on regions of nucleosome depletion. Early studies in *Drosophila* have shown that the 5' regions of active genes are hypersensitive to DNase I, which is evidence for nucleosome destabilisation.

Regions of the genome that are sensitive to cleavage by DNase I are known as DNase I hypersensitive sites (DHSs), and were discovered in the 1980s (Gross and Garrard, 1988). The resulting fragments of DNA after digestion by DNase I can be measured using tiling arrays (Crawford et al., 2006; Sabo et al., 2006) or sequencing (Hesselberth et al., 2009; Song and Crawford, 2010). Quantification of these DNA fragment can provide a genome-wide view of genome accessibility. The DNase I method consists of the following steps:

1. low concentration of DNase I agent added to DNA
2. DNase I cleaves DNA at regions that are accessible
3. capturing of the resulting DNA fragments
4. ligate DNA fragments to sequencing adapters
5. quantification of DNA fragments using microarrays or sequencing

The DHS regions of the genome are known to be in more accessible regions, which strongly indicates the presence of regulatory elements, making these regions of high interest to study (Thurman et al., 2012). Interestingly, DHS regions themselves are also under the influence of genetic regulation (dsQTLs) (Degner et al., 2012).

1.7.2 *Formaldehyde-Assisted Isolation of Regulatory Elements*

Another method of identifying nucleosome-depleted DNA is Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE), which is derived from CHiP protocols (Giresi et al., 2007). When crosslinking DNA with formaldehyde, the formaldehyde will be cross-linked with the chromatin. After shearing the DNA by sonication, only DNA fragments that are not bound by formaldehyde will be extracted. In contrast, for the non-treated sample, all DNA regions are extracted. Depending on whether sequencing or microarrays are used, the DNA fragments are quantified to discover both chromatin-bound and open chromatin regions. The FAIRE method consists of the following steps:

1. crosslinking of DNA with formaldehyde
2. sonication of DNA
3. phenol-chloroform extraction of non formaldehyde bound DNA fragments
4. quantification of DNA fragments using microarrays or sequencing

1.7.3 *Assay for Transposase-Accessible Chromatin*

The latest technology for determining the chromatin status is Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013). The significant advantages of ATAC-seq in comparison with the FAIRE and DNase I methods are that the duration of preparation time and quantity of cells required are massively reduced, from approximately two days to a couple of hours and less than 50,000 cells.

In this method, Hyperactive Tn5 transposase loaded with sequencing adapters is added to a sample. Transposase will insert itself into regions of open chromatin. After this insertion,

Transposase will integrate the sequencing adapters to these regions and can subsequently be amplified by PCR and sequenced. Buenrostro showed that ATAC-seq, although using far less input material, yields highly similar results in GM12878 for DNase I-seq (reported R of 0.79 and 0.81) (Buenrostro et al., 2013), making it a likely candidate for future epigenetic studies.

The ATAC method consists of the following steps:

1. hyperactive Tn5 transposase loaded with sequencing adapters is added to a sample
2. transposase inserts into open chromatin regions
3. integration of sequencing adapters
4. PCR multiplication of regions with sequencing adapters
5. quantification using microarrays or sequencing

1.7.4 *Comparison of chromatin accessibility assays*

Because nucleosome depletion is a broad aspect of active chromatin regulation, the FAIRE assay provides an efficient method to assay regulatory elements. Previous studies have shown that FAIRE largely overlaps regions from other chromatin assays like DNase-seq and ChIP-seq, marking transcription factors, histone marks, and chromatin accessibility (Giresi and Lieb, 2009; Song et al., 2011). Another benefit of the FAIRE assay in comparison with, for example, DNase I, is that it is technically easy to use this method and it does not require a large quantity of input material. Because of this latter feature, the FAIRE assay scales up easily if there is a need to measure many samples. Also, as FAIRE is sensitive to many of the biochemical activities occurring at the level of DNA, it can be used as a proxy measurement to discover regions of interest. However, due to this behaviour, it is difficult to attribute chromatin accessibility solely by using a FAIRE measurement. For this reason, if the aim of a study is

the reliable measurement of chromatin accessibility, DNase is the preferred assay. Additionally, the initial ATAC-seq results indicate that both FAIRE and DNase as assays for chromatin accessibility are becoming obsolete, in favour of the former.

1.8 DISCOVERY OF QUANTITATIVE HERITABLE TRAITS

Since the early days of population genetics and by methods developed by pioneers like Fisher (Section 1.1.2), scientists have been conducting association studies between quantitative traits and genetic marks by using quantitative trait loci (QTL) studies and genome-wide association studies (GWAS). Arguably, even common diseases are most likely not a binary trait (sick versus healthy) (Wellcome Trust Case Control Consortium et al., 2007), as in many typical GWAS studies, but can be categorized within a conditional range and thus actually represent a quantitative trait that can be studied in a liability model as has been successfully shown. For example, for type 2 diabetes (Rung et al., 2009), glucose fasting (Bouatia-Naji et al., 2008) and C-reactive protein (Elliott et al., 2009) a quantitative approach has yielded striking results.

Much effort has been invested in finding molecular traits that associate with specific genetic variants, where the genetic component can be a polymorphism (SNP or SNV), a small insertion or deletion (indel), a structural variance (SV), or larger haplotype blocks. In quantitative traits loci (QTL) studies, genetic differences between individuals are associated with quantifiable molecular or whole-body phenotypes that are characterized by a continuous range of values.

1.8.1 *History of QTL mapping*

Particularly in the field of plant breeding, QTL studies have been performed for many decades (see review (Kearsey and

Farquhar, 1998; Paran and Zamir, 2003; Young, 1996)). In the early days, in humans these QTL studies used so-called linkage maps (Lander and Botstein, 1989). The power of identification of these early linkage studies in humans was due to using pedigree information and phenotypes that were known to be heritable (e.g. cystic fibrosis) (Groman et al., 2002; Tsui et al., 1985). More recent successful QTL studies on humans have used whole-body phenotypes such as height (Gudbjartsson et al., 2008; Visscher, 2008) or body mass index (Wang et al., 2011). On the molecular level, many successful studies have identified QTLs for gene expression (Montgomery et al., 2010; Veyrieras et al., 2008), DNase I hypersensitivity (Degner et al., 2012), and protein levels (Melzer et al., 2008). For many more examples of complex human traits in genome-wide studies, refer to the review by Stranger (Stranger et al., 2011).

1.8.2 *Molecular QTLs*

The easiest to understand, and thus best-studied, molecular effect under the influence of the genetic variation is gene expression. These so-called expression-QTLs (eQTL) have been successful, even when applied to low numbers of LCLs (e.g. (Göring et al., 2007; Monks et al., 2004; Schadt et al., 2003)), and in particular when restricted to cis-based variants (Montgomery et al., 2010; Veyrieras et al., 2008). Other eQTL discovery studies have shown eQTLs to be specific for splicing events (Lalonde et al., 2011), that some are highly tissue-specific (Zhang et al., 2009) and are involved in mechanisms regulating allele-specific expression (Pickrell et al., 2010).

The effect of an eQTL is sometimes easy to understand, especially, when the eQTL is nearby the gene it alters. In these cases, the eQTL can be located inside a promoter region or can alter the binding of a transcription factor. However, for most eQTLs there is no easy explanation regarding why this genetic change leads to transcriptional changes. Large eQTL meta-analysis iden-

tifies trans-eQTLs that are known to be associated with diseases (Schadt et al., 2005; Westra et al., 2013), and some studies suggest that these trans-eQTLs are regulating intermediate genes involved with these diseases (Fehrmann et al., 2011). However, only by fully understanding what the causal effect of a genetic alteration in a non-coding region or in trans is, can we achieve a more systematic understanding of these eQTLs. Most likely, these eQTLs are not directly affecting transcription, but are altering chromatin status, histone modifications, or the binding of transcription factors that in turn regulate transcription. Mapping these eQTLs to other molecular effects will be essential to understand and fully utilise the potential of many GWAS studies.

1.8.3 QTL mapping

The discovery, or mapping, of QTLs is based on the assumption that a phenotype (P) is determined by the sum of the combination of genetics (G_i) and the environment (E), see equation 1.

$$P = G_i + E \quad (1)$$

However, parents pass on genes and not phenotypes and thus heritability is the ratio of phenotypic variance (V_p) that can be attributed to genetic variance (V_g). This type of heritability is known as broad-sense heritability, see equation 2.

$$H^2 = \frac{V_g}{V_p} \quad (2)$$

The genetic variance can be divided further into additive genetic effects (V_a), dominance genetic effects (V_d) (i.e. interactions between alleles at the same locus), and epistatic genetic effects (V_e) (i.e. interactions between alleles at different loci), see equation 3.

$$V_g = V_d + V_a + V_e \quad (3)$$

Because each parent transfers only a single copy of each gene to their offspring (e.g. in offspring, for each locus there is a paternal and maternal allele), even relatives will most often only share a single gene copy. For this reason, genetic effects like dominance and epistatic genetic effects that rely on transfer of both copies of a single allele from a single parent cannot influence the phenotypic resemblance. Narrow-sense heritability is determined by considering additive genetic (V_a) effects solely, and thus is commonly used to determine heritability, see equation 4 (Houle, 1992).

$$h^2 = \frac{V_a}{V_p} \quad (4)$$

When a phenotype is additive, there will be a linear correlation between the three genotype possibilities (AA, Aa and aa) and the phenotype. Therefore, for an additive trait, the assumption is that for each allelic difference, a similar change in phenotype will be observed. To model a dominant versus recessive effect in a full model, an additional parameter is needed. For the discovery, or mapping, of QTLs, classically the choice is between an additive model and a full model, or in other words, between a one-parameter test and a two-parameter test.

When using an additive model, the genotypes are converted into numeric values (AA = 0, Aa = 1, aa = 2) so that they can be used in a statistical method such as Pearson's Correlation or Spearman's Rank. As Pearson's Correlation assumes linear relationships and Spearman's Rank assumes a monotonic relationship, the better method to use will depend on the data normalisation steps performed before these further analyses. After data normalisation, results obtained with Pearson's Correlation and Spearman's Rank are highly similar. However, the ranking step within Spearman's Rank negatively affects the power to discover correlations. In the literature, both Spearman's Rank (e.g. (Montgomery et al., 2010)) and Pearson's Correlation (e.g. (Degner et al., 2012)) are used for QTL mapping.

1.8.4 *Multiple testing problem*

Sample size is an important consideration when performing a QTL (or any normal GWAS) study (Spencer et al., 2009). As there are a large number of tests for association performed, a simple p-value threshold of 0.05 for each test is not sufficient to conclude significance. Because of multiple testing considerations, finding significance requires a more stringent standard. For example, when performing a single statistical test, it might be sufficient to use a simple p-value threshold like < 0.05 . This p-value tells you that there is a 5% chance that this result would occur by random chance. However, if, for example, you were to perform 20 tests, in parallel, with the same p-value threshold, one of these tests could show a positive result purely by chance. In QTL studies, there are hundreds of thousands of tests performed, and the p-value for judging significant findings has to be corrected. A potential solution is the Bonferroni correction, in which the desired p-value threshold is divided by the number of tests performed (Bonferroni, 1935). However, the downside of the Bonferroni correction is that it is extremely conservative, which results in many false negatives. By the introduction of false discovery rates (FDR), an estimation can be made about the expected percentage of false positives among the claimed positives (Benjamini and Hochberg, 1995; Hochberg and Benjamini, 1990). FDRs have subsequently been successfully applied in many fields (e.g. microarrays (Reiner et al., 2003)).

Another way of dealing with many statistical tests is permutations. Using permutation, it is possible to determine empirical p-values. By performing permutations, it is possible to draw a null distribution. By ranking the real p-value in this null distribution, the confidence of the found association is determined. Another method that addresses multiple testing, without expensive computational permutations, involves determining the adjusted p-values (q-values) (Storey and Tibshirani, 2003), in which the characteristics of the p-value distribution (specifically,

the region of flattening of the p-value distribution) are used to set the expected false positive rate. When applying a robust mapping approach, even small populations tend to be effective in discovering QTLs by optimising the study design (see reviews (Flint and Mackay, 2009; Mackay et al., 2009)). However, small population studies will suffer from a high false negative rate, making population size one of the most significant limiting factors of the QTL mapping rate (Evans and Purcell, 2012).

The constant reduction of costs for tiling arrays and next-generation sequencing (Wetterstrand, 2013) make it feasible to study many additional molecular assays in QTL-like studies. Combining this fact with the already-available cell line resources, such as LCLs, there is enormous future potential for large-scale systematic QTL mapping.

1.9 SCOPE OF THIS THESIS

The present thesis is divided into three parts. In the first part, the discovery and characteristics of FAIRE-QTLs is presented. A tiling array is designed to measure 10,507 regulatory elements for FAIRE activity. Using the HapMap v3 genotypes, FAIRE-QTLs are discovered. By overlapping these with other QTL and ENCODE datasets, some of the informative characteristics of these QTLs are discussed.

In the second part, the molecular assays and QTLs, performed in a human population, are integrated. The human population used for these experiments is composed of 60 individuals having northern and western European ancestry (CEU). On these 60 individuals, several assays are performed: FAIRE-chip, CTCF-seq, RNA-seq and DNA-seq. Each of the molecular assays is associated with each of the other assays to discover phenotype-to-phenotype correlations. Furthermore, QTL data are used to dissect the causality for these phenotype-to-phenotype correlations.

The third part presents a comprehensive view of CTCF binding on the X chromosome, and its implications for X-chromosome inactivation. The transcription factor CTCF is known for its multifunctionality. There is not a single main function for CTCF that is observed, as its function depends on cell state, cell type, co-factors, or location. In the population of LCLs discussed, a novel X chromosome-wide CTCF effect is observed. Using the gender of each of the cell lines, observations are made about which CTCF sites are dosage compensated, active on both chromosomes, or are only bound in females.

Part II

DISCOVERY OF QUANTITATIVE TRAITS

FINDING QTLS FOR CHROMATIN ACCESSIBILITY

This chapter describes the results of discovery and the characterisation of Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) regions as Quantitative Trait Loci (QTL). Data from this chapter will later be re-used in Chapter 3 and Chapter 4.

2.1 INTRODUCTION

By measuring FAIRE (Giresi et al., 2007) for a small population of 60 lymphoblastoid cell lines (LCLs) that are part of HapMap, it is possible to identify FAIRE-specific QTLs (FAIRE-QTLs). Having a list of FAIRE-QTLs is useful for the interpretation of genome-wide association studies (GWAS). If a FAIRE-QTL is also known as an expression-QTL (eQTL), this might provide a causal reason for the change in expression driven by this genotype. However, because of the small sample size, there will be many unidentified FAIRE-QTLs. In addition, when this study was designed in 2009, the most cost-effective way to measure this quantity of cell lines using a molecular assay was by using tiling arrays. The tiling array was designed in such a way that probes are targeted for known regulatory features. By design, the tiling array is most likely enriched for regions with FAIRE activity. However, as this process only assays known regulatory features, it provides a limited view of how FAIRE acts genome-wide. In essence, this should be interpreted as a pilot study, indicating the potential for further discovery of FAIRE-QTLs, and thus should be repeated using sequencing and a larger sample size to achieve a more definitive result.

2.2 RESULTS

To understand the genetic regulation of epigenetic marks, I conducted a study using FAIRE as a quantitative phenotype of chromatin accessibility. This was performed in LCLs representing 60 unrelated individuals from northern and western European ancestry (CEU) from the International HapMap Project ([International HapMap Consortium, 2003, 2005, 2007](#)) and the 1,000 Genomes ([1000 Genomes Project Consortium et al., 2010, 2012](#)). I further address variance in this type of study design and explore the relationships of FAIRE-QTLs with other molecular and disease QTLs.

For this study, I performed all of the analysis. The growing of cells and performing the tiling arrays was performed by the group of Jason Lieb at University of North Carolina.

2.2.1 *FAIRE intensity data for 60 CEU individuals*

Sites of accessible chromatin were identified by FAIRE-chip using a custom Nimblegen probe design. The FAIRE-chip targets 10,507 regulatory regions that were previously identified as regulatory regions using DNase I, and that showed inter-individual variance for CTCF binding in two trios ([McDaniell et al., 2010](#)). Each identified regulatory region was ranked on inter-individual variance, which enhances the likelihood of discovery of FAIRE-QTLs.

The FAIRE-chip does not have genome-wide coverage. To determine whether the measured FAIRE signal is only active for the targeted regulatory element, two probes tiling the target region were added. Using the readout of these two probes, I was able to ensure that the measured FAIRE signal is specific to the target region rather than a much larger region. The control probes are approximately 100bp up- and downstream of the target region. For example, for a regulatory region of interest, our FAIRE-chip targets the signal over two control probes and

several target probes (Figure 2.1). The target probes are summarised into a probeset (FAIRE-peak). The FAIRE-peak signal, for each target region, is the ratio of FAIRE-treated sample and control sample. I used the statistical method Robust Multichip Average (RMA) (Irizarry et al., 2003) to average all probes into probesets, resulting in a final summary statistic (Section 2.5.1). SNPs are only considered FAIRE-QTLs when the genotype association is significant for the probeset but not for the control probes.

To discover probe-specific variation in the FAIRE-chip, two individuals are assayed five times as replicates. In this way, both inter- and intra-individual variance can be determined for each probe, and probes that exhibit too much inter-individual variance can be removed.

In summary: There are 60 individuals assayed using our FAIRE-chip, and two of these individuals are assayed five times as replicates. All 60 individuals are genotyped, 56 in HapMap v3, and 46 in Phase1 of the 1,000 Genomes Project.

2.2.2 *Filtering out intra-individual variance*

One of the most significant challenges when working with microarrays is hybridisation variation of the probes. The DNA molecule in the sample may not hybridise equally well to each probe for each individual sample.

For cell lines GM12891 and GM12892, replicates were grown and harvested for FAIRE-chip. The comparison of replicates indicated that the probe intensities of the last replicate were negatively correlated. This could indicate a swap in CYS3 and CYS5 for these samples (e.g. the background was labelled with the wrong dye). To ensure data integrity, I decided to remove these samples from the analyses.

To examine the role of the microarray design in introducing signal variance, I filtered the most variable FAIRE-probes. The variability for each probe was determined by taking the aver-

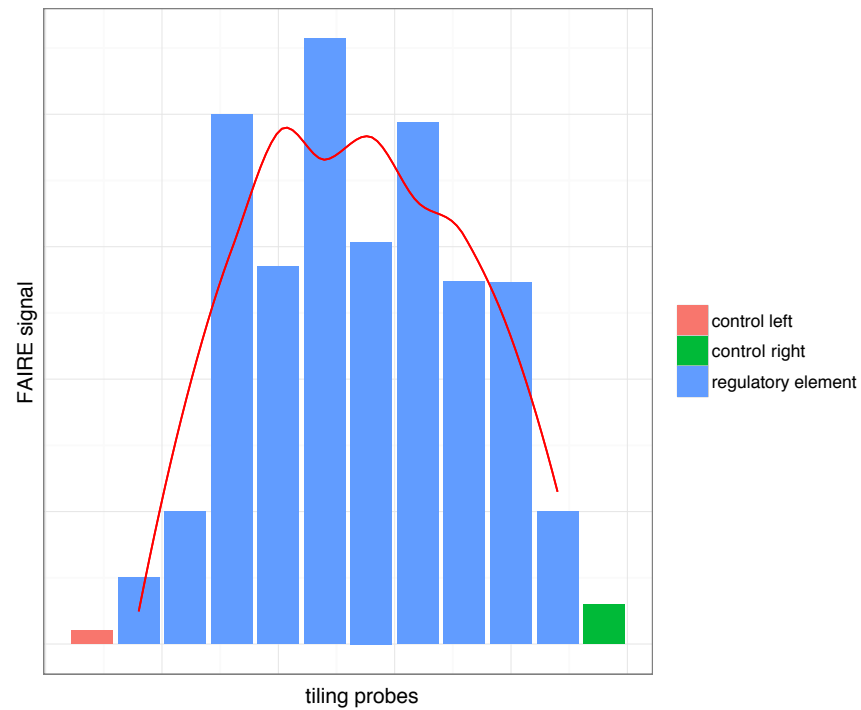


Figure 2.1: FAIRE tiling array design. For each regulatory element of interest, a set of probes is designed to measure the FAIRE signal (blue probes). To ensure that measured effects are due to this specific regulatory element, a background control probe is added on the left (red) and right (green) side.

age between the estimated standard deviation for both intra-individual replicates sets. Using this method, individual specific probe variance will bias the filtering less significantly, as only probes that are highly variable in both individuals will be filtered. The top 1%, 5% and 10% ranking variable FAIRE-probes were filtered out and used to determine the optimal filtering.

For the remaining four replicates, the intra-individual variance of replicates for both individuals was lower than the inter-individual variance, as indicated by reduced correlation in the inter-individual samples (Pearson $R^2 = 0.49 - 0.57$ for within GM12891 (Table 2.1) or GM12892 (Table 2.2), compared with $R^2 = 0.41 - 0.47$ for between GM12891 and GM12892 (Table 2.3)).

Examining the effect on the Pearson correlation coefficients (R^2) for both the inter-individual (Table 2.3) and intra-individual cases (Table 2.1, GM12891; Table 2.2, GM12892) comparisons indicated that I could reduce the overall noise and improve correlation by filtering the top 10% most variable probes across all data sets.

The comparison of probe intensities between GM12891 replicate 1 against GM12891 replicate 2 is shown (Figure 2.3). The colouring indicates which probes remain after filtering the top 1, 5 and 10% of intra-individual variance. The Pearson correlation coefficient between the two replicates increases from $R^2 = 0.49$ to $R^2 = 0.57$. It is clear that obvious outliers between the two replicates have been removed, and that the probe intensities of the probes after filtering are far more consistent between the replicates.

To study the inter-individual variance, I compare the probe intensities of GM12891 replicate 1 against GM12892 replicate 1 (Figure 2.2). The Pearson correlation coefficient between the two replicates increases from $R^2 = 0.41$ to $R^2 = 0.47$, indicating that the results are more comparable. Importantly, filtering improves intra-individual variance more than it removes inter-individual variance. This is significant because the inter-individual variance is the signal of interest.

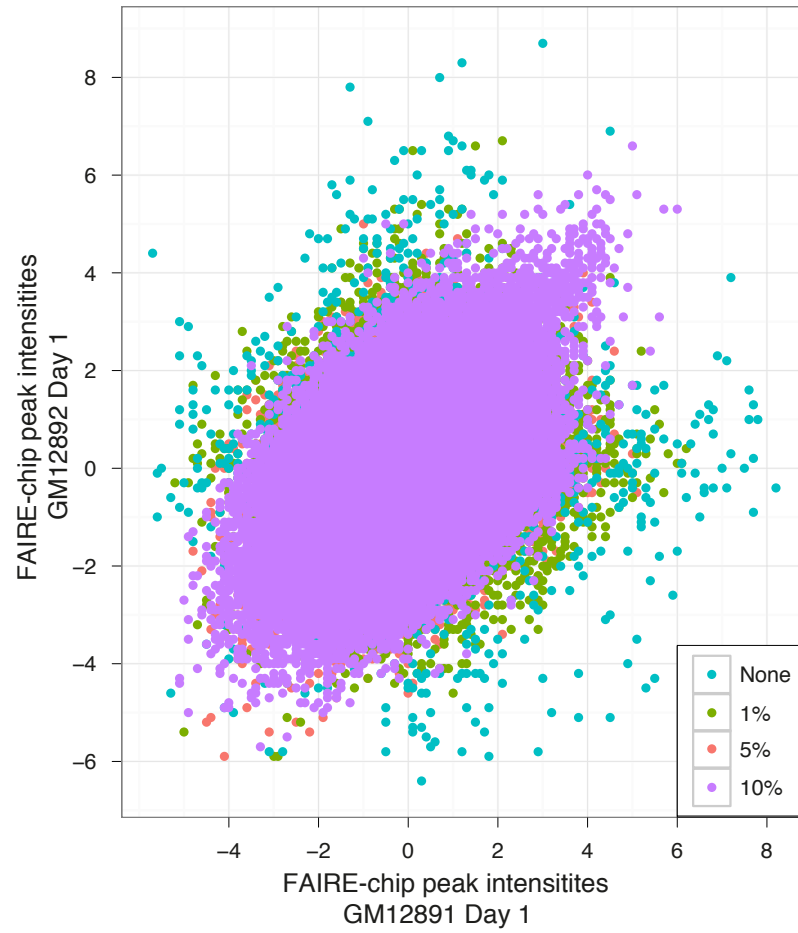


Figure 2.2: Effect of filtering FAIRE probes on GM12891 replicate 1 against GM12892 replicate 1 as an example of inter-individual variance. Probes are ranked by variance and filtered accordingly. Effect for filtering none, top 1%, top 5% and top 10% is shown in colour coding.

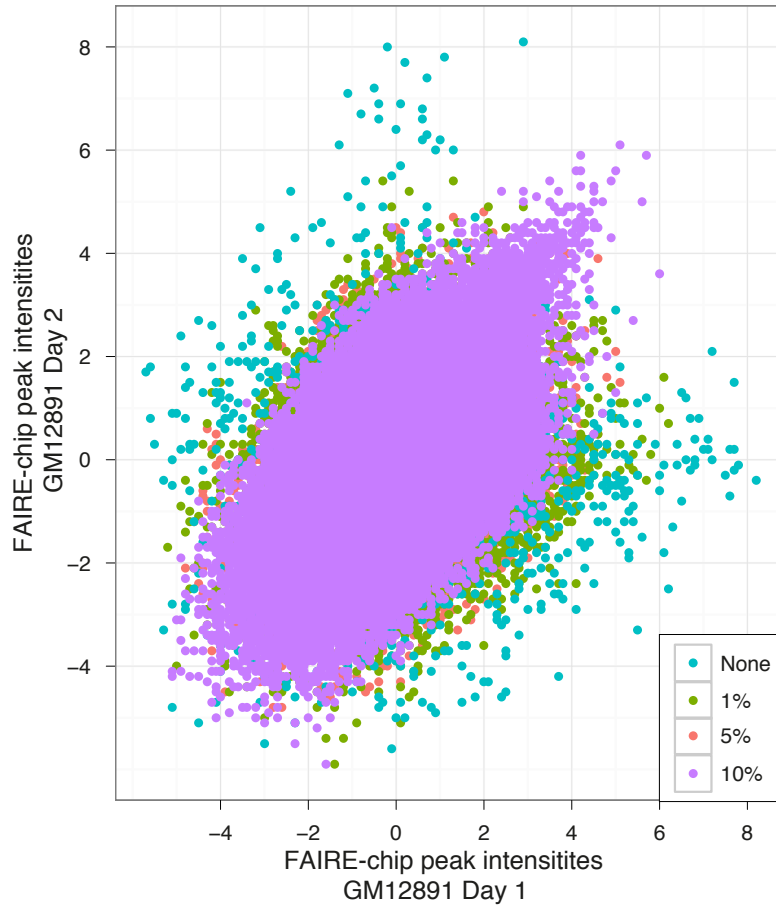


Figure 2.3: Effect of filtering FAIRE probes on GM12891 replicate 1 against replicate 2 as an example of intra-individual variance. Probes are ranked by variance and filtered accordingly. Effect for filtering none, top 1%, top 5% and top 10% is shown in colour coding.

Table 2.1: Pearson correlation coefficient for the intra-individual variance and the effect of filtering for replicate GM12891. The average estimated standard deviation for the probes is given after each filtering step. Probes are filtered in order of their respective variance.

GM12891	SD	Replicate		
		1 vs 2	1 vs 3	1 vs 4
No filter	0.78	0.49	0.52	0.47
1%	0.74	0.51	0.53	0.48
5%	0.66	0.54	0.56	0.51
10%	0.62	0.57	0.58	0.55

Table 2.2: Pearson correlation coefficient for the intra-individual variance and the effect of filtering for replicate GM12892. The average estimated standard deviation for the probes is given after each filtering step. Probes are filtered in order of their respective variance.

GM12892	SD	Replicate		
		1 vs 2	1 vs 3	1 vs 4
No filter	0.63	0.51	0.53	0.57
1%	0.61	0.53	0.56	0.59
5%	0.57	0.57	0.59	0.63
10%	0.54	0.60	0.62	0.65

Table 2.3: The results of the Pearson correlation coefficient for the inter-individual variance. The resulting correlation coefficient for different levels of filtering is given.

GM12891 vs GM12892	Replicate			
	1 vs 1	2 vs 2	3 vs 3	4 vs 4
No filter	0.41	0.55	0.54	0.66
1%	0.42	0.56	0.55	0.68
5%	0.45	0.58	0.57	0.69
10%	0.47	0.59	0.59	0.70

2.2.3 Filtering polymorphic probes

Since the design of the FAIRE-chip, new genotype datasets have been published with novel SNPs and indels. SNPs and indels are a known source for probe measurement variability in microarray experiments when they are within the microarray probes. To

remove this potential bias, I excluded 14,728 FAIRE-probes that overlap with SNPs from the high-coverage trios from the CEU 1,000 Genomes pilot ([1000 Genomes Project Consortium et al., 2010](#)). At the time of this study, this dataset best resembles the population of this study.

2.2.4 *Properties of the filtered FAIRE probes*

Variance in hybridisation can be caused by differences in GC content ([Kuo et al., 2002](#)). In a comparison of GC content between filtered and non-filtered probes, there is no evidence that explains the variance observed in the filtered probes (Figure [2.4](#) and [2.5](#)), and the probe filtering did not exclude any of the regions from the analysis (i.e. at least one probe for each regulatory region remained).

Furthermore, comparison of the QTL analysis after removal of these highly variable probes indicated that the removal of these probes resulted in more QTLs (Section [2.2.5](#) and Figure [2.6a](#)).

Therefore, for the discovery of FAIRE-QTLs, I applied a probe-filtering step prior to calculating the FAIRE-peak signals. The top 10% of FAIRE-probes sorted by variability (measured by standard deviation) in the 4-day replicates dataset were filtered. From these highly variable FAIRE-probes, 1,497 overlap the probes that are filtered due to being polymorphic. Filtering both the highly variable and the polymorphic FAIRE-probes removes 26,992 FAIRE-probes (19.6% of the FAIRE-probes (Table [2.7](#))). After filtering, the probes are summarised into probesets using RMA, which I refer to as FAIRE-peaks.

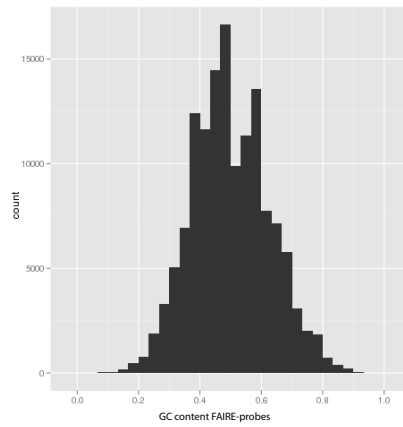


Figure 2.4: GC content for the FAIRE probes that are non-polymorphic and have low FAIRE variance.

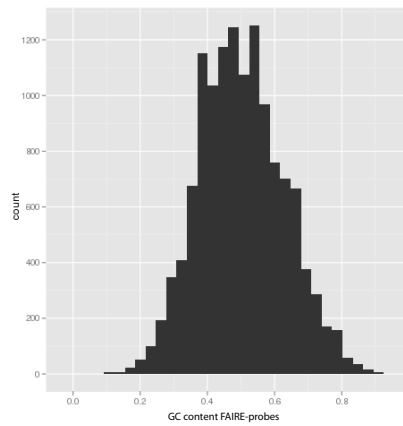


Figure 2.5: GC content for FAIRE probes that are polymorphic or are have high FAIRE variance.

2.2.5 Finding proximal QTLs

In this section, I will use the 10,506 FAIRE-peaks and HapMap genotypes to discover FAIRE-QTLs. Analogous to eQTL studies (Stranger et al., 2007b), I identified associations by calculating the Spearman's Rank correlation. I correlate the signal between each FAIRE-peak and proximal SNPs. I associate all SNPs within 500kb up and downstream of the FAIRE-peak (1Mb window with FAIRE-peak in the middle), because with the current sample size, only QTLs showing large effects can be discovered, and these are mostly proximal. Empirical p-values under the null distribution were determined by permuting the FAIRE annotation of the samples (N=1000, Section 2.5.4 for this method). I consider SNPs within this window as being cis-acting. Associations were calculated for SNPs genotyped in 56 of the individuals in HapMap v3. Both of the replicate cell lines (GM12891 and GM12892) used for filtering the FAIRE-probes were excluded from the QTL analysis to avoid potential bias due to the applied filtering method.

To study the impact on the QTL analysis of filtering the highly variable probes identified from the replicate analysis, I compared associations calculated with and without variable probe filtration (Figure 2.6a). Comparisons of the Spearman's Rank p-values by using a Quantile-Quantile-plot (QQ-plot) indicate that the unfiltered dataset revealed few, if any, significant associations. Upon filtering, the association results deviated from the normal distribution at observed p-values smaller than 1.95×10^{-6} , providing a strong indication of significance above this level. The top ranking association showed genome-wide significance at a p-value of 3.7×10^{-11} .

The associations based on the filtered data have lower p-values and pass a more stringent empirical p-value threshold (Figure 2.6b, Figure 2.6c). This leads to more significant association results after permutations when filtering is applied. In total, significant QTLs can be found for 42 FAIRE-peaks (empirical

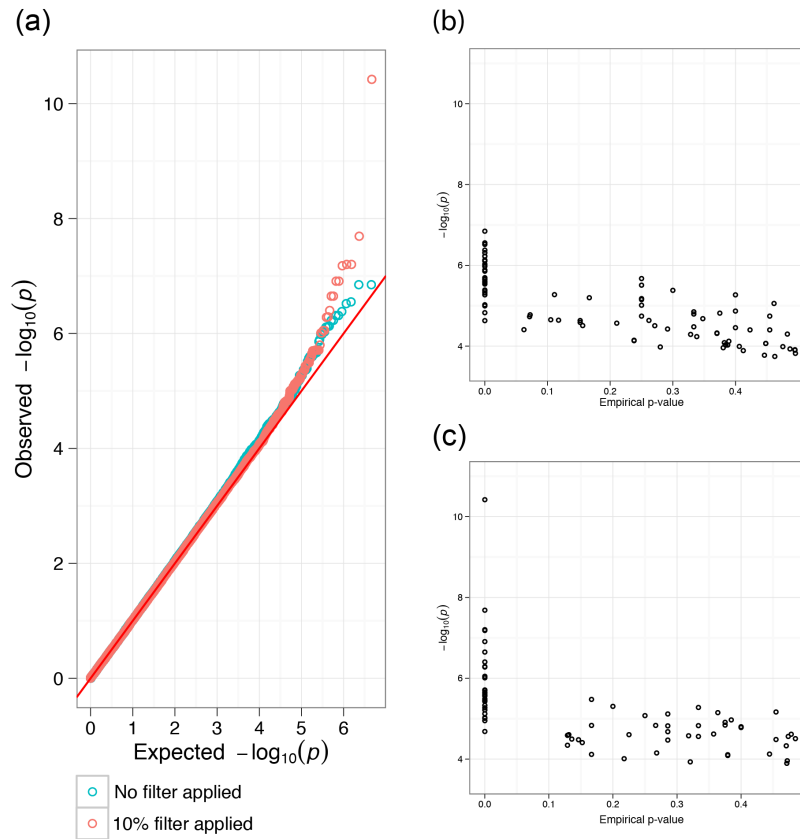


Figure 2.6: FAIRE-QTLs QQ plot. For filtered and unfiltered FAIRE-peaks, QTLs are mapped and empirical p-values are calculated. In the QQ-plot, the filtered FAIRE-peaks result in more significant results (a). The permutation empirical p-value approach demonstrates that there are fewer significant FAIRE-QTLs before filtering (b) than after filtering (c).

p-value ≤ 0.5). Several QTLs are significantly associated with multiple FAIRE-peaks, resulting in a total of 173 FAIRE-QTLs (Table 2.4 brief overview FAIRE-QTLs, Supplementary Table B.1 all FAIRE-QTLs). Given the empirical p-value threshold for significance that was used, there is a considerably high level of potential false discoveries. In the discussion section of this chapter, I will address this concern in more detail.

Table 2.4: Brief overview FAIRE-QTLs. The number of FAIRE-QTLs, FAIRE-peaks and overlap with eQTLs is given for different empirical p-values.

Empirical p-value	FAIRE-peaks	FAIRE-QTLs	Overlap with eQTLs
<0.1	20	48	1
<0.2	24	94	19
<0.3	28	107	21
<0.4	31	121	24
=<0.5	42	173	45

2.2.6 Functional analysis of FAIRE-QTLs

To understand the functionality of the discovered FAIRE-QTLs, a more detailed analysis of these associations is performed. First, I examine a FAIRE-QTL example. Second, I overlap FAIRE-QTLs with other regulatory elements. Third, I overlay FAIRE-QTLs with whole-body phenotype GWAS results. Finally, I overlay FAIRE-QTLs with other molecular QTLs.

2.2.6.1 Example FAIRE-QTL

To illustrate how a FAIRE-QTL might look, I picked the following example. For FAIRE-peak 6090, located on chromosome 9 (138,444,546 to 138,446,041), the best association is with rs2590504 (Spearman's Rank p-value of 6.3×10^{-8}) (Figure 2.7a). Our probe design permitted me to confirm that the association is specific to the targeted region and not part of a larger genomic feature. I also tested the significant SNP for correlation with the left and the right control probe (Figure 2.7b and Figure 2.7d). In both cases, the correlations were not sufficiently significant to pass the permutation threshold (empirical p-value > 0.5) (Figure 2.7c, p-values are 0.065 and 0.007, respectively). This contrasts with the target region, in which the FAIRE-peak is significantly correlated with the SNP (Figure 2.7d).

The location of the most significant SNP for each FAIRE-peak is studied to determine if there is any location preference with respect to the FAIRE-peak. Half of the FAIRE-QTLs are within 157,158bp distance of the targeted FAIRE-peak; this quantity is significantly larger than expected (Figure 2.8, Fisher Exact Test p-value = 0.001). This result is similar to eQTLs studies in which SNPs proximal to the gene or transcript are most often the eQTLs that are discovered.

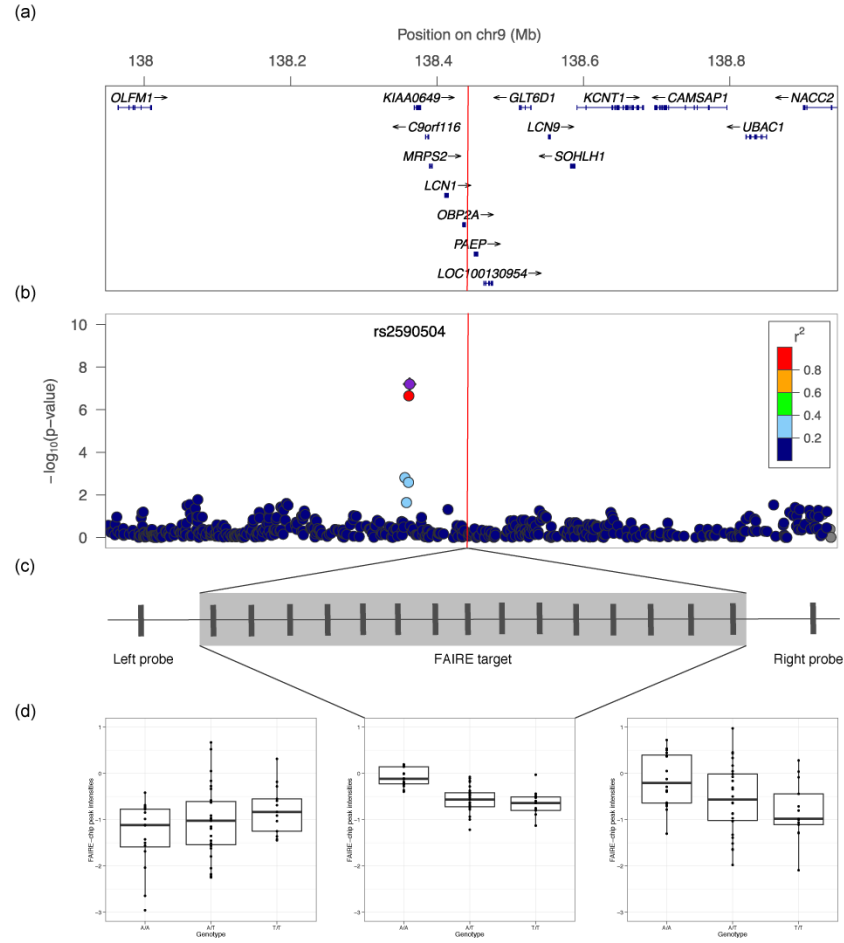


Figure 2.7: Example FAIRE-QTL for FAIRE-peak 6090, on chromosome 9 from 138,444,546 to 138,446,041. a) The genes surrounding the FAIRE-peak. b) All proximal FAIRE-peak SNPs that have been tested for association with rs2590504 are coloured red; our FAIRE-QTL and all other SNPs are coloured by their LD with respect to rs2590504. c) Schematic overview of FAIRE-peak probes and flanking left and right probe. d). Boxplots with associations of left flanking probe, the FAIRE-peak and the right flanking probe with rs2590504.

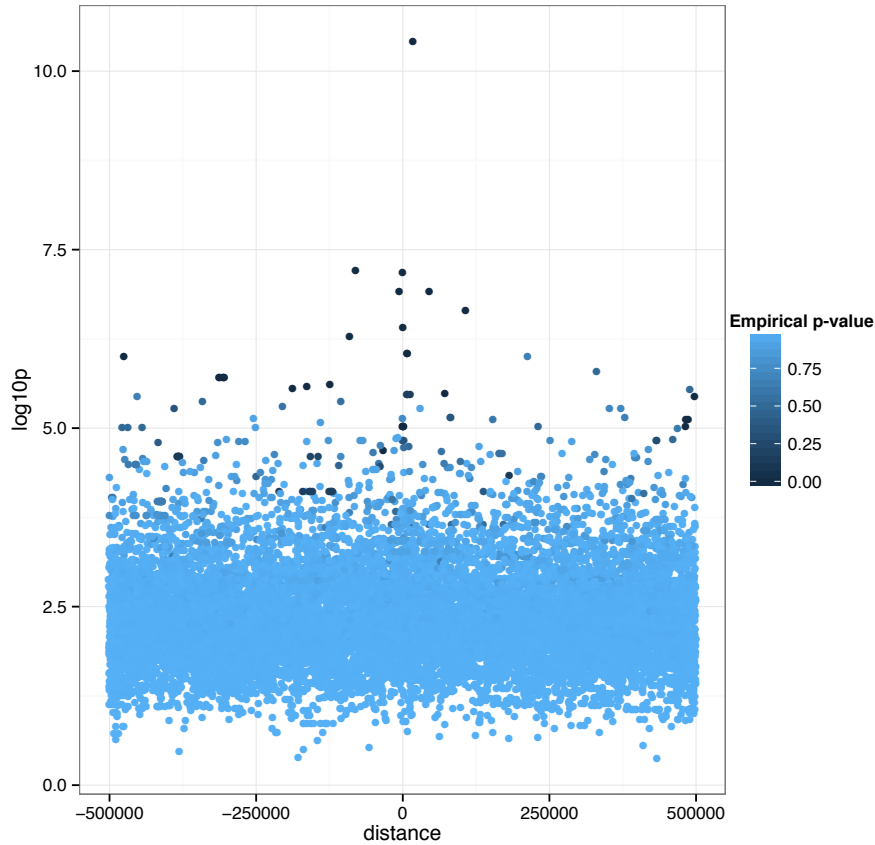


Figure 2.8: Distance of the strongest FAIRE-QTL to its FAIRE-target. The p-value of the FAIRE-QTL is given on the Y-axis and the distance on the X-axis. Colour coding indicates the empirical p-value under the null distribution. Half of the significant associations are within 157,158bp of the FAIRE-peak.

2.2.6.2 Regulatory enrichment of FAIRE-QTLs

To understand what other regulatory information is known about the FAIRE-QTLs, I evaluated the data published by ENCODE ([ENCODE Project Consortium, 2012](#)). I selected all 173 significant FAIRE-QTLs (empirical p-value ≤ 0.5) and overlapped them with the regions identified by ENCODE as being functional, using the Regulome DB tool ([Boyle et al., 2012](#)). The Regulome DB tool will return a score indicating the level of regulatory data that overlaps each SNP. A Regulome DB score is a combination of a number and a letter. The number is the

main score, where a score of 1 indicates that the SNP is an eQTL and is overlapping transcription factor (TF) binding, while a 2 indicates there is an overlap with a TF but not with an eQTL. Furthermore, the scoring varies depending on whether the SNP overlaps with a DNase footprint, DNase peak and TF motif. The letter portion of the score provides information about how important the overlapping dataset is; for example, a TF motif is more important than being located within a DNase peak. For a complete description of the Regulome DB scoring scheme, refer to Section [2.5.7](#)

For five SNPs, Regulome DB reported that the SNPs might be tri-allelic in some other population than CEU, and thus excludes them. For 36 of the SNPs, Regulome DB reported the lowest score, indicating no overlap with ENCODE data. More interestingly, 28 SNPs were assigned the highest score, indicating an overlap with at least one eQTL, and 9 had a score of 2, indicating an overlap with a transcription factor binding region and at least 2 other features (Table [2.5](#)). The highest-scoring SNP, rs10814581, was identified with a score of 1b. This score indicates that this SNP is an eQTL, inside a TF binding, a DNase Footprint, and a DNase peak. Refer to Section [2.5.7](#) for RegulomeDB scoring.

I identified 54 FAIRE-QTLs that map within ENCODE FAIRE peaks. There are 71 FAIRE-QTLs overlapping DNase-seq peaks and that are identified as regions involved in chromatin structure. This gives a strong indication of the validity of these FAIRE-QTLs as being directly involved in rearranging chromatin accessibility. For nine of our FAIRE-QTLs, there is an overlap with CTCF binding sites. For 50 FAIRE-QTLs, there is an overlap with one or multiple motifs. These results permit me to have high confidence that the identified FAIRE-QTLs are SNPs that have a regulatory function.

Table 2.5: Enrichment of FAIRE-QTLs in ENCODE using Regulome DB. The score indicates with how many regulatory datasets the FAIRE-QTL overlaps and whether a given SNP is an eQTL. A Regulome DB score of 1 indicates that the SNP is an eQTL and overlaps transcription factor (TF) binding, while a 2 indicates that there is an overlap with a TF but not with an eQTL. Furthermore, the scoring varies depending on whether the SNP overlaps with a DNase footprint, DNase peak and TF motif. For a complete description of the Regulome DB scoring scheme, refer to Section 2.5.7.

Regulome DB score	Number of FAIRE-QTLs
1b	1
1d	3
1f	24
2a	1
2b	8
3a	2
4	11
5	32
6	50
7	36
Excluded	5

2.2.7 *Comparison of FAIRE associations to other association studies*

I selected known associations for autoimmune-related diseases including Crohn's disease, ulcerative colitis and Type 1 diabetes. I used the most recent list of known associations from the Wellcome Trust Case-Control Consortium (WTCCC) ([Wellcome Trust Case Control Consortium et al., 2007](#)) and the meta-analysis results for Crohn's disease and ulcerative colitis ([Jostins et al., 2012](#)).

Because both our study and many GWAS were designed to genotype functional regions of the genome, this will lead to a bias. It is to be expected that a random selection of genotypes with a phenotype associated to them (e.g. a GWAS hit) will be more likely in non-coding but regulatory regions of the genome, and thus are potentially biased for the FAIRE signal. By adding height as an additional non-disease phenotype as a control that is, most likely, non-immune related, the normal enrichment for FAIRE within GWAS hits can be assessed.

For each SNP associated with a GWAS phenotype, I identified SNPs that were also tested in this FAIRE association study. As any SNP can be associated with several FAIRE-peaks, there can be several FAIRE-QTL p-values for each SNP associated with a disease. I plotted the FAIRE-peak p-values as a QQ-plot for each disease association study and non-disease phenotype to gain insight into potential enrichment of FAIRE-QTLs for other phenotypes.

Coincidental overlap is a major point of concern when assessing the direct overlap of QTLs with different GWAS studies. In the case of a coincidental overlap, the lead QTL or GWAS-hit for both associations is not in Linkage Disequilibrium (LD) with each other, and thus the association is not driven by the same genetic effect. To address this issue for each of the overlapping SNPs, I determined what the LD between the overlapping SNP and the lead FAIRE-QTL for that FAIRE-peak is. To accomplish this, I used PLINK and the 1,000 Genomes CEU population.

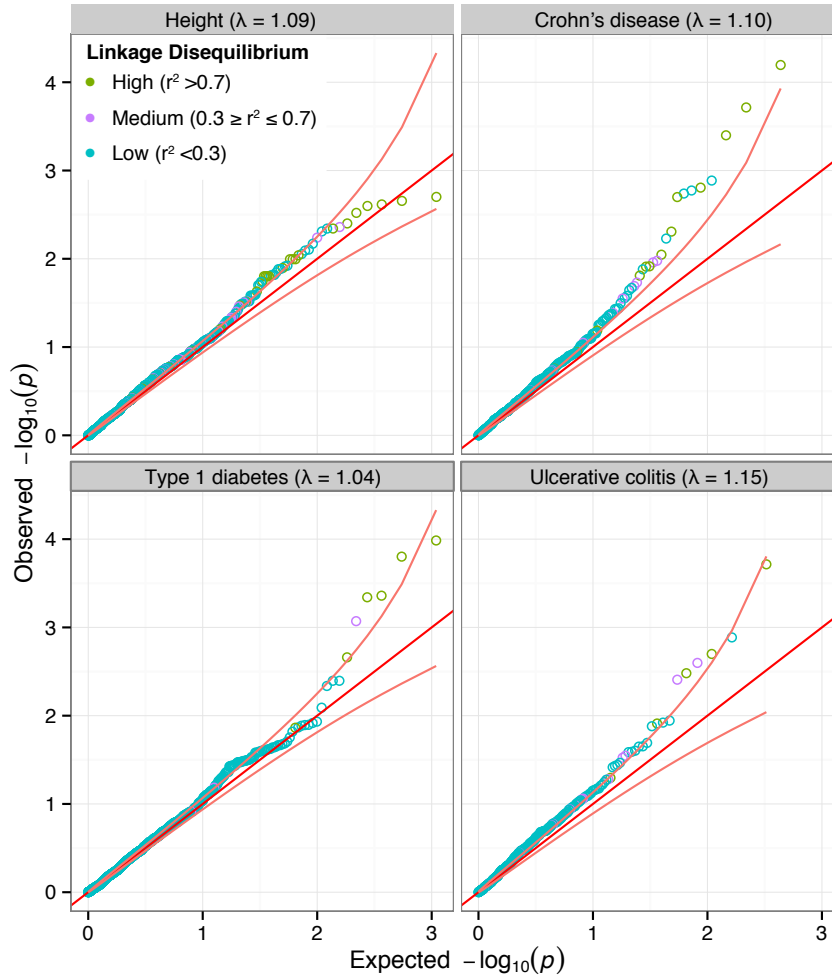


Figure 2.9: FAIRE-QTLs results when genotypes are selected that are identified in GWAS studies. Shown are overlapping FAIRE p-values for Height, Crohn's disease, Type 1 diabetes and Ulcerative colitis. The genomic inflation factor (λ) is given for each QQ-plot; the red lines indicate the 95% confidence interval. All disease phenotypes show FAIRE-QTL enrichment, while height shows no abbreviation from the expected distribution.

When the LD is high (measured in r^2), there is a high confidence that both SNPs are a non-coincidental overlap. I classified each of the overlapping SNPs as low LD ($r^2 < 0.3$), medium LD (r^2 between 0.3 and 0.7) and high LD ($r^2 > 0.7$).

For several of the autoimmune-related diseases, the LCL FAIRE-QTL p-values deviated from the expected uniform distribution, indicating that the disease-associated SNPs at the tail of the uniform distribution are likely to also influence chromatin accessibility. In the other non-immune-related GWAS phenotypes, including height, the QQ-plot does not deviate from the expected uniform distribution, suggesting that the FAIRE-QTLs p-values are uniformly distributed, and that none of the LCL FAIRE-QTL SNPs show evidence of being enriched (Figure 2.9).

However, for none of the disease-associated SNPs I was able to identify significant FAIRE-QTLs (empirical p-value > 0.5). As our sample numbers are small, this is most likely due to the effect size of the chromatin signatures nearby the selected SNPs being too small to be detected as significant.

2.2.7.1 *Expression QTLs*

Several studies have examined the effect of noncoding variants on a variety of molecular traits, including RNA expression, chromatin accessibility, and transcription factor binding. To better understand the interplay of these molecular traits and to identify the extent of FAIRE-QTL involvement in expression variation, I examined the overlap of the FAIRE-QTLs identified here with other molecular trait datasets.

RNA-sequencing has been performed on the same 60 individuals. In these data, expression-QTLs (eQTLs) have been identified for whole-transcript and exon expression levels using the same HapMap genotypes (Montgomery et al., 2010). In Chapter 3 and Chapter 4, I will discuss the direct correlation of FAIRE to gene expression. Presently, I overlap the FAIRE-QTLs with the eQTLs from Montgomery et al. The eQTLs have been published with three different levels of permutation FDRs (< 0.001 : 103 eQTLs,

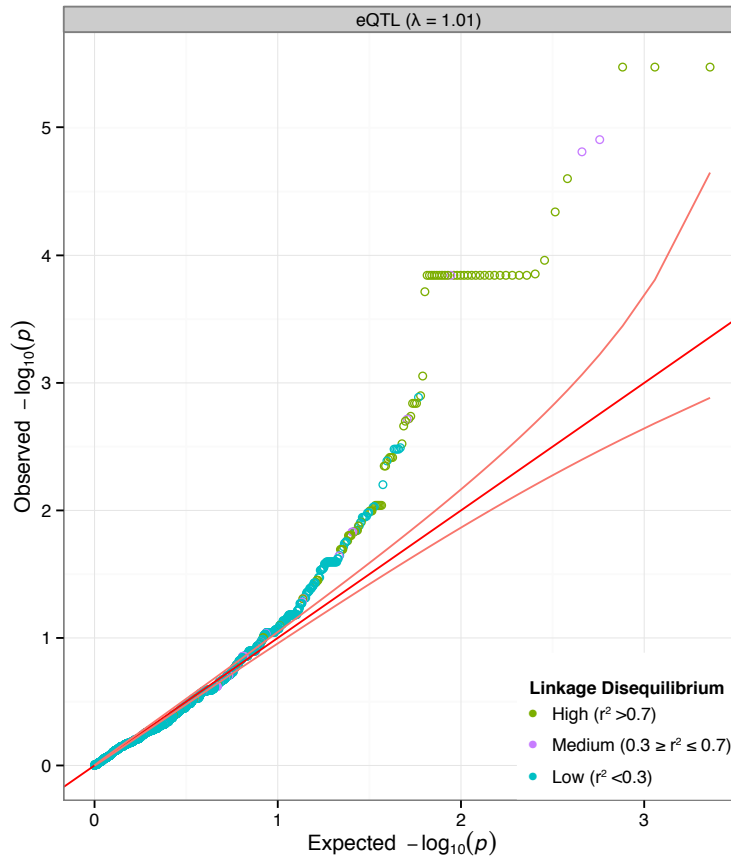


Figure 2.10: FAIRE-QTLs show enrichment for expression-QTLs (Montgomery et al., 2010). The genomic inflation factor (λ) is given and the red lines indicate the 95% confidence interval. The colour coding indicates the Linkage Disequilibrium (LD) between the lead QTL SNPs for both phenotypes when there is an overlap. The LD shows that SNPs with better p-values are nearly all non-coincidental overlaps.

< 0.01 : 836 eQTLs and < 0.05 : 3,258 eQTLs). From the set of eQTLs at $FDR < 0.05$, there are 45 that are also FAIRE-QTLs.

I selected all SNPs that are eQTLs ($FDR < 0.05$) and tested them for being a FAIRE-QTL. If an eQTL has been associated with several FAIRE-peaks, only the best FAIRE-QTL is used. Similarly to the overlap with GWAS phenotype, I determined whether the overlapping SNPs are coincidentally overlapping or not. Most of the SNPs overlapping are in low LD (3,096), and only a minority is in medium (60) or in high LD (102). In the QQ-plot, if I actively selected for known eQTLs in our study, I observed a strong signal for significant FAIRE-QTLs (Figure 2.10). This indicates that many of the genotypes involved in regulating expression are also likely to be involved in regulating chromatin signatures. The plot also indicates that a group of overlapping QTLs are in tight LD with each other. This is revealed by the sudden horizontal trend in the QQ-plot. The FAIRE-QTLs with the highest p-values tend to be in high LD with lead eQTL, indicating that they are driven by the same, or similar, genetic effects.

2.2.7.2 DNase I QTLs

Chromatin accessibility can also be measured using DNase I hypersensitivity. In a recent study, Degner et al. have applied DNase-seq on 70 individuals from YRI descent (Yoruba in Ibadan, Nigeria) (Degner et al., 2012). By using DNase-seq as a quantitative molecular trait, 8,902 DNase I-sensitive QTLs (dsQTLs) were identified using HapMapv3 genotypes imputed with the 1,000 Genomes Project genotypes. Because the sample size of the dsQTL study is larger and the YRI population has more genetic variation than the CEU population (1000 Genomes Project Consortium et al., 2012), the dsQTL study has greater statistical power to detect QTLs.

As I only use HapMap genotypes, I remove all non-HapMap genotypes, leaving 6,464 dsQTLs. For each of the 6,464 dsQTLs, I select the p-value from the FAIRE association. If a dsQTL has

been associated with several FAIRE-peaks, only the FAIRE-QTL with the lowest empirical p-value is used. For many of the dsQTLs, there is no significance to any of the tested nearby FAIRE-peaks.

For each FAIRE-QTL that is with a dsQTL, I determine if the overlap is coincidental or not by determining the LD. Most of the SNPs overlapping are in low LD (6,227) and only a minority is in medium (176) or in high LD (61). Despite the two studies being performed in different populations, the QQ-plot of the FAIRE-QTLs that overlap indicates that there is an overlap of chromatin accessibility QTLs (Figure 2.11). More importantly, the overlapping QTLs with a low p-value (high in $-\log_{10}$) tend to be the FAIRE-QTL or in high LD (green dots) with the FAIRE-QTL and thus are non-coincidental overlaps.

I also noticed that there is a stronger deviation for the eQTLs for FAIRE-QTLs identified in the matched European population (Figure 2.10) than for the dsQTLs identified in the Yoruban population, suggesting that there are haplotypes specific to each population that are drivers for both expression and chromatin accessibility.

2.2.8 *Low coverage sequencing data to improve QTL discovery*

The availability of tag-SNP genotypes from the HapMap Project and the low coverage sequence genotypes from the 1,000 Genomes Project provided the opportunity to model the results achievable using genome sequence-level data. I had originally selected the entire set of the samples to be part of the expected Phase1 release of the 1,000 Genomes Project set. However, currently only 46 of the samples have been sequenced and published. Here, I refer to the genotypes derived from the 1,000 Genomes Phase1 release as the "denser" genotypes. Using these denser genotypes requires an increased number of statistical tests for the same window than when using HapMap geno-

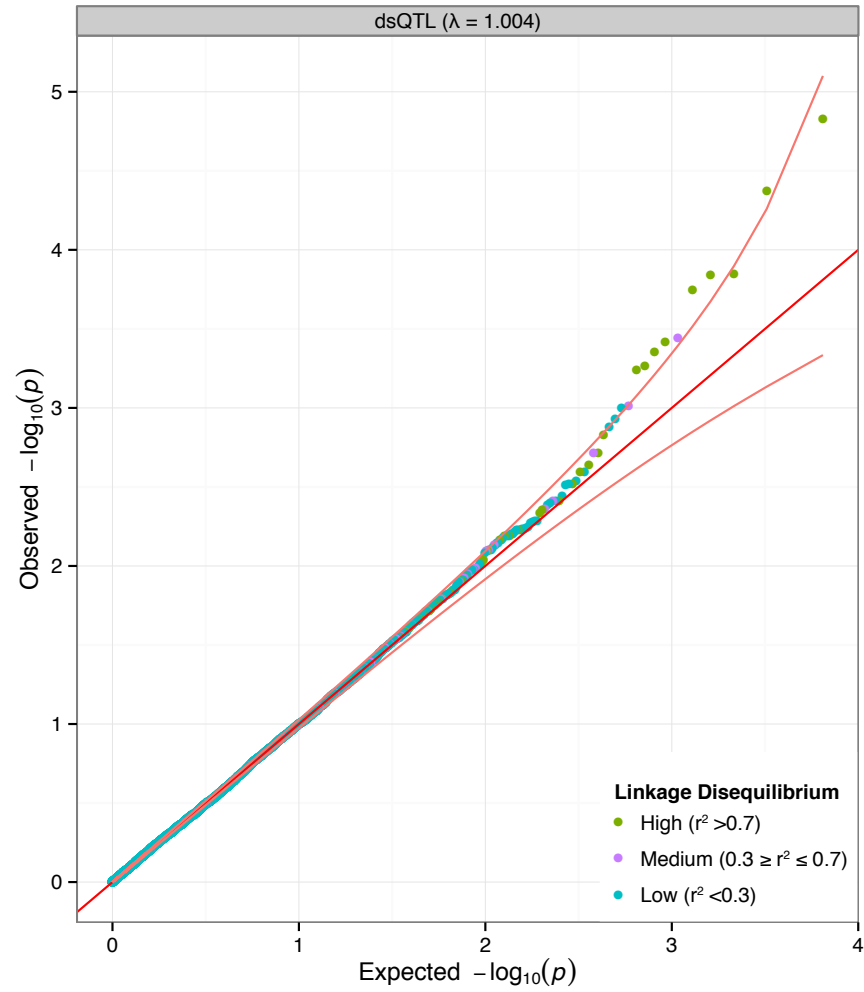


Figure 2.11: FAIRE-QTLs show enrichment for dsQTLs. There are 6,464 dsQTLs overlap with SNPs that are tested for being a FAIRE-QTL. There is enrichment evident, indicating that some of the dsQTLs are also FAIRE-QTLs. The light red lines represent the 95% confidence interval for the expected distribution of p-values. The colours of the points indicate the Linkage Disequilibrium (LD) between the overlapping SNP and the lead FAIRE-QTL for that FAIRE-peak. Overlapping SNPs with a high p-value are nearly all in high LD with the lead FAIRE-QTL for their respective FAIRE-peak.

types. Because of this, lower p-values are required to pass the equivalent empirical p-value threshold.

Nevertheless, comparison of the associations identified using both genotypes showed that in many cases it is beneficial to have denser genotypes available. This is because the causal genetic variation, or a SNP in stronger LD with the causal genetic variation, was present in the denser genotypes and thus easier to detect even with lower sample size. To illustrate this, I selected two FAIRE-QTL examples.

The first example involves FAIRE-peak 6090. I was able to reconfirm the FAIRE-QTL to the same locus with the denser genotypes. However, when the same SNP was tested for association, the observed p-value was worse due to the lower sample size (Figure 2.12). In addition, due to the greater genotype density, the permutation-derived empirical p-value will be lower because of the larger null distribution.

The second example involves FAIRE-peak 12062. To identify the FAIRE-QTL using the original HapMap genotypes, the SNP is not in enough LD with the causal locus to obtain a significant p-value (Figure 2.13). When I re-analysed the same FAIRE-peak with the 1,000 Genomes SNPs, I was able to obtain a significant FAIRE-QTL. Consequently, while the statistical power was lower due to the more limited sample size, new FAIRE-QTLs could be identified.

A similar number of FAIRE-QTLs are discovered by using HapMap as with the 1,000 Genomes genotypes. However, when considering our empirical p-value cut-off, I identified more FAIRE-QTLs with the HapMap genotype set than the 1000 Genomes genotype set. For this reason, I focused the main analysis on FAIRE-QTLs using the HapMap genotypes.

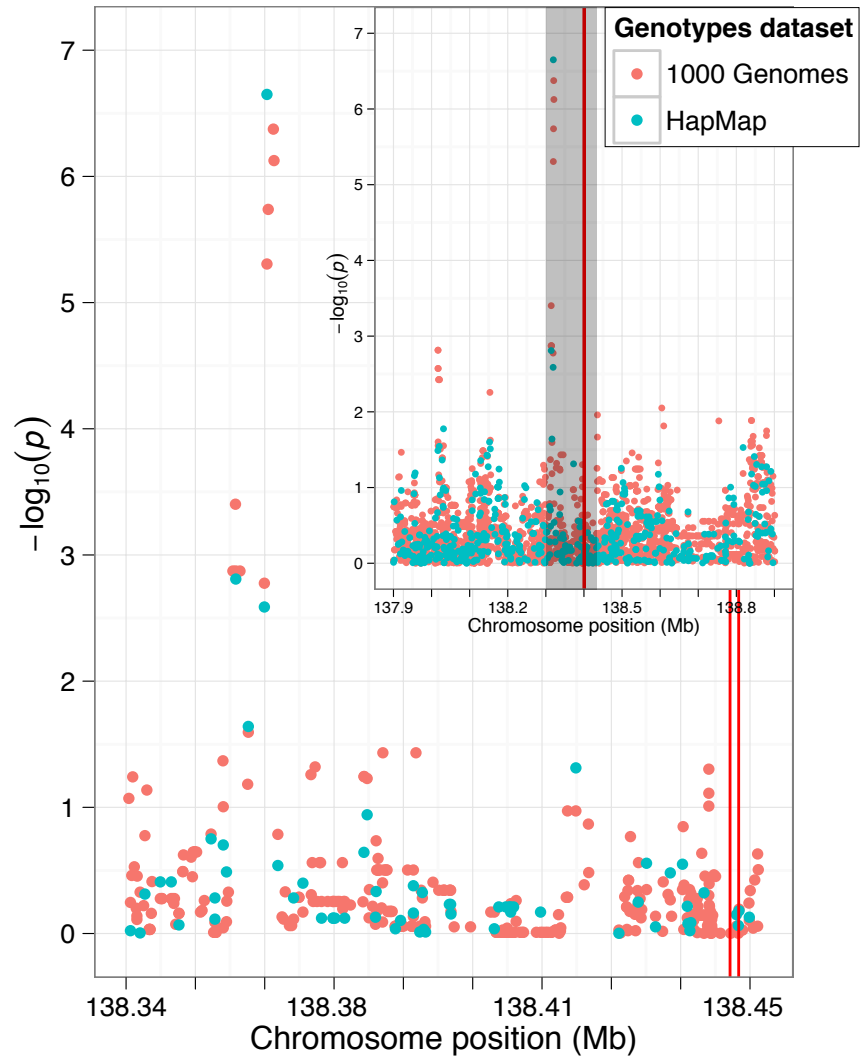


Figure 2.12: Denser genotypes reconfirm the previously-identified QTL. Due to lower sample numbers, when the same SNP is tested for association, the p-value is worse for 1,000 Genomes genotypes. Inset is the zoomed-out FAIRE-target region. The red lines outline the FAIRE-peak target.

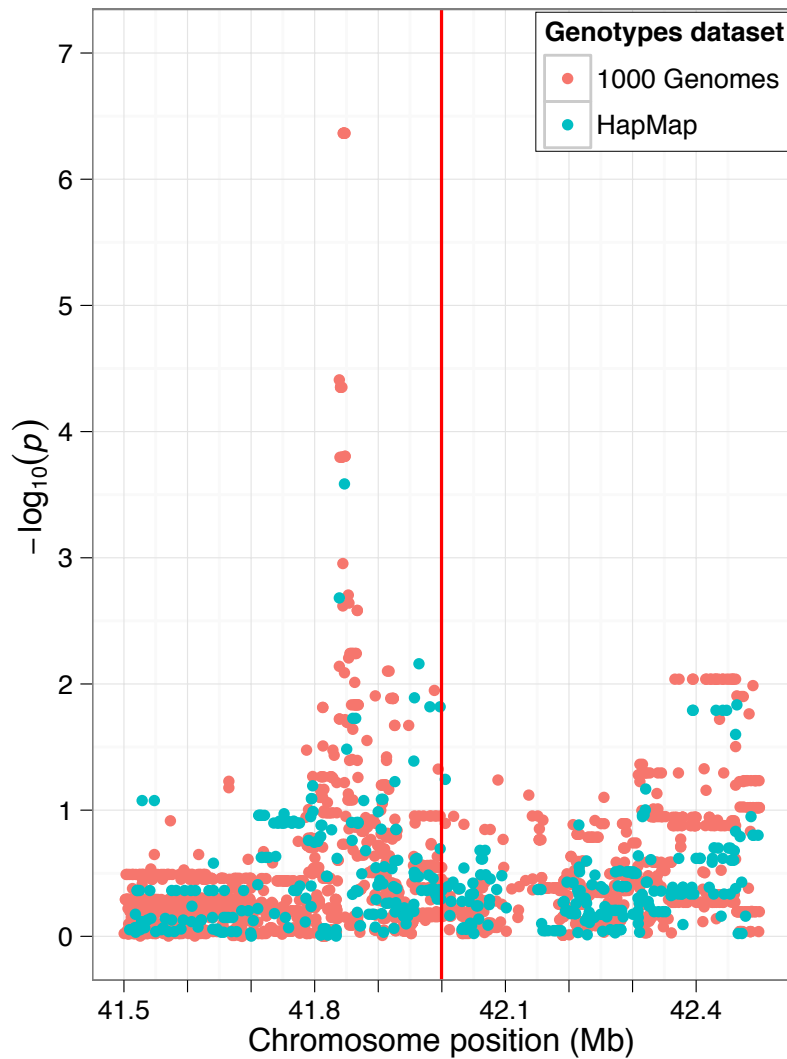


Figure 2.13: Denser genotypes are beneficial in identifying greater number of QTLs. When denser genotypes are available, these can help to identify new significant QTLs that would not have been discovered with less dense genotypes. In this example, the dense genotypes from 1,000 Genomes enable identification of additional QTLs. The red lines outline the FAIRE-peak target.

2.3 CONCLUSION

In this study, I utilised the FAIRE signal to identify chromatin accessibility QTLs in a small population. Targeting 10,506 genome locations with known regulatory properties using tiling arrays, I was able to identify 173 FAIRE-QTLs (empirical p -value < 0.5) for 41 unique FAIRE sites. Furthermore, I will show that denser genotypes from the 1,000 Genomes Project can be beneficial. Using a denser genotype set, I identify additional FAIRE-QTLs. However, because of the increased multiple testing penalties and lower sample size, I used the HapMap genotypes for the identification of FAIRE-QTLs. One in four of the identified FAIRE-QTLs were previously also identified as an eQTL.

Despite targeting only a small portion of the genome, using a relatively small sample size, I identified strong associations of SNPs with chromatin accessibility assayed by the FAIRE method. This leads to new insights for several SNPs around the genome in non-coding regions and suggests the function of regulating chromatin directly or indirectly. Specific genotypes can lead to the increase and decrease of accessibility, and thus are a major force in regulating expression on a molecular level. As I have shown, many of our identified FAIRE-QTLs are involved with or related to biochemical events in the genome identified by the ENCODE project. This strengthens evidence that these FAIRE-QTLs are directly involved in adjusting and regulating gene expression. However, without measuring additional assays in the same individuals, it is impossible to ascertain the real molecular function associated with these SNPs.

An important lesson for study design from this work is the addition of replicates when working with cell lines. I determined that a small number of probes tend to have intra-individual variation that can give misleading inter-individual signals for the targeted region. By removing the top 10% most variable probes showing intra-individual variation, I was able to boost the sensitivity of our study, enabling application of more stringent em-

pirical p-value thresholds without losing any associations. This knowledge strengthens the case for ensuring that population studies include multiple replicates for a few individuals.

In comparison with early tiling array eQTL studies, I found fewer associations. The Pritchard group, using Illumina whole-genome expression arrays (Stranger et al., 2007a), were able to identify associations for 6.5% of targeted genes (Veyrieras et al., 2008), compared with our results identifying a significant association for 1.6% of the targeted functional ENCODE elements at the least stringent empirical p-value threshold. This difference in yield between studies can be explained by our study design focusing only on targeted regions of the genome, in which I might not cover FAIRE regions of interest. However, the most important reason is that in our study there is a much smaller number of samples. Potentially, this also shows the limitation of using tiling arrays for the readout. With the vast amount of data published by the ENCODE consortium, a more optimal tiling array could likely be designed that includes additional regions showing inter-individual variance. Furthermore, with the decreasing costs of next-generation sequencing, FAIRE-sequencing of all samples could be considered to increase the coverage of FAIRE-QTLs on the human genome.

Many GWAS studies result in large numbers of associations with non-coding genotypes, leaving the molecular explanation unaddressed. The expectation is that many of these non-coding genotypes are in regulatory regions where they, directly or indirectly, influence gene expression. I found that many SNPs identified in GWAS show non-significant enrichment for FAIRE-QTL, and that larger sample sizes could identify the significant SNPs as FAIRE-QTLs. It is likely that these SNPs contribute to the control of gene expression or transcription factor binding either by directly altering protein-binding sites or perhaps through regulating the accessibility of chromatin. Identification of these SNPs as FAIRE-QTLs suggests plausible molecular mechanisms to explain the association with disease. These results strengthen

the idea of medical epigenetics, in which disease samples are not sequenced and genotyped only to find genetic variations, but also include the study of several epigenetic marks and chromatin accessibility. Our data suggest that FAIRE, or any other chromatin accessibility assay, would be a good assay to include in such a setting.

One in four of the FAIRE-QTLs is also an eQTL for transcript or exon expression levels. As these eQTLs are identified using the same individuals, this indicates a close biochemical connection between open chromatin and gene expression. For many of the cases in which an SNP is both an eQTL and a FAIRE-QTL, I found a strong correlation between the transcript levels and the FAIRE-peak intensities, which will be discussed further in Chapter 3 and Chapter 4.

Having established the benefits of a chromatin accessibility assay in an association study, the choice of FAIRE in this study is likely not optimal. Earlier publications have shown that there is a high amount of overlap between DNase I hypersensitive regions and FAIRE regions, which would make their results complimentary (Song et al., 2011). In addition, it is known that FAIRE has lower sensitivity at a similar sequence coverage than DNase I (Kumar et al., 2013), making it the less favourable assay in current practice. These readouts will create bias in the analysis and lead to false positives in the association study. At the moment, no DNase I sequencing data exists for the same cell lines as in this study. Only a couple of cell lines (mainly GM12878) are part of the central ENCODE cell lines, and thus have DNase-seq available. However, these are not nearly sufficient to perform a similar study as ours, because the sample number is too low to reach any level of significance. Instead, these DNase I data could be used to assess whether the signal is consistent with the FAIRE-chip signal.

In later chapters (Chapter 3 and 4.) it is established that a better signal for the FAIRE-chip is obtained when using each FAIRE-probe independently as a measurement of chromatin

accessibility. Using the 1,000 Genomes imputed genotypes, also used for mapping CTCF-QTLs (Chapter 3), I am able to identify 16,972 FAIRE-probe QTLs with a significant empirical p-value of < 0.01 . This indicates that for inter-individual differences that can be measured by FAIRE, they are most often located in more narrowly-defined regions. Hence, using the independent FAIRE probes will increase the amount of FAIRE-QTLs tremendously and would make subsequent analysis potentially different than their current presentation. It is noteworthy that when a FAIRE-peak QTL has been discovered, the same genotype is also associated with at least one of the individual FAIRE-probes. Consequently, the FAIRE-probe QTL set is a superset of the FAIRE-peak QTL set. For publication of the present chapter, it will be best to use only the imputed genotypes associated with the FAIRE probes as our main finding. It is still of interest to report that some larger regions of the FAIRE signal are associated with a QTL, in contrast with the FAIRE-probe QTLs.

2.4 DATA

2.4.1 *Genotype data*

I used HapMap Phase III version 2 genotype data for 57 of the individuals in our main analysis ([International HapMap Consortium, 2003, 2005, 2007](#)). For 46 of these individuals, I used the June 2011 release genotypes of the 1,000 Genomes Project as our denser genotypes data ([1000 Genomes Project Consortium et al., 2010, 2012](#)). All individuals are from the CEU population. Genotypes with a minor allele frequency (MAF) < 0.05 were filtered.

2.4.2 FAIRE-chip data

I used FAIRE intensity levels that were measured in lymphoblastoid cell lines (LCLs) from 60 unrelated individuals of the CEU HapMap (HapMap individuals of European descent) population. The FAIRE intensities used were measured with our custom-designed Nimblegen chip array (FAIRE-chip). Regions of interest were selected by using the ENCODE pilot data. The targeted regulatory regions have shown a high level of variation for DNaseI or CTCF signals between four individuals ([McDaniell et al., 2010](#)). The targets are evenly distributed across the genome and were prioritized by the presence of an allele-specific signal from the four individuals, indicative of a possible functional genetic polymorphism. For each target, a dense set of between 10 and 21 probes of variable length (40-60bp, average 46bp) with, on average, 110bp intervals were designed to measure the FAIRE signal.

2.4.3 Expression-QTLs

For exons and whole transcripts, eQTLs were obtained using the mRNA fraction of the transcriptome ([Montgomery et al., 2010](#)). These eQTLs are mapped within same 60 LCLs as my study using 37-base pairs paired-end Illumina sequencing. In total, 4,781 eQTLs are reported.

2.4.4 GWAS and molecular QTL datasets

I selected different sets of molecular QTLs, whole-body QTLs, and immune-disease GWAS results from various sources (Table [2.6](#)).

2.5 METHODS

2.5.1 *Robust Multi-array Average*

I used the NimbleScan software version 2.6 www.nimblegen.com/products/software/nimblescan/index.html to obtain Robust Multi-array Average (RMA) (Irizarry et al., 2003) processed expression values for our FAIRE-chip microarrays. RMA of microarrays consists of the following steps:

1. The raw intensity probe values are background-corrected
2. The corrected probe values are Log₂-transformed
3. The Log₂ probe values are quantile-normalized
4. A linear model is fit to the normalized data to obtain an intensity measure for each probe set on each array.

2.5.2 *Filtering FAIRE-probes*

For filtering and quality control purposes, five replicates have been generated for GM12891 and GM12892. This yields five measurements for each replicate (10 samples in total). Due to potential sample swap, one replicate is dropped. Accordingly, there are four usable replicates (8 samples in total).

For each probe, the average standard deviation in both individuals is determined. The probes are ranked descending according to the standard deviation of the replicates. The 10% most variable probes are then removed from our dataset, prior to RMA processing. Next, probes with known SNPs in the 1,000 Genomes Project are filtered (1000 Genomes Project Consortium et al., 2010, 2012). Sufficient probes remain after filtering to be able to determine a signal for each FAIRE target (Table 2.7).

Table 2.6: SNPs that have been considered in the overlap with FAIRE-QTLs and their corresponding reference

Phenotype	Source	Number of SNPs
Crohn's disease	WTCCC and Jostins et.al. (Jostins et al., 2012; Wellcome Trust Case Control Consortium et al., 2007)	73
Ulcerative colitis	WTCCC and Jostins et.al. (Jostins et al., 2012; Wellcome Trust Case Control Consortium et al., 2007)	47
Height	Open GWAS (Johnson and O'Donnell, 2009)	277
Type 1 diabetes	NHGRI GWAS (Hindorff et al., 2009)	387
Expression-QTLs	Montgomery et al. (Montgomery et al., 2010)	4,781
DNaseI-QTLs	Degner et al. (Degner et al., 2012)	8,902

Table 2.7: Effect of filtering FAIRE probes per chromosome and background. Polymorphic probes are probes that overlap SNPs in the low coverage samples of the 1,000 Genomes and are subsequently filtered. The top 10% variable probes (the most intra-individual variable probes) in our replicates are filtered. In both, all of the polymorphic and intra-individual variable probes are filtered.

Chromosome	FAIRE-targets	Original probes	Polymorphic probes	Variable probes	Polymorphic + variable probes
1	1006	12689	11424	11405	10277
2	766	9128	8181	8141	7296
3	641	7564	6808	6735	6049
4	422	4716	4213	4154	3717
5	504	11442	10296	10034	9025
6	596	7043	6325	6181	5552
7	515	6062	5428	5367	4805
8	416	4781	4270	4241	3793
9	468	5812	5233	5148	4635

Continued on next page

Table 2.7: FAIRE probes

Chromosome	FAIRE-targets	Original probes	Polymorphic probes	Variable probes	Polymorphic + variable probes
10	476	7033	6334	6089	5502
11	608	7696	6957	6904	6233
12	518	6183	5573	5527	4989
13	211	2288	2057	2040	1834
14	381	4516	4080	4027	3633
15	324	4106	3716	3663	3310
16	396	5328	4800	4799	4333
17	534	7075	6393	6332	5720
18	188	2118	1883	1860	1652
19	590	7684	7029	6865	6286
20	313	3860	3462	3436	3075

Continued on next page

Table 2.7: FAIRE probes

Chromosome	FAIRE-targets	Original probes	Polymorphic probes	Variable probes	Polymorphic + variable probes
21	115	1411	1257	1241	1116
22	250	3356	2998	2974	2654
X	269	3100	2804	3100	2804
BACKGROUND	0	2613	2322	2613	2322

2.5.3 *Testing for associations*

I used Spearman's Rank correlation to perform the pairwise associations between the FAIRE-probesets or FAIRE-probes and the genotypes. For each association, the Spearman's Rank Rho and p-value are stored. Within R, the function `spearman.test` from the package `pspearman` was used for this purpose. Similar to eQTL experiments with comparable sample sizes, I restricted the discovery window to a 1Mb window, 500kb upstream and 500kb downstream, for each FAIRE-peak/FAIRE-probe. The associations are performed using all of the HapMap and all of the 1,000 Genomes genotypes. All locations are converted to hg18 to compare the result of both associations.

2.5.4 *Determining significance of associations*

To address the multiple testing problem, I applied permutations to determine the empirical p-value under the null distribution. The empirical p-value is determined per FAIRE-peak, independently giving a different empirical p-value for each FAIRE-peak. Our empirical p-value approach consists of the following 5 steps:

1. For each FAIRE-peak, I randomly swap the sample identities 1,000 times
2. Each random set of annotations is used to calculate a new Spearman's Rank correlation p-value
3. For each permutation, the lowest p-value is stored in the FAIRE-peak specific null distribution
4. Each association is ranked within the FAIRE-peak specific null distribution
5. Each rank is divided by the number of p-values in the null distribution

The rank in the null distribution divided by the number of p-values in the null distribution is defined as our empirical p-value (5).

$$\text{Empirical p-value} = \frac{\text{rank p-value in null distribution}}{\text{number of permutations} + 1} \quad (5)$$

2.5.5 *Identifying of FAIRE-QTLs previously identified in other phenotypes*

After identifying FAIRE-QTLs in the current dataset using the HapMap genotypes, I address their potential functionality by overlapping with SNPs implicated in known GWAS phenotypes from previously-published studies obtained from the NHGRI GWAS or Open GWAS catalogs. Also, for Crohn's disease and ulcerative colitis, I used the most recent list of known GWAS results from the Wellcome Trust Case-Control Consortium (WTCCC) ([Wellcome Trust Case Control Consortium et al., 2007](#)) and a large meta-analysis study ([Jostins et al., 2012](#)). Because the FAIRE-QTLs were identified in lymphoid cells, I assumed that the most likely overlap enrichment would be with SNPs implicated in immune-related diseases, as opposed to other phenotypes. I selected known associations for immune-related diseases including Crohn's disease, ulcerative colitis and Type 1 diabetes, as well as for height (Table 2.6).

2.5.6 *Identifying known biochemical ENCODE regions using RegulomeDB*

I selected all of the identified FAIRE-QTLs with an empirical p-value of 0.5 and lower. I loaded this set of SNPs into the RegulomeDB ([Boyle et al., 2012](#)). Two SNPs were discarded because of possible tri-allelic annotation, and 1 SNP could not be found for unknown reasons. After removal of those SNPs, there were 163 SNPs that could be overlapped with known

ENCODE regions. For each SNP, a Regulome DB score was assigned (Table 2.8).

2.5.7 Regulome DB score

For Regulome DB the given score (Table 2.8) for each SNP indicates the overlap with regulatory data (Boyle et al., 2012).

Table 2.8: Regulome DB score

Score	Supporting data
1a	eQTL and TF binding and matched TF motif and matched DNase Footprint and DNase peak
1b	eQTL and TF binding and any motif and DNase Footprint and DNase peak
1c	eQTL and TF binding and matched TF motif and DNase peak
1d	eQTL and TF binding and any motif and DNase peak
1e	eQTL and TF binding and matched TF motif
1f	eQTL and TF binding or DNase peak
2a	TF binding and matched TF motif and matched DNase Footprint and DNase peak
2b	TF binding and any motif and DNase Footprint and DNase peak
2c	TF binding and matched TF motif and DNase peak
3a	TF binding and any motif and DNase peak
3b	TF binding and matched TF motif
4	TF binding and DNase peak
5	TF binding or DNase peak
6	other
7	none

2.5.8 *Data visualisation*

Within R ([Team, R Development Core, 2005](#)), the package Ggplot2 ([Wickham, 2009](#)) was used to visualise all data.

For visualisation of an example FAIRE-QTL plot, I used LocusZoom ([Pruim et al., 2010](#)). In a LocusZoom plot the genomic surrounding of the FAIRE-peak and the genotypes tested for FAIRE-QTL are shown. In addition, for all genotypes the LD towards the best scoring FAIRE-QTL is given.

Part III

DISSECTING EPIGENETIC CORRELATIONS USING GENETICS

In this part of my thesis, I dissect how molecular regulatory mechanisms are intertwined. This is accomplished by combining molecular and genetic population datasets. The datasets combined in this part are from the Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) chapter (Chapter 2), a collaborative project involving CTCF ChIP-seq (described below), gene expression data from gEUVADIS (Lappalainen et al., 2013), and transcription factor and histone modification data from the ENCODE project (ENCODE Project Consortium, 2012).

For this section of my thesis research, I worked together in a larger collaboration that resulted in a single publication about CTCF biology, on which I am the joint first author (submitted). To avoid confusion, I will first describe the general purpose of the CTCF project and subsequently describe my portion of the project.

In a population overlapping the individuals from the FAIRE study, CTCF ChIP-seq experiments have been performed on 51 LCLs from CEU descent to discover Quantitative Trait Loci (QTLs or CTCF-QTLs) and to study CTCF biology.

Zhihao Ding from the Wellcome Trust Sanger Institute, Hinxton, United Kingdom, performed peak calling, finding CTCF-QTLs and characterising these. In addition, from the 51 individuals, only 41 are part of the phase 1 release of the 1,000 Genomes Project. To be able to use denser genotypes, Zhihao imputed SNPs and indels for the missing 10 cell lines.

YunYun Ni from the University of Texas at Austin, Austin, United States, performed the ChIP-seq experiments and defined the allele-specific behaviour of CTCF in these lines.

I performed the discovery of insulating CTCF-loops by utilising the population data (Chapter 3), defining causality of QTLs on insulating loops (Chapter 4), and investigating CTCF behaviour on the X-chromosome (Chapter 5).

CORRELATING MOLECULAR PHENOTYPES

This chapter describes the results obtained from studying the integration of FAIRE, CTCF and gene expression on the autosomes of 60 individual LCLs. Section 3.2.6, which addresses CTCF-to-CTCF correlations, is part of a submitted paper on which I am the joint first author. The other sections have been written for this thesis.

3.1 INTRODUCTION

For even the simplest biological process, there is a complex myriad of interactions between many molecules. For example, let us consider the change of gene expression due to a genomic variant. Though the observation of the change of genotype seems to infer a change in expression, this effect is almost never a binary effect (i.e. active versus inactive). On the DNA level, there is a complex system at play that combines dozens of transcription factors (TF), histone modifications (HM), and chromatin accessibility changes. Each of these has to be ordered in a specific way to accomplish the eQTL regulatory function. Although the eQTL is correlated with the gene expression, to fully comprehend the reason for this, an improved view of the epigenome and its regulation is essential. This becomes even more vital in understanding disease GWAS results, which are commonly found in non-coding regions overlapping or nearby regulatory elements ([Schaub et al., 2012](#)).

There is a great interest in the research community in increasing the understanding of how different molecular phenotypes interact. By acquiring a view of the epigenome in a group of individuals, it will be possible to improve our understanding of the

interplay between components of the epigenome. In this chapter, I integrate three datasets for a population of CEU LCLs. First, I have the FAIRE dataset, as described in Chapter 2. Second, I have a CTCF ChIP-seq dataset. Finally, I have a RNA-seq dataset containing transcript abundances (Lappalainen et al., 2013).

There are several potential reasons that correlations exist between different molecular phenotypes. I am interested in the correlations caused by the biochemical link between molecules. However, I must first address other sources of variance that could introduce correlations. For example, the gender of the individual or technical variances could create a systematic bias leading to correlation. Another type of systematic variation can be due to the cell lines themselves. Because each of the LCLs is a mixture of B-cells, there can be differences in measured signals due to a different ratio of B-cell subtypes (e.g. different ratio of immature and mature B-cells).

To remove these sources of variation, different strategies may be applied. First, by removing several components from a Principal Components Analysis (PCA), the larger systematic effects can be removed. Second, the molecular assays are performed in different laboratories, using different cell growths, and thus correlations cannot be driven by experimental factors specific to a lab. Third, by performing permutations, I determine an empirical p-value with which it is possible to determine significance at a higher threshold than when using the Bonferoni correction, while addressing the multiple testing problem. After these steps are completed, the correlations that remain are most likely to be due to biochemical interactions.

3.1.1 *Molecular interactions*

The different measured molecular phenotypes are known to have the ability to regulate each other. CTCF is known to be bound slightly upstream of open chromatin regions and often changes the chromatin status (Boyle et al., 2008). In turn, open

chromatin is known for its transcription activity (Gilbert et al., 2004). Another molecular function of CTCF is to act as an insulator protein. By forming a loop between two independent CTCF sites, it can block accessibility to enhancers, promoters and genes by changing the chromosomal 3D structure. By insulating these regions, CTCF can regulate gene expression (Kim et al., 2007).

By utilising individual differences for each of the molecular phenotypes, I can discover the correlation between them (Figure 3.1). For example, if different CTCF sites have a correlating profile of intensities across our individuals, these might be involved in forming a CTCF-to-CTCF loop, or any other larger biochemical mechanism. By doing this pair-wise, for all sites, with all molecular phenotypes available for my population, I can identify correlations between the assays and thus, potentially, biochemical interactions. The combined discovery of all correlations will help to create a limited epigenome view. In Chapter 4, I will further study the epigenome by determining the direct and indirect causal effects of QTLs.

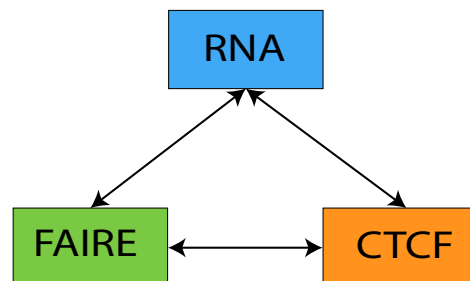


Figure 3.1: Schematic overview of the integration of the different molecular assays analysed in CEU LCLs. Pairwise correlations are performed between each assay.

3.2 RESULTS

In this chapter, I will identify correlations between CTCF-RNA pairs, CTCF-FAIRE pairs, FAIRE-RNA pairs, and CTCF-CTCF pairs. To exclude any potential sex-chromosomal effects, only the autosomes are used in this chapter.

3.2.1 *Correlating assays*

To identify whether molecular assays are correlated, pairwise correlations are performed. To test for correlation, I use Spearman's Rank correlation. Furthermore, when applicable, permutations are performed to calculate empirical p-values.

3.2.1.1 *Discovery of significant associations*

In order to address the multiple testing concern, permutations (1,000) are performed on the pairwise correlations to determine empirical p-values. With empirical p-values, it will be possible to determine which correlations are significant and are not occurring due to chance. The empirical p-values approach is similar to the one used for FAIRE-QTL discovery in Chapter 3.

I will briefly explain the reason for using empirical p-values and how this is implemented. For each pair tested for association, I seek to determine whether they are significantly correlated with each other. However, problems might arise, as some sites will be correlated, at some level, by pure chance. The simplest solution would be the application of the Bonferroni correction. However, the Bonferroni correction is, with the number of correlations performed, too strict and will lead to an unhelpful number of false negatives. The applied permutations technique that will help to determine the empirical p-value is more sensitive and, although computationally intensive, is used in this thesis.

Empirical p-values are determined on a site-by-site basis. For each site, all of the other sites that are tested for association are permuted 1,000 times. In each permutation, the annotation of the individuals is randomly swapped for one site and correlated against the other, original site. From each random permutation round, the lowest p-value is stored. The collection of lowest p-values is used to form a null distribution. By determining the position of the original associated p-value in the null distribution, the chance of being able to find the same result by chance is obtained. This is our empirical p-value. To illustrate an example, consider correlating 500 A sites with 500 B sites. Site A1 is correlated with 500 B sites. Each of these correlations is permuted 1,000 times. For each permutation round, the best p-value is stored in our null distribution (in this case, our null distribution will contain 1,000 p-values). To obtain the site-specific empirical p-value, the p-value of this correlation is ranked in the null distribution (Equation 6).

$$\text{Empirical p-value} = \frac{\text{rank p-value in null distribution}}{\text{number of permutations} + 1} \quad (6)$$

3.2.2 CTCF to RNA

As the first component of dissecting the epigenome, I evaluate CTCF-RNA pairs. There are 37,245 RNA transcripts and 55,166 CTCF sites with CTCF ChIP-seq binding affinity on the autosomes. To limit the search space, only intra-chromosomal pairs are tested for correlation. In total, there are 107,989,952 intra-chromosomal CTCF-RNA pairs that are associated with each other.

3.2.2.1 Results CTCF-RNA pairs

For each chromosome, a QQ-plot shows the results of these tests (Figure 3.2). The colours of the points indicate the empirical p-value threshold for that specific correlation. For CTCF-RNA

pairs, it was possible to permute the CTCF sites or RNA expression. I decided to permute the RNA expression 1,000 times. In the QQ-plot, I noted several points of interest, especially those having p-values $< 1 \times 10^{-6}$ (green line) that deviated from the expected uniform p-value distribution. For example, for chromosome 7 there is a clear group of points with empirical p-value < 0.01 that derives from the line. Depending on the empirical p-value threshold, there are between 7,297 (empirical p-value < 0.1) and 711 (empirical p-value < 0.01) significant CTCF-RNA pairs (Table 3.1). In the remainder of this section, I consider CTCF-RNA pairs significant at empirical p-value < 0.1 .

Table 3.1: Overview statistics results for CTCF to RNA correlations. Number of CTCF-RNA pairs per chromosome is reported. Number of significant results per chromosome is given for empirical p-values 0.1, 0.05 and 0.01.

Chr	Combinations	<0.1	<0.05	<0.01
1	18744880	712	347	74
2	10780536	531	276	55
3	7649994	468	232	48
4	4756536	411	190	30
5	5797274	421	209	44
6	6421618	408	195	45
7	5928228	417	212	45
8	3903540	297	142	36
9	3872790	291	153	27
10	4162995	366	185	34
11	7113106	369	169	29
12	5892201	382	199	32
13	1088898	156	79	20
14	3115534	330	161	22
15	2641272	263	117	22

16	3129300	250	120	27
17	5498295	321	158	35
18	840180	178	89	14
19	3985185	314	168	31
20	1381408	176	76	16
21	282862	69	30	9
22	1003320	167	83	16
Total	107989952	7297	3590	711

3.2.2.2 *Close-by correlations have positive Rho*

Because CTCF is a known transcription factor (TF), the expectation is that a change in binding of CTCF corresponds to a potential change in transcription levels. To test this assumption, I used all 7,297 significant CTCF-RNA pairs and binned them according to the distance between the CTCF site and the gene. There is a strong bias for significant CTCF-RNA pairs to be in close proximity ($< 10\text{kb}$) to each other (Chi-square test p-value $< 1.8 \times 10^{-9}$, Figure 3.3). The most significant CTCF-RNA pairs are observed when the CTCF site is overlapping with a gene.

One of the interesting questions that can be answered with this correlation approach is whether there is a generally positive or negative effect (and thus correlation) between CTCF and RNA. The Spearman's Rank correlation Rho value indicates the direction of the correlation. For example, if the Rho value is positive, this is due to the phenomenon that when CTCF binding is higher, there is higher gene expression. I took all of the Spearman's Rank correlation Rho values of both significant and insignificant pairs. I separated them over several distance bins (Figure 3.4). The reported distance is the distance between the CTCF site and the gene. There is a strong trend for overlapping CTCF-RNA pairs and CTCF-RNA pairs within 10kb distance between each other to have a positive correlation (Chi-square

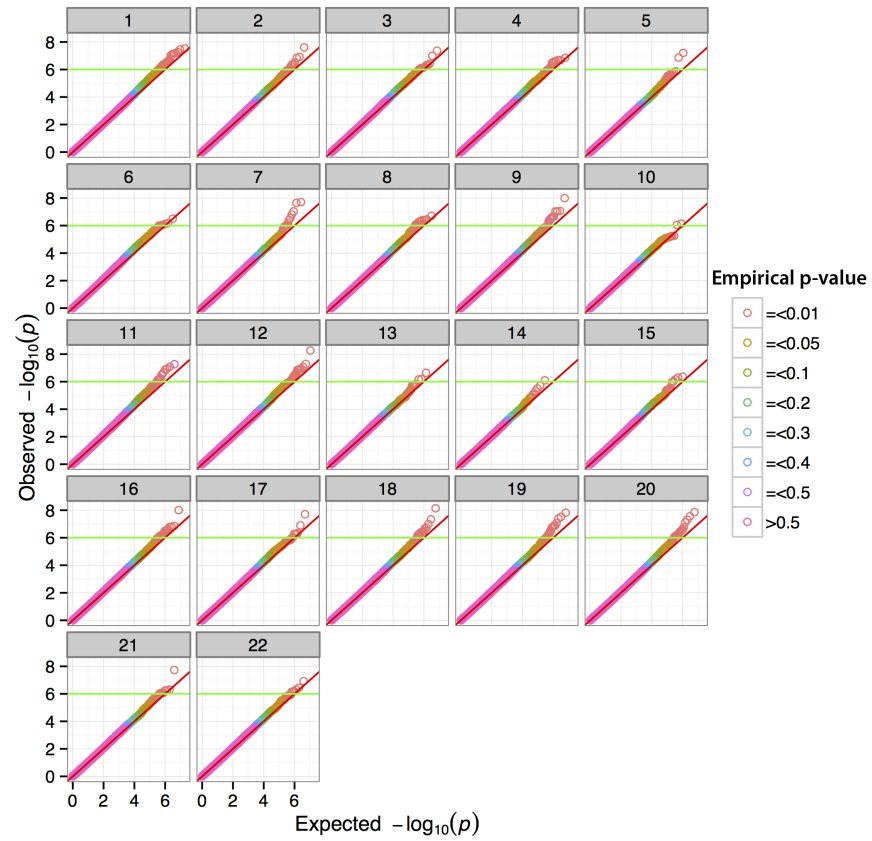


Figure 3.2: QQ-plots for CTCF-RNA pairs on all human autosomes. Each sub-plot represents a different chromosome. Empirical p-value colour coding is based on 1,000 permutations of the RNA sites.

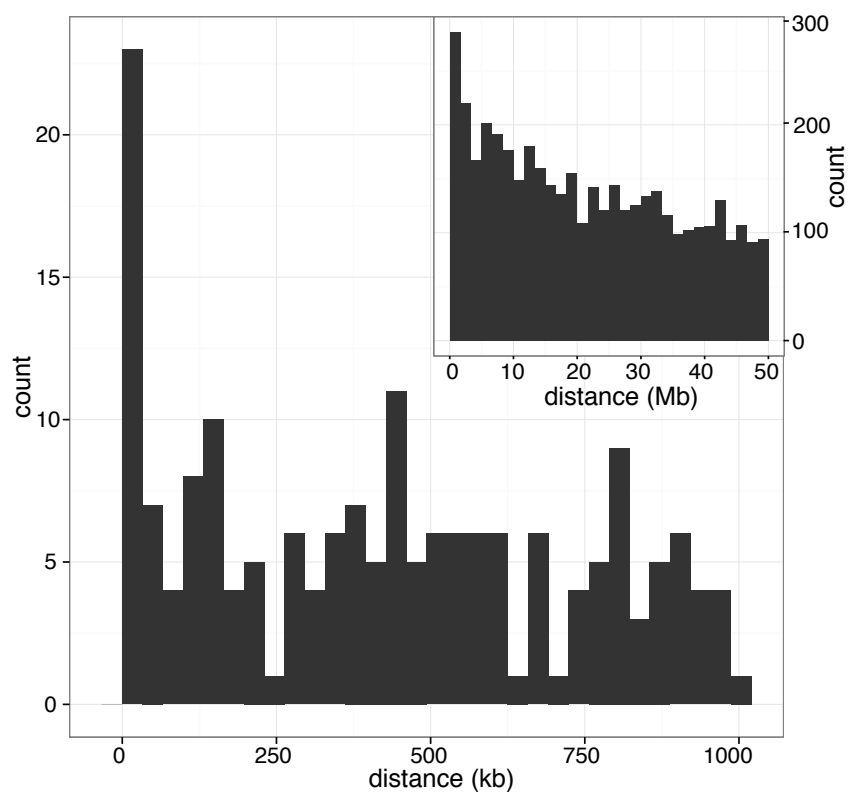


Figure 3.3: Distance plot for 7,297 CTCF-RNA pairs on all human autosomes where empirical p-value < 0.1 . The main plot shows all pairs within 1Mb distance between each other. The smaller plot is an overview of all pairs within 50Mb distance between each other.

test p -value = 0.004). However, given the nature of correlations, no direction of causality can be determined by the Rho value. Hence, correlations alone are not sufficient to determine whether CTCF binding is changing gene expression or vice versa.

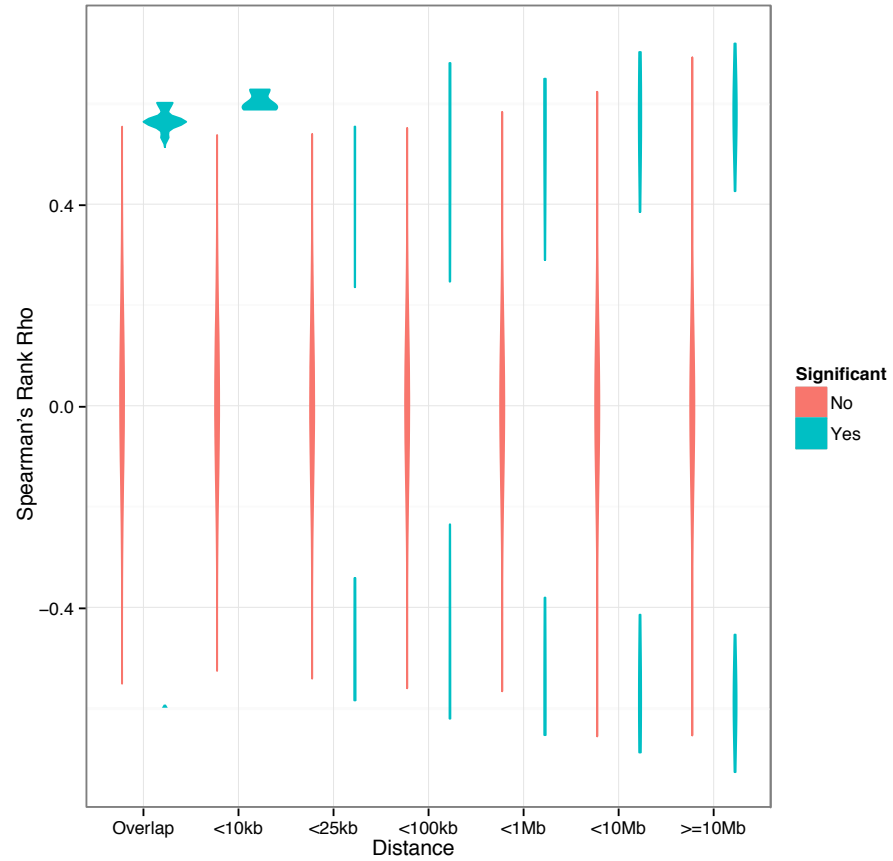


Figure 3.4: Kernel smoothed density of Rho values CTCF to RNA (violin plot). For a given distance, the Spearman's Rank correlation Rho value distribution is given. For each distance bin, the Rho distribution for the significant pairs (empirical p -value < 0.1) and the non-significant pairs are given. Overlapping and proximal ($< 10\text{kb}$) significantly correlated pairs have mainly positive Rho values.

3.2.3 CTCF to FAIRE

In this portion of dissecting the epigenome, I investigate the relation between CTCF binding and chromatin accessibility measured by FAIRE. There are 55,166 CTCF sites with ChIP-seq binding affinity and 10,509 FAIRE sites with FAIRE-chip expression on the autosomes. In total, there are 29,782,408 intra-chromosomal CTCF-FAIRE pairs that are tested for association using Spearman's Rank correlation.

3.2.3.1 Results CTCF-FAIRE pairs

As shown for CTCF-RNA pairs, I create a QQ-plot for each chromosome (Figure 3.5). For CTCF-FAIRE pairs permutations, I permute the FAIRE sites 1,000 times to obtain empirical p-values. In the QQ-plot, I noted several points of interest, especially with Spearman's Rank p-value $< 1 \times 10^{-6}$ (green line) that deviate from the expected uniform p-value distribution. For example, for chromosome 3 there is a clear group of points with empirical p-value < 0.01 that are deviated from the expected p-value trend. Depending on the empirical p-value threshold, there are between 1,593 (empirical p-value < 0.1) and 147 (empirical p-value < 0.01) significant results (Table 3.2). In the remainder of this section, I consider significance at empirical p-value < 0.1 .

Table 3.2: Overview statistics results for CTCF-to-FAIRE correlations. Number of CTCF-FAIRE pairs per chromosome is reported. Number of significant results per chromosome is given for empirical p-values 0.1, 0.05 and 0.01.

Chr	Combinations	<0.1	<0.05	<0.01
1	5042072	151	69	12
2	3156686	114	46	11
3	2396699	112	53	9
4	1218736	40	17	3
5	1596305	65	29	11

6	1910776	101	50	10
7	1501740	83	37	10
8	1073280	41	20	4
9	1088568	73	36	9
10	1290285	84	43	11
11	1779008	96	40	6
12	1562806	83	40	5
13	301519	43	20	4
14	779907	54	25	5
15	640548	46	24	2
16	752400	64	27	3
17	1380390	105	48	8
18	239324	29	16	3
19	1240770	101	48	11
20	490784	49	21	3
21	69805	17	7	1
22	270000	42	24	6
Total	29782408	1593	740	147

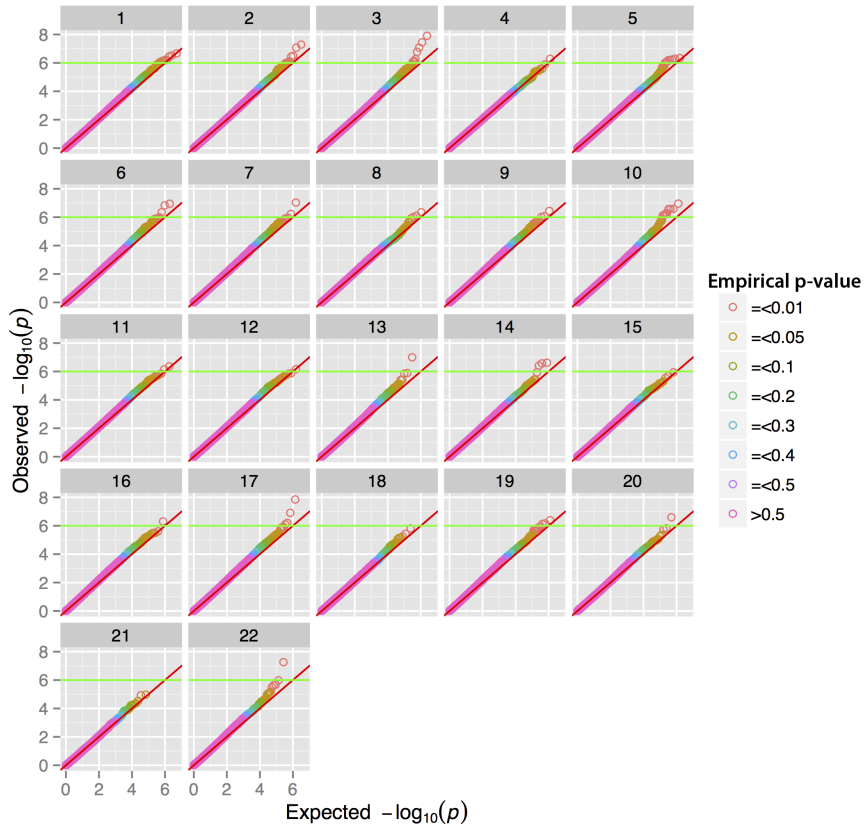


Figure 3.5: QQ-plots for CTCF-FAIRE pairs on all human autosomes. Each sub-plot represents a different chromosome. Empirical p-value colour coding is based on permuting the CTCF sites 1,000 times.

3.2.3.2 *Direct overlap of CTCF and FAIRE does not drive correlation*

For molecular phenotypes such as CTCF and FAIRE, a strong correlation would normally be expected based on the proximity of each of the sites. For example, when chromatin is open, it would be more likely that CTCF could be bound. In our data set, there is no indication that such an effect is present. However, when looking at FAIRE sites that overlap CTCF sites (sometimes multiple ones), there is no evidence in our association results to support such a theory (Figure 3.6). A reason for this could be that the actual accessible chromatin effect is only visible on the probe level and not on the probeset level.

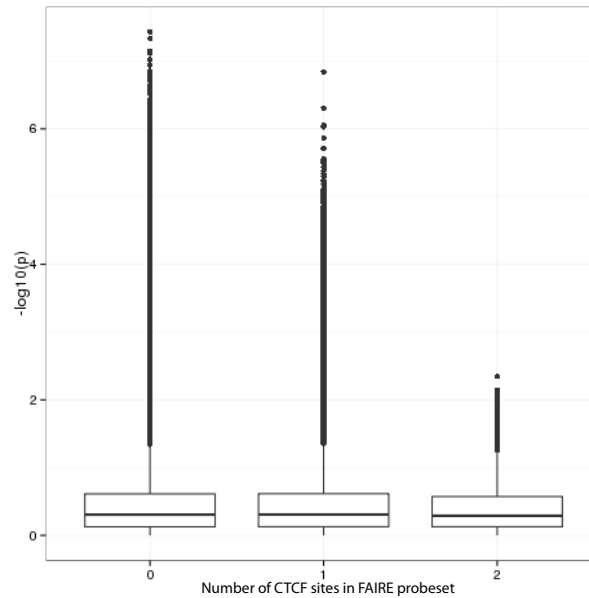


Figure 3.6: CTCF sites that are overlapping FAIRE-targeted regions. There is no difference in observed Spearman's Rank correlation p-values between CTCF sites that overlap the FAIRE probeset and are not overlapping.

3.2.4 CTCF to FAIRE probes

Regions targeted by FAIRE probesets are much larger than the CTCF peaks. Because of this, direct changes derived from CTCF binding might be missed in the averaging of all FAIRE probes into a probeset. This might be the reason for the lack of difference in results for sites that directly overlap (Section 3.2.3.2). To avoid this problem, I run the same analysis again, but this time using each individual FAIRE probe as an independent measurement of the FAIRE signal.

In total, there are 131,891 FAIRE probes. There are 384,023,280 intra-chromosomal CTCF-FAIRE probe pairs on the autosomes that are tested for correlation using Spearman's Rank correlation. Using probes in comparison with the probesets yields an additional 14,815 significant correlations (empirical p-value < 0.1), and much better p-values are observed (Figure 3.7, Table 3.3).

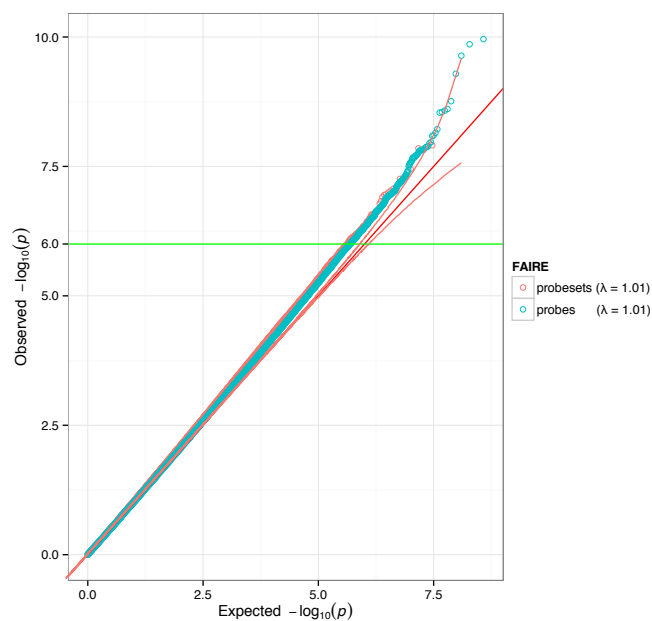


Figure 3.7: QQ-plot CTCTF-FAIRE: difference for using FAIRE probes versus FAIRE probesets. The light red lines represent the 95% confidence interval for the expected distribution of p-values. The genomic inflation factor (λ) is given for each dataset. The FAIRE probes yield more correlations that deviate from the expected trend than the FAIRE probesets do.

3.2.4.1 Results CTCF-FAIRE probe pairs

For each chromosome, a QQ-plot shows the results of the correlations (Figure 3.8). The colours of the points indicate the empirical p-value threshold for that specific correlation. In this QQ-plot, I noted several points of interest, especially Spearman's Rank correlation p-values $< 1 \times 10^{-6}$ (green line) that deviate from the expected line. For example, for chromosome 9 there is a clear group of points with empirical p-value < 0.01 that deviate from the expected p-value distribution. Depending on the empirical p-value threshold, there are between 16,408 (empirical p-value < 0.1) and 1,563 (empirical p-value < 0.01) significant results (Table 3.3). In the remainder of this section, I consider significance at empirical p-value < 0.1 .

Table 3.3: Overview statistics results for CTCF to FAIRE probe correlations. Number of CTCF-FAIRE probe pairs per chromosome is reported. Number of significant results per chromosome is given for empirical p-value thresholds of 0.1, 0.05 and 0.01.

Chr	Combinations	<0.1	<0.05	<0.01
1	63597268	1543	725	113
2	37516096	1119	551	109
3	28281796	1075	518	113
4	13619808	513	257	52
5	36168162	1252	623	105
6	22579858	958	486	107
7	17676792	819	407	82
8	12334980	575	283	48
9	13496146	803	418	81
10	19024265	819	394	71
11	22518496	1116	526	109
12	18654111	727	370	66

13	3269552	285	124	27
14	9244252	515	245	46
15	8117562	457	227	45
16	10123200	624	295	55
17	18288875	920	411	94
18	2696214	269	129	28
19	16159452	1017	497	125
20	6052480	427	218	39
21	856477	150	68	14
22	3624480	425	197	34
Total	383900322	16408	7969	1563

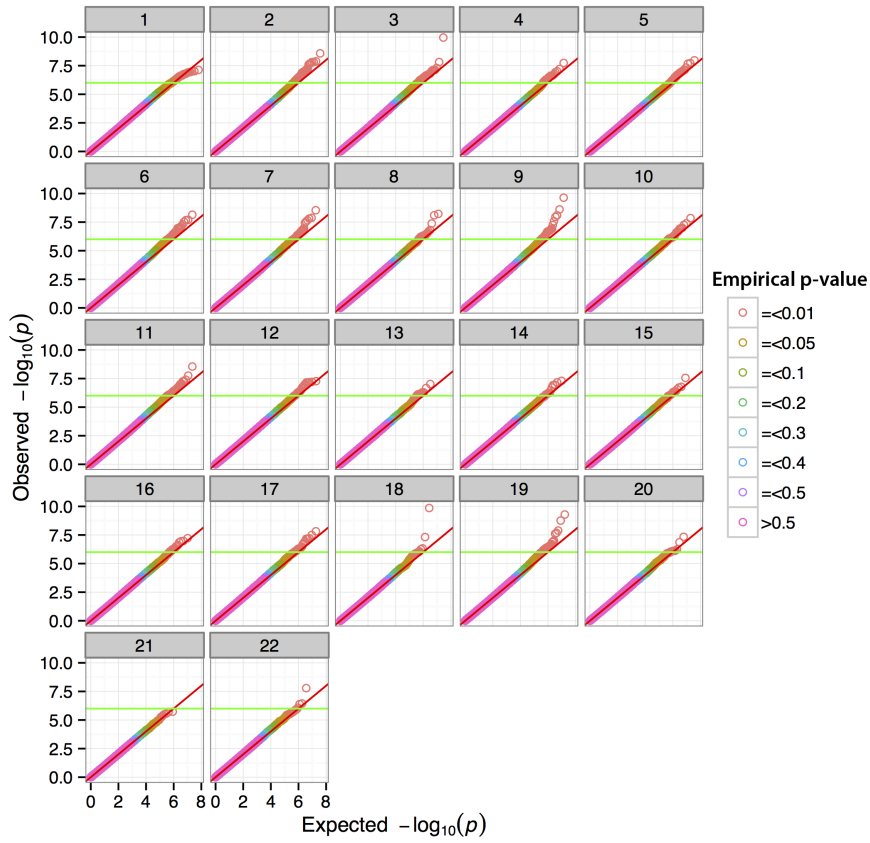


Figure 3.8: QQ-plots for CTCF-FAIRE probe pairs on all human autosomes. Each sub-plot represents a different chromosome. Empirical p-value colour coding is based on permuting the CTCF sites 1,000 times.

The expectation is that CTCF and FAIRE probes are more likely to correlate than FAIRE probesets (Section 3.2.3). To study this, I took all CTCF-to-FAIRE probe pairs that had an empirical p-value < 0.1 . Of these significant pairs, there were 12 pairs that overlapped. From these 12 pairs, only 1 had a negative Rho value (-0.6), while all other pairs had a Rho larger than 0.5 (rounding these Rho values, 9 had Rho 0.5 and 2 had Rho 0.6). This suggests that in the cases in which both CTCF and FAIRE are measured for a particular location, there is a positive correlation, suggesting a potential biochemical link.

Although the number of significant pairs that overlap is lower than in CTCF-RNA pairs (Figure 3.3), there is a strong tendency to be in close proximity with each other (Figure 3.9). CTCF-FAIRE probe pairs within 100kb distance of each other are significantly more often correlated to each other than more distant pairs (Chi-square test p-value $< 1.4 \times 10^{-7}$).

I took all of the Spearman's Rank correlation Rho values of both significant and insignificant pairs and separated them over several distances between the CTCF site and the FAIRE probe bins (Figure 3.10). There is a strong trend for pairs within 10kb distance of each other to have a positive correlation. This suggests that when CTCF sites are correlated and overlapping or proximal ($< 10\text{kb}$) to a FAIRE probe, they have a positive correlation (Chi-square test p-value = 0.04).

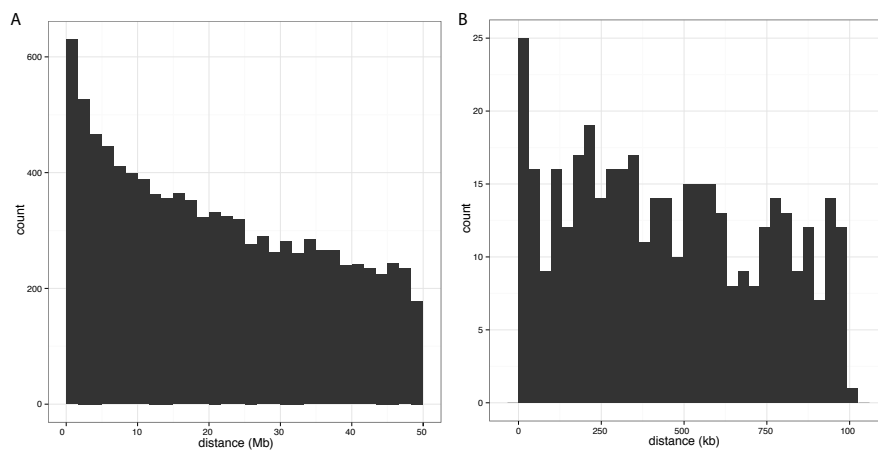


Figure 3.9: Histogram CTCF-FAIRE probe distance for all significant (empirical p-value < 0.1) pairs for all within 50Mb between each other (A), and all pairs within 1Mb distance between each other (B).

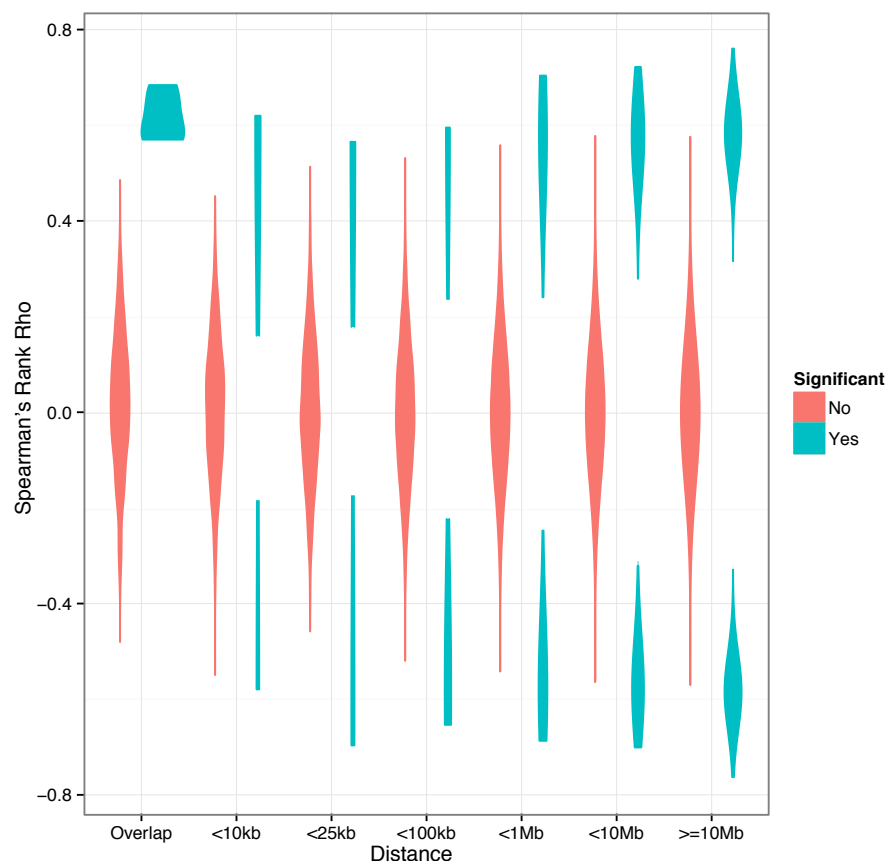


Figure 3.10: Kernel-smoothed density of Rho values for CTCF to FAIRE probes (violin plot). For a given distance, the Spearman's Rank correlation Rho value distribution is given. For each distance bin, the Rho distributions for the significant (empirical p-value < 0.1) pairs and the non-significant pairs are given. Overlapping significant pairs have only positive Rho values. Proximal pairs (< 10kb) are significantly positively correlated to each other (Chi-square test p-value = 0.04).

3.2.5 FAIRE probes to RNA

In this portion of my exploration of the epigenome, I investigate FAIRE to RNA expression. When testing the correlation between CTCF and FAIRE, it was determined that for examining molecular phenotype correlations, it is preferable to analyse the probes instead of the probesets (Section 3.2.4). In this section, I will, for this reason, use FAIRE probes. In total, there are 261,574,637 intra-chromosomal FAIRE-RNA pairs that are tested.

3.2.5.1 Results FAIRE-RNA pairs

For each chromosome, a QQ-plot shows the results of the correlations (Figure 3.11). The colours of the points indicate the empirical p-value threshold, obtained by permuting the RNA. In the QQ-plot, I noted several points of interest, especially below the Spearman's Rank correlation p-value $< 1 \times 10^{-6}$ (green line) from the expected distribution. For example, refer to chromosome 19, where there is a group of FAIRE-RNA pairs with empirical p-value < 0.01 that deviate. Depending on the threshold, there are between 17,058 (empirical p-value < 0.1) and 1,880 (empirical p-value < 0.01) significant FAIRE-RNA pairs (Table 3.4). In the remainder of this section, I consider FAIRE-RNA pairs significant at empirical p-value < 0.1 .

Table 3.4: Overview statistics results for FAIRE probes to RNA expression correlations. Number of FAIRE-RNA pairs per chromosome is reported. Number of significant results per chromosome is given for empirical p-values 0.1, 0.05 and 0.01.

Chr	Combinations	<0.1	<0.05	<0.01
1	47456860	1867	917	205
2	23878848	1093	548	102
3	15475944	1049	553	144
4	7767252	568	294	65

5	20984628	1479	746	149
6	14107129	1013	544	122
7	12324046	727	336	69
8	7233653	590	294	58
9	9676980	762	387	98
10	10823787	974	496	94
11	18708976	1062	528	116
12	12075399	902	469	110
13	1743456	313	146	37
14	6873352	552	263	47
15	5485616	537	289	66
16	8775216	670	334	82
17	15048525	757	371	85
18	1397880	252	131	31
19	14561180	836	434	93
20	3400660	554	274	51
21	657526	124	66	13
22	3117724	377	202	43
Total	261574637	17058	8622	1880

I used the distance between the FAIRE and RNA for each FAIRE-RNA pair to study whether proximal or overlapping sites are more likely to correlate. Of the significant FAIRE-RNA pairs, there are 91 pairs that overlap. From these 91 pairs, only 1 has a negative Rho value (-0.6), while all other pairs have a Rho larger than 0.5 (rounding these Rho values, 50 have Rho 0.5, 35 have Rho 0.6, and 5 have Rho 0.7). This result suggests that in cases where there is an overlap, FAIRE is positively correlated with RNA abundance (Chi-square p-value $< 1.0 \times 10^{-24}$). For the non-overlapping significant pairs, there is a strong tendency to

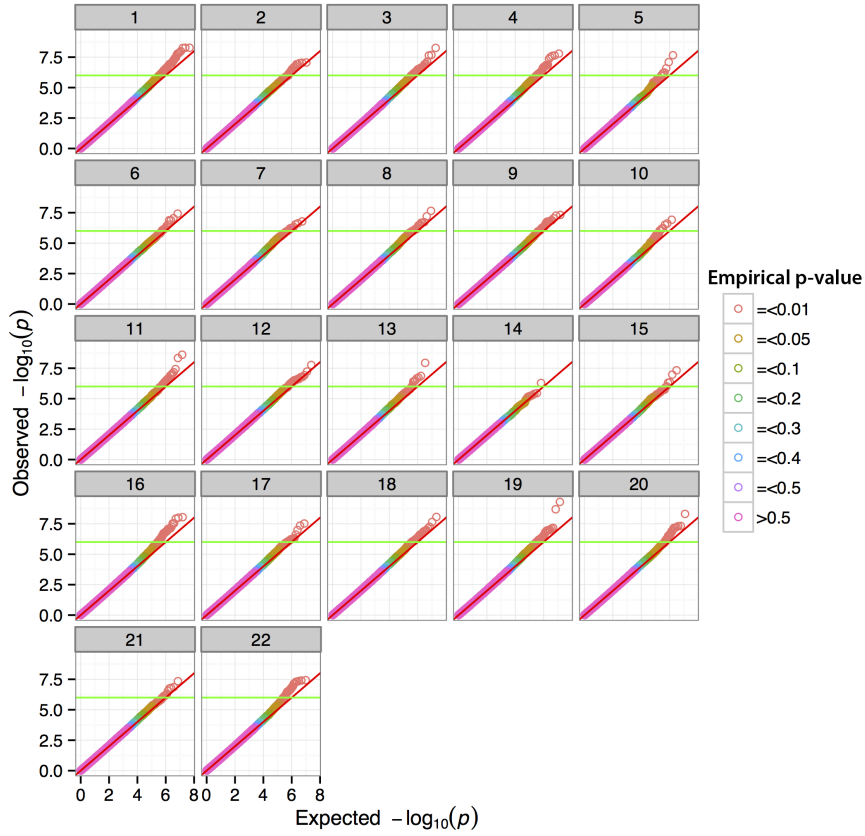


Figure 3.11: QQ-plots for FAIRE-RNA pairs on all human autosomes. Each sub-plot represents a different chromosome. Empirical p-value colour coding is based on permuting the RNA sites 1,000 times.

be in close proximity ($< 10\text{kb}$) between each other (Chi-square $p\text{-value} < 1.0 \times 10^{-14}$, Figure 3.12).

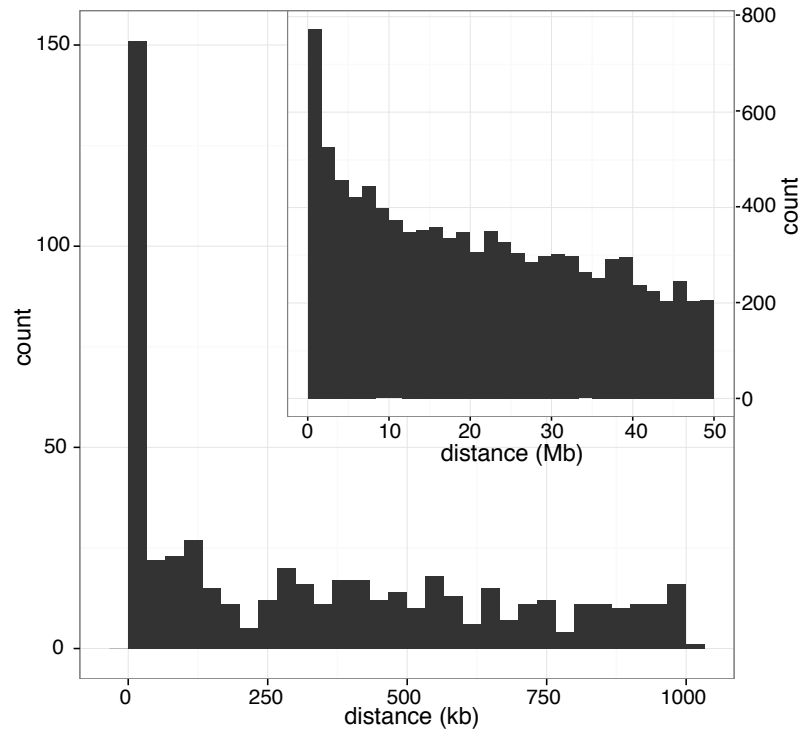


Figure 3.12: Histogram of FAIRE-RNA distance for all significant (empirical $p\text{-value} < 0.1$) pairs. The main plot shows all pairs within 1Mb distance between each other. The smaller plot is an overview of all pairs within 50Mb distance between each other.

FAIRE probes show mainly positive correlation with RNA. Nearly all significant FAIRE-RNA pairs that are overlapping or proximal have a positive Spearman's Rank Rho (Figure 3.13). In comparison with CTCF-FAIRE pairs and CTCF-RNA pairs, the FAIRE-RNA pairs have a positive Spearman's Rank correlation Rho value for all distances. For FAIRE-RNA pairs, I observed that most significant correlations within 100kb have a positive Spearman's Rank correlation Rho value (Chi-square $p\text{-value} < 1.0 \times 10^{-16}$).

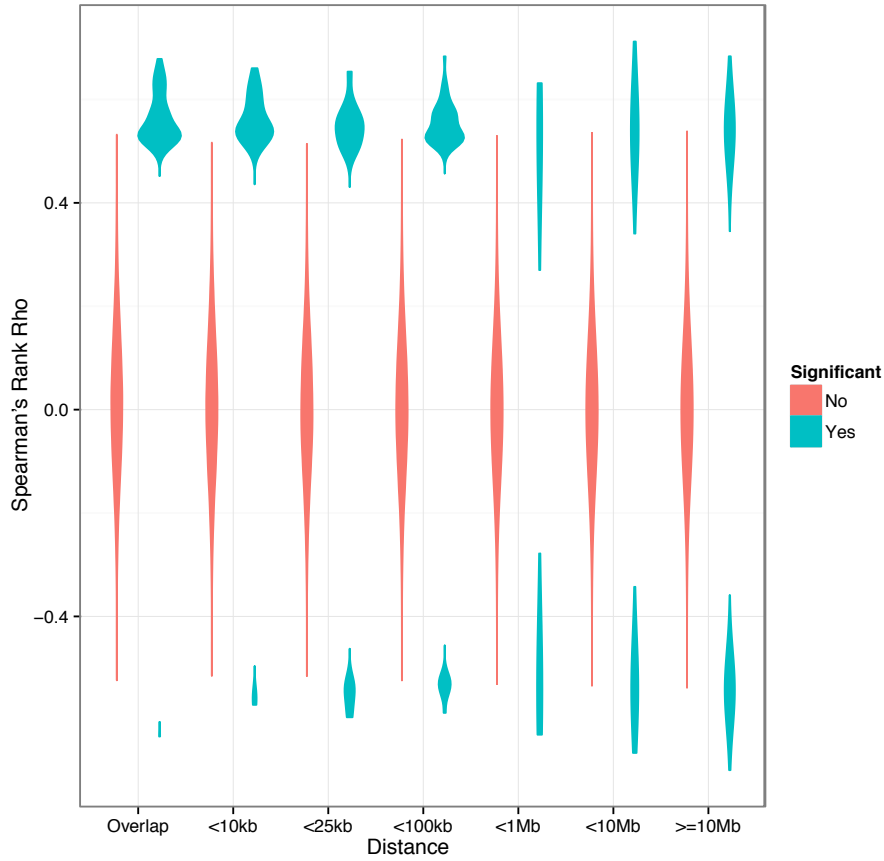


Figure 3.13: Kernel smoothed density of Rho values FAIRE-RNA (violin plot). For a given distance, the Spearman's Rank correlation Rho value distribution is provided. For each distance bin, the Rho distribution for the significant pairs and the non-significant pairs are given. Proximal (< 100kb) and overlapping significant pairs have mainly positive Rho values (Chi-square p-value < 1.0×10^{-16}).

3.2.6 *CTCF to CTCF*

There are several biochemical models in which CTCF sites form a loop. To study the correlating properties of CTCF, I performed Spearman's Rank correlation for each pair of CTCF sites on autosomes. I call these correlations CTCF-CTCF pairs. As for CTCF-CTCF pairs both sides of the pair are from the same dataset, permutation-based empirical p-values are not feasible. Therefore, a single Spearman Rank correlation p-value threshold was used to establish significant pairs. There are 57,167 CTCF sites, and in total there are 80,329,389 CTCF-CTCF pairs tested for association using Spearman's Rank correlation.

3.2.6.1 *CTCF insulating loops are common*

For the 80,329,389 combinations of intra-chromosomal CTCF-CTCF pairs, I found that 8,060 are nominally significantly correlated (Spearman's Rank correlation p-value $< 1.0 \times 10^{-6}$). The fraction of CTCF sites that are significantly correlated are similar for each chromosome (Table 3.5), and the p-value distribution for each chromosome is similar (Figure 3.14). Some CTCF sites are involved in multiple CTCF loops, but the vast majority of CTCF sites is only associated with a single other CTCF site (Figure 3.15). The observation that, for most CTCF-CTCF pairs, the CTCF sites involved are only involved in one nominally significant pair, supports the idea that there is no larger systematic bias driving the correlations.

The distance between significant CTCF-CTCF pairs is highly variable (Figure 3.16). The distance ranges from proximal sites (410bp distance) to distal sites (237Mb distance). In general, most significant CTCF-CTCF pairs are proximal to each other. The Spearman's Rank correlation Rho values range for negative Rho values between -0.78 and -0.62, and for positives between 0.62 and 0.97. The number of significant CTCF-CTCF pairs with positive or negative Rho values is skewed towards positive correlation, with 559 negatively-correlated pairs and 7,501

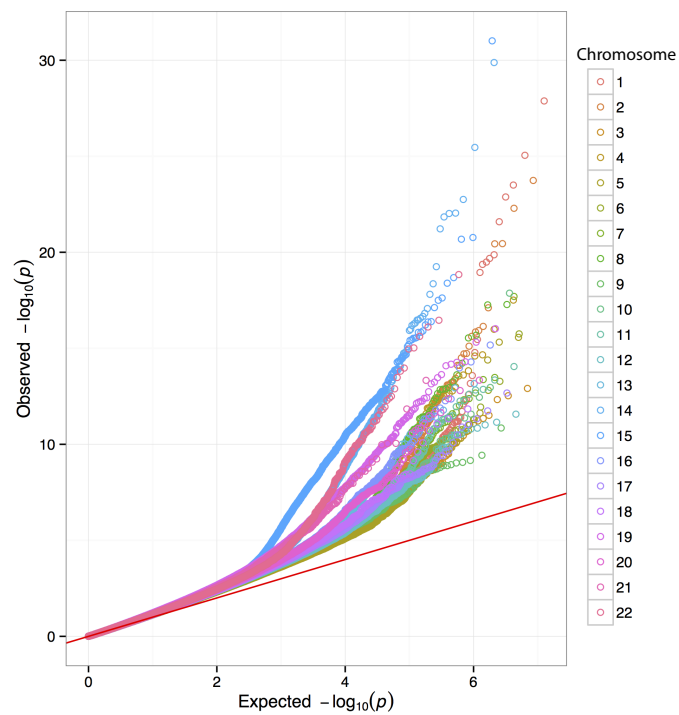


Figure 3.14: QQ-plot for all CTCTF-to-CTCTF correlations for all autosomal chromosomes. Each autosome is coloured differently.

positively-correlated pairs. This is as expected, because for a CTCF insulating loop to be present, both CTCF sites should have an increased CTCF signal.

To gain a better understanding about how the CTCF-CTCF pairs are correlated, I compared the Spearman's Rank correlation Rho values for the significant pairs versus the insignificant pairs (Figure 3.17). Irrespective of the distance between the CTCF sites in a CTCF-CTCF pair, the significantly correlated pairs have a positive Rho. This indicates that both sites realise similar additional CTCF binding. This could be an indication of a CTCF loop being formed between the two CTCF sites directly, or indirectly through another biochemical complex.

Table 3.5: Overview results CTCF to CTCF correlations. For each chromosome, the total number of tested and significant (Spearman's Rank p -value $< 1.0 \times 10^{-6}$) CTCF-pairs is given.

Chromosome	All combinations	Significant pairs
1	12557566	578
2	8489260	463
3	6988191	417
4	4168828	273
5	4994380	159
6	5137615	367
7	4250070	271
8	3326910	235
9	2703975	129
10	3657160	239
11	4279275	370
12	4549636	254
13	1020306	115
14	2094081	736
15	1953276	1675

16	1804050	258
17	3339820	225
18	809628	67
19	2210253	802
20	1228528	171
21	183921	51
22	582660	205
Total	80329389	8060

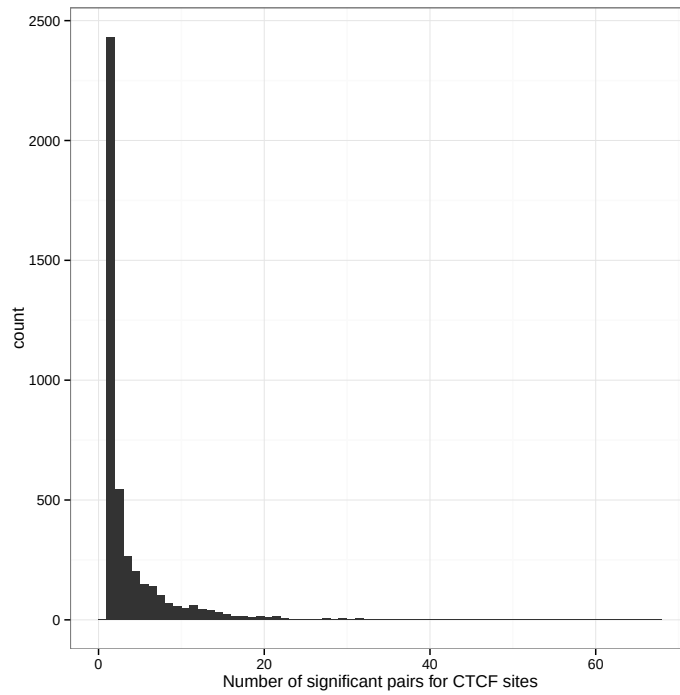


Figure 3.15: Histogram showing the number of significant associations that are observed for each CTCF site. Most CTCF sites with a significant correlation (Spearman's Rank p -value $< 1.0 \times 10^{-6}$) are only associated with a few sites.

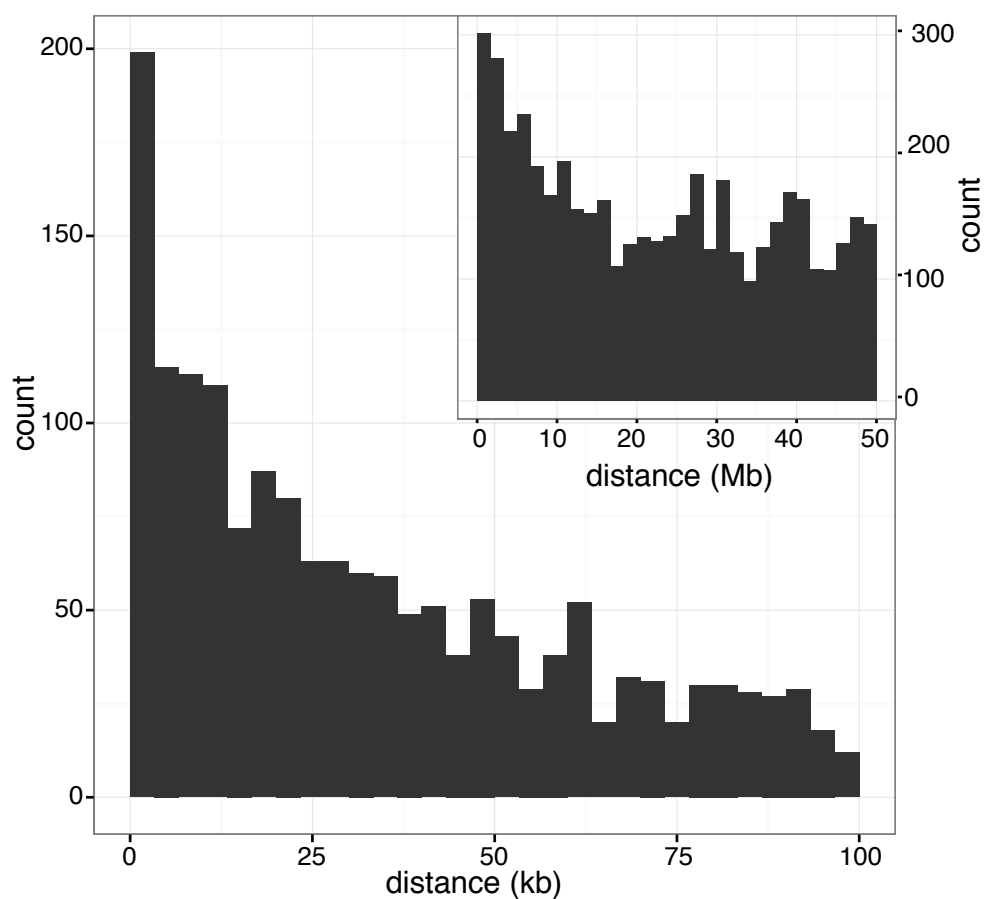


Figure 3.16: Histogram distance for all 8,060 significant CTCF-CTCF pairs on all autosomes. Main plot shows CTCF-CTCF pairs that are within 100kb distance of each other. The smaller plot is an overview of all pairs within 50Mb distance of each other. Most significant CTCF-CTCF pairs are proximal to each other.

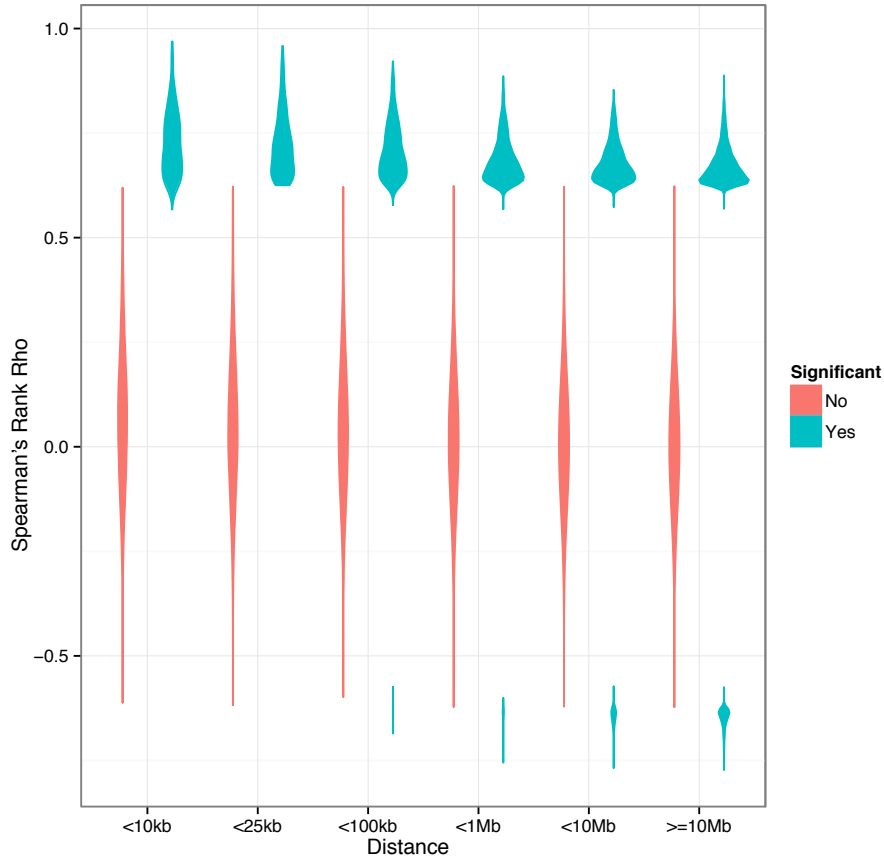


Figure 3.17: Kernel-smoothed density of Spearman's Rank Rho values for associated CTCF-CTCF pairs. For a given distance, the Spearman's Rank correlation Rho value distribution is given. For each distance bin, the Rho distributions for the significant pairs and the non-significant pairs are given. For any given distance, the significant pairs are mainly positively correlated.

3.2.7 *Validating CTCF-CTCF pairs using ChIA-PET*

To gain additional insight into the significance of the observed CTCF-CTCF pairs, I explored the overlap with CTCF ChIA-PET data. I used three CTCF ChIA-PET datasets from ENCODE ([ENCODE Project Consortium, 2012](#)) measured in MCF-7 (in replicate) and in K562.

As the ENCODE datasets are not performed in LCLs and use a slightly different chemistry, they explore different CTCF sites than those investigated in my study. To overcome this disparity between the datasets, I focus only on the CTCF-CTCF pairs and ChIA-PET CTCF-CTCF pairs that could have been identified in both studies. To do this, I overlap all of the CTCF sites from my study with the CTCF sites in the ChIA-PET data. Only if both CTCF sites in my study are found as a pair or independently in the ChIA-PET data do I consider them. In this way, both studies can identify the same CTCF-CTCF pairs. For this reason, only the CTCF-CTCF pairs on chromosomes 18 and 20 are considered.

The distribution of Spearman's Rank Rho values for each set of overlapping CTCF-CTCF pairs is given for ChIA-PET dataset MCF-7 replicate 1 (Figure 3.18a), MCF-7 replicate 2 (Figure 3.18b) and K562 (Figure 3.18c) in violin plots. The left violin plot is for CTCF-CTCF pairs from which both CTCF sites are found in a ChIA-PET pair, but they are not paired by ChIA-PET. The right violin plot shows the CTCF-CTCF pairs that are also identified using ChIA-PET. There are more significant CTCF-CTCF pairs that overlap ChIA-PET, as is visualised in the higher density of the right violin plot for significant Spearman's Rank Rho values. This tests whether the distribution of the CTCF-CTCF pairs that overlap ChIA-PET are significantly different from the non-overlapping ones. I use the KolmogorovSmirnov (KS) test for this analysis. With the KS test, it is possible to compare two different distributions and test how equal they are (goodness-of-fit test). The KS test can be used to determ-

ine whether the Spearman's Rank p-values distributions for ChIA-PET and non ChIA-PET pairs are equal, because as shown earlier, both distributions are different and the KS test can assess if these difference are significant. I use the KS test to determine whether the overlap is greater than the non-overlap. Both MCF-7 replicates (KS p-value = 1.5×10^{-19} and KS p-value = 1.7×10^{-10}) and the K562 (KS p-value = 2.8×10^{-6}) are significantly greater. These results indicate that for all ChIA-PET datasets, the significant CTCF-CTCF pairs are significantly more often also a ChIA-PET CTCF pair. Given the size of the data set, the KS test is most likely not strict enough. This becomes evident, as the subtle differences in the distributions do not significantly affect the KS test results.

In this thesis, I use Spearman's Rank p-value $< 1.0 \times 10^{-6}$ as a threshold for nominal significance for CTCF-CTCF pairs. At this p-value threshold, the Spearman's Rank correlation Rho value is at least 0.62 for positive correlations. However, at this threshold, there are only a few CTCF-CTCF pairs that overlap with ChIA-PET CTCF pairs. For this reason, for the FDR calculations, I use a lower threshold, which is based on the Rho value rather than the p-value. The FDR discovery rate is determined for CTCF-CTCF pairs with a Rho > 0.5 (Figure 3.18, red line). The FDR is determined after transforming the non-overlapping CTCF-CTCF pairs to match the size of the overlapping CTCF-CTCF pairs. Both MCF-7 replicates give a FDR of 0.05 at Spearman's Rank Rho > 0.5 . For K562 at Spearman's Rank Rho > 0.5 , the FDR is 0.11. As these FDRs are determined for a higher Spearman's Rank correlation p-value than used previously, it is reasonable to say that our significant results are all at least significant with a FDR < 0.05 .

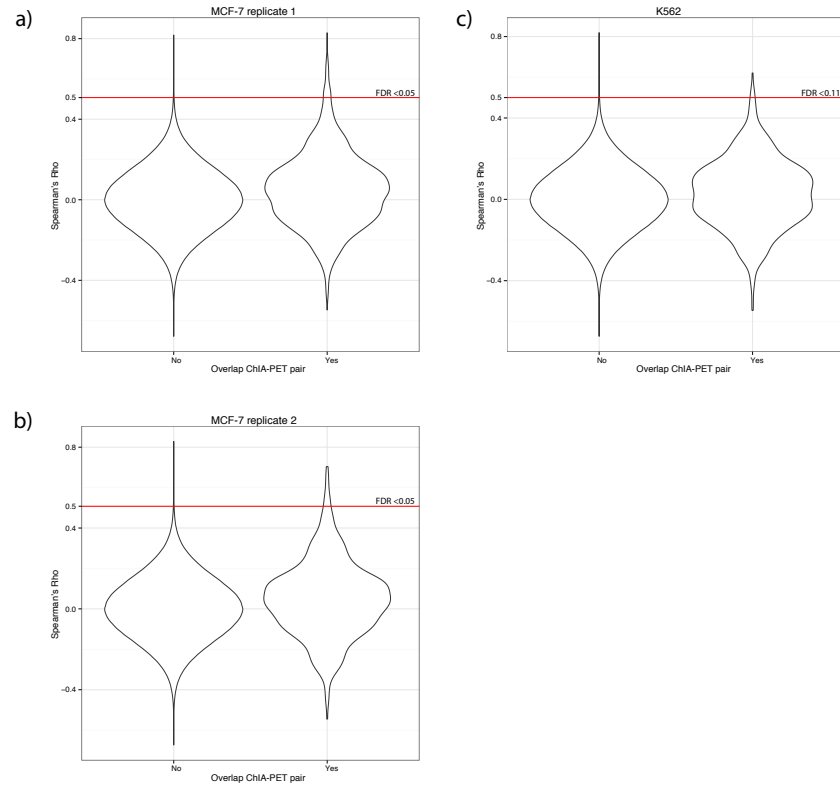


Figure 3.18: CTCF-CTCF pairs overlapping ChIA-PET. A) Overlap with K562, B) overlap with MCF-7 replicate 1, and C) overlap with MCF-7 replicate 2. Only MCF replicate 1 has overlap with inter-chromosomal CTCF-CTCF pairs. For all ChIA-PET datasets, the overlapping pairs have significantly higher Rho values. False Discovery Rates (FDR) are determined for pairs that are positively correlated (Spearman's Rank Rho values > 0.5).

Both the KS test results and the FDR data indicate that the CTCF-CTCF pairs identified by correlations are found significantly more often in ChIA-PET data. However, my method of detection of CTCF-CTCF pairs only discovers pairs that have individual differences. Therefore, many CTCF-CTCF pairs that are biochemically linked without these abundant differences can only be found in the ChIA-PET data.

3.2.7.1 *CTCF-loops show specific TFs and HMs profile*

CTCF loops are known to have many different molecular functions and roles depending on cell type and environment (Chapter 1.3). Many of these functional differences require that CTCF must interact with different transcription factors, histone marks, and different chromosome statuses to perform its function. To study this with respect to the length of the CTCF loops, all significant CTCF-CTCF pairs (Spearman's Rank p-value $< 1.0 \times 10^{-6}$) on chromosomes 18 and 20 are binned based on the distance between the two CTCF sites of a pair. Each CTCF site that is part of the CTCF-CTCF pair, thus both the "left" and the "right" CTCF site, is overlapped with available ENCODE data measured on GM12878. When replicates are available, multiple overlaps for the same factor or histone mark are present. For each overlap, the signal intensity is averaged for the CTCF-pair and scaled per column to enable direct intensity comparison between the ENCODE datasets (Figure 3.19).

There is a strong difference in overlapping intensity "profiles" for CTCF-CTCF pairs when the distance is shorter than 1Mb. For example, the transcription factors YY1, ZNF143 and ELF1 overlap strongly with longer loops ($> 1\text{Mb}$), while there is hardly any overlap with proximal loops (Students t-test p-value < 0.05). There is no obvious known biological reason for having this distinctive difference in overlap pattern for these specific transcription factors. It is interesting to note that Handoko et al. suggest that in mice, a distinct difference in overlap of TFs and HMs is observed for loops longer than 200kb (Handoko

[et al., 2011](#)). One of the most commonly known involvements of CTCF is the cohesin complex. There is no difference in overlap profile with members of the cohesin complex, RAD21 and SMC3, depending on the distance between the CTCF sites. For the long-distance linkages, a possible reason for the overlap with many other TFs and HMs can be found in the fact that these are mainly in hotspot regions or that these are all part of transcriptional factories ([Ling et al., 2006](#)).

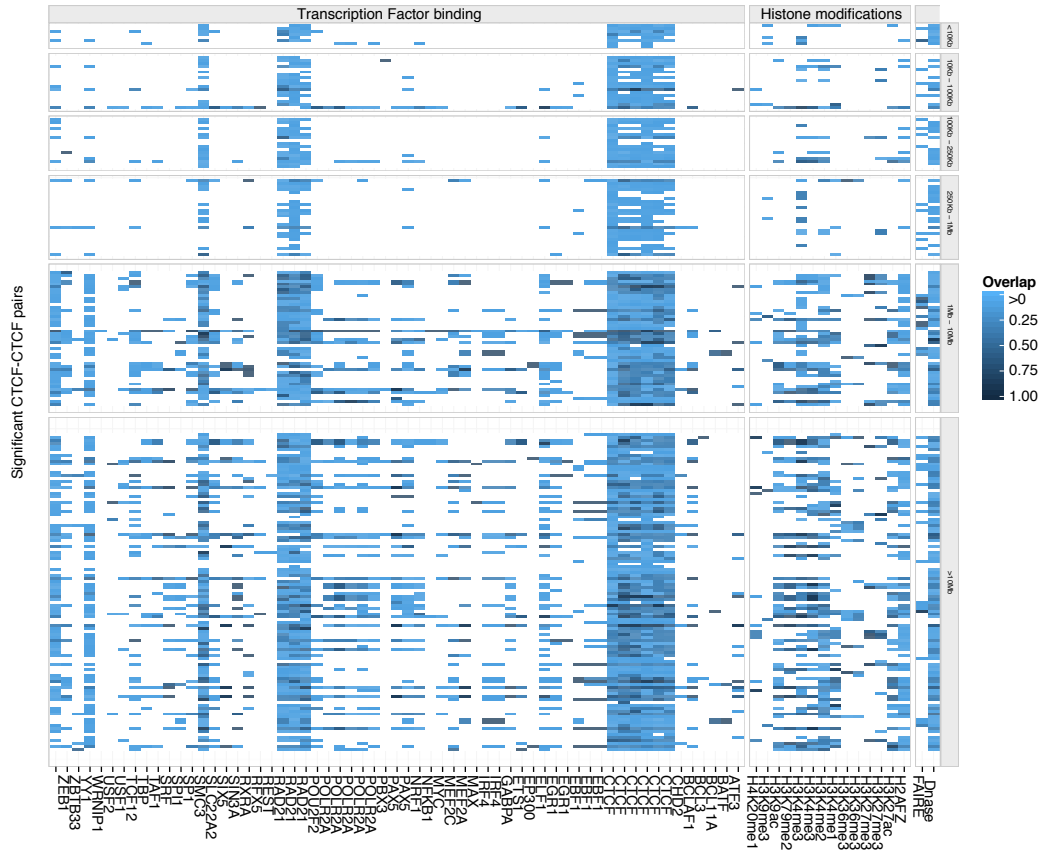


Figure 3.19: CTCF-loops show distinctively different profiles for lengths. The Y-axis divides CTCF-CTCF pairs based on distance, from proximal (top) to being located on different chromosomes (bottom). The average ENCODE signal overlapping both CTCF sites is given. The ENCODE signal intensity is scaled to enable direct comparisons. A distinctive difference in overlap patterns can be observed for CTCF loops smaller than 1Mb (Students t-test p-value < 0.05 for YY1, ZNF143 and ELF1).

3.2.8 *Molecular assays are highly correlating*

Combining all of the previously-discovered significant associations into a single system could identify larger molecular systems at play. To discover these systems, I overlap all significant FAIRE-CTCF pairs, FAIRE-RNA pairs, and CTCF-RNA pairs. There are 345 instances where each molecular assay is significantly associated with another.

I have selected an example to illustrate such a system. Located on chromosome 1, there is a system involving a FAIRE probe (chr1:226591906-226591950), CTCF site (chr1:227015770-227016147) and gene (ENSG00000203685, chr1:226736501-226796915) that are correlated (Figure 3.20). The regulatory features are proximal to the gene (144kb and 279kb). The measured RNA originates from a gene located in middle of the CTCF site and FAIRE probe (Figure 3.20A). The FAIRE probe and gene expression are positively correlated (Spearman's Rank p-value 3.8×10^{-7} , Rho 0.62). The CTCF site and the FAIRE probe (Spearman's Rank p-value 2.2×10^{-7} , Rho -0.66) and the CTCF site and the gene expression (Spearman's Rank p-value 1.7×10^{-4} , Rho -0.32) are negatively correlated (Figure 3.20B). This system provides a strong example of how the final gene expression is correlated with the status of FAIRE and CTCF. Although both FAIRE and CTCF can independently indicate something about the expected gene expression (or vice versa), it is only when both are considered that the gene expression can most accurately be predicted. A potential explanation for this would be that, in this case, the binding of CTCF is regulating the accessibility of the proximal chromatin, enabling potential gene expression. This could be accomplished, for example, by CTCF insulating the promoter region. However, given the current data, this hypothesis is not statistically testable. To further study this observed system of correlations, I overlapped the location of the FAIRE probe, CTCF site, and gene with FAIRE, DNase I, CTCF, SMC3, RAD21 and ATF3 data from ENCODE measured in GM12878

(ENCODE Project Consortium, 2012) (Figure 3.20C). I also included the average FAIRE probe intensity; this highlights the regions that are measured with the FAIRE tiling array. The CTCF site overlaps with both SMC and RAD21; this indicates a potential involvement of CTCF in cohesin. CTCF in a cohesin complex is known for its regulatory function, and thus this could indicate the reason for the observed correlations. The FAIRE probe overlaps a region with higher DNase I activity. However, not visible on this plot, the actual probe does not overlap with a DNase I peak. The FAIRE-probe is, however, overlapping with an Activating transcription factor (ATF3) peak. This could be the reason why this FAIRE probe is showing correlation with both the CTCF and the gene expression, as it might be a proxy measurement for the actual ATF3 activity. Noteworthy is that the FAIRE-seq in GM12898 for this entire region does not show any FAIRE signal. This lack of signal strengthens the idea that the FAIRE-probe is actually based on ATF3 activity.

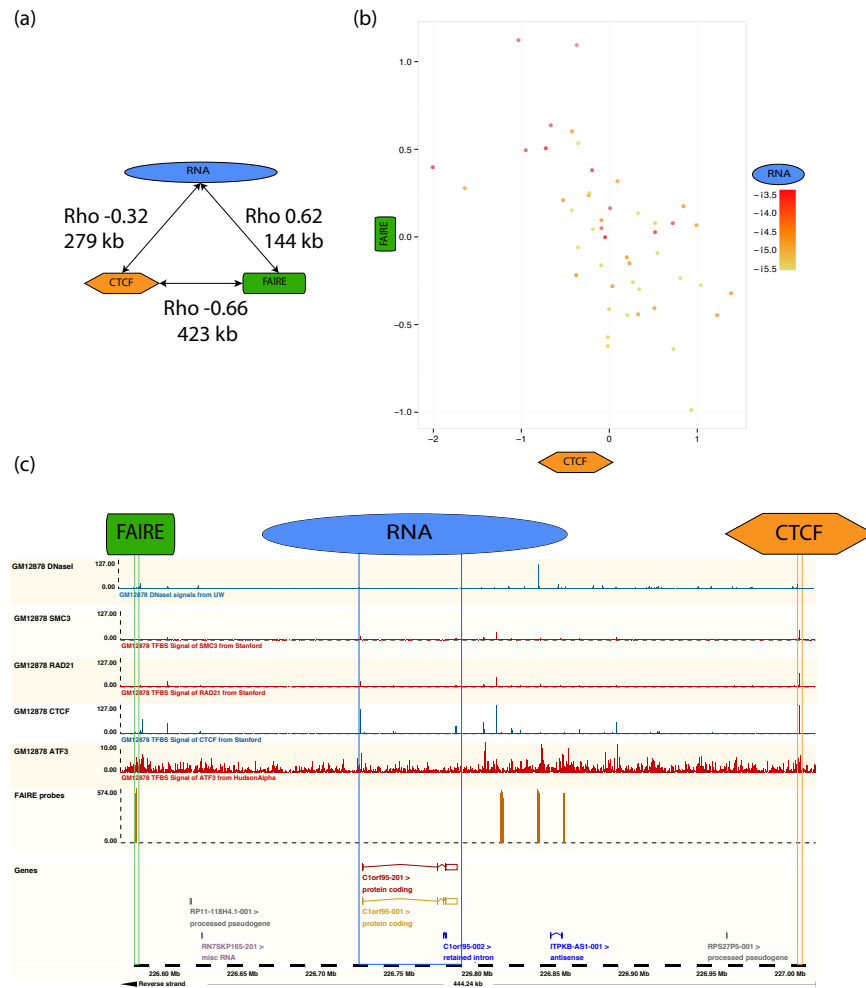


Figure 3.20: Example FAIRE-CTCF-RNA correlation. A) The FAIRE probe, CTCF site and gene are all proximal. All three assays are associated with each other using Spearman's Rank correlation. The Rho value indicates the direction of the correlation. B) The FAIRE probe and the CTCF site have a negative correlation that affects the gene expression that is indicated by the colour. C) The region overlapped with several ENCODE datasets measured in GM12878 (ENCODE Project Consortium, 2012). The entire region shows no evidence of FAIRE-seq enrichment. The FAIRE probe is adjacent to a DNase I-enriched region and overlaps a ATF3 peak. The CTCF site overlaps a RAD21 and SMC3 peak; this potentially indicates that the CTCF site is part of the cohesin complex.

3.3 CONCLUSIONS

The correlation results in this chapter suggest that, even when considering only a small population of cell lines, there is a great abundance of significant correlations. Though some assays are more often associated with one another, there is no great imbalance between the results achieved, even when only considering intra-chromosomal autosome pairs. There are, in total, 8,060 significant CTCF-CTCF pairs, 16,408 CTCF-FAIRE pairs, 7,297 CTCF-RNA pairs, and 17,058 FAIRE-RNA pairs (empirical p -value < 0.1).

Independent of which assays are associated with each other, most significant pairs are proximal to each other. This is strong evidence that these correlations are being driven by local biochemical interactions rather than global cellular features. The different pairs show for overlapping and proximal pairs that Spearman's Rank correlation ρ is nearly always positive. This is expected, and could indicate that chromatin accessibility is responsible for gene expression or vice versa (Boyle et al., 2011; Hoffman et al., 2012; Thurman et al., 2012). There is scientific evidence for both cases, and my data and analysis can only identify these correlations, but not determine causality. The CTCF-CTCF pairs in this result most likely indicate that these different CTCF sites are biochemically linked (directly in loops or within larger complexes). When ChIA-PET CTCF data for a single or multiple LCLs becomes available, this could be used to validate my results.

3.3.1 *The molecular pathway regulating transcription*

Combining all of the significant phenotype-to-phenotype correlations can provide a useful indication of how intertwined these assays are on a functional level. There are 345 examples of systematic correlations between FAIRE, CTCF and gene expression (i.e. all assays are correlated with each other). These systems

show how the different assays interact with each other and that there is, potentially, a strong link of causality between them. However, given that I only use molecular assays at a single time point, no direction of causality can be determined. As such, it is not possible, for example, to conclude whether chromatin accessibility enables transcription factor binding, or vice versa.

However, the associations suggest that most assays are positively correlated when in close proximity. While causality from DNA towards transcription is a complex combination of biochemical events, our assays, and their correlations, suggest that this is far from a random process. I have population data for chromatin accessibility (FAIRE), transcription factor binding (CTCF), and gene expression. The correlations between all of these data suggest that gene expression is mostly positively correlated with chromatin accessibility (i.e. when chromatin is more accessible, there is more gene expression). Also, gene expression is positively correlated with CTCF binding. Furthermore, chromatin accessibility is, for proximal sites, positively correlated with CTCF binding. This combined knowledge suggests a far-from-random association between each of the phenotypes. Unfortunately, using the current combination of only molecular assays, it is not possible to determine causality. However, as QTLs are observed for each molecular assay, I will discuss in the next chapter how I can use QTLs to determine direct and indirect causality.

3.3.2 *Lengths of CTCF-CTCF pairs suggest different roles*

For all of the CTCF-CTCF pairs found, a more detailed examination was performed for those on chromosomes 18 and 20. To gain insight into the role of CTCF loops, the significant CTCF-CTCF pairs are overlapped with a wide variety of TFs, HMs and whole-genome assays measured on GM12878 by the ENCODE consortium. By clustering the CTCF-CTCF pairs on the length of the potential loops (the distance between both sites), a clear

distinctive profile is found. CTCF-CTCF pairs that are within 1Mb distance of each other show less overlap with HMs or TFs. Rad21 and SMC3, the only members of the cohesin complex that were measured, show no bias for short or long loops. The longer CTCF-CTCF pairs show a much stronger overlap profile, with many of the measured histone modifications as well as several transcription factors. Most interesting is the difference between the overlap of factors like YY1, ZNF143 and ELF1, although no obvious biological functional reason for this is known. However, mediator is a known transcriptional cofactor involved in binding two cohesin units to form DNA loops ([Kagey et al., 2010](#)), and I could not study these mediator-enabled DNA loops as there is no mediator ChIP data available that was measured on GM12878.

When partially reconstructing the epigenome of a small population using only FAIRE, CTCF, and transcript information, a high amount of correlation between the different measurements can be observed. There are thousands of significant correlations, and when these data sets are combined, there is a strong suggestion that there is a particular way the system works to regulate transcription. The FAIRE data suggest that the chromatin has to be in accessible status for CTCF to bind and for transcription to be increased. By correlating CTCF sites to each other, I discovered many potential CTCF loops. There appears to be a significant difference in biochemical activity depending on the length of the loops, which indicates a divergence in functionality.

3.4 DATA

3.4.1 CTCF

I obtained ChIP-sequencing data for 51 unrelated individuals from the CEU HapMap population. On each of the sequencing files, peaks were called. To address biological and technical

variance that had not been accounted for, the data were quantile-normalised, centred, and the first PCA component was removed.

Out of 63,753 CTCF sites identified, Zhihao found 2,898 CTCF sites falling in repeat sequences or regions with complex structures in the genome. We reasoned that a much larger sample size is required to resolve the complex structures in these regions, and that they should be removed from our analysis. In detail, 2,578 CTCF sites lie completely within repeat sequences marked by a merged set consisting of Repeat Masker, Segmental Duplication, or Simple Repeats data sets from the table browser of the UCSC Genome Browser. 35 CTCF sites lie within the Immunoglobulin heavy chain locus (chr14:106053226-106330470) and 285 fall in the Major histocompatibility complex (MHC) region (chr6:28477797-33448354). We also removed 3,427 CTCF sites showing very little signal.

The number of CTCF sites before and after filtering is consistent with the length of each chromosome and ranges, after blacklisting filtering, from 33 for chromosome Y, to 5012 for chromosome 1 (Table 3.6).

Table 3.6: Number of CTCF sites per chromosome

Chromosome	All	After blacklisting
1	5935	5012
2	4970	4121
3	3981	3739
4	3113	2888
5	3429	3161
6	3664	3206
7	3210	2916
8	2839	2580
9	2600	2326
10	2925	2705
11	3070	2926

12	3156	3017
13	1503	1429
14	2247	2047
15	2242	1977
16	2066	1900
17	2728	2585
18	1369	1273
19	2237	2103
20	1616	1568
21	646	607
22	1196	1080
X	2782	1968
Y	229	33

3.4.2 FAIRE

I obtained FAIRE-chip data for 60 individuals from the CEU HapMap population. For details, refer to Chapter 2.4.2. When using the FAIRE peaks, the first PCA vector was removed. The first three PCA components are removed when using FAIRE probes to reduce systematic variance.

3.4.3 RNA-seq

I obtained RNA-sequencing data for all individuals of the CEU HapMap population for whom we have FAIRE or CTCF data from the Geuvadis consortium (Lappalainen et al., 2013). Paolo Casale determined a per-gene abundance using BitSeq 0.4.3 (Glaus et al., 2012) for all protein-coding genes. The Ensembl *Homo sapiens* GRCh37.69 gene build is used as the reference

genome (Flicek et al., 2012). The first PCA component is removed to reduce systematic variance.

3.4.4 *ENCODE datasets on GM12878*

GM12878 is a lymphoblastoid cell line produced from the blood of a female donor with northern and western European ancestry (CEU) by EBV transformation. This cell line has been part of the original HapMap consortium and has since been featured as a central cell line for ENCODE (ENCODE Project Consortium, 2012; ENCODE Project Consortium et al., 2007) as well as for deep sequencing in HapMap (International HapMap Consortium, 2003, 2005, 2007) and 1,000 Genomes (1000 Genomes Project Consortium et al., 2010, 2012). Due to the study design of ENCODE, there is a large amount of molecular data available for this cell line, and due to its nature, it is a good reference cell line for our CTCF and FAIRE panel. For GM12878, I obtained data for transcription factors (Table 3.7), histone modifications (Table 3.8), and genome-wide sequencing assays (Table 3.9).

As within the ENCODE consortium, the following abbreviations are used to indicate the source lab/university for each dataset:

- Broad is Broad Institute of MIT and Harvard
- Duke is Duke University
- Haib is HudsonAlpha Institute for Biotechnology
- OpenChrom is Duke University, The University of North Carolina at Chapel Hill, and The University of Texas at Austin
- Sydh is Stanford University, Yale University, University of Southern California, and Harvard University
- Uw is University of Washington

Table 3.7: Transcription Factors in ENCODE

Lab	Transcription Factor	Number of peaks
Haib	ATF3	1884
Haib	BATF	24600
Haib	BCL11A	13256
Haib	BCL3	6147
Haib	BCLAF1	3422
Sydh	BRCA1	339
Sydh	CHD2	4036
Broad	CTCF	40006
OpenChrom	CTCF	42492
Sydh	CTCF	42808
Uw	CTCF	35011
Broad	CTCF	61525
Stanford	CTCF	79957
Haib	EBF1	27622
Haib	EBF1	27655
Sydh	EBF1	25043
Haib	EGR1	1236
Haib	EGR1	13662
Haib	ELF1	20528
Haib	EP300	3687
Sydh	EP300	716
Haib	ETS1	2879
Sydh	FAM48A	11
Sydh	FOS	1744
Haib	GABPA	5095
Sydh	IRF3	711

Continued on next
page

Table 3.7: Transcription Factors in ENCODE

Lab	Transcription Factor	Number of peaks
Haib	IRF4	14229
Haib	IRF4	14229
Sydh	JUND	1715
Sydh	KAT2A	26
Sydh	MAX	2087
Haib	MEF2A	16694
Haib	MEF2C	6777
OpenChrom	MYC	2415
Sydh	NFE2	713
Sydh	NFKB1	10073
Sydh	NFKB1	10390
Sydh	NR2C2	1054
Sydh	NRF1	5042
Haib	PAX5	19618
Haib	PAX5	14340
Haib	PBX3	7431
Haib	POLR2A	20091
Haib	POLR2A	34699
OpenChrom	POLR2A	12781
Sydh	POLR2A	21446
Sydh	POLR2A	9040
Sydh	POLR3A	150
Haib	POU2F2	14441
Haib	RAD21	36015
Sydh	RAD21	27295
Stanford	RAD21	23947

Continued on next

page

Table 3.7: Transcription Factors in ENCODE

Lab	Transcription Factor	Number of peaks
Haib	REST	2541
Sydh	RFX5	3069
Haib	RXRA	2965
Haib	SIN3A	10504
Haib	SIX5	4442
Haib	SLC22A2	14592
Sydh	SMC3	25470
Haib	SP1	13139
Haib	SPI1	35821
Haib	SRF	2412
Haib	SRF	9118
Sydh	STAT1	9
Sydh	STAT3	124
Haib	TAF1	5169
Sydh	TBP	9766
Haib	TCF12	15028
Haib	USF1	7074
Sydh	USF2	6794
Sydh	WRNIP1	1138
Haib	YY1	22261
Sydh	YY1	1821
Haib	ZBTB33	1934
Haib	ZEB1	8304
Sydh	ZNF143	18310
Sydh	ZZZ3	740

Table 3.8: Histone modifications in ENCODE

Lab	Histone modification	Number of peaks
Broad	H2AFZ	40693
Broad	H3K27ac	30501
Broad	H3K27me3	10100
Broad	H3K36me3	37061
Broad	H3K4me1	41605
Broad	H3K4me2	45162
Broad	H3K4me3	24575
Broad	H3K79me2	34750
Broad	H3K9ac	20320
Broad	H3K9me3	26162
Broad	H4K20me1	6024
Uw	H3K27me3	18085
Uw	H3K36me3	44425
Uw	H3K4me3	24112

Table 3.9: Genome-wide sequencing assays in ENCODE

Lab	Assay	Number of peaks
OpenChrom	FAIRE	235562
Duke	Dnase	183953

3.4.4.1 CTCF ChIA-PET ENCODE datasets

In the ENCODE datasets, there are no CTCF ChIA-PET assayed in LCLs available. However, for 2 other cell lines, CTCF ChIA-

PET has been performed ([ENCODE Project Consortium, 2012](#)). K562 is a female cell line of a patient with myelogenous leukemia ([Lozzio and Lozzio, 1975](#)). MCF-7 is a female cell line of a patient with breast cancer ([Soule et al., 1973](#)). For MCF-7, a replicate CTCF ChIA-PET is also available. In [Table 3.10](#), the number of ChIA-PET CTCF pairs is provided for chromosomes 18, 20 and between 18 and 20.

Table 3.10: ENCODE CTCF ChIA-PET dataset

Cell line	Chr18 - Chr18	Chr18 - Chr20	Chr20 - Chr20
K562	226		487
MCF-7	439	8	1175
MCF-7	133		561

3.5 METHODS

3.5.1 *Testing for associations*

I used Spearman's Rank correlation to perform the pairwise associations between each of the datasets. Within R, the function `spearman.test` from the package `pspearman` was used for this purpose. For CTCF-to-CTCF, CTCF-to-FAIRE, CTCF-to-RNA, and FAIRE-to-RNA, I limited the associations to all pairs on the same autosomes. For each pairwise association, the p-value, Rho, and distance between the sites are stored for further analysis.

3.5.2 *Testing associations for significance*

To address the multiple testing issue, empirical p-values are determined for all associations, excepting the CTCF-to-CTCF associations. To accomplish this, permutations are performed. The empirical p-value is determined by ranking the original p-value within a site-specific null distribution. The empirical p-value is obtained using the following steps:

1. For each region, I randomly swap the annotations 1,000 times
2. Each random set of annotations is used to calculate a new p-value

3. For each permutation, the lowest p-value is stored as our-region specific null distribution
4. Each association is ranked within this null distribution
5. Each rank is divided by the number of p-values in the null distribution

The rank in the null distribution divided by the number of p-values in the null distribution is considered our empirical p-value (7).

$$\text{Empirical p-value} = \frac{\text{rank p-value in null distribution}}{\text{number of permutations} + 1} \quad (7)$$

For CTCF-FAIRE and CTCF-RNA interactions, I considered each CTCF site as a tested location. For FAIRE-RNA, I considered each FAIRE peak, or FAIRE probe, as a tested location.

3.5.3 Comparing CTCF-CTCF pairs with ChIA-PET data

To compare the distribution of identified CTCF-CTCF pairs that are also CTCF ChIA-PET pairs, I used the KolmogorovSmirnov (KS) test. CTCF-CTCF pairs are considered to be overlapping if both CTCF sites overlap a ChIA-PET CTCF pair with at least 1bp.

The KS test is a nonparametric method to compare two samples for differences in the shape and location of both distributions. By testing for differences between the two groups of CTCF-CTCF pairs using the KS test, I am able to determine whether ChIA-PET pairs are more likely to be significantly correlated with the population data.

3.5.4 Overlapping assays

To determine which other molecular phenotypes overlap with the significant pairs, I perform an overlap based on the peaks

called in ENCODE. I consider an overlap to exist if at least 1bp is overlapping.

3.5.5 *Data visualisation*

Within R ([Team, R Development Core, 2005](#)), the package Ggplot2 ([Wickham, 2009](#)) was used to visualise all data.

THE DRIVER OF EPIGENETIC CORRELATIONS

This chapter expands on the results obtained in Chapter 3 by integrating Quantitative Trait Loci (QTLs).

4.1 INTRODUCTION

The nature of the epigenomic landscape was assessed with the combination of the Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE), CTCF binding and gene expression techniques (Chapter 3). The high correlation of data between these molecular traits, driven by their functional and biochemical dependency, does not allow determination of the chain of causality. Genotypes, especially the germ-line variations, are unique in that they offer a mostly fixed reference point to analyse the correlations. Assuming that there are no somatic mutations in the cell lines, each experiment is performed on a fixed set of genotypes that are tagged in HapMap and 1,000 Genomes. Given this assumption, the genetic backgrounds for the cells used for the FAIRE, CTCF and gene expression experiments are identical. Therefore, any individual differences in the epigenome are, when the environment is identical, at some level, most likely due to a causal regulatory genotypic effect. Obviously, some of the observed epigenome differences can be attributed to population substructures rather than a single genotype. However, this study has been performed on unrelated individuals, which have been selected within HapMap to avoid population structures. Another potential pitfall is that for some of the epigenome correlations, with genotypic regulation, the regulation could be attributed to coincidentally-overlapping loci rather than coincidental loci. In these cases, the given genotypic regulation is not in

linkage disequilibrium (LD) with the real genotypic regulation, and is thus not causal.

4.1.1 *Inference of causality*

When assessing association studies and their causality, there is a chance that the association is confounded and thus false. Within the field of epidemiology, a method has been established to avoid such confounding issues. Mendelian randomisation is the idea, first described by Katan ([Katan, 1986](#)), that within meiosis each allele has an equal chance being parental and maternal ([Katan, 1986](#)), and thus the germline genetic variations can be used as a proxy for environmental changes ([Wheatley and Gray, 2004](#)) ([Lawlor et al., 2008](#)). This random assortment of alleles, and thus genotypes, gives rise to a situation similar to a randomised trial ([Davey Smith and Ebrahim, 2005](#)). As such, the knowledge of known effects of genetic variation can be used to understand what variance can be expected from other environmentally-influencable phenotypes, such as diseases. The incorporation of this knowledge of variation in combination with confounder effects provides a much more reliable method to affirm genetic association studies that claim causality. When Mendelian randomisation is applied within population studies, it has demonstrated its effectiveness. For example, in coronary disease, it has been shown that not C-reactive protein ([C Reactive Protein Coronary Heart Disease Genetics Collaboration \(CCGC\) et al., 2011](#)) but rather lipoprotein(a) is causal ([Kamstrup et al., 2009](#)).

Another method to understand causality is structural equation modelling (SEM) ([Haavelmo, 1943](#); [Wright, 1921](#)). Using SEM, it is possible to model and study how phenotypes relate to each other through direct and indirect genetic influence ([Gianola and Sorensen, 2004](#)). When incorporating all genotype-to-phenotype and phenotype-to-phenotype interactions in a SEM, it would be possible to predict the effect of a genotype change or the effect

of an environmental change. However, given the significant number of unknown phenotypic relations and the effect that genotypes have on them, constructing an SEM is challenging.

4.1.2 *Direct versus indirect causality*

In this chapter, I introduce genotypes and analyse the effect on the phenotype-to-phenotype correlations using assay-specific QTLs. The interactions between FAIRE, CTCF and gene expression from the previous chapter are used; however, now the analysis incorporates the corresponding QTL effects (Figure 4.1). To explore the order of these interactions, I will start from the genetic effects and infer their direct and indirect molecular functions. The results of this chapter can be used as an indicator of these kinds of interactions and their molecular functionality. However, as previously stated, usage of SEM or Mendelian randomisation would be beneficial to establish true causality.

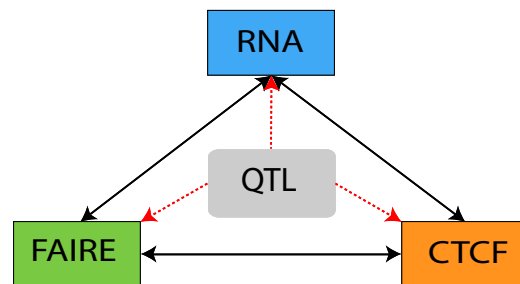


Figure 4.1: Causality of a QTL inside a complex integration of FAIRE, CTCF and RNA. A QTL can have an effect on each of these molecular events independently, but most likely, there is a single direct causal effect and the rest are indirect causal effects.

When two sites correlate and one of them has a QTL, this QTL will also, at some level, correlate with the other site. For example, if FAIRE correlates with CTCF and there is a known CTCF-QTL, this QTL will likely show similar significance when tested as FAIRE-QTL (Figure 4.2). Although both of the genotypes can

be significantly associated with two different phenotypes, there is a risk that both QTLs are coincidentally overlapping and thus are not actually driven by the same genetic effect (e.g. the best QTLs for both phenotypes are in high LD, but are actually associated with both phenotypes). When this is not the case, and both phenotypes are driven by the same genetic effect, it is interesting to study which processes the QTL is driving the most.

Obviously, the causality can be similarly dissected for the same situations holding for CTCF-to-CTCF, CTCF-to-FAIRE, FAIRE-to-RNA, and CTCF-to-RNA.

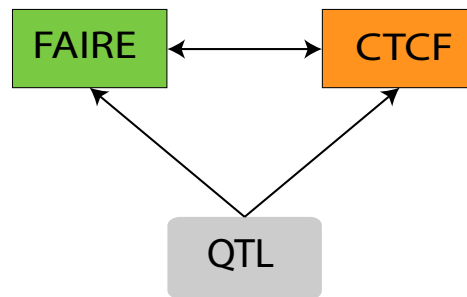


Figure 4.2: Causality of a QTL. When a QTL is both associated with, for example, FAIRE and CTCF, the question remains: what is the real causal effect and what is the indirect effect?

In this chapter, I will use the terms direct and indirect causality. For each QTL, there is, possibly, a single direct causal effect, while all other observed changes are indirect (Figure 4.3). For the population of interest, I have measured chromatin status, CTCF binding and transcription. These molecular assays were shown in the previous chapter to be highly correlated. However, given the limited dataset, there will be other active biochemical pathways that are unmeasured in these correlations. For this reason, for not all discovered correlated phenotypes, the direct causal effect was measured and thus effects are indirect.

Molecular assays are an indirect measurement of the actual phenotype. Because of this, there might be crosstalk between

the actual phenotype that is influencing the measurement. Furthermore, the assay could, for various reasons, achieve more precise quantifications at different genomic locations. For example, there will be some crosstalk between FAIRE and CTCF, as these two biochemical events overlap. For example, in order for CTCF to bind, the chromatin has to be accessible, and thus FAIRE can be proxy for CTCF binding.

To dissect whether a QTL is affecting phenotype A or phenotype B, several correlation were performed. A naive view of discovery of the most likely causal pathway would be to assess the strength of the correlations (given by effect size). In addition, it is also possible to assess conditional correlation. For each QTL, I calculate its effect using Spearman's Rank correlation and store the p-value for significance and Rho value for the effect size. For the first correlation, the original phenotype is used. For the second correlation, I regress out the other molecular phenotype first. After this, I compare the Spearman's Rank correlation p-values (in $-\log_{10}$ scale) to determine the effect of the regression for both molecular phenotypes (I call this delta). The delta p-value can be used to determine which phenotype is most affected by the QTL.

To illustrate this, I provide the following example (Figure 4.4). QTLx is correlated with phenotype A and phenotype B, both have a similar Spearman's Rank p-value (8) and Rho value (0.5). When phenotype B is regressed out of phenotype A, the Spearman's Rank correlation between QTLx and the residuals gives p-value 6 and Rho 0.4. When phenotype A is regressed out of phenotype B, the Spearman's Rank correlation between QTLx and the residuals gives p-value 1 and Rho 0.1. The delta in p-value indicates that regressing out phenotype A has the largest effect. Hence, for this particular example, the direct causal effect of QTLx is most likely phenotype A, making phenotype B the indirect effect. This method will provide a good indication about the causality of the QTL in the given epigenetic correlation. The

delta will be used to rank all pairs; in this way, common patterns in causality can be discovered.

A potential pitfall of this method is that the compared p-values are influenced by the quality of the assay that is used. For example, if an assay has more internal variation, this will, by nature, lead to worse p-values and thus a potential bias towards this assay being less-often causal. I also split the result in each of the plots based on the QTL of discovery; in this way, it is possible to determine that not all directions of causality are in the direction of the assay in which the QTL was discovered originally. To avoid major differences in p-values between the assays caused by different sample sizes, only individuals are used that have been evaluated by all three assays. Throughout this chapter, I will use the delta of Spearman's Rank p-value to discriminate between direct and indirect causality.

An issue when overlapping different molecular QTLs is that sometimes overlapping QTLs are in the same region, but not being driven by the same underlying locus (i.e. coincidental overlap of QTLs). In these cases, although a QTL is overlapping, this is not the real or true QTL. One way to address this problem is to determine the LD between the best QTL for each of the overlapping molecular phenotype. In this way, only QTLs that are in high LD should be considered for causality determination, because only for these cases is the same locus (and thus genetic regulation) used. Unfortunately, there is no genome-wide QTL data (each study focussed only on nearby variants for QTL mapping) for the different molecular assays, and thus we cannot assess whether the overlapping QTL is also the best QTL genome-wide. Especially, because many phenotype-to-phenotype correlations are distal, the QTL data does not overlap. For this reason, some of the conditional regressions performed in this chapter can be attributed to phenotypes being driven by two distinct QTLs. This will most likely be true for many of the pairs that are located far away from each other. For this reason, I have only showed examples, if possible, when

the distance between the molecular phenotypes is small and the LD between the best QTLs is high ($r^2 > 0.8$). However, for many of the cases that I do not show (but which are part of the distributions), I cannot determine the LD between the best QTLs, because the distance between both molecular phenotypes is too great, and there is no way of knowing if a distal QTL is actually the best QTL.

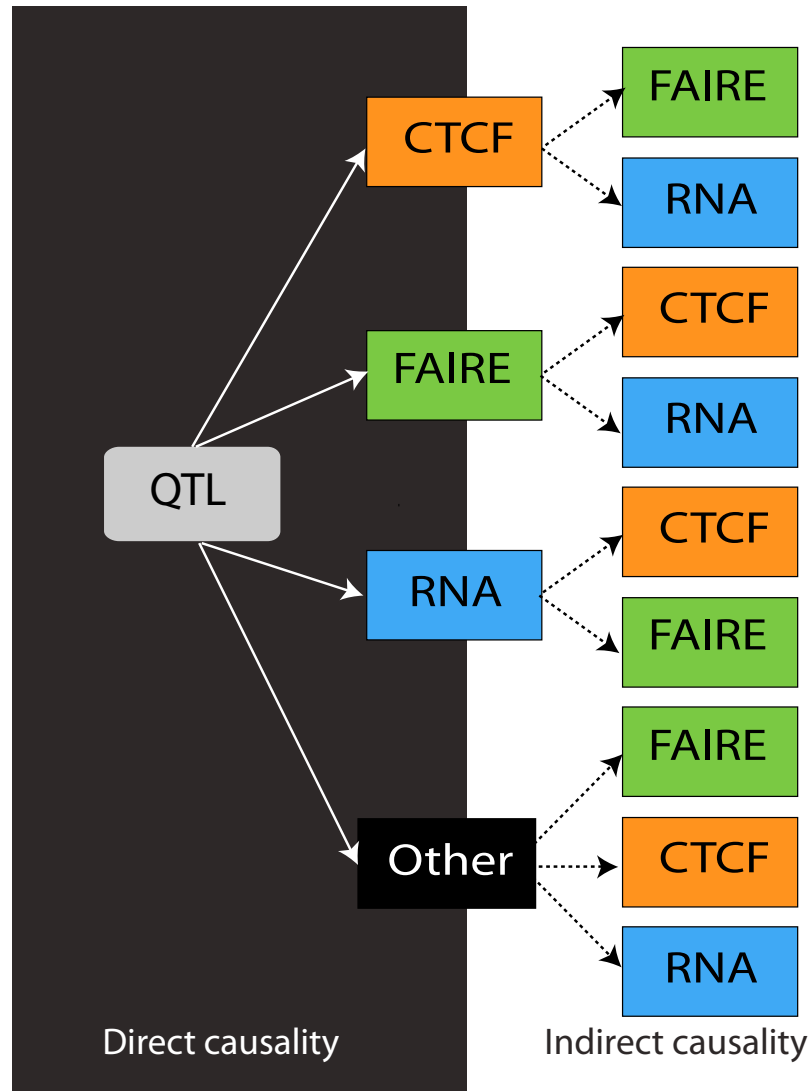


Figure 4.3: Direct versus indirect causality. Although a genotype might have been observed to be associated with a specific phenotype such as a QTL, this does not necessarily imply a direct causal effect. Most likely, for many QTLs, the only observation made will be indirect.

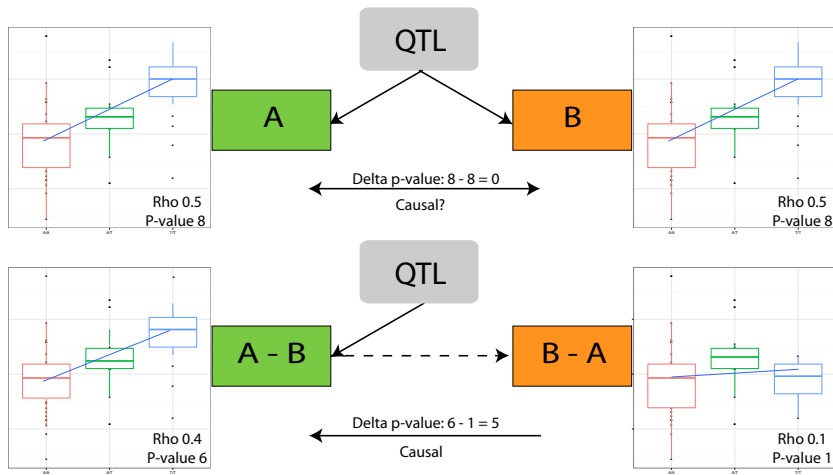


Figure 4.4: Direct versus indirect causality example. A given genotype is associated with two different molecular phenotypes as being a QTL with similar Spearman's Rank p-value and Rho. The difference in p-value (delta p-value) is 0, and thus there is no indication of this QTL being specific for phenotype A or B. After regressing out phenotype A out of B and B out of A, the same correlations are performed, and the difference in p-values is calculated. In this case, the delta p-value is 5. This indicates that, for this example, phenotype A is the direct causal effect and phenotype B the indirect causal effect.

4.2 RESULTS

The high number of associations between FAIRE, CTCF and gene expression requires a limitation of the potential overlap between their results. In this chapter, only the significant pairs are taken into consideration. For the CTCF-CTCF pairs, significance is defined as a Spearman's Rank p-value $< 1.0 \times 10^{-6}$. For the CTCF-RNA pairs, CTCF-FAIRE probes pairs, and FAIRE probes-RNA pairs, empirical p-values have been determined using permutations (1,000 times). Only pairs with an empirical p-value < 0.1 are considered significant.

There are 8,060 significant CTCF-CTCF pairs, 16,408 CTCF-FAIRE probes pairs, 7,297 CTCF-RNA pairs, and 17,058 FAIRE probes-RNA pairs (Chapter 3). In total, there are 26,554 CTCF-QTLs, 164 FAIRE-QTLs (Chapter 2), and 4,781 RNA-QTLs (Montgomery et al., 2010). Combined, there are 31,406 unique molecular QTLs. As discussed in Chapter 2, subsequently to this analysis I have performed QTL analysis using the FAIRE probes instead of the FAIRE probesets. However, the analyses in this chapter are based only on using QTLs for the FAIRE probesets.

4.2.1 *Random direction of causality*

As stated in the introduction of this chapter, a potential pitfall of my method for determining causality involves differences in the variance of the assays. I have performed a simple test to test the effect of the variance on the causality method. I randomly selected 1,000 significant pairs of CTCF-RNA, CTCF-FAIRE and FAIRE-RNA. I then constructed 500 random genotype assignments. These random genotypes can only be associated to these phenotype-to-phenotype pairs by chance. Given this fact, when using my method for causality determination, if assays have similar levels of variance, the resulting distribution of the delta p-values should be uniformly distributed and not skewed towards any of the assays used.

The random assessment of causality using the generated genotypes indicates no skew towards any of the used molecular assays (Figure 4.5). In each case, there are similar numbers of pairs driven by phenotype A as by phenotype B. Because of this result, I can assume that when using QTLs instead of generated genotypes, any given skew in the resulting delta p-values can be attributed to the QTL and not to a difference in variance of the molecular assays.

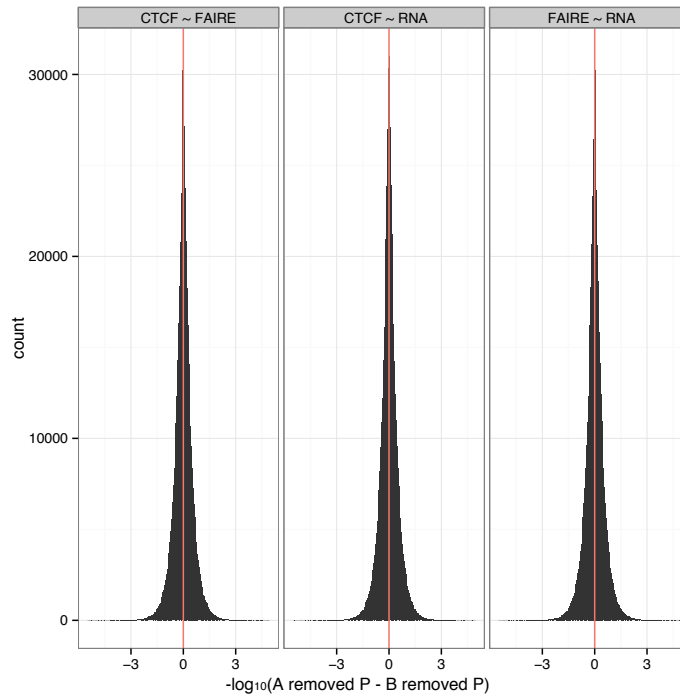


Figure 4.5: For 1,000 randomly-selected significant CTCF-FAIRE, CTCF-RNA and FAIRE-RNA pairs (empirical p-value < 0.01), the causality has been determined for 500 randomly-generated genotypes. For each pair of phenotypes, two Spearman's Rank tests were performed to determine the association between the genotype and the first phenotype (A) after removing the second phenotype (B), and vice versa. The distribution of difference in p-values between the two Spearman's Rank tests is shown (in $-\log_{10}$). None of the assays shows evidence for having more variance, and thus potentially influencing further results.

4.2.2 *CTCF to FAIRE*

There are 16,408 significant CTCF-FAIRE probes pairs on the autosomes. When overlapping both the CTCF and the FAIRE probes with the molecular QTLs, 538 pairs have a QTL. All 538 molecular QTLs are in origin CTCF-QTLs. The reason for this observation is the low number of original FAIRE-QTLs identified.

Similar to the strategy explained in the introduction of this chapter, I used regression to determine causality by comparing the Spearman's Rank correlation p-values (Figure 4.6). For the pairs with a QTL, the comparison of p-values of the regression tests indicates that 79% of the pairs are driven by CTCF. Hence, this analysis suggests that, most often, the CTCF is the driver for the FAIRE probe. Below, I provide an example of when CTCF is driving (Section 4.2.2.1) and when the FAIRE probe is driving (Section 4.2.2.2) the causality.

4.2.2.1 *CTCF driving FAIRE*

To illustrate how a CTCF site drives the triangle of correlations, I selected the following example (Figure 4.7A, p-value $<1.8 \times 10^{-5}$). The CTCF site (chr22:42928810-42929321) is correlated with a FAIRE probe (chr22:42612893-42612941) 315.9kb away (Figure 4.7B). The QTL rs299322 (chr22:42941390) is significantly associated with the CTCF site (Figure 4.7E, p-value = 0.0019) and the FAIRE probe (Figure 4.7C, p-value = 0.008). To understand whether the CTCF or the FAIRE probe is driven by the molecular QTL, the same correlations were performed after regressing out the other molecular assay.

For example, for the CTCF-QTL, the FAIRE probe effect is first regressed out (Figure 4.7F, p-value = 0.009), and for the FAIRE-QTL, the CTCF effect is regressed out (Figure 4.7D, p-value = 0.97). By eliminating the effect of each assay independently, the QTL effect can be compared. The removal of the FAIRE effect from the CTCF signal has almost no consequence (delta Rho

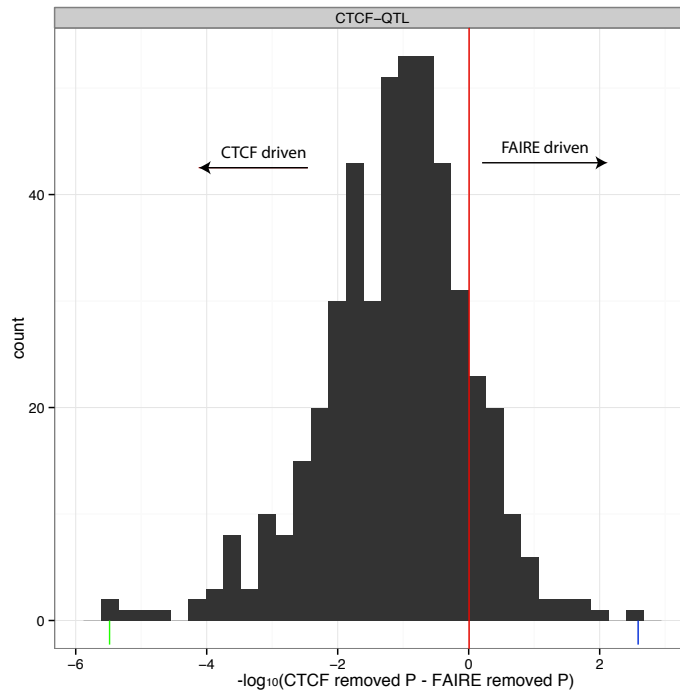


Figure 4.6: CTCF-FAIRE probe pairs QTL effect. The division on the X-axis ranks the effect (delta p-value) between regressing out CTCF and FAIRE. The majority of pairs are CTCF-driven. The blue tick is an example showing FAIRE driving CTCF [4.2.2.2](#). The green tick is an example showing CTCF driving FAIRE [4.2.2.1](#)

0.06), indicating that, for this QTL, the actual FAIRE levels are not important. When removing the CTCF effect from the FAIRE signal, the QTL effect disappears (delta Rho 0.24). This indicates that the most likely driver for this correlation is CTCF.

Although FAIRE probes as a measure has a higher variability and can be influenced by several components, the regression results indicate that, for this QTL, the CTCF has a major impact and thus is most likely the causal effect. Given that no FAIRE probe QTLs have been established, I cannot determine whether this overlap is coincidental. Because of this, the observed disparity in causality could be due to the FAIRE probe not being driven by the same genetic effect as CTCF, but by another that is coincidentally overlapping with the CTCF signal.

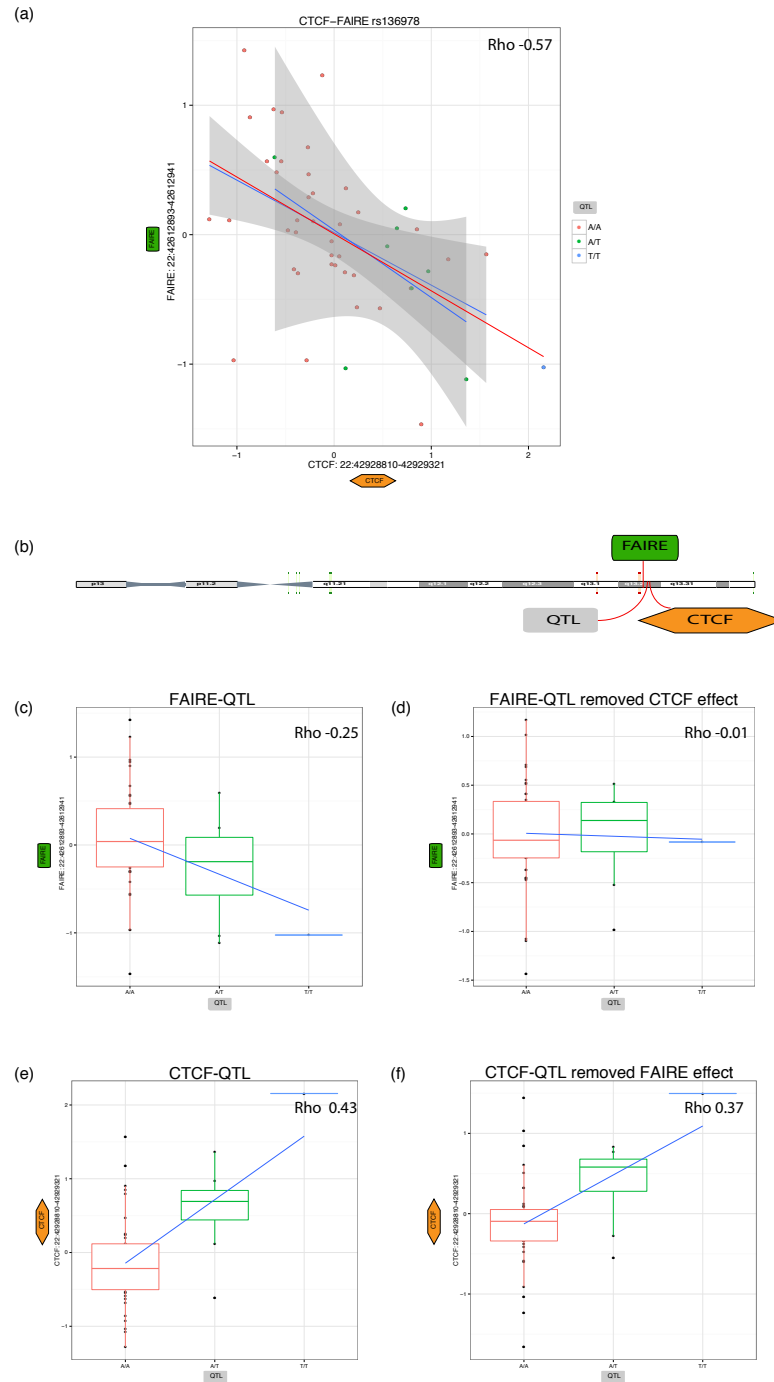


Figure 4.7: CTCF-FAIRE probe pair with QTL that is driven by the CTCF. In this example, the proximal (b) CTCF-FAIRE probe pair are tightly correlated and are under the influence of a QTL; individuals are coloured based on their genotype (a). The QTL is both a cis CTCF-QTL (e) and a cis FAIRE-QTL (c). By regressing out the CTCF, the FAIRE-QTL is removed (d). By regressing out the FAIRE probe, the CTCF-QTL is only slightly impacted (f).

4.2.2.2 FAIRE driving CTCF

For most CTCF-FAIRE pairs with a molecular QTL, the CTCF is driven by the QTL, and the FAIRE is indirectly driven by the CTCF. However, there are examples where the opposite is true (Figure 4.8A, p-value $<5.4 \times 10^{-8}$). The CTCF site (chr10:31040756-31041163) and FAIRE probe (chr5:31041026-31041076) are correlating and overlapping (Figure 4.8B). The QTL rs10826861 is significantly associated with the CTCF site (Figure 4.8E, p-value $<7.6 \times 10^{-9}$) and the FAIRE probe (Figure 4.8C, p-value $<1.6 \times 10^{-13}$). The reason why this QTL was not discovered as a FAIRE-QTL, is because for mapping FAIRE-QTLs, I only examined the probe-sets and not each individual probe. To understand if the CTCF or the FAIRE probe is more important for the molecular QTL, the same correlations are calculated, but now after regressing out the other molecular assay. So, for the CTCF-QTL, the FAIRE probe effect is first regressed out (Figure 4.8F, p-value = 0.13) and for the FAIRE-QTL, the CTCF effect is regressed out (Figure 4.8D, p-value $<4.5 \times 10^{-4}$). The removal of the CTCF effect from the FAIRE has a small effect (delta Rho = 0.34), indicating that for this QTL, the actual CTCF levels are less important. When removing the FAIRE effect from the CTCF signal, the QTL effect disappears (delta Rho = 0.50). This indicates that the FAIRE signal is more important in driving this correlation. Given that the QTL overlaps the peak called for the CTCF binding and only 48bp away from the FAIRE probe, the location would have suggested causality in the other direction. In addition, because of the proximity of the molecular signals, the results are less likely to be influenced by coincidental overlapping and are both caused by the same locus.

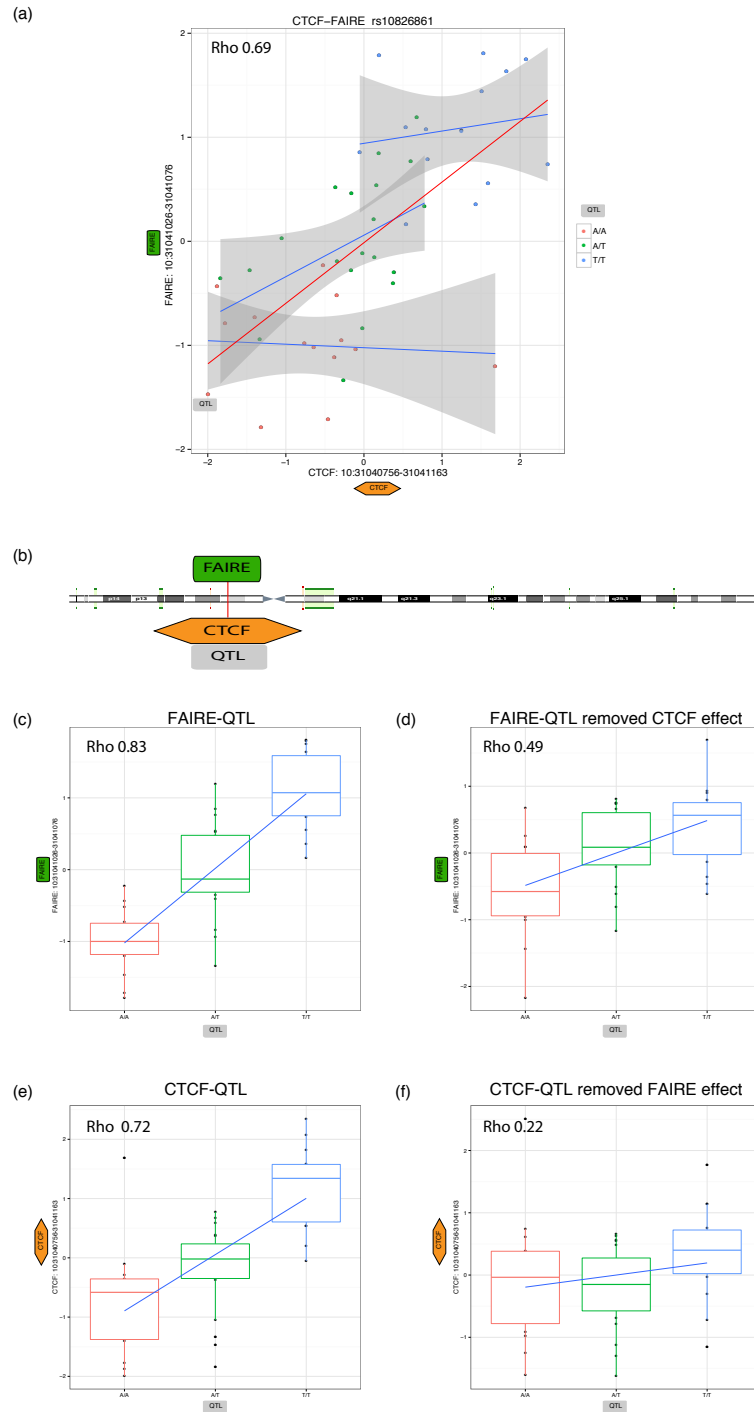


Figure 4.8: CTCF-FAIRE probe pair with QTL that is driven by the FAIRE. In this example, the overlapping (b) CTCF-RNA pair is tightly correlated and is under the influence of an overlapping QTL; individuals are coloured based on their genotype (a). The QTL is both a cis CTCF-QTL (e) and a trans eQTL (c). By regressing out the CTCF, the FAIRE-QTL is only slightly impacted (d). By regressing out the FAIRE signal, the CTCF-QTL is removed (f).

4.2.3 CTCF to RNA

There are 7,297 significant CTCF-RNA pairs on the autosomes. There are 467 CTCF-RNA pairs that have a molecular QTL. For 199 CTCF-RNA pairs, the molecular QTL is originally a CTCF-QTL, and for 287 CTCF-RNA pairs, the molecular QTL is originally a RNA-QTL. There are 10 CTCF-RNA pairs for which there is a QTL for both the RNA and the CTCF; however, the QTL for the RNA and CTCF are different genotypes but in LD ($r^2 > 0.8$). Regarding both the CTCF-QTLs and RNA-QTLs, only proximal genotypes were tested; these were not discovered, as they were outside of the discovery window.

As before, I compare the change in Spearman's Rank correlation p-value after regression to determine causality (Figure 4.9). The molecular QTLs are divided by the molecular trait of discovery. There is a distinctive difference in effect between the CTCF-QTLs and eQTLs. For the CTCF-QTLs, 80% of QTLs are influenced by removing the CTCF effect, as might be expected. However, regressing out the CTCF or the RNA similarly influences the eQTLs (42% CTCF and 58% RNA). Hence, this analysis suggests that most often the CTCF is the driver for the RNA, even when the QTL was discovered as an eQTL. I show an example for when CTCF is the causal effect (Section 4.2.3.1) and when RNA is the causal effect (Section 4.2.3.2)).

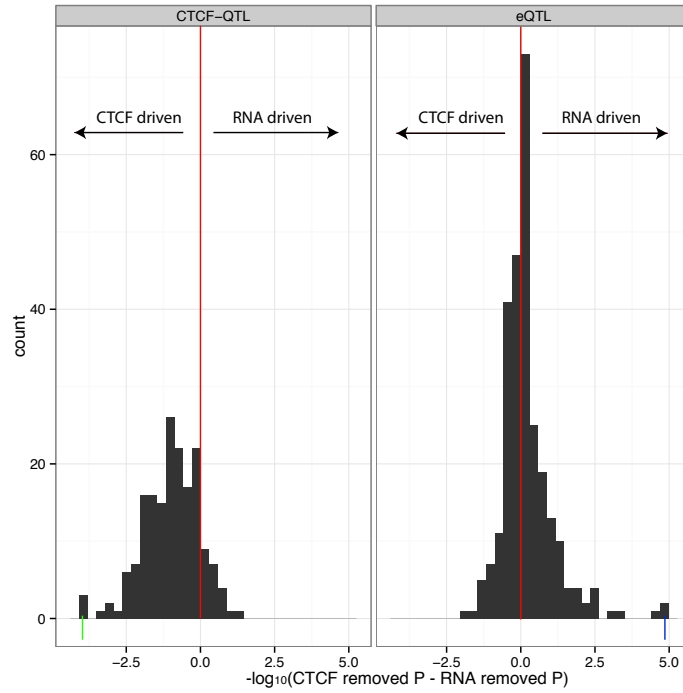


Figure 4.9: CTCF-RNA pairs QTL effect. The division on the X-axis compares the effect (measured in p-value) between regressing out CTCF and RNA. The majority of pairs are CTCF-driven. The blue tick is an example showing CTCF driving RNA 4.2.3.1. The green tick is an example showing RNA driving CTCF 4.2.3.2).

4.2.3.1 CTCF driving RNA

In the following example, I show a CTCF-RNA pair that is driven by a CTCF site (Figure 4.10A, p-value $< 7.0 \times 10^{-5}$). The CTCF site (chr22:42928810-42929321) is correlated to RNA expression (chr22:24655897-24659732) 329 kb away (Figure 4.10B). The QTL rs12483950 (chr22:24296175) is significantly associated with the CTCF site (Figure 4.10E, p-value $< 7.2 \times 10^{-4}$) and the RNA expression (Figure 4.10C, p-value $< 5.2 \times 10^{-3}$). The regressing out of the CTCF (Figure 4.10D, p-value = 0.75, delta Rho = 0.45) and the RNA (Figure 4.10F, p-value = 0.007, delta Rho = 0.09) suggests that CTCF is the most likely direct causal effect of the QTL, and that the changes in gene expression are indirect.

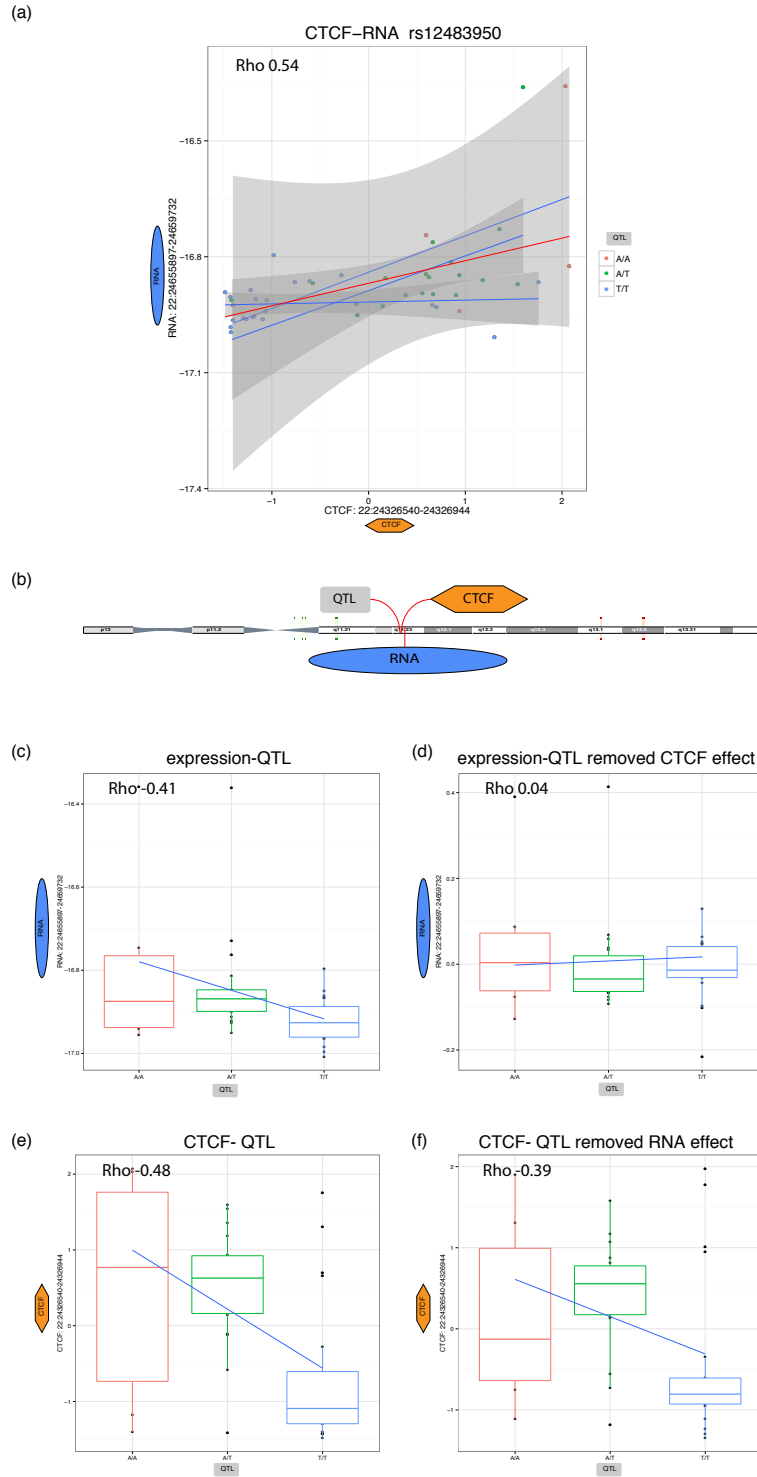


Figure 4.10: CTCF-RNA pair with QTL that is driven by the CTCF. In this example, the distal (b) CTCF-RNA pair are tightly correlated and are under the influence of a QTL; individuals are coloured based on their genotype (a). The QTL is both a cis CTCF-QTL (e) and a trans eQTL (c). By regressing out the CTCF the eQTL is removed (d). By regressing out the RNA the CTCF-QTL is only slightly impacted (f).

4.2.3.2 RNA driving CTCF

In contrast to the CTCF-RNA pairs that are driven by CTCF, the pairs that are driven by RNA are mostly distal from each other (86% of pairs, > 250kb). The reason for this could be that all of the observed effects on this side of the distribution are indirect, operating via a trans regulatory effect, and that the real biochemical event is unmeasured. In addition, there is a significant chance that many of the overlaps are coincidental and thus not caused by the same locus. For this reason, the example shown is for a distal CTCF-RNA pair (Figure 4.11A, p-value < 2.8×10^{-5}). The CTCF site (chr12:122467093-122467538) is correlated to RNA expression (chr12:6560856-6575683) 115,9Mb away (Figure 4.11B). The QTL rs2041385 (chr12:6562293) is significantly associated with the CTCF site (Figure 4.11E, p-value < 4.4×10^{-4}) and the RNA expression (Figure 4.11C, p-value < 8.9×10^{-7}). Similar to earlier examples, both the CTCF (Figure 4.11D, p-value < 5.7×10^{-6} , delta Rho = 0.11) as well as the RNA expression (Figure 4.11F, p-value = 0.45, delta Rho = 0.39) are regressed out to determine direct versus indirect causality. In this example, when only considering these two molecular phenotypes, the RNA is a direct causal effect of the QTL. However, given the large distance between both phenotypes, this is under the condition that this QTL effect is caused by coincidental overlap.

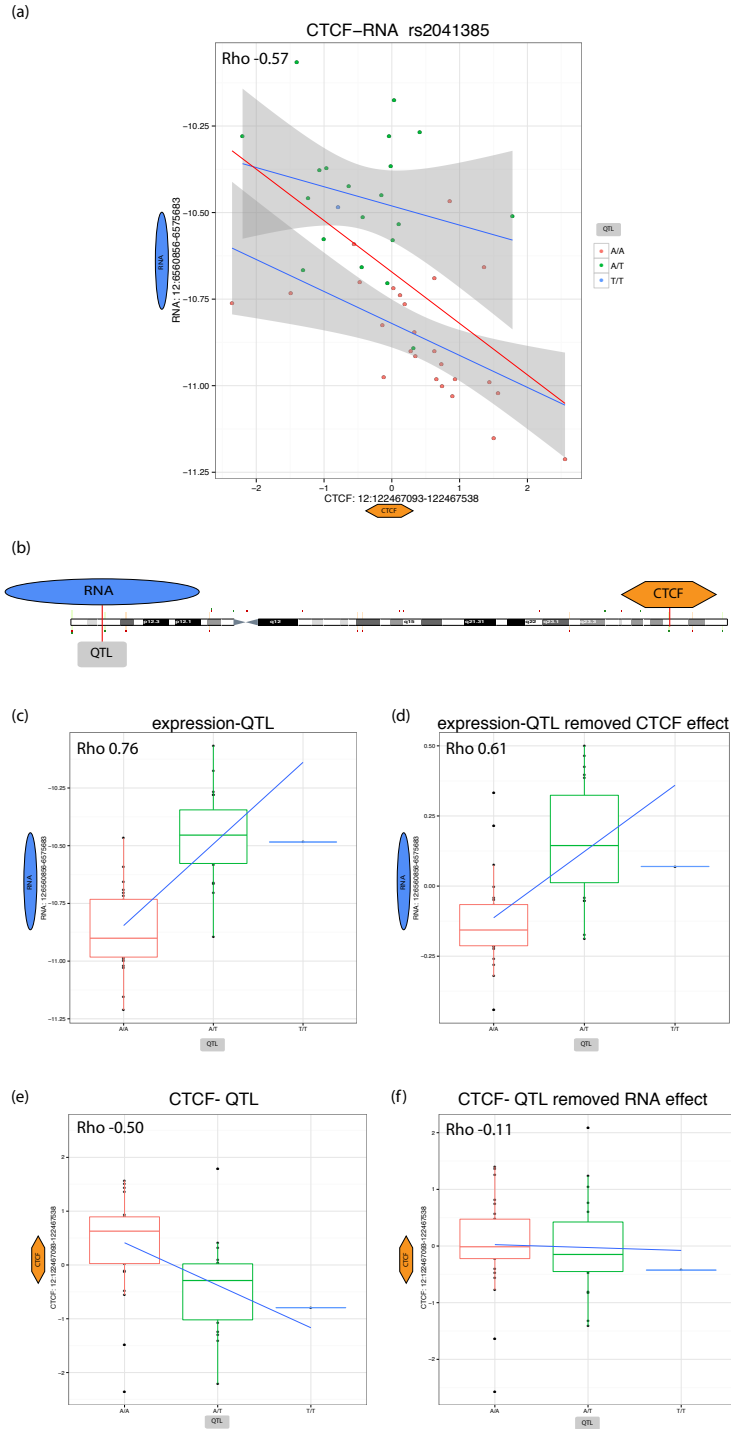


Figure 4.11: CTCF-RNA pair with QTL that is driven by the RNA. In this example, the distal (b) CTCF-RNA pair are tightly correlated and are under the influence of a QTL; individuals are coloured based on their genotype (a). The QTL is both a cis eQTL (e) and a trans CTCF-QTL (c). By regressing out the CTCF the eQTL is only slightly impacted (d). By regressing out the RNA the CTCF-QTL is removed (f).

4.2.4 CTCF to CTCF

There were 8,060 significant CTCF-CTCF pairs discovered on autosomes, and 186 CTCF-CTCF pairs determined to be under the influence of a QTL.

4.2.4.1 *The effect of CTCF-QTLs on CTCF-CTCF pairs*

The existence of CTCF-QTLs shows that CTCF binding can be altered by genotypic differences. However, the effect these QTLs have on CTCF-CTCF pairs is largely unknown. Therefore, similarly as for the other phenotype-to-phenotype pairs, I regress out the genotypic effect to determine how CTCF-CTCF pairs are affected by CTCF-QTLs.

When a QTL affects a CTCF site within a CTCF-CTCF pair, this can result in three different potential effects. First, there is no effect and the CTCF-CTCF pair is not affected. Second, due to the alteration of a CTCF site, the association between the CTCF sites disappears. Third, the direction of the association is inverted (e.g. the Rho value is changed from positive to negative). For each of these three effects, I selected an example to illustrate how a positively correlated CTCF-CTCF pair can be under the influence of a QTL (Figure 4.12).

In the first example, two CTCF sites are correlated and are independently under the influence of a QTL (Figure 4.12A1). Both axes represent a CTCF site, and each point is an individual. The colour of the points indicates the alleles of the individuals: AA, AT or TT. The red line indicates the direction of the correlation (Rho) between the two CTCF sites. In a second plot, the QTL effect is removed (Figure 4.12A2). In this example, the correlation between the CTCF sites remains intact after removal of the QTL effect. This result suggests that the CTCF-CTCF pair correlation is not driven, or altered, by the QTL, but rather by other unmeasured factors.

The second example illustrates another correlated CTCF-CTCF pair (Figure 4.12B1). In contrast to the first example, the

removal of the QTL effect leads to the removal of the CTCF-CTCF pair correlation (Figure 4.12B2). This result suggests that this CTCF-CTCF pair is driven by the QTL.

In a final example, the CTCF-CTCF pair is again positively correlated (Figure 4.12C1). However, after removing the QTL effect, the actual direction of association (Spearman's Rank Rho) is reversed (Figure 4.12C2), though not significant (Spearman's Rank p-value > 0.05).

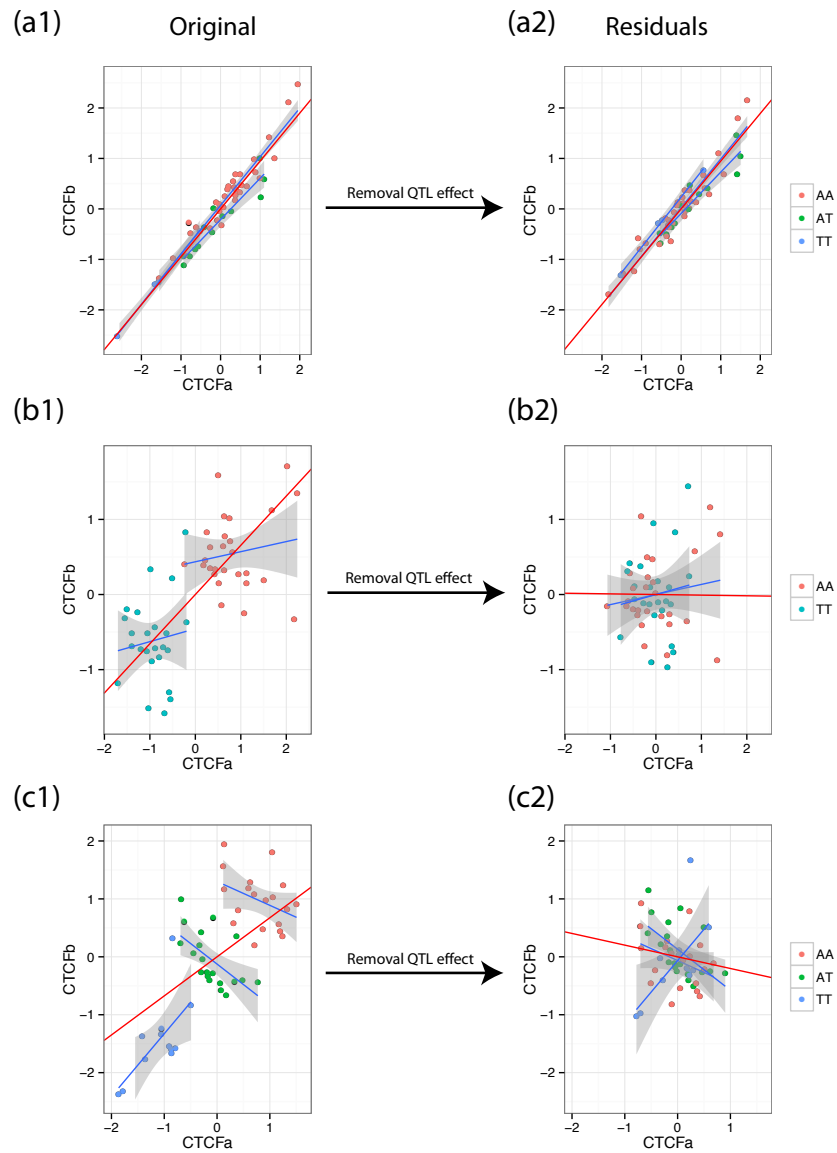


Figure 4.12: CTCF-CTCF pairs that are positively correlated with each other and have a shared QTL. By regressing out the QTL effect on both CTCF sites, the causal effect of the QTL on the discovered CTCF-CTCF pair is visualised. Individuals are coloured based on their genotype. There are QTLs that do not affect the CTCF-CTCF pair correlation (a), QTLs that reduce the correlation (b), and inverse the correlation (c).

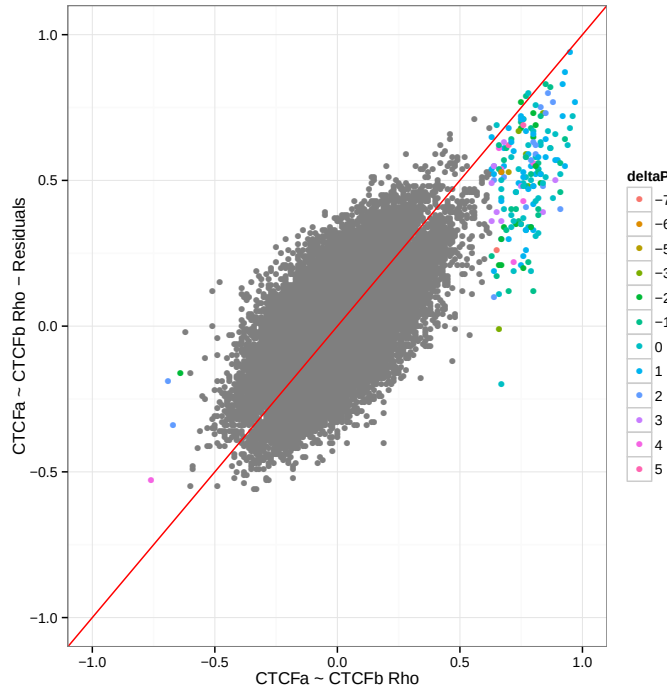


Figure 4.13: CTCF-CTCF pairs with shared QTL. Each CTCF-CTCF pair in this figure has a single QTL that is affected by both CTCF sites simultaneously. On the x-axis is the original Spearman's Rank Rho value between both CTCF sites. On the y-axis is the Spearman's Rank Rho value after regressing out the QTL effect. The colour indicates which CTCF site in the pair is causal. If the pair is driven by CTCFa the deltaP is negative, if the pairs is driven by CTCFb the deltaP is positive.

4.2.4.2 The causal CTCF site

In the previous sections, I attempted to establish the direct effect of a QTL within a phenotype-to-phenotype correlation. However, for CTCF-CTCF pairs it is clear that, at least within the pair, CTCF must be the driver. Although the direction is often not interesting, there is interest in determining whether there is an imbalance in the effect a CTCF-QTL has on CTCF-CTCF pairs. There are 186 significant CTCF-CTCF pairs under the influence of a QTL, which are predominantly positively correlated (Figure 4.13, coloured points are significant pairs). The Y-axis shows

the effect of removing the QTL effect on the CTCF-CTCF pair. Noteworthy is that there are no cases where the QTL negatively affects the CTCF-CTCF pair. There are 55 CTCF-CTCF pairs in which there is a strong change in correlation (> 0.3 Spearman's Rank Rho change), indicating that the left (negative difference p-value) or the right (positive difference p-value) is the causal CTCF site for this QTL. For the remaining 131 CTCF-CTCF pairs, it is not possible to determine whether a single CTCF site is mostly affected by the given QTL. This could indicate that, in these cases, the QTL affects both CTCF sites equally or that the CTCF-CTCF pair is driven by another unmeasured factor.

To illustrate a CTCF-CTCF pair in which the CTCF-QTL is driving one CTCF site, I selected an example of two proximal CTCF sites that correlate (Figure 4.14). For this example, I label one CTCF site CTCFa (chr14:106800840-106803801) and the other CTCFb (chr14:106790714-106791186). CTCFa is correlated to CTCFb (p-value $< 5.3 \times 10^{-14}$), 13kb away (Figure 4.14B). The QTL rs11623092 (chr14:106822043) is significantly associated (CTCFa-QTL p-value $< 2.5 \times 10^{-10}$, CTCFb-QTL p-value $< 3.5 \times 10^{-10}$) with both CTCF sites (Figure 4.14C,E). To understand whether CTCFa or CTCFb is the actual CTCF site for the CTCF-QTL, the same correlations are calculated after regressing out the different CTCF effects. Accordingly, for the CTCFa-QTL, the CTCFb effect is regressed out (Figure 4.14F, p-value = 0.02, delta Rho = 0.44), and for CTCFb-QTL, the CTCFa effect is regressed out (Figure 4.14D, p-value 0.94, delta Rho = 0.76). The results of these regressions indicate that, for this specific CTCF-CTCF pair under the conditions of this QTL, the CTCF loop mostly depends on CTCFb. This is shown in the following way: by regressing out CTCFb, the association is removed, while this is not the case when CTCFa is removed. Therefore, the QTL is most likely causal for variations of CTCFb binding and not CTCFa binding.

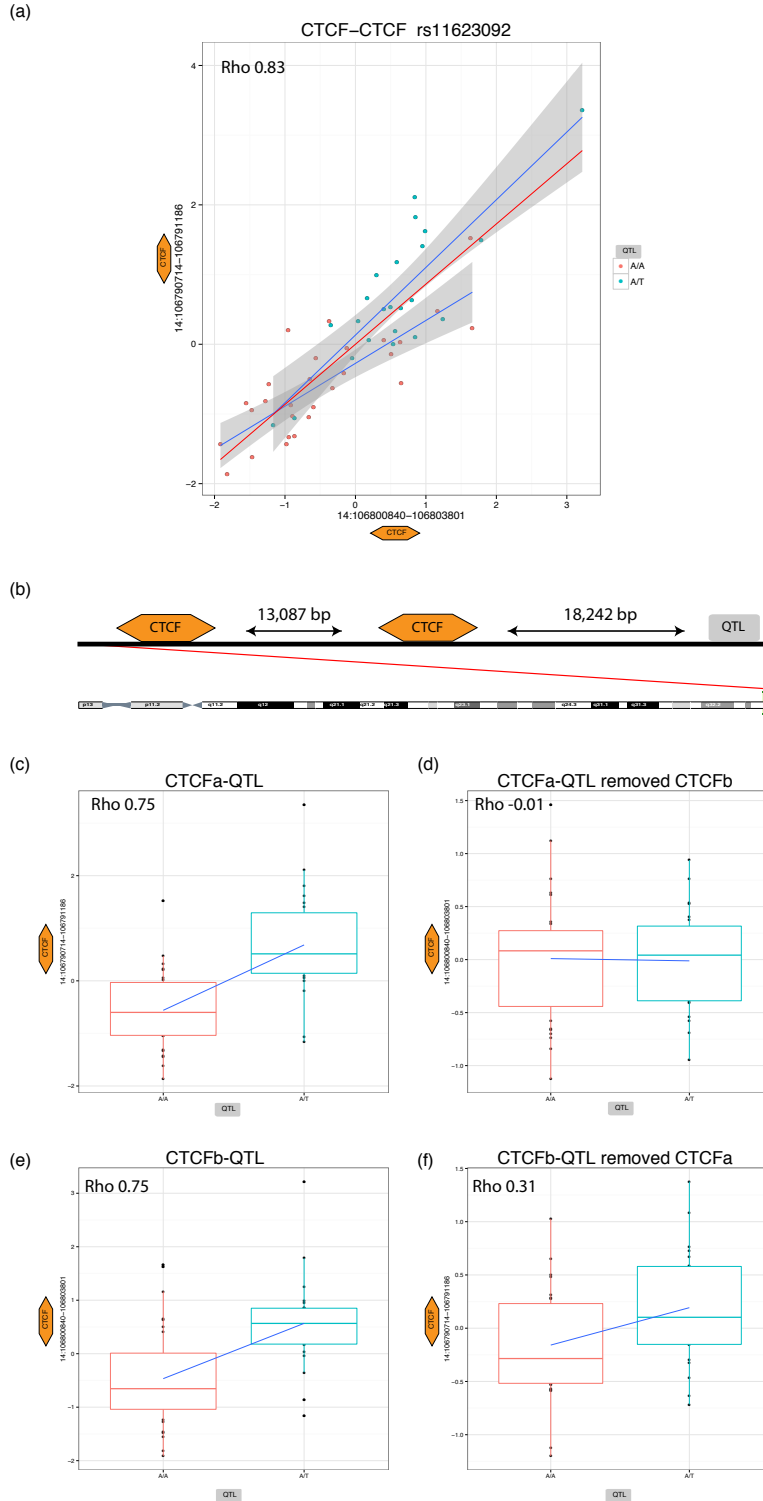


Figure 4.14: CTCF-CTCF pair with QTL that is driven by one CTCF directly. In this example, the proximal (b) CTCF-CTCF pair are tightly correlated and are under the influence of a QTL; individuals are coloured based on their genotype (a). The QTL is for CTCF "a" a cis CTCFa-QTL (c) and for CTCF "b" a cis CTCFb-QTL (e). By regressing out the CTCFa effect, the CTCFb-QTL is only slightly impacted (f). By regressing out the CTCFb effect, the CTCFa-QTL is removed (d).

4.3 CONCLUSIONS

The results of this chapter, combined with those of the correlations chapter, (Chapter 3) provide a great insight into the complex interplay of genetics and epigenetics.

4.3.1 *Common direction of causality*

Building on the previously-established phenotype-phenotype correlations, QTLs can provide great insight into the regulation of these correlations. Incorporating the QTLs within the phenotype-to-phenotype correlation enables the dissection of causality. Doing this, it becomes clear that for many correlations there is a favourable direction of causality. However, for some correlations, no distinction of causality can be made. This lack of a significantly direct causal effect most likely indicates that both measured phenotypes are under the same genetic influence. Alternatively, a potential explanation for this could be that both measured phenotypes are actually under the regulation of an unmeasured third component that is actually regulated by the QTL. It would have been of great interest to model each of these outcomes and to distinguish them. However, given the current method based on differences in p-values, this is desirable but not feasible. I have shown that the same analysis performed with random genotypes yields a different result; however, these provide no easy starting point for determining a fixed threshold to permit distinguishing each of the pairs. For this reason, in this thesis I only use the difference in p-values to rank each of the pairs to highlight general trends and select some examples. Subsequently, no claims about significance are made for each of the causality claims.

4.3.1.1 *CTCF is the most common driver*

Correlation between CTCF and the FAIRE signal is common. For many CTCF-FAIRE pairs, incorporating the genetic effect of a QTL can permit us to distinguish between direct and indirect causality (i.e. driven by CTCF or FAIRE). The most observed effect, under condition of a QTL, is that a CTCF site is driving the FAIRE site or that both are equally affected by the QTL (Figure 4.6). In contrast, there are several CTCF-FAIRE pairs in which the FAIRE site is the direct causal effect; nevertheless, nearly all of these cases are related to distal CTCF and FAIRE sites, making the result biologically difficult to interpret. However, proximity is not essential for chromatin looping interactions, and long-range interactions are commonly found (Sanyal et al., 2012).

Similarly, CTCF is the driver in many of the CTCF-RNA pairs. In the comparison between CTCF-QTLs and eQTLs, within these pairs, the removal of the molecular effect is striking. For eQTLs, the removal of CTCF or RNA can lead to any outcome, while for CTCF-QTLs, the direction is strongly biased towards CTCF being causal (Figure 4.9).

This result is a strong indication of how, irrespective of the molecular assay of QTL discovery, CTCF is most likely changing the chromatin accessibility and thus gene expression. For example, CTCF is changing chromatin into an accessible state so that other transcription factors can bind and transcription can occur. These results are complementary to the results found by Kilpinen et al.: they established that genetic changes are driving TF binding and histone modification, which in turn are the main determinant for gene regulation (Kilpinen et al., 2013). There is evidence that, in this context, histone modifications are merely reflecting earlier regulatory events (Kilpinen et al., 2013), and are often correlated with genetic changes in TF binding sites (McVicker et al., 2013). However, many of these effects will be locus-specific, and there is evidence that epigenetic changes are

both causal and a result of gene expression. For example, in yeast the interplay of TF binding and chromatin are responsible for regulating gene expression, essential for the phosphate response pathway (Lam et al., 2008). In contrast, another study in yeast shows how even after fixing histone modifications that are known to be essential for gene expression, the gene expression abundance is unaltered (Zhang et al., 2014). Although my results suggest that CTCF is a driver of gene expression and chromatin status, some additional experiments are desirable. It would be of great interest to determine the effect of alternating CTCF binding in the discovered phenotype-to-phenotype pairs that are driven by CTCF.

4.3.2 *CTCF loops*

There is a notable difference between CTCF-CTCF pairs and any of the other phenotype-to-phenotype pairs tested in this chapter. Nearly all CTCF-CTCF pairs are positively correlated to each other. Given the biological nature of CTCF-CTCF pairs, i.e. being biochemically linked in so-called loops, this is to be expected, and adds new weight to the discovered CTCF-CTCF pairs. CTCF-CTCF pairs under the influence of a QTL demonstrate that there are three distinctive QTL effects. First, they do not change, and it seems that the individual QTL has a minimal effect on the CTCF loop. Second, the CTCF-CTCF pair disappears, making the QTL a master switch in enabling and disabling the loop. Third, the loop effect is reversed, which entails a positive correlation being converted into a negative one.

When considering CTCF-CTCF pairs under influence of a QTL, there is a considerable number of pairs in which the QTL drives, or alters, both CTCF sites simultaneously (Section 4.2.4). Because of this, the removal of the effect of a QTL quite often leaves the CTCF-CTCF pair correlation intact. Another large group of CTCF-CTCF pairs demonstrates how a QTL affects only a single CTCF site (Section 4.2.4.2). This result suggests

that a single QTL can regulate the binding of a single CTCF site that in turn affects the entire CTCF loop and its potential biological regulatory function.

4.3.2.1 *False discoveries*

Similar to when discovering molecular QTLs, or performing phenotype-to-phenotype correlations, larger sample sizes will yield better results. With increased sample size, the less-strong effects could become significant, which will help to create a more comprehensive epigenomic view. Furthermore, inclusion of additional molecular phenotypes will help to dissect epigenetic causality. For these reasons, it would be ideal to assay as many molecular traits in the same individuals as possible. In this way, the full cascade of causality can be discovered, as well as the direct effect of molecular QTLs that regulate gene expression.

4.4 DATA

4.4.1 *Molecular QTLs*

A combination of QTL datasets is created for this chapter. QTLs mapped for gene expression, FAIRE and CTCF, are used. Combined, there are 31,406 unique molecular QTLs. There are 26,554 CTCF-QTLs (Chapter 3), 164 FAIRE-QTLs (Chapter 2), and 4,781 RNA-QTLs (Montgomery et al., 2010).

4.4.2 *Karyotypes*

In the example plots, I use karyotypes to indicate the location of the correlating molecular phenotypes. Karyotypes and chromosomal locations of the CTCF sites, FAIRE sites, genes and QTLs were obtained from the Ensembl genome browser (Flicek et al., 2012).

4.5 METHODS

4.5.1 *Testing for associations*

I used Spearman's Rank correlation to perform the pairwise associations between each of the phenotypes and the genotypes. For each association, the Spearman's Rank Rho and p-value are stored. Within R, the function `spearman.test` from the package `pspearman` was used for this analysis.

4.5.1.1 *Direction of causality*

A simplistic model was created to determine the most likely direct and indirect causal effect of a QTL, given a phenotype-to-phenotype correlation. In this model, I assumed the variance for each phenotype to be similar.

For a given phenotype-to-phenotype correlation, under the influence of a QTL, I performed the following steps. I independently regressed out the effect of each molecular phenotype on the phenotype-to-phenotype correlation. This can be represented as 8, where A and B are the different phenotypes, $\mathbf{1}$ is a vector of ones, i_A and i_B are the intercepts for A and B, respectively, s_A and s_B are their slopes, and the error is an N by 2 matrix (N = number of samples). Between the residuals of the regression and the QTL, a Spearman's Rank correlation is performed to test the after-regression association. For these two correlations, the p-value and Rho are stored, and the difference in p-value is used to determine the causality.

$$[A, B] = \mathbf{1} * [i_A, i_B] + A * [s_A, s_B] + \text{error} \quad (8)$$

For example, in a CTCF-FAIRE pair, I regress the CTCF signal out of the FAIRE signal. Then, the residuals of this regression are correlated with the genotype using Spearman's Rank correlation, and the p-value is stored. Subsequently, the same process is performed, but now the FAIRE signal is regressed out before

applying Spearman's Rank correlation. The difference in p-value in $-\log_{10}$ scale of these regressions is calculated and is called the delta p-value. The delta p-value is broadly able to classify three different classes. A negative delta p-value indicates that the first phenotype has the largest effect (in the example, this would be CTCF). A positive delta p-value indicates that the second phenotype has the largest effect. Finally, all delta p-values around zero indicate that both phenotypes affect the pair similarly. In R code, this was done by the following steps:

```

1 #load R package for Spearman's Rank test
2 library("pspearman")
3
4 #data object with both CTCF, FAIRE abundances, and the
  genotypes (QTL)
5 data = cbind(CTCF, FAIRE, QTL)
6
7 #regression of both phenotypes
8 data.minusCTCF = residuals(lm(data~CTCF))
9 data.minusFAIRE = residuals(lm(data~FAIRE))
10
11 #correlation on the residuals
12 cor.FAIRE = spearman.test(data.minusCTCF$FAIRE,data$QTL,
  approximation="t-distribution");
13 cor.CTCF = spearman.test(data.minusFAIRE$CTCF,data$QTL,
  approximation="t-distribution");
14
15 #calculation of the delta p.value
16 delta = -log10(cor.FAIRE$p.value) - -log10(cor.CTCF$p.value)

```

4.5.1.2 Causality of random associations

To assess whether the difference in variance of each molecular assay is attributed to a bias in causality direction, I performed the same regression method using randomly-generated genotypes.

I generated a set of 500 random genotype assignments. Each genotype assignment consists of a random combination of A/A,

A/T and T/T. As these genotypes are randomly-assigned, any significant association is purely due to chance. Because of this, each genotype has a similar chance to be correlated with any of the molecular phenotype pairs.

In a subsequent step, I sampled 3,000 significant CTCF-RNA, CTCF-FAIRE and FAIRE-RNA pairs (empirical p-value <0.1; 1,000 pairs each) and used these within the regression model 4.5.1.1. The delta p-values (in $-\log_{10}$) of these random tests were stored. The delta p-values can be used to determine any bias derived from variance introduced by the different molecular assays or signals.

4.5.2 *Data visualisation*

Within R (Team, R Development Core, 2005), the package Ggplot2 (Wickham, 2009) was used to visualise all data.

Part IV

X-CHROMOSOME BIOLOGY

CTCF INVOLVEMENT IN X-CHROMOSOME INACTIVATION

This chapter describes the results obtained from studying CTCF binding on the X chromosome to characterise the role of CTCF in X-chromosome inactivation. When studying the biochemical mechanistic link between CTCF sites by correlating CTCF sites binding to each other, I observed striking difference for CTCF-CTCF pairs on the X chromosome. Many CTCF sites on the X chromosome are correlated with each other, which is driven by the gender of the cell lines. This chapter is part of a paper on which I am the joint first author and that is has been submitted for journal publication.

5.1 INTRODUCTION

In humans, the sex chromosomes consist of the X and Y chromosomes. Females will have two copies of X (XX), and males will have one X copy and one Y copy (XY). As females have two copies of the X chromosome, this could lead to double the amount of gene expression in a similar situation than in males if both X chromosome are transcribed. This poses a problem: effectively, because of this discrepancy, males and females can be considered to be two different organisms. Hence, dosage compensation between the genders is essential. Humans and other mammals use a mechanism of dosage compensation in which one of the female X chromosomes becomes inactivated; in this way, females have one active (X_a) and one inactive (X_i) X chromosome. This dosage compensation mechanism is an extraordinary piece of epigenetic biology that has been studied

extensively (see review (Payer and Lee, 2008; Schulz and Heard, 2013)).

While CTCF is known for being involved in X-chromosome inactivation at the Xist locus, the scale of its involvement has not been fully explored (Chao et al., 2002; Pugacheva et al., 2005; Spencer et al., 2011). In this chapter, I show that there are three classes of CTCF sites on the X chromosome. First, CTCF sites are dosage-compensated, i.e. CTCF binding in males and females is equal. Second, other CTCF sites are not dosage-compensated, and there is binding to both X chromosomes. Third, a few other CTCF sites are directly involved in regulating the transcription of noncoding-RNAs (ncRNAs) in females that show evidence of involvement in X-chromosome inactivation.

5.2 RESULTS

When studying the correlation of CTCF-CTCF pairs (Chapter 3.2.6), I noticed a striking difference in results for the X chromosome. The association between the 1,968 X chromosomal CTCF sites was significant between almost all sites, a clear indication of some hidden population structure or systematic bias. As the CTCF dataset was normalised and the first Principal Components Analysis (PCA) component was removed, there was no indication that a large systematic variable remained in our dataset that could explain this behaviour. Also, all other chromosomes showed far fewer significant associations, indicating that this might be an X chromosome-specific behaviour. The striking number of highly-associated CTCF sites were determined to be driven by the gender of the individual samples.

5.2.1 *CTCF binding on chromosome X has a gender-specific profile*

To understand the reason for the large number of significant associations of CTCF binding sites on the X chromosome, I

performed a more detailed exploration of the CTCF affinity. Examining the distribution of CTCF binding affinities, divided per individual, there is a significant distinction between the male (average binding 0.85) and female (average binding 1.34) LCLs (Student's t-test p-value $< 2.1 \times 10^{-10}$, Figure 5.1). This is a strong indication that most of the observed significant correlations are due to gender. Although the majority of CTCF sites on the X chromosome tend to be highly correlated, there is still a large group of CTCF sites that do not show any difference in CTCF binding affinity between the sexes.

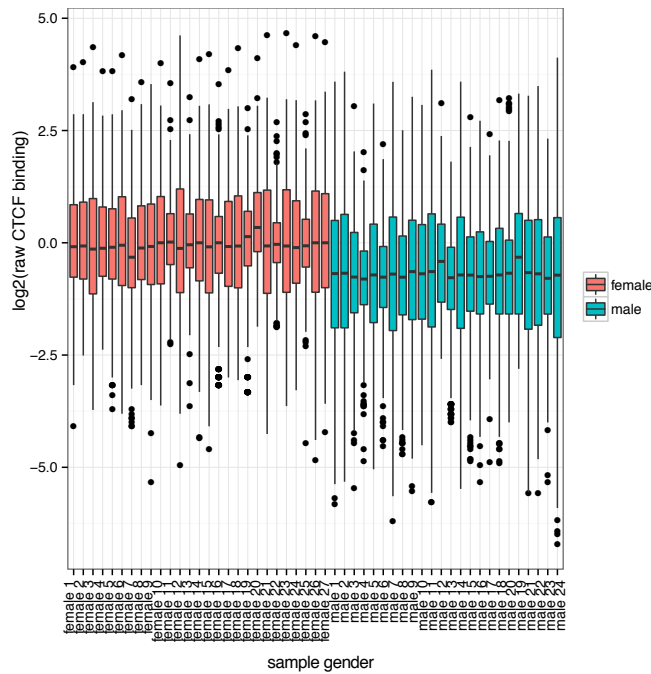


Figure 5.1: Boxplot for CTCF binding affinity on the X chromosome for each LCL. (a) Individuals are ordered and coloured by gender. There is a significant difference in binding affinity between sites in male and female lines (Student's t-test p-value $< 2.1 \times 10^{-10}$).

5.2.2 *Overall lack of dosage compensation for CTCF*

To separate CTCF binding sites on the X chromosome that are gender-specific from the ones that are not, I applied a site-by-site Mann Whitney U (MWU) test. In this test, the distribution of the male samples is compared with the female samples, and a significant MWU means that the distributions have different median values.

The gender-specific CTCF binding pattern that was observed could be due to CTCF binding to both the Xa and the Xi chromosomes, which results in females having twice the amount of CTCF binding. To investigate this scenario, I obtained messenger RNA (mRNA) and non-coding RNA (ncRNA) data for the same cell lines from the gEUVADIS project ([Lappalainen et al., 2013](#)) and studied their transcript behaviour. As CTCF does not seem to be dosage-compensated for the majority of sites, another key question is whether this behaviour is specific to CTCF, or whether it is a more widespread effect that can also be found in other molecular assays. At the time of my study, there had been no other molecular assays performed on a similar scale that overlapped my samples. The closest available dataset, from a QTL study by Degner et al., comprises DNaseI-sequencing data for 90 YRI HapMap LCLs ([Degner et al., 2012](#)). Although, this DNaseI-QTL study was performed on a different population, our focus is on an essential phenotype that this is unlikely to be highly specific to a population.

The results of the site-by-site MWU test on mRNA, ncRNA, DNase I and CTCF were reported by an autosomal group and an X chromosome group (Figure 5.2). The test shows that there are almost no gender-specific mRNA and ncRNA transcripts. Furthermore, there is no significant difference in reported p-value distributions between the autosomes and X chromosome for mRNA (Student's t-test p-value = 0.08) and ncRNA (Student's t-test p-value = 0.87). In contrast, for the DNase I sites, there is a significant increase of gender-specific sites on the X chromo-

some (Student's t-test p-value $< 2.1 \times 10^{-12}$). Likewise, the test indicates that the majority of CTCF sites on the X chromosome have a gender-biased binding affinity (Student's t-test p-value $< 2.8 \times 10^{-70}$), strongly suggesting that a large number of CTCF sites are influenced or driven by gender.

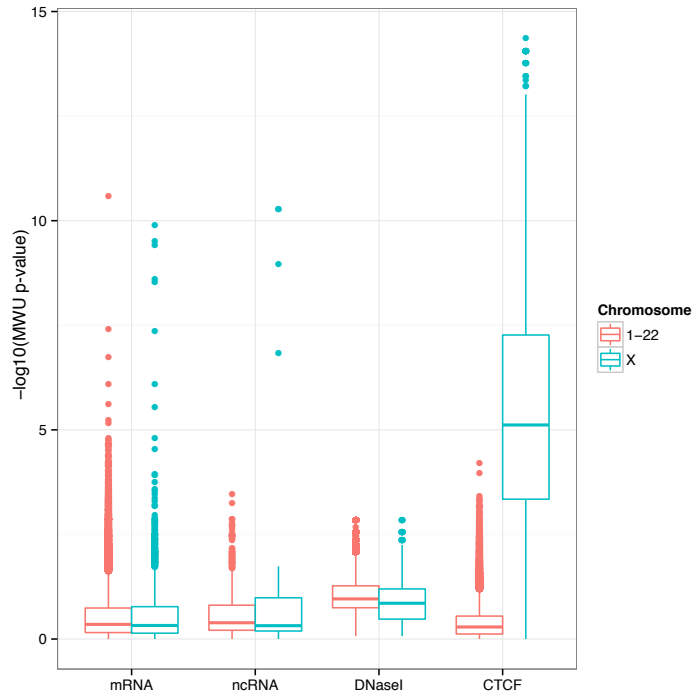


Figure 5.2: MWU results for autosomes and X chromosome for mRNA, ncRNA, DNase I and CTCF. The p-value indicates the degree of gender bias in each of the datasets. The mRNA and ncRNA transcripts show very few gender-specific transcripts. For both DNase I and CTCF, there is an indication that sites on the X chromosome are gender-biased.

To highlight the differences for the X chromosome, all mRNA, ncRNA, DNase I, and CTCF sites outside the pseudoautosomal region (PAR) regions are classified as being gender-specific or not (Figure 5.3). In total, 87% of CTCF sites on the X chromosome outside the PAR show a strong gender-specific effect (MWU P-value < 0.01), while for the other assays, this number is much

lower. A gender-specific effect occurs in only 3% of mRNAs, 9% of ncRNAs, and 4% of DNase I sites.

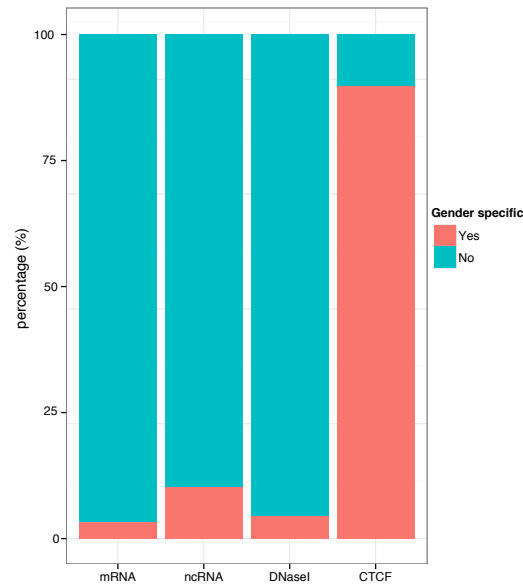


Figure 5.3: Number of mRNA, ncRNA transcripts, and DNase I, CTCF sites that are significantly biased against gender (MWU p-value < 0.01).

5.2.3 CTCF site classification

The CTCF binding affinity difference between males and females can be classified in gender and non-gender driven sites using the MWU test. Comparing the average binding of males and females suggests that there is also a third class of CTCF sites, which only binds in females (Figure 5.4, note that the female binding is divided by two). Taking this into consideration, there are two main classes of CTCF sites and one additional subclass. The first main class comprises CTCF sites that have similar binding behaviour in males and females, and are therefore not significant in the MWU test. I call these CTCF sites "dosage-compensated". The second main class includes CTCF sites that show twice as much binding in females than in males, and are thus significant in the MWU test (p-value < 0.01). This class of CTCF sites can

be further subdivided into CTCF sites that have only binding in females, which I refer to as "female-specific" sites, and "both-active" sites. This classification divides the X chromosome CTCF sites into 1,692 CTCF sites that are both-active, 253 CTCF sites that are dosage-compensated, and 23 CTCF sites that show a female-specific binding affinity.

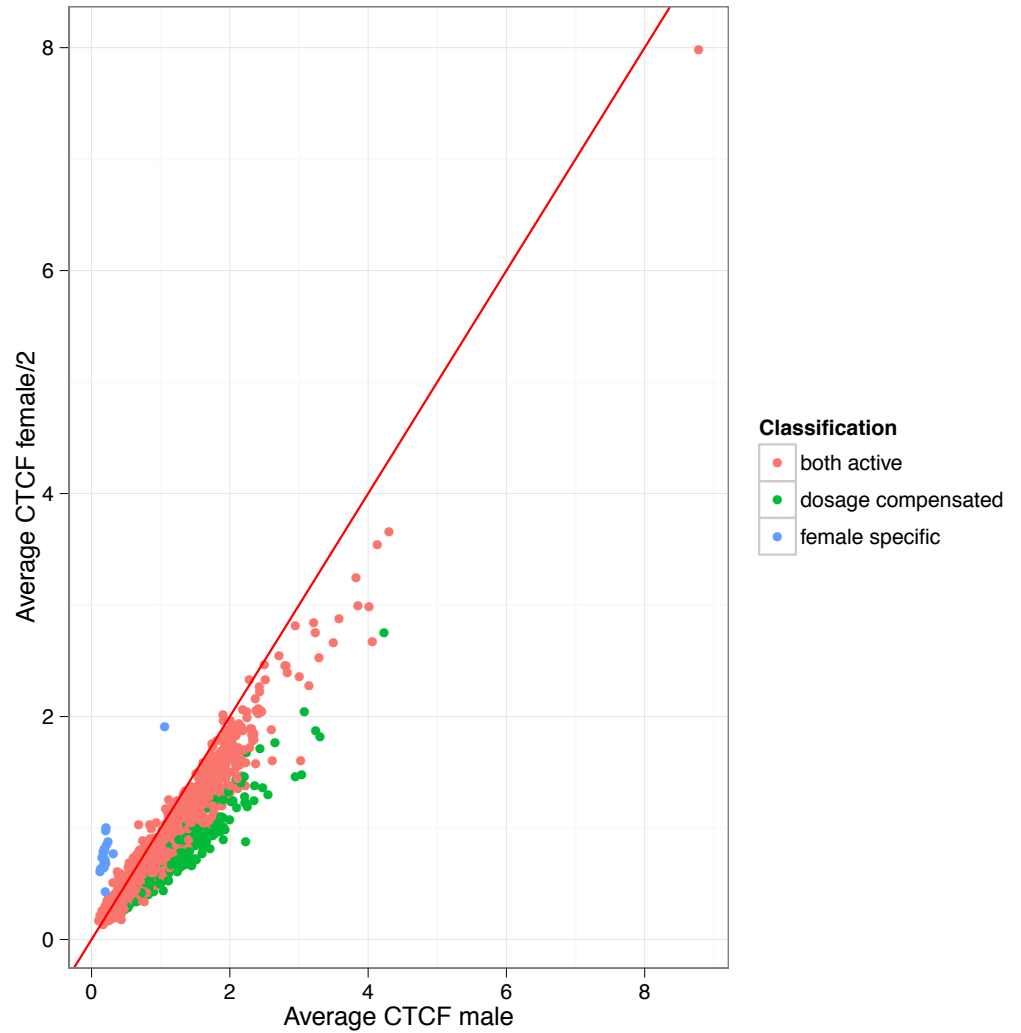


Figure 5.4: CTCF sites on the X chromosome classified based on their MWU p-value and the ratio between the average female/average male. In an effort to aid the visual separation of classes, the female signal is shown as a single X chromosome and not as a pair. In this way, both-active CTCF sites in females have a similar binding affinity to CTCF sites in males (on red diagonal). The red line represents when CTCF sites are both-active. This classification results in 1,692 both-active, 253 dosage-compensated, and 23 female-specific CTCF sites.

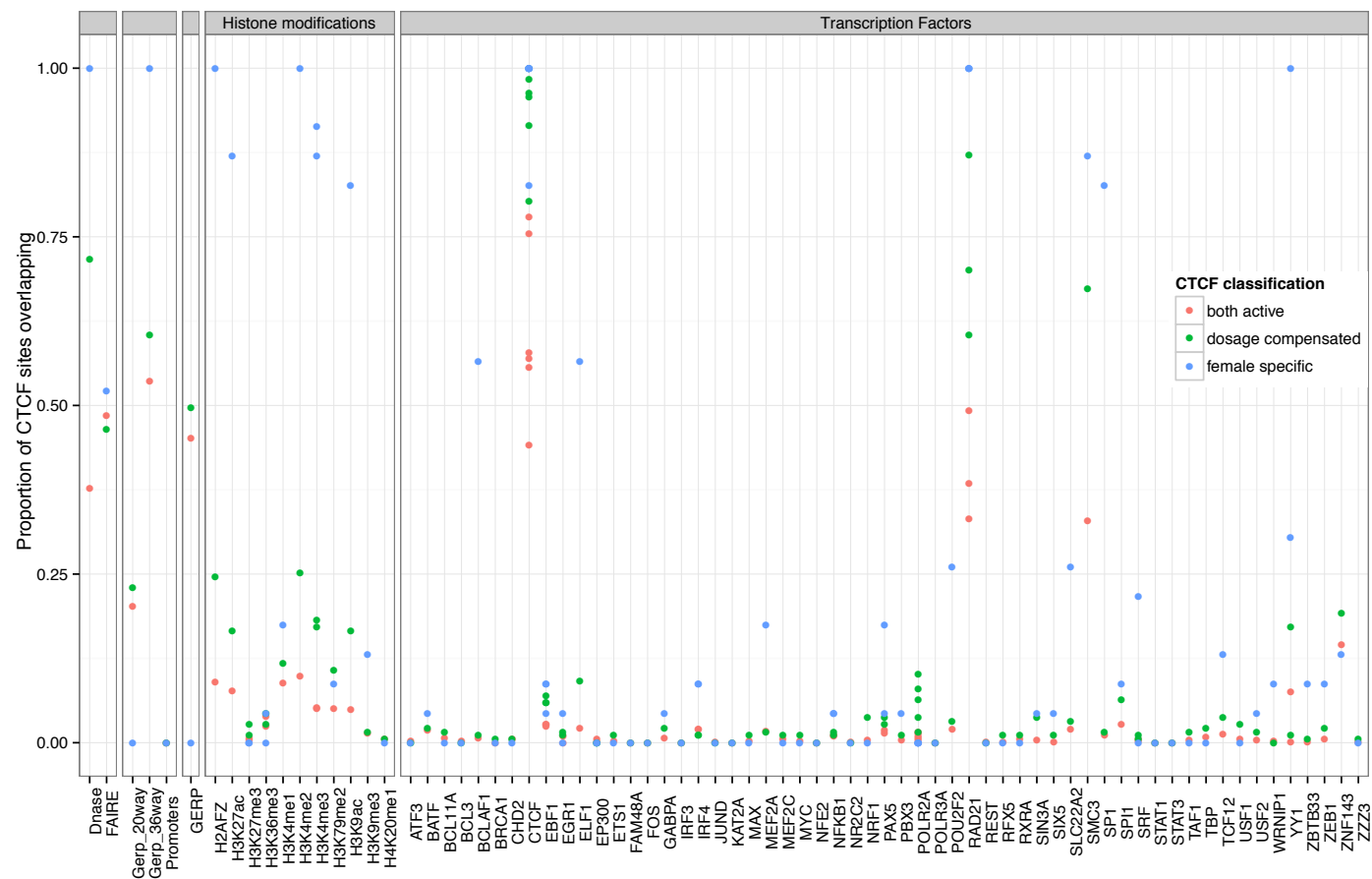


Figure 5.5: CTCF sites that are classified as dosage-compensated, both-active and female-specific are overlapped with ENCODE and Ensembl data. The number of sites overlapping is represented in a proportion overlapping. For some datasets, there are replicates from different labs resulting in multiple data points (CTCF for example). All 3 classes of CTCF sites show enrichment for several HM and TF, though some class-specific overlaps can be observed.

5.2.4 *Biochemical enrichment of different CTCF classes*

In order to study, and potentially understand, the functional differences between the different classes of CTCF sites, I investigated the overlap of these sites with transcription factor (TF) binding and histone modification (HM) from the ENCODE consortium ([ENCODE Project Consortium, 2012](#)). Additionally, I analysed an overlap of CTCF sites with evolutionary constraint regions identified by GERP ([Cooper et al., 2005](#)) and promoter sites ([Flicek et al., 2012](#)), also in GM12878, to understand the genomic context. To ensure that differences in TF binding and HMs are not due to sites being located inside a promoter, only CTCF sites that are not located in a promoter region were analysed.

The overlap was performed both a) on all sites and b) on the sites with the highest signal (top 50%), in which I expected better classification. I present here the top 50% figures; except for noise due to lower numbers, the overlaps using "all sites" are almost identical to the top 50% values.

CTCF sites that are both-active, dosage-compensated, or female-specific are all highly conserved, indicating a strong functional selection of all sites. All of the female-specific sites overlap the 36 eutherian mammal GERP data ([Flicek et al., 2012](#)), while the both-active and dosage-compensated sites respectively overlap 62% and 53%, well in excess of the genome-wide coverage of GERP, which is 5.5% ([Cooper et al., 2005](#)).

The data show a strong tendency for sites to overlap with some key elements of CTCF biology (Figure 5.5). To avoid any promoter bias, all CTCF sites overlapping a promoter were removed for this analysis. First, the members of the cohesion complex, SMC3 and Rad21 (part of the ENCODE data), show a large overlap for both the both-active (SMC3 > 25%, Rad21 > 25%) and the dosage-compensated (SMC3 > 60%, Rad21 > 65%) CTCF sites. Although there are only 23 female-specific CTCF sites, these sites seem to suggest that this class of CTCF sites are in regions with high amounts of biochemical activity.

For FAIRE and DNase I, sequencing was performed within ENCODE on the female LCL GM12878. Therefore, it is reasonable to expect that any essential biochemical activity necessary for female X chromosome biology is present in these overlaps. Irrespective of the classification, approximately 50% of the CTCF sites overlap a FAIRE region. Interestingly, DNase I shows much more disparity, with 100% female-specific, 72% dosage-compensated, and only 38% both-active CTCF sites. Further analysis of DNase I in this chapter will demonstrate that DNase I and CTCF show a high amount of overlap in classification (Section 5.2.5.2).

A Chi-square test was performed to identify which datasets show significantly different overlapping between dosage-compensated and both-active CTCF sites. To visualise these results, the ratio between the overlap for both-active and dosage compensation is plotted and coloured for the p-value results of the Chi-square test (Figure 5.6). Dosage-compensated sites tend to overlap more often with other HMs and TFs than both-active sites (all data points below ratio 1). This could suggest that dosage-compensated CTCF sites have a stronger regulatory role and are consequently more likely to overlap with other factors or marks. However, some sites show higher overlap with both-active CTCF sites. Unfortunately, these overlaps are not significant (Chi-square p-value > 0.05) and are biased because of the low number of sites overlapping with HMs or TFs (red points at the top of the plot).

The dosage-compensated sites overlap significantly more with DNase I sites (Chi-square p-value $< 2.4 \times 10^{-6}$). This suggests that chromatin is more accessible for dosage-compensated CTCF sites, which is something that is known to be essential for regulatory regions. Although CTCF sites that are inside known promoter regions are filtered out, there is still a significant overlap with POLR2A for dosage-compensated sites (Chi-square p-value < 0.011). POLR2A is the largest subunit of RNA polymerase II that transcribes DNA into RNA. All of these results

are consistent with the hypothesis that dosage-compensated sites are in fact CTCF sites that have a regulatory function.

For CTCF to form insulating loops, it needs to interact with protein members of the cohesin complex. This complex machinery consists of several proteins bound together in a single complex (Section 1.3.1). In ENCODE, only SMC₃ and RAD21 are quantified using ChIP-seq. If a site is active for SMC₃, RAD21, and CTCF, this provides a convincing indication of the presence of the cohesin complex. Both SMC₃ and RAD21 are significantly (Chi-square p-value $< 5.0 \times 10^{-5}$) enriched in the dosage-compensated sites compared with the both-active CTCF sites. This suggests that dosage-compensated sites are more often involved in the cohesin complex and are subsequently more likely to form insulating loops.

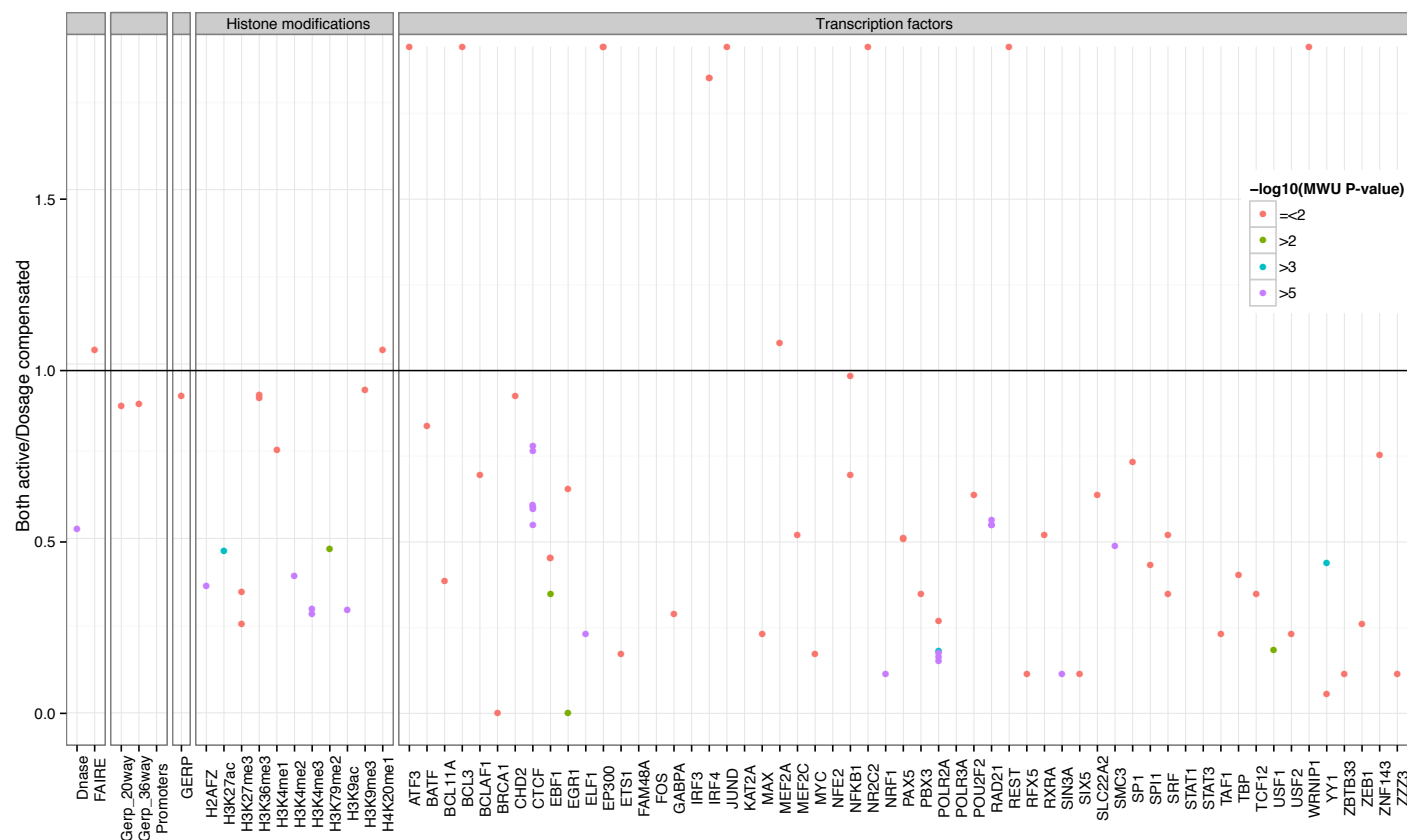


Figure 5.6: CTCF sites that are classified as dosage-compensated or both-active-were overlapped with ENCODE and Ensembl data. On the Y-axis is the ratio of overlap, meaning that at ratio 1, both dosage-compensated and both-active CTCF sites have a similar overlapping profile. The points are coloured by the Chi-square test p-value (in $-\log_{10}$), indicating a significant difference in overlap between both groups. For most datasets, there is no difference in overlap between both groups (red points). In general, CTCF sites that are dosage-compensated, overlap significantly more often with other marks and factors than both-active sites (Chi-square p-value < 0.03).

5.2.5 *Binding signal and spacing of TF and HM*

As shown previously (Section 5.2.4), there is a significant disparity regarding which TFs and HMs overlap with CTCF sites that are classified as both-active and dosage-compensated. To study the exact difference in binding affinity, aggregate plots were created based on the CTCF locations for both classes. Aggregation plots were created for the following HMs: H₃K₄me₃, H₃K₄me₁, H₃K₂₇ac and H₃K₂₇me₃ (figure 5.7A). Interestingly, H₃K₂₇me₃ did not show any increased binding for the dosage-compensated or the both-active CTCF sites. This was surprising, as H₃K₂₇me₃ is known to be involved in X-chromosome inactivation (Brinkman et al., 2006; Marks et al., 2009). However, active CTCF sites were investigated, and thus potential CTCF sites that are repressed by H₃K₂₇me₃ were not part of this comparison. The HMs for the dosage-compensated CTCF sites show strong evidence for accessible chromatin.

Another set of aggregation plots was created for the nucleosome spacing around CTCF sites (Figure 5.7B). The nucleosome spacing of both classes of CTCF sites is useful to determine whether both CTCF sites are actually used. The spacing is similar for both classes, and the both-active CTCF sites even seem to have a more distinctive profile than the dosage-compensated CTCF sites. This evidence is compatible with the dosage-compensated sites being actively bound and being functional. The depletion of nucleosome spacing at the CTCF site suggests that both classes of sites are bound by CTCF at the centre and organise the nucleosomes around them.

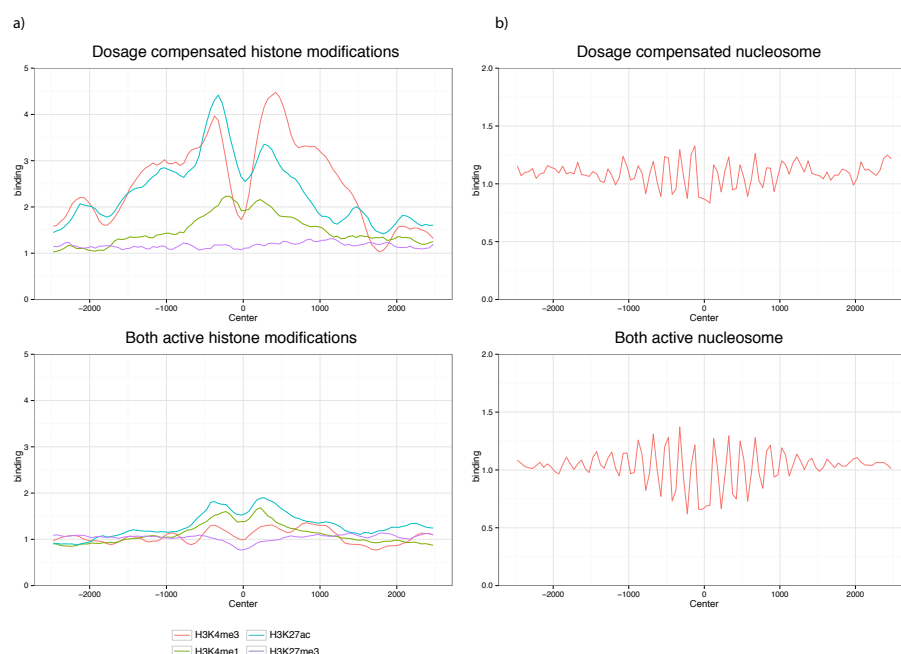


Figure 5.7: CTCF sites that are classified as dosage-compensated or both-active are aggregated with ENCODE data for GM12878. The top 50% abundant CTCF sites that are not classified as promoters are considered. A) Aggregate signal for several histone modifications. Dosage-compensated CTCF sites show higher level of HMs activity, excepting H3K27me3. This suggests accessible chromatin for the dosage-compensated sites. B) Aggregate signal for nucleosome spacing. Irrespective of the CTCF classification, the CTCF is bound and the nucleosomes are organised accordingly.

5.2.5.1 *X chromosome signal for mRNA, ncRNA, DNase I and CTCF*

In a next step to understand whether other molecular datasets are dosage-compensated, a comparison plot was created comparing mRNA, ncRNA, DNase I and CTCF (Figure 5.8). To create comparable datasets, each of the values was transformed to values ranging between 0 and 1. The actual signal intensities are not important for this plot, as the main interest is the trends. The datasets are shown in a comparison between the haploid female signal (Y-axis, average female signal divided by 2) and the gender difference (X-axis, average female signal divided by 2, minus the average male signal). This transformation of the data emphasises the difference between dosage compensation and both-active sites. Accordingly, there are 3 groups to be distinguished: first, the sites that are dosage-compensated are approximately on the red diagonal line. Next, the both-active sites are approximately on the green vertical line. Finally, there are female-specific sites on the right of the green vertical line.

The mRNA panel shows that the majority of mRNAs are, as expected, dosage-compensated. The ncRNA are mostly dosage-compensated, while the different Xist transcripts have only transcription for the females. For the chromatin structure information provided by DNase I, the sites are spread out between dosage-compensated and both-active. In the CTCF panel, there is a spread of sites between being dosage-compensated and active on both chromosomes. Also, there is a rather large group of sites that appear to be female-specific. In comparison with CTCF, for DNase I there are far fewer sites that are active on both chromosomes, and the majority of sites are closest to being dosage-compensated. This is a good indication that CTCF is more important for the both-active regions than DNase I is. Note that there are some regions in DNase I that might indicate male-specific behaviour, but these are not further analysed in this thesis.

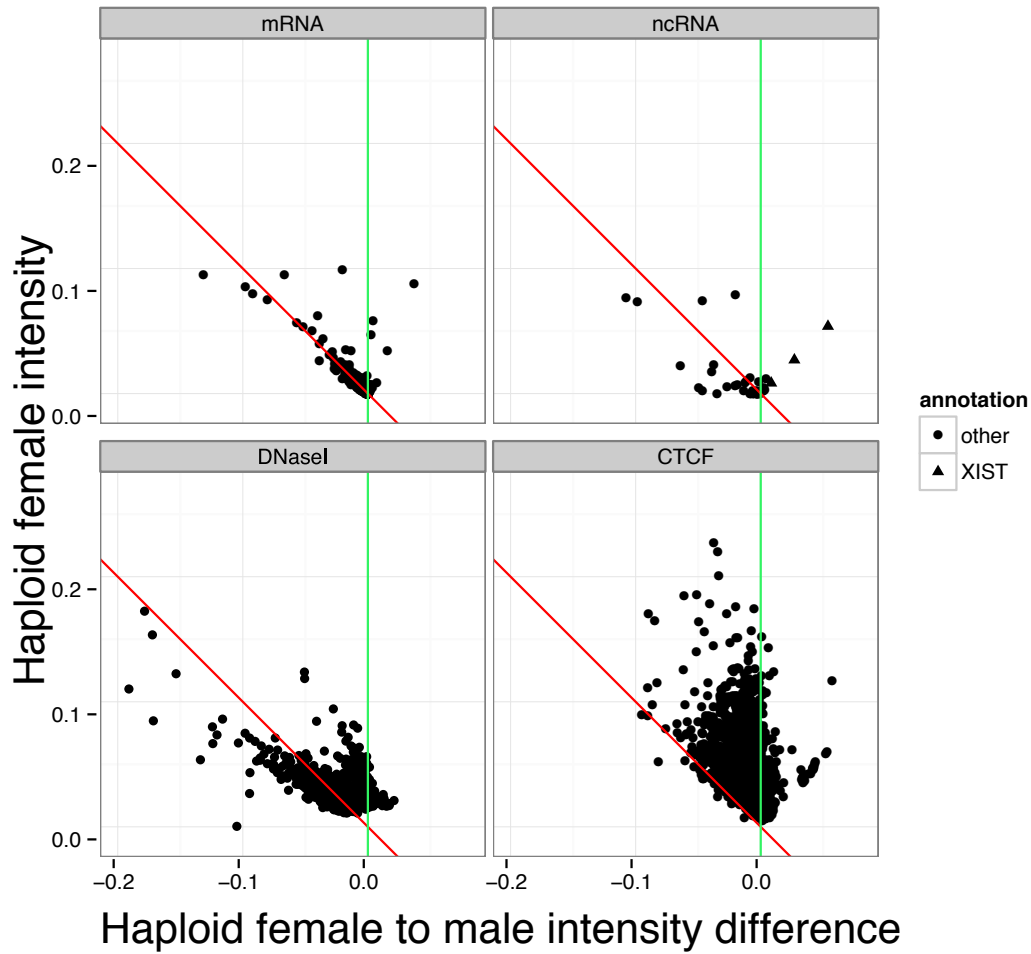


Figure 5.8: Trends in signal for mRNA, ncRNA, DNase I and CTCF on chromosome X. The signal difference between the expected level of signal when both X chromosomes are active and when one is inactivated is plotted. The Y-axis shows the average female signal divided by 2, which would be the normal dosage-compensated signal. On the X-axis, I have the same average female signal divided by 2, but then I subtract the average male signal; this gives an indication of the signal being biased towards males or females. In the ncRNA panel, I assigned the Xist transcripts a different symbol. The red line indicates the trend for dosage-compensated sites. The green line indicates the trend for both-active sites.

5.2.5.2 *Overlap of CTCF classes with DNase I*

I observed that DNase I has more dosage-compensated regions than CTCF. However, there are DNase I sites that act similarly to CTCF in being active on both X chromosomes in females. The most confident of 5% DNase I sites on the X chromosome were compared for the haploid female DNase I intensity against the haploid female to male DNase I intensity difference (Figure 5.9A, same as in Figure 5.8). I then coloured by the classification of the CTCF when there is an overlap. The CTCF classification is a good prediction for the DNase I intensity, because the sites that overlap show similar behaviour.

The DNase I sites with a CTCF classification are compared by the average DNase I intensity for males and females for both-active and dosage-compensated classification (Figure 5.9B). The dosage-compensated DNase I sites have similar average intensities for both male and female (Student's t-test p-value > 0.05). For the DNase I sites that overlap the both-active CTCF-classified sites, females have double the amount of DNase I intensity than males and are significantly different (Student's t-test p-value < 2.3×10^{-8}). This indicates that these DNase I sites have a similar behaviour as the overlapping CTCF sites.

These results show that the CTCF classification predicts the behaviour of DNase I sites. This is despite the fact that these are different cell lines, from a different population, grown in a different laboratory, and analysed using a very different assay. Interestingly, the female-specific CTCF sites have no overlapping DNase I sites. This could indicate that the female-specific sites identified are, most likely, more closely related to CTCF biology than chromosome structure. There are, however, some other DNase I sites that show a slight tendency to be female-specific, but this trend is not as strong as the trend observed in mRNA, ncRNA or CTCF.

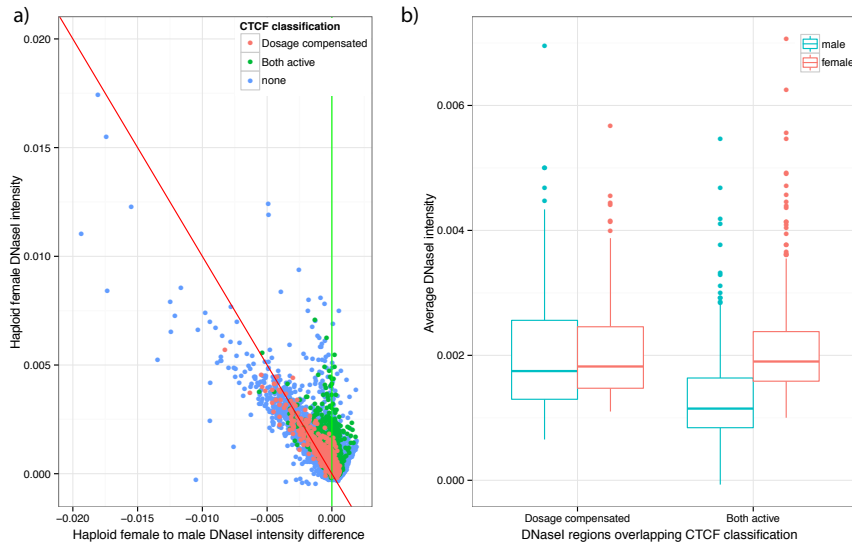


Figure 5.9: The CTCF classification is imposed on the most confident 5% DNase I sites on the X chromosome. a) Haploid female DNase I intensity is compared against the haploid female-to-male DNase I intensity difference. For DNase I sites overlapping CTCF sites, the CTCF classification is used for colouring. The CTCF classification is able to predict most of the DNase I effect for being dosage-compensated (red line) or both-active (green line). b) The DNase I sites overlapping the CTCF sites are compared on the CTCF classification. For the dosage-compensated sites, the average male intensity is similar to the average female intensity (Student's t-test p -value > 0.05). In contrast, for the both-active sites, the average female intensity is significantly twice as high as the average male intensity (Student's t-test p -value $< 2.3 \times 10^{-8}$).

5.2.6 *X-chromosome inactivation centre*

The X-chromosome inactivation centre (XIC) plays a central role in the X-chromosome inactivation pathway. In this region is the *Xist* gene that is transcribed into the *Xist* ncRNA, as well as the antisense *Tsix* ncRNA. *Xist* is the major executor in X chromosome inactivation; it coats and silences the X chromosome. The average *Xist* transcript intensities, for each cell line, are compared between male and female LCLs (Figure 5.10). There is a clear, and expected, difference between females and males for *Xist*, where obviously the female lines show much higher *Xist* transcript levels (Student's t-test p-value $< 1.3 \times 10^{-8}$). Although this result is not novel, it is a positive control of our transcript dataset, which ensures that, in the processing and normalisation pipeline, gender-specific signals like *Xist* are retained and not removed as being classified as an artefact.

A more detailed investigation of the XIC region was conducted to study the signal trends of both CTCF and DNase I sites. The RNA transcript abundance data shows a gender-specific signal. In close proximity to the XIC region, I found three CTCF sites and two DNase I sites. Previous studies have suggested that CTCF, together with the TF YY1, are essential for activation or silencing of the *Xist* transcript (Donohoe et al., 2007; Jeon and Lee, 2011; Moseley et al., 2012; Thorvaldsen et al., 2011). Both the CTCF and the YY1 GM12878 ENCODE signal tracks were selected (GM12878 is a female cell line) (ENCODE Project Consortium, 2012). To minimise lab-specific bias, different tracks from different labs were selected (Figure 5.11). Interestingly, the YY1 track peak does not overlap any CTCF site, but overlaps two DNase I sites (green box). The DNase I sites overlapping the YY1 binding are female-specific. The three CTCF sites in the XIC show no evidence of being gender-specific (Figure 5.12, Student's t-test p-value > 0.05), suggesting that the CTCF sites that are used in this study are not involved with maintaining X chromosome inactivation. In contrast, both DNase I sites overlap

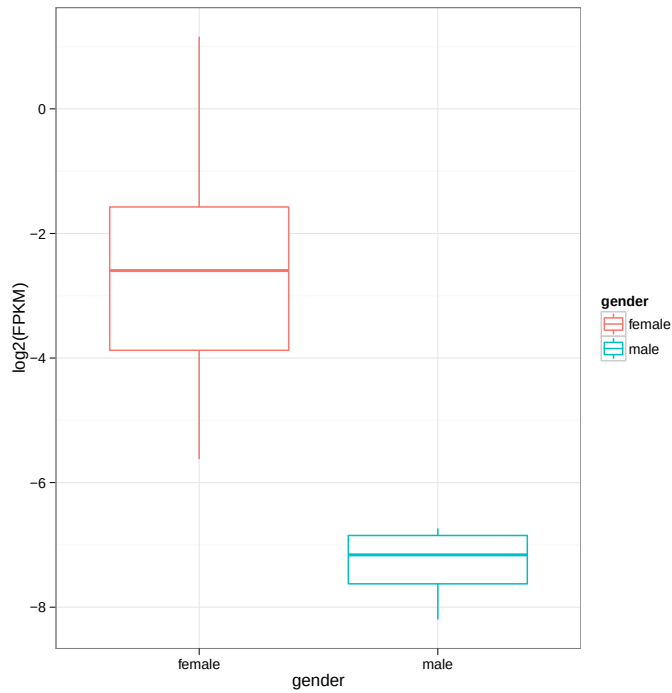


Figure 5.10: Xist average transcript abundance for male and female cell lines. The female cell lines have significantly higher transcript abundance for Xist (Student's t-test p-value < 1.3×10^{-8}).

with YY1 and are gender-specific (MWU p-value < 0.05). This observation suggests that the XIC region, and in particular this region in proximity to the Xist transcript, is being regulated on a chromatin accessibility basis, or by another factor such as YY1, but most likely not by CTCF.

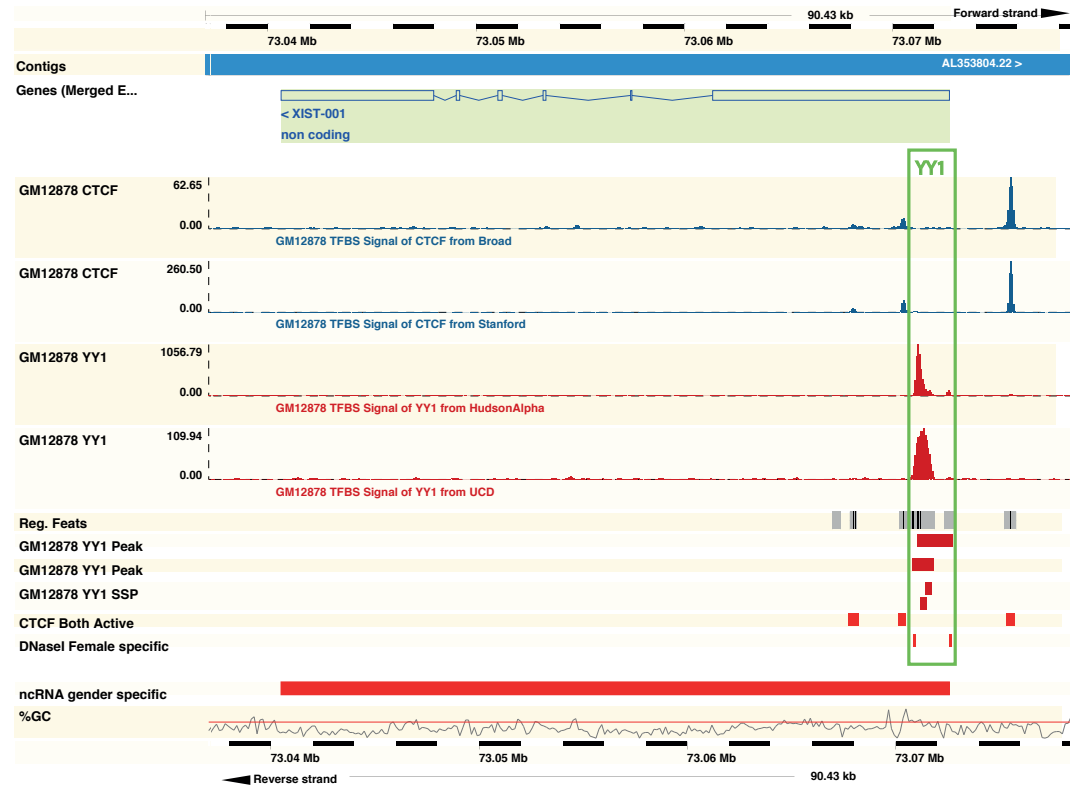


Figure 5.11: Screenshot Ensembl (Flicek et al., 2012) of the XIC region. Classification of our CTCF sites is shown (only "both-active" in current view), classification of mRNA and ncRNA (both gender-specific, MWU p-value < 0.05) and classification of DNase I sites (female-specific, MWU p-value < 0.05 and no male signal). From the ENCODE consortium, the CTCF and YY1 signal track is shown. The green box is drawn to illustrate the overlap between YY1 and DNase I sites.

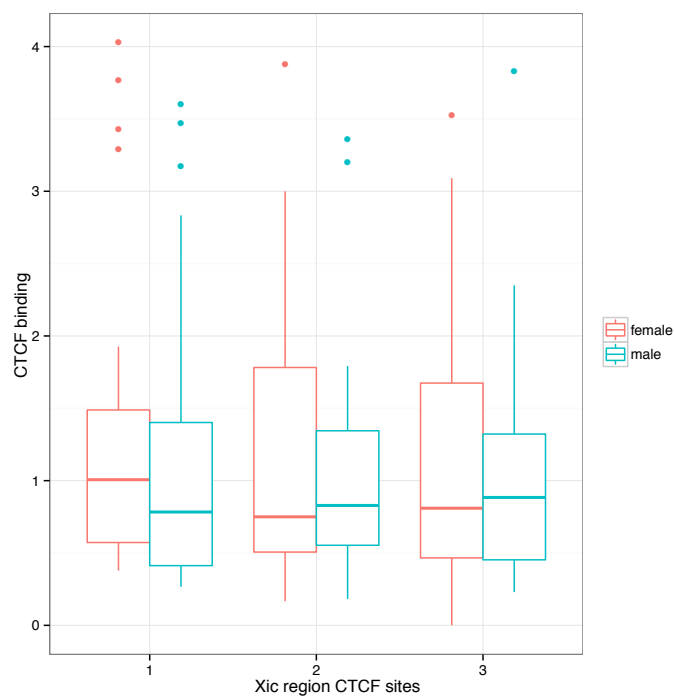


Figure 5.12: CTCF sites in the XIC region nearby the Xist transcript binding affinity are shown. For each site, the male-versus-female binding is portrayed in box plots. None of the CTCF indicates a gender-specific bias (Student's t-test $p\text{-value} > 0.05$).

5.2.7 Female specific CTCF sites

There are 23 CTCF sites with a distinctive female-only binding affinity on the X chromosome. Interestingly, these sites do not overlap XIST nor the XIC region, while such overlap might have been expected. These 23 sites cover two different X chromosome regions. Each of these two regions exactly overlaps a different ncRNA (X56 and X130) that is without extensive annotation. Both ncRNAs have been shown in previous studies to be altered in X-chromosome inactivation (Horakova et al., 2012). It has been shown that CTCF sites that overlap with YY1 are involved in X-chromosome inactivation (Donohoe et al., 2007; Jeon and Lee, 2011; Moseley et al., 2012; Thorvaldsen et al., 2011). However, when inspecting the ENCODE Consortium YY1 peaks called for GM12878, all of the female-specific sites are enriched for YY1 (Section 5.2.4 , Figure 5.5). In addition, strong enrichment for histone marks, H2AFZ, H3K27ac, H3K4me2 and H3K4me3 are present in all sites and H3K9ac in 80% of the sites. Another TF that is present in all sites is POU2F2, although there is no obvious hypothesis to explain why this is the case. Adding that all of these sites are conserved regions in the genome according to the GERP data from Ensembl, the most likely conclusion on functionality for these CTCF sites is that they are, most likely, vital for the regulation of these two ncRNAs.

In previous studies, in female cells, both of these ncRNAs were expressed on the Xa. To study the transcript behaviour of these 2 ncRNAs more comprehensively, I asked our collaborators at the Wellcome Trust Sanger Institute to perform fluorescence *in situ* hybridisation (FISH) for these ncRNAs in male and female cells. Furthermore, FISH was conducted on the XIST transcript as a positive control. Results from the RNA-FISH experiment show that for each of these 2 ncRNAs, the signal is mainly female-specific and originates from the Xa (Figure 5.13). Interestingly, the male cells also show a signal for both X56 and X130, although with a lower intensity than the female nuclei.

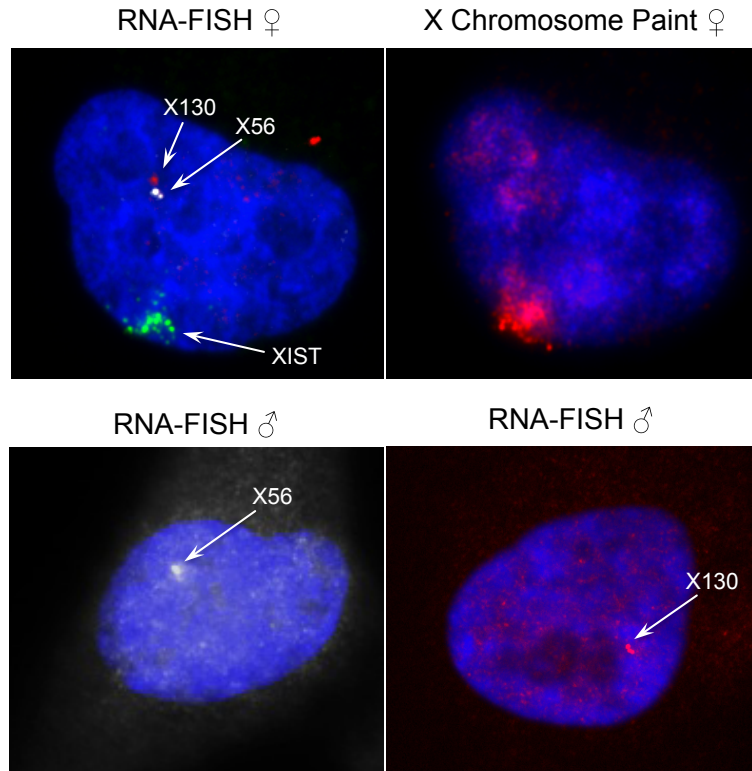


Figure 5.13: RNA-FISH results for female-specific ncRNAs X56, X130 and XIST. A) Representative RNA-FISH image of X56 (white signal) and X130 (red signal) expression relative to XIST RNA (green signal) in female nuclei (counterstained with DAPI) and sequential DNA-FISH representative image with X chromosome paint probe (red signal). The staining shows that X56 and X130 are transcribed from the active X chromosome nuclear territory. B) Representative images of X56 (white signal) and X130 (red signal) expression by RNA-FISH in male nuclei (counterstained with DAPI).

5.3 CONCLUSION

The transcription factor CTCF is known for its multifunctionality. In Chapter 3 I described that CTCF sites correlate with each other, suggesting biochemical binding between these sites. In this chapter, I show that CTCF has a strong involvement in X-specific behaviour. Notably, the CTCF behaviour found in this chapter is not related to Xist or the XIC region.

CTCF sites on the X chromosome tend to have three different behaviours. First, there are the dosage-compensated CTCF sites that behave similarly to mRNA. Second, there are both-active CTCF sites that have double the amount of binding as dosage-compensated sites. Third, there are CTCF sites that are only bound in female lines. By using this distinction, I classified the X chromosome CTCF sites into 1,692 sites that are both-active, 253 sites that are dosage-compensated, and 23 sites that are female-specific.

A reasonable hypothesis is that these distinctively different behaviours of CTCF are due to differences in function. For the dosage-compensated CTCF sites, I expect that these are normal regulatory CTCF sites. The both-active sites I expect to be involved in enabling the 3D structure of the X chromosomes. The chromosome structure is important for both Xi and Xa, hence these sites are bound on both the Xi and the Xa. The female-only CTCF sites are most likely involved in X-chromosome inactivation processes. This study presents a simple and robust method to identify the difference between regulatory (dosage-compensated), structural (both-active) sites and X-chromosome inactivation related (female-specific) sites.

The LCL GM12878 has been studied in depth within ENCODE and was used to determine co-binding of the CTCF sites. Each class of CTCF sites shows a unique pattern in the factors that are co-binding and the histone modifications that overlap. However, GM12878 is female, and for this reason it is impossible to indicate for each of the found overlaps if these are gender-

biased. Ideally, at least one male LCL would have been available to study in similar depth as GM12878. In this way, for each of the overlaps, a rough ratio between binding in male and female could be determined. This ratio could provide an indication of how these overlapping factors are gender-biased and, potentially, similar to CTCF involved in X chromosome inactivation. Aggregation plots of the histone marks H3k4me3, H3k4me1 and H3k27ac show clear distinctive differences between both-active and dosage-compensated sites. H3k27me3, which is known to be involved in X-chromosome inactivation, appears to be indifferent between both classes, and the signal intensity seems to be only that of a background level.

It will be interesting to determine in the near future, when additional similar projects are published, how other molecular assays behave on the X chromosome. There are, most likely, several other transcription factors and histone modifications that are involved with the X-chromosome inactivation process. The identification of male- or female-specific sites is of great interest to enhance our understanding of X chromosome biology and the different functional roles that sites can have. Combining the knowledge of other assays performed in a population could help to create better distinctive classifications of CTCF sites. This would help to identify the actual functional difference between each of the classes. In particular, population data for the TFs and HMs, which were shown to be significantly different in overlap in the GM12878 cell line, are of great interest for further X chromosome studies. It is my expectation that there are clear, distinctively different roles for CTCF sites, where some are involved in the chromosome structure and some are more regulatory.

While our study has been performed on human cell lines, there is no reason to believe that the same methods cannot be applied on other mammals with similar X-chromosome inactivation. The human CTCF sites are highly conserved, and it would be interesting to see how CTCF binding in mice would behave,

as well as whether homologous sites have similar CTCF binding differences between males and females.

The two identified female-specific ncRNAs have homologues in mice. To understand the exact role of these two transcripts, different knockout mice could be created. Using these knockout mice, the functional role of these two ncRNAs could be studied. Although Xist is known to be the most important ncRNA in the X-chromosome inactivation process, these two transcripts could be involved with other essential biochemical mechanisms.

5.4 DATA

5.4.1 CTCF

Refer to Chapter 3.4.1. Differently than in previously chapters, the CTCF data in this chapter is utilised without removing the first PC from the PCA analysis. Consequently, no significant gender bias is removed by this normalisation step.

5.4.2 DNase I

I obtained DNase I sequencing data for 70 individuals from the YRI HapMap population (Degner et al., 2012). To ensure that none of our results are biased due to normalisation, the author of the study kindly provided us with GC-corrected DNase I sensitivities for all of the sites they call. By using data that has not been normalised, any significant gender bias that may be present remains intact.

5.4.3 RNA-seq

I obtained RNA-sequencing data for 51 individuals from the CEU HapMap population from the Geuvadis consortium (Lap-

[palainen et al., 2013](#)). The dataset consists of FPKM values for all mRNAs and ncRNAs.

5.5 METHODS

5.5.1 *Classification of CTCF sites*

To ensure that our normalisation would not introduce any bias, I used the raw CTCF binding intensities. Our dataset consists of 27 females and 24 males. After blacklisted regions were removed, there were 1,968 CTCF binding regions on the X chromosome.

For each CTCF site, I assessed gender specificity for the binding region by a Mann-Whitney U (MWU) test between the male and female samples. The MWU test evaluates for significant difference between the medians of both groups (male versus female). A significant result in the MWU test ($p\text{-value} < 0.01$) results in a CTCF site being classified as gender-biased, i.e. there is significant disparity in binding between males and females. The sites that are not gender-biased (non-significant in the MWU test) are termed "dosage-compensated".

In a next step, for each gender-biased CTCF site, the ratio between the average CTCF binding in males and females is calculated. If this ratio indicates a five-fold or greater difference in binding in favour of the female cell lines, they are classified as "female-specific". The remaining gender-biased sites are classified as "both-active". There were no sites that were binding in favour of male; otherwise, those would have been classified as "male-specific".

For the mRNA, ncRNA, and DNase I data, the same MWU analysis was performed. These classifications were used to indicate the gender bias within the Ensembl browser.

5.5.2 *Biochemical co-activity with CTCF sites*

For the dosage-compensated and both-active sites, I analysed the overlap of each category of sites with the ENCODE TFs and histone modifications datasets for the female CEU LCL GM12878. Consequently, for each CTCF site, I was able to determine which other biochemical binding takes place. As a definition of an overlap, I consider a partial overlap an overlap (1bp minimal).

The Chi-square test was used to identify TFs or histone modifications that are significantly more or less overlapping with each of the CTCF classifications. Arbitrary p-value thresholds were set to identify the "level" of significance of disparity of these overlaps.

5.5.3 *Aggregation plots*

For the dosage-compensated and both-active sites, I analysed the overlap of each category of sites with the ENCODE TF and histone modification datasets for the female CEU LCL GM12878. To avoid bias introduced by unequal distribution of promoter sequences between the classes, I removed all sites that overlap with promoters identified in GM12878.

Signal aggregation of each of the classes of sites for histone modification and TF ChIP-seq data, and micrococcal nuclease cleavage, was calculated using the ACT toolkit (<http://act.gersteinlab.org/>, (Jee et al. 2011)) with the parameters nbins=50 mbins=0. Only CTCF sites that are in the 50% of bound sites were used in this aggregation. ENCODE bedGraph files for both TF and histone modifications were obtained from ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/signal/jan2011/bedgraph/ and converted into signal files that are used as input for ACT.py.

5.5.4 *Data visualisation*

Within R ([Team, R Development Core, 2005](#)), the package Ggplot2 ([Wickham, 2009](#)) was used to visualise all data.

Part V

CONCLUSIONS

FINAL REMARKS AND FUTURE DIRECTIONS

For this thesis, three research projects were designed and conducted. The first project entailed discovering molecular Quantitative Trait Loci (QTLs): in this case, chromatin status QTLs using Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)-chip data. The second project involved integrating molecular assays and discovering complex interactions using a systems genetics approach. The third project studied the unique behaviour of CTCF on the X chromosome and its involvement in the inactivation of the X chromosome.

6.1 CHROMATIN AS A HERITABLE TRAIT

For the first project, a novel dataset was generated to measure chromatin accessibility with FAIRE in 60 humans. The FAIRE signal was quantified using an in-house designed tiling array targeting 10,507 known regulatory genome regions. I have demonstrated that it is feasible to use FAIRE to identify QTLs. Considering that I used a relatively small population, used only HapMap tagging SNPs, and used tiling arrays, this project proved to be quite successful with the mapping of 167 FAIRE-QTLs when using probesets. In addition, when using each independent FAIRE-probe, a total of 16,972 FAIRE-QTLs were discovered. The FAIRE-QTLs tended to be proximal to the measured FAIRE signal and were not restricted to the promoters of genes. Several of our FAIRE-QTLs were also previously identified as eQTLs; this indicates that, in some cases, the chromatin is directly correlated with gene expression. Also, using our current samples, the FAIRE-QTLs show a higher association with autoimmune-related diseases than other genome-wide association studies

(GWAS) phenotypes such as height. These overlaps may augment our understanding of the molecular effect of GWAS SNPs for complex diseases.

The main limitation to discovery of FAIRE-QTLs at this stage is the small sample size. If a larger group of individuals was measured using the FAIRE-chip, or even more preferably, using FAIRE-seq, a more comprehensive collection of FAIRE-QTLs could be discovered. Also, the introduction of a different background population could help to distinguish between population-specific FAIRE-QTLs and the more universal FAIRE-QTLs. Also, overlap with GWAS catalogues indicates that having genome-wide FAIRE data, or any other chromatin assay, for some immune-related diseases might be of great interest.

I am able to identify a reasonable number of FAIRE-QTLs using microarrays. However, using next-generation sequencing (NGS) instead of microarrays would most certainly help to discover a larger set of FAIRE-QTLs, due to its genome-wide coverage and lower readout noise. As FAIRE assays how accessible chromatin is, it will quite often discover QTLs that are only circumstantial for FAIRE. The real biochemical reason for the discovery of these QTLs is due to the transcription factor (TF) binding or histone modifications (HM) that are actually controlling the accessibility of chromatin.

6.2 SYSTEM GENETICS

When partially reconstructing the epigenomes of a small population using only FAIRE, CTCF, and gene expression, a high degree of correlation between the different assays was observed. This in itself indicates how complex and intertwined our epigenome is. There are thousands of significant correlations between each of the molecular phenotypes, and when all of these correlations are integrated, the results suggest that there is a particular preferred way to regulate gene transcription. As the actual direction of causality cannot be determined by correlations, con-

clusions drawn from correlations are susceptible to different interpretation. The CTCF-to-FAIRE correlations suggest that CTCF binding increases the FAIRE signal, which indicates that the chromatin is in a more accessible state. Both the FAIRE-to-gene expression and the CTCF-to-gene expression correlations suggest that an increase in either FAIRE or CTCF is increasing gene expression.

By correlating CTCF sites to each other, I discovered many potential CTCF interactions. These CTCF interactions can be involved in insulating essential regulating features through looping interactions. In addition, there appears to be a significant difference in biochemical activity depending on the length of the loops, which indicates a divergence in functionality. However, a clear functional reason for this disparity in biochemical activity is not clear. One potential reason might be that these longer CTCF interactions are different CTCF sites that are pulled together into transcription factories.

As each of the molecular phenotypes can have a biochemical effect on other phenotypes, determining the direct and indirect causal effects can be difficult (e.g. is CTCF changing FAIRE, or is FAIRE changing CTCF?). However, by incorporating genetic information, i.e. QTLs, a starting point for causality is introduced. Knowing what is the direct or the indirect effect of a QTL will help to unravel molecular regulatory mechanisms. QTLs can drive epigenetic changes that, as a result, can change gene expression; however, obviously, this causality cannot occur the other way around (genotypes are fixed and cannot be changed). With the incorporation of QTLs, it is possible to dissect these strong epigenetic correlations and determine what is the direct causal effect. I discovered that QTLs are the main drivers of CTCF binding. For example, CTCF affects FAIRE or RNA, and not the other way around. The CTCF binding is the strongest indication of transcription and the accessibility of chromatin. In conclusion, I have established the distinction between in-

direct and direct causality for many phenotype-to-phenotype correlations using QTLs.

For CTCF-CTCF pairs that are under the influence of a CTCF-QTL, it was discovered that the genetic control often adjusts only one of the CTCF sites, rather than affecting both equally. Furthermore, when the QTL effect is removed for CTCF-CTCF pairs, the correlation can be removed or sometimes even reversed.

Another main conclusion from this chapter is that eQTLs are quite often proxy-QTLs. This means that, although a QTL can initially be identified as an eQTL, this change in expression is due to more upstream biochemical processes. In essence, a molecular QTL most likely indirectly affects the gene expression through a regulatory feature, and thus is more likely to be a CTCF-QTL, FAIRE-QTL, or any other molecular-QTL. These results correspond to findings in Yoruba in Ibadan, Nigeria (YRI) LCLs, where it is shown how DNase I and gene expression QTLs commonly overlap and are sometimes located within a transcription factor binding region (McVicker et al., 2013). Complementary to my findings, there are studies showing how TF binding and histone modifications are the main determinants for gene expression (Kilpinen et al., 2013). This result is a strong indication of how, irrespective of the molecular assay of QTL discovery, CTCF most likely changes the chromatin accessibility and thus gene expression. For example, CTCF may change chromatin into an accessible state so that other transcription factors can bind and transcription can occur. However, it is too early to reach a final conclusion about this topic. There are examples in which the epigenetic state is causal for expression (Lam et al., 2008) as well as for the result of changes in gene expression (Zhang et al., 2014). Thus, although my results suggest that CTCF is a driver of gene expression and chromatin status, some additional experiments are desirable. It would be of great interest to determine the effect of alternating CTCF binding in the discovered phenotype-to-phenotype pairs that are driven by CTCF.

I show that the independent mapping of molecular QTLs for each molecular phenotype is not comprehensive enough to fully grasp the genetic effects. There is a high level of overlap, and thus understanding the real direct causal effect of a QTL is not easy. This is especially true when considering that some TFs might have both proximal and distal effects (Sanyal et al., 2012). Because of this, it is of great importance to understand what a molecular QTL actually affects on a molecular level. This would be particularly useful if the results are used to improve the annotation of disease GWAS results (He et al., 2013). Although software tools such as Sherlock can successfully map a molecular QTL to a disease GWAS, if the true molecular effect of the QTL is unknown, these results might not improve the understanding of the disease and its pathways. Understanding the molecular mechanisms regulated by QTLs will provide assistance to identify new hypotheses for potential drug targets. This will add value to and use personal genomes in everyday medicine.

6.3 CTCF BINDING ON THE X CHROMOSOME

CTCF is known to be a multifunctional protein that is involved in a wide variety of tasks for both regulatory and chromosomal structural functions. On the X chromosome, CTCF is more than just a regulator that is involved with the X-chromosome inactivation center (XIC). The population data suggest that CTCF binding can be classified into three distinct classes. Most CTCF sites are bound (or active) on both the inactive X (Xi) as well the active X (Xa) chromosome. Another large subset of CTCF sites is dosage-compensated (bound on Xi or, most likely, Xa). A minority of CTCF sites are only bound in females. I show that each of these classifications could be associated with different functional roles.

In the present analysis, the data suggest that the CTCF sites that are active on both chromosomes are most likely involved

in a chromosomal structural role. On the other hand, CTCF sites that are only bound on one chromosome show evidence of a more regulatory role. Most certainly, the regulatory CTCF sites are only bound on the Xa chromosome. The female-specific CTCF sites regulate the expression of two different ncRNAs (X56 and X130) and are most likely involved in, or controlled by, X-chromosome inactivation processes. RNA-fluorescence *in situ* hybridization (FISH) experiments on these two ncRNAs show, similar to Xist, expression from mainly the inactivated X chromosome. All of these results suggest yet another major biological role for CTCF.

Using the CTCF site classification, I was able to predict the corresponding DNase I signal for overlapping genomic regions. In the current limited view, it appears that CTCF is a stronger indicator for both active sites than DNase I is. However, by increasing the number of molecular assays performed on my population of cell lines, a more comprehensive epigenomic view of the X chromosome could be attained. There are, without doubt, other molecular phenotypes that are, similarly to CTCF, essential for the X chromosome. In particular, now that the different behaviours of CTCF on the X chromosome have been revealed, it would be interesting to determine whether this result can be used to train a model to predict whether CTCF sites on other chromosomes are regulatory or structural. Using overlap with transcription factors and histone modifications from ENCODE, a model could be trained for this purpose. However, the ENCODE cell line GM12878 is a female line, and hence it is difficult to predict how the overlap would look for a male cell line.

6.4 INTEGRATION OF EPIGENETIC DATA

In this thesis, I was successful in discovering a large number of correlations between different epigenetic traits. I also showed that when genetic information is added, a better understanding

of the causality cascade in epigenetic trait correlations can be achieved. However, this study would have benefited considerably from a larger population size, as the discovery of causal loci is mainly limited by sample size (Spencer et al., 2009). With an increased population size, QTLs that are further away could also be included for QTL mapping. For the projects described in this thesis, I focussed only on coding genes; however, this excludes important ncRNAs that are known to have regulatory function (Ambros, 2001). Including those genes would improve the overall quality of the created epigenome, as it is almost certain that ncRNAs will influence the molecular phenotypes that were included in the epigenome. Also, obviously, the expansion to more epigenetic phenotypes would help to construct a better epigenome that, in turn, could be used to improve the annotation of disease GWAS results.

If the number of correlations observed in this thesis is a good representation of the kinds of biochemical interactions that occur in healthy individuals, it would be of great interest to create the same correlations for disease samples and to compare the results. Differential interactions analysis could help to distinguish between housekeeping and disease-specific biochemical interactions. A great example of how an epigenetic phenotype can help to explain a disease GWAS result can be found for Type 2 diabetes (T2D). For this disease, a region of islet-selective accessible chromatin overlaps a SNP that is strongly associated with T2D; this links genetics, epigenetics and environment together (Drong et al., 2012; Gaulton et al., 2010; McCarthy, 2010)).

6.4.1 *Single molecular level*

To date, nearly all studies, including those in my thesis, have been performed using a mixture of cells, because for the assays to work, a large quantity of DNA is required. Furthermore, it is nearly impossible to ensure that cells have equivalent epigenetic state and genetic background. There has been some success

with single-nucleus sequencing for breast cancer, indicating how diverse the genetics within a tumour can be (Navin et al., 2011). Obviously, single-cell sequencing is also applicable for the transcriptome (Yan et al., 2013) and other epigenetic marks. These single-cell studies have shown that the transcriptome profile is unique for cell state and cell type (Jaitin et al., 2014; Ramsköld et al., 2012; Shalek et al., 2013). The success of single-cell sequencing for the genome and transcriptome indicates that there is need to expand these methods towards regulatory features. As I have shown, using a population of cells (even from a cell line) necessitated the removal of highly-variable FAIRE probes from my experiments. Most likely, this variance is an indication of an epigenetic differentiation at the specific regulatory feature that could be studied using single-cell approaches.

New methods such as Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) reduce the amount of input required, from approximately two million cells to less than 50,000 cells (Buenrostro et al., 2013). These reductions will yield more robust results, as there will be less internal variance due to the great number of different cells. Furthermore, this lower number of required cells means that ATAC-seq could be used to identify regulatory regions in disease samples or can be used in a diagnostic setting. Moreover, if the quantity of input material could be lowered even further, the assessing the accessibility of chromatin could be used on a single-cell level, which would greatly improve our understanding of the regulation of the epigenome. Early studies have been able to explain how DNA methylation drives embryonic differences (Lorthongpanich et al., 2013). And the first single-cell DNA methylation methods are capable of measuring 10M CpGs in mice, which is nearly half of the CpGs sites known (Smallwood et al., 2014).

Studying the direct interaction between TF binding and chromatin is useful for understanding their relationship. Methods like DNase I-released fragment-length analysis of hypersensitivity (DNase-FLASH) show promising results for directly map-

ping the interaction between TF binding and the accessibility of the chromatin (Vierstra et al., 2014). By comparing the length of the DNase I-released fragments, an indication can be given about DNA fragments being located in accessible chromatin (~200kb fragments length) or being flanked by the nucleosome (~500kb fragments length).

Understanding the chromatin landscape is essential for unravelling the human epigenome. However, all of the current methods have their limitations, and even when assaying naked DNA, differences in chromatin will be measured, due to unknown digesting preferences from DNase I rather than real epigenetic changes (He et al., 2014). These disparities obviously mean that some of the identified regulatory features will need to be reevaluated.

The additional information of the epigenome on a single-cell level would help to address the variance of biochemical interactions. Even more, these data would be powerful when integrated into a more complex structural equation modeling (SEM) system. Attaining a clear overview of the normal, healthy epigenome is a starting point for exploring environmental effects or diseases.

6.5 THE EPIGENOME ERA

In recent years, several large-scale epigenome projects have been initiated. For example, the NIH Roadmap Epigenomics Mapping Consortium was launched with the goal of producing a public resource of human epigenomic data to catalyse basic biology and disease-oriented research (Bernstein et al., 2010; Chadwick, 2012). Other projects have more specific goals: for example, the BLUEPRINT consortium is interested in establishing a full epigenome view of haematopoietic cells (Martens and Stunnenberg, 2013), and the International Cancer Epigenome Project is, broadly speaking, interested in mapping cancer epigenomes (Beck et al., 2012). These consortia have in common

that they use a mixture of healthy cell lines and disease cell lines. Potentially more importantly, LCLs are slowly being replaced in research projects by induced pluripotent stem (IPS) cells (Takahashi and Yamanaka, 2006). IPS has the great benefit, in comparison with the lymphoblastoid cell lines, that any cell type can be used to create an IPS cell line. This means that, for studying whole-body phenotypes in the current research environment, a more fitting cell type (or several) can be used instead.

Evidently, these kinds of improvements in cell line technique and large-scale consortia will increase our understanding of the human epigenome. However, many consortia focus mainly on disease samples or a wide range of different cell types, and thus their findings will not be as generally applicable as the results presented in this thesis. In order to improve our epigenome view, it will be essential to have studies that are not focussed on comparing different cell lines with some replicates, but rather to emphasize studies of populations of a specific cell type. Because of this issue, integrating the data is difficult, as cell types and unknown population structure will drive many correlations and will make it nearly impossible to determine what is causing the correlation. These consortia, using a mixture of cell types, will be more applicable to extending our knowledge; for example, I used ENCODE in this thesis to confirm my hypothesis and to provide an indication of what might be occurring in particular regions of the genome.

One important piece of future work will be the extensive exploration of inter-individual and intra-individual (e.g. between cell types or sick and healthy) differences in epigenomes. These studies would elucidate the, so-far, poorly characterised genetic regulation of the whole epigenome. Furthermore, increased understanding of the epigenome will be applicable within the field of personal medicine.

Part VI

APPENDIX

INDIVIDUALS USED IN THIS THESIS

A.O.1 *Overlap assays*

The CEU population used in this thesis has been assayed for a variety of molecular assays. However, there is some slight disparity for what molecular assay has been performed for each individual (Table A.1).

Table A.1: Overview individuals and the molecular assays performed on them

Individual	RNA	FAIRE	CTCF
NA06986	yes	yes	yes
NA07037	yes	yes	yes
NA07051	yes	yes	yes
NA07346	yes	yes	yes
NA07357	yes	yes	yes
NA11829	yes	yes	yes
NA11830	yes	yes	yes
NA11831	yes	yes	yes
NA11832	yes	yes	yes
NA11840	yes	yes	yes
NA11881	yes	yes	yes
NA11894	yes	yes	yes
NA11918	yes	yes	yes
NA11920	yes	yes	yes
NA11931	yes	yes	yes

Continued on next
page

Table A.1: Molecular assays performed on individuals

Individual	RNA	FAIRE	CTCF
NA11992	yes	yes	yes
NA11993	yes	yes	yes
NA11994	yes	yes	yes
NA11995	yes	yes	yes
NA12003	yes	yes	no
NA12004	yes	no	yes
NA12005	yes	yes	yes
NA12006	yes	yes	yes
NA12043	yes	yes	yes
NA12045	yes	yes	yes
NA12144	yes	yes	yes
NA12154	yes	yes	yes
NA12155	yes	yes	yes
NA12156	yes	yes	yes
NA12234	yes	yes	yes
NA12249	yes	yes	yes
NA12287	yes	yes	yes
NA12489	yes	yes	yes
NA12749	yes	yes	yes
NA12750	yes	yes	yes
NA12751	yes	yes	yes
NA12760	yes	yes	yes
NA12761	yes	yes	yes
NA12762	yes	yes	yes
NA12763	yes	yes	yes
NA12776	yes	yes	yes
NA12812	yes	yes	yes

Continued on next
page

Table A.1: Molecular assays performed on individuals

Individual	RNA	FAIRE	CTCF
NA12813	yes	yes	yes
NA12814	yes	yes	yes
NA12815	yes	yes	yes
NA12828	yes	yes	no
NA12872	yes	yes	yes
NA12873	yes	yes	yes
NA12874	yes	yes	yes
NA12891	no	yes	no
NA12892	no	yes	no
NA11919	yes	no	no
NA12044	yes	no	yes
NA12414	yes	no	no
NA10847	yes	no	yes
NA06985	yes	no	yes
NA06994	yes	yes	yes
NA12717	yes	no	yes
NA07000	yes	no	yes
NA12716	yes	no	yes
NA07347	yes	no	yes
NA10851	yes	no	yes

OVERVIEW ALL FAIRE-QTLS

Table B.1: Overview FAIRE-QTLs. The FAIRE target, FAIRE target genomic location, associated SNP, Spearman's Rank p-value ($-\log_{10}$) and Rho, and empirical p-value are given. The Spearman's Rank result for when using HapMap genotypes and 1,000 Genomes (1kg) genotypes are given.

FAIRE-peak	FAIRE-peak tion	posi-	SNP	SNP posi-	distance	Rho	p-value	empirical p-value	p-value 1kg
target_6940	1:172631563-172632321		rs12060864	173064777	432456	-0.54	4.82	0.29	2.60
target_6940	1:172631563-172632321		rs12075195	173065003	432682	0.54	4.82	0.29	2.60
target_10251	2:127829389-127830505		rs12475915	127902187	71682	-0.58	5.49	0.00	4.34
target_10251	2:127829389-127830505		rs2118508	127909259	78754	0.58	5.43	0.00	4.34

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak tion	posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_10251	2:127829389-127830505		rs3943703	127897320	66815	0.56	5.21	0.00	4.11
target_10251	2:127829389-127830505		rs2118506	127897503	66998	-0.56	5.21	0.00	4.11
target_11148	2:120048442-120048900		rs1530562	120048729	0	0.54	4.83	0.33	4.76
target_11148	2:120048442-120048900		rs4848546	120051015	2115	0.54	4.83	0.33	4.76
target_11912	2:153045957-153046723		rs1806557	153056846	10123	0.54	4.74	0.50	3.89
target_3655	3:192616451-192617016		rs12485900	192623779	6763	0.58	5.48	0.17	4.06
target_3655	3:192616451-192617016		rs9869202	192625190	8174	0.58	5.48	0.17	4.06
target_3655	3:192616451-192617016		rs1948721	192629447	12431	-0.58	5.48	0.17	4.06
target_3655	3:192616451-192617016		rs4687368	192631810	14794	-0.55	4.91	0.38	2.95
target_3655	3:192616451-192617016		rs4687367	192631798	14782	0.54	4.81	0.50	2.99
target_8329	3:132377918-132379799		rs9854387	132862516	482717	0.56	5.12	0.00	4.85

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_8329	3:132377918-132379799	rs13079918	132866503	486704	-0.56	5.12	0.00	4.85
target_8841	3:108125294-108125916	rs936266	108125614	0	0.65	7.18	0.00	6.22
target_8841	3:108125294-108125916	rs4855651	108111663	-13631	0.56	5.17	0.45	4.21
target_8841	3:108125294-108125916	rs4855559	108115036	-10258	0.56	5.17	0.45	4.21
target_8841	3:108125294-108125916	rs4855650	108109031	-16263	-0.56	4.97	0.38	4.21
target_1980	4:118055550-118055749	rs1948223	118545387	489638	-0.58	5.54	0.50	5.30
target_10519	5:133223715-133224381	rs244693	133455153	230772	0.56	5.03	0.50	2.54
target_11459	5:18466219-18466561	rs7728989	18460758	-5461	-0.64	6.91	0.00	6.63
target_11459	5:18466219-18466561	rs10942236	18511790	45229	-0.64	6.91	0.00	6.63
target_11459	5:18466219-18466561	rs1527374	18459977	-6242	-0.63	6.29	0.00	6.63
target_11459	5:18466219-18466561	rs10042810	18482995	16434	-0.57	5.27	0.00	4.35

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak tion	posi-	SNP	SNP posi-	distance	Rho	p-value	empirical p-value	p-value 1kg
target_11459	5:18466219-18466561		rs2962949	18424199	-42020	0.53	4.56	0.47	4.34
target_8236	5:35920892-35921444		rs12657378	35608310	-312582	0.59	5.71	0.00	4.49
target_8236	5:35920892-35921444		rs7720593	35608763	-312129	0.59	5.71	0.00	4.49
target_8236	5:35920892-35921444		rs12514518	35614503	-306389	0.59	5.71	0.00	4.49
target_8236	5:35920892-35921444		rs12657936	35616027	-304865	-0.59	5.71	0.00	4.49
target_8236	5:35920892-35921444		rs11951251	35616805	-304087	0.59	5.71	0.00	4.49
target_8236	5:35920892-35921444		rs10512624	35585671	-335221	-0.59	5.70	0.00	4.47
target_8236	5:35920892-35921444		rs10050597	35585954	-334938	-0.59	5.70	0.00	4.47
target_8236	5:35920892-35921444		rs10057637	35586010	-334882	0.58	5.65	0.00	4.47
target_8236	5:35920892-35921444		rs1508650	35618486	-302406	0.54	4.83	0.17	3.72
target_8236	5:35920892-35921444		rs1508652	35618540	-302352	0.54	4.83	0.17	3.72

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_8236	5:35920892-35921444	rs1508653	35618552	-302340	0.54	4.83	0.17	3.72
target_8236	5:35920892-35921444	rs12652967	35596972	-323920	0.54	4.68	0.29	3.48
target_8236	5:35920892-35921444	rs4703415	35589089	-331803	0.53	4.58	0.32	3.36
target_8689	5:108017046-108017283	rs7737443	108016782	-264	-0.55	5.02	0.00	3.23
target_8689	5:108017046-108017283	rs7703883	108017262	0	0.55	5.02	0.00	3.23
target_8689	5:108017046-108017283	rs10477928	108018082	799	0.55	5.02	0.00	3.23
target_8689	5:108017046-108017283	rs11952230	108018395	1112	0.55	5.02	0.00	3.23
target_11554	6:3187079-3187396	rs6912511	3187454	58	0.62	6.40	0.00	6.28
target_11554	6:3187079-3187396	rs9392465	3162378	-24701	0.57	5.45	0.00	4.89
target_12051	6:30461304-30462395	rs2523989	30078275	-383029	0.53	4.60	0.13	3.74
target_12051	6:30461304-30462395	rs2239529	30078330	-382974	0.53	4.60	0.13	3.74

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak tion	posi-	SNP	SNP posi-	distance	Rho	p-value	empirical p-value	p-value 1kg
target_12051	6:30461304-30462395		rs2249099	30079307	-381997	-0.53	4.60	0.13	3.74
target_12051	6:30461304-30462395		rs2517598	30080274	-381030	-0.53	4.60	0.13	3.74
target_12051	6:30461304-30462395		rs2844793	30080496	-380808	-0.53	4.60	0.13	3.74
target_12051	6:30461304-30462395		rs2523985	30081334	-379970	-0.53	4.60	0.13	3.74
target_12051	6:30461304-30462395		rs2523984	30082003	-379301	0.53	4.60	0.13	3.74
target_12051	6:30461304-30462395		rs2523735	30122657	-338647	0.53	4.59	0.13	3.65
target_12051	6:30461304-30462395		rs3094034	30363351	-97953	0.53	4.50	0.14	3.00
target_12051	6:30461304-30462395		rs3130782	30914843	452448	0.52	4.48	0.15	3.42
target_12051	6:30461304-30462395		rs3131934	30931844	469449	-0.52	4.48	0.15	3.42
target_12051	6:30461304-30462395		rs3131783	30932068	469673	-0.52	4.48	0.15	3.42
target_12051	6:30461304-30462395		rs3132649	30321057	-140247	-0.52	4.41	0.15	3.46

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_12051	6:30461304-30462395	rs3094061	30321189	-140115	0.52	4.41	0.15	3.46
target_12051	6:30461304-30462395	rs3130374	30321336	-139968	0.52	4.41	0.15	3.46
target_12051	6:30461304-30462395	rs3094627	30321482	-139822	-0.52	4.41	0.15	3.46
target_12051	6:30461304-30462395	rs3130375	30321732	-139572	-0.52	4.41	0.15	3.46
target_7145	6:163601983-163602667	rs7742763	164084488	481821	-0.55	5.02	0.00	4.62
target_7145	6:163601983-163602667	rs9458878	164073368	470701	0.55	4.95	0.00	3.53
target_7145	6:163601983-163602667	rs9364696	164065525	462858	0.54	4.84	0.38	3.27
target_6533	7:127670510-127672406	rs11971935	127422539	-247971	-0.52	4.33	0.47	2.58
target_6970	7:100197836-100198529	rs35111986	100000274	-197562	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs1000215	100004543	-193293	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs3900792	100004577	-193259	0.50	4.12	0.17	2.81

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_6970	7:100197836-100198529	rs7783159	100017454	-180382	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs28490152	100020982	-176854	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs6975729	100028187	-169649	-0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs1073	100031612	-166224	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs11761725	100039815	-158021	-0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs11761784	100039970	-157866	-0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs28578163	100044754	-153082	-0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs2406255	100053690	-144146	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs7794485	100068936	-128900	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs4472444	100070758	-127078	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs7809801	100072031	-125805	0.50	4.12	0.17	2.81

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_6970	7:100197836-100198529	rs11766752	100073292	-124544	-0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs11559117	100076614	-121222	-0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs2406253	100077273	-120563	-0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs11760536	100078754	-119082	-0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs11763511	100081944	-115892	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs11769700	100090049	-107787	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs7783550	99987805	-210031	-0.50	4.12	0.17	2.06
target_6970	7:100197836-100198529	rs11769886	99998607	-199229	0.50	4.12	0.17	2.81
target_6970	7:100197836-100198529	rs34989573	100097895	-99941	-0.50	4.01	0.22	2.81
target_6970	7:100197836-100198529	rs7787620	100089234	-108602	-0.49	3.93	0.32	2.81
target_7342	8:120428095-120429658	rs3134058	119954108	-473987	0.60	6.01	0.00	4.00

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak tion	posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_7342	8:120428095-120429658	rs3134057	119953468	-474627	-0.55	5.01	0.00		4.10
target_7342	8:120428095-120429658	rs1564861	119965909	-462186	-0.53	4.56	0.33		4.10
target_10690	9:37966185-37966913	rs1887455	37488725	-477460	0.55	5.01	0.50		4.42
target_10690	9:37966185-37966913	rs6476635	37498334	-467851	-0.55	5.01	0.50		4.42
target_10690	9:37966185-37966913	rs10973394	37522672	-443513	-0.55	5.01	0.50		4.42
target_3019	9:37465999-37466571	rs10758440	37573330	106759	0.63	6.65	0.00		4.41
target_3019	9:37465999-37466571	rs7027222	37651804	185233	0.54	4.83	0.27		2.92
target_3019	9:37465999-37466571	rs10814581	37584687	118116	-0.53	4.62	0.48		2.81
target_3019	9:37465999-37466571	rs4599892	37603046	136475	-0.53	4.62	0.48		2.81
target_3019	9:37465999-37466571	rs12380266	37606701	140130	0.53	4.62	0.48		2.81
target_3019	9:37465999-37466571	rs10814586	37610198	143627	0.53	4.62	0.48		2.81

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_3019	9:37465999-37466571	rs10973438	37615548	148977	0.53	4.62	0.48	2.81
target_3019	9:37465999-37466571	rs2209561	37620873	154302	-0.53	4.62	0.48	2.81
target_3019	9:37465999-37466571	rs10814588	37621667	155096	0.53	4.62	0.48	2.81
target_6090	9:138444546-138446041	rs2590504	138365495	-79051	0.65	7.20	0.00	6.37
target_6090	9:138444546-138446041	rs12238018	138365624	-78922	-0.65	7.20	0.00	6.12
target_6090	9:138444546-138446041	rs2777323	138364389	-80157	-0.63	6.65	0.00	5.31
target_12170	10:108030713-108031514	rs1556379	107998037	-32676	0.55	4.68	0.00	2.43
target_12461	11:62264440-62265021	rs953894	62762422	497401	-0.57	5.44	0.00	3.01
target_12461	11:62264440-62265021	rs2187384	62761161	496140	-0.57	5.33	0.00	3.01
target_2934	11:46721763-46722968	rs7107778	46860855	137887	0.50	4.11	0.38	3.81
target_2934	11:46721763-46722968	rs7929294	46776176	53208	-0.51	4.09	0.38	3.81

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak tion	posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_2934	11:46721763-46722968		rs10734548	46787573	64605	-0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs7940578	46812227	89259	-0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs10769208	46822240	99272	-0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs10734549	46822686	99718	-0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs7101374	46824241	101273	-0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs1007738	46849360	126392	-0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs7480140	46856335	133367	0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs1551744	46865000	142032	-0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs10769211	46872873	149905	-0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs7943866	46881954	158986	0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs2306026	46889713	166745	-0.49	3.96	0.47	3.81

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak tion	posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_2934	11:46721763-46722968		rs3816614	46890165	167197	0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs2290884	46890388	167420	0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs11039014	46895378	172410	-0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs964551	46897995	175027	0.49	3.96	0.47	3.81
target_2934	11:46721763-46722968		rs2306027	46907827	184859	0.49	3.96	0.47	3.81
target_3094	12:775816-776770		rs2535426	686076	-89740	0.61	6.28	0.00	5.28
target_3094	12:775816-776770		rs2535436	698634	-77182	0.57	5.30	0.20	4.68
target_3094	12:775816-776770		rs2607924	700675	-75141	-0.56	5.12	0.29	4.61
target_3094	12:775816-776770		rs3809263	773456	-2360	0.53	4.50	0.48	3.59
target_4346	12:12716524-12717414		rs16907810	12327655	-388869	0.57	5.28	0.33	3.23
target_4346	12:12716524-12717414		rs11054701	12289156	-427368	0.56	5.08	0.50	3.03

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak tion	posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_4346	12:12716524-12717414		rs16907786	12314602	-401922	0.56	5.08	0.50	3.03
target_4346	12:12716524-12717414		rs2300231	12315880	-400644	0.56	5.08	0.50	3.03
target_7442	13:22423199-22423877		rs9550818	22754120	330243	-0.60	5.80	0.50	4.96
target_1190	14:53172966-53174397		rs17725312	53065714	-107252	-0.52	4.48	0.45	3.49
target_2278	14:90135472-90135995		rs4904547	89794530	-340942	0.57	5.37	0.50	3.46
target_12381	17:47856476-47857395		rs1867087	47440420	-416056	0.54	4.80	0.40	3.83
target_12381	17:47856476-47857395		rs365038	47967675	110280	-0.54	4.78	0.40	3.08
target_2715	17:37906275-37907227		rs4795405	38088417	181190	0.52	4.34	0.13	3.77
target_2715	17:37906275-37907227		rs4378650	38080865	173638	-0.51	4.15	0.27	3.39
target_3368	17:18854435-18854773		rs1859083	18843306	-11129	0.49	3.84	0.50	3.17
target_506	17:47491425-47492976		rs596783	47329705	-161720	-0.58	5.58	0.00	3.98

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_506	17:47491425-47492976	rs757554	47345847	-145578	0.55	5.03	0.50	3.82
target_506	17:47491425-47492976	rs16947546	47324353	-167072	-0.53	4.62	0.36	3.16
target_7683	17:40266513-40267923	rs2033675	40110764	-155749	-0.53	4.60	0.22	3.23
target_7683	17:40266513-40267923	rs12602950	40123829	-142684	0.53	4.60	0.22	3.23
target_7683	17:40266513-40267923	rs4796744	40096473	-170040	0.49	3.89	0.47	3.23
target_7683	17:40266513-40267923	rs4594300	40098187	-168326	0.49	3.89	0.47	3.23
target_7683	17:40266513-40267923	rs4796750	40114544	-151969	-0.49	3.89	0.47	3.23
target_11408	19:50957982-50959368	rs3745504	50771609	-186373	-0.58	5.56	0.00	4.77
target_6274	19:52598247-52599273	rs10410908	52680938	81665	0.56	5.15	0.36	3.95
target_6274	19:52598247-52599273	rs10411335	52681110	81837	-0.56	5.15	0.36	3.95
target_9523	19:13999997-14001221	rs12610739	14461417	460196	0.55	4.84	0.50	3.82

Continued on
next page

Table B.1: Overview FAIRE-QTLs

FAIRE-peak	FAIRE-peak tion	posi- tion	SNP	SNP posi- tion	distance	Rho	p-value	empirical p-value	p-value 1kg
target_373	20:49942253-49943223		rs1474843	49950741	7518	0.60	6.05	0.00	3.86
target_373	20:49942253-49943223		rs6021118	49950753	7530	0.60	6.05	0.00	3.86
target_373	20:49942253-49943223		rs4811159	49955717	12494	0.59	5.70	0.00	4.41
target_10733	X:115028350-115029418		rs6608481	115046966	17548	0.75	10.42	0.00	7.83
target_10733	X:115028350-115029418		rs11091032	115021945	-6405	0.67	7.69	0.00	5.70
target_10733	X:115028350-115029418		rs11091036	115023111	-5239	-0.56	5.08	0.25	4.62
target_1817	X:115085080-115085589		rs6608481	115046966	-38114	0.52	4.47	0.29	3.60
target_1817	X:115085080-115085589		rs11091032	115021945	-63135	0.50	4.12	0.44	3.35
target_6542	X:12842980-12843491		rs4316308	12719036	-123944	0.58	5.61	0.00	2.62

REFERENCES

- 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319): 1061–1073, Oct. 2010.
- 1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov. 2012.
- V. G. Allfrey, R. Faulkner, and A. E. Mirsky. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 51:786–794, May 1964.
- F. W. Alt, E. M. Oltz, F. Young, J. Gorman, G. Taccioli, and J. Chen. VDJ recombination. *Immunology today*, 13(8):306–314, Aug. 1992.
- V. Ambros. microRNAs: tiny regulators with great potential. *Cell*, 107(7):823–826, Dec. 2001.
- L. H. Augenlicht and D. Kobrin. Cloning and screening of sequences expressed in a mouse colon tumor. *Cancer research*, 42(3):1088–1093, Mar. 1982.
- O. T. Avery, C. M. MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of experimental medicine*, 79(2):137–158, 1944.

- J. F. Ayroles, M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman, M. M. Magwire, S. M. Rollmann, L. H. Duncan, F. Lawrence, R. R. H. Anholt, and T. F. C. Mackay. Systems genetics of complex traits in *Drosophila melanogaster*. *Nature genetics*, 41(3):299–307, Mar. 2009.
- D. Bachtrog. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature reviews Genetics*, 14(2):113–124, Feb. 2013.
- A. J. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, Mar. 2011.
- M. L. Barr and E. G. Bertram. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature*, 163(4148):676–676, Apr. 1949.
- A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, May 2007.
- S. Beck, B. E. Bernstein, R. M. Campbell, J. F. Costello, D. Dhanak, J. R. Ecker, J. M. Greally, J.-P. Issa, P. W. Laird, K. Polyak, B. Tycko, P. A. Jones, and AACR Cancer Epigenome Task Force. A blueprint for an international cancer epigenome consortium. A report from the AACR Cancer Epigenome Task Force. *Cancer research*, 72(24):6319–6324, Dec. 2012.
- Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.
- D. R. Bentley. Whole-genome re-sequencing. *Current opinion in genetics & development*, 2006.

- B. E. Bernstein, A. Meissner, and E. S. Lander. The mammalian epigenome. *Cell*, 128(4):669–681, Feb. 2007.
- B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*, 28(10):1045–1048, Oct. 2010.
- A. Bird, M. Taggart, M. Frommer, O. J. Miller, and D. Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91–99, Jan. 1985.
- A. P. Bird. CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067):209–213, May 1986.
- E. Birney. Chromatin and heritability: how epigenetic studies can complement genetic approaches. *Trend in Genetics*, 27(5):172–176, May 2011.
- C. E. Bonferroni. *Il calcolo delle assicurazioni su gruppi di teste*. Tipografia del Senato, 1935.
- N. Bouatia-Naji, G. Rocheleau, L. Van Lommel, K. Lemaire, F. Schuit, C. Cavalcanti-Proença, M. Marchand, A.-L. Hartikainen, U. Sovio, F. De Graeve, J. Rung, M. Vaxillaire, J. Tichet, M. Marre, B. Balkau, J. Weill, P. Elliott, M.-R. Jarvelin, D. Meyre, C. Polychronakos, C. Dina, R. Sladek, and P. Froguel. A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. *Science*, 320(5879):1085–1088, May 2008.
- A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, 132(2):311–322, Jan. 2008.

- A. P. Boyle, L. Song, B.-K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford, and T. S. Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome research*, 21(3):456–464, Mar. 2011.
- A. P. Boyle, E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski, K. J. Karczewski, J. Park, B. C. Hitz, S. Weng, J. M. Cherry, and M. Snyder. Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, 22(9):1790–1797, Sept. 2012.
- A. B. Brinkman, T. Roelofsen, S. W. C. Pennings, J. H. A. Martens, T. Jenuwein, and H. G. Stunnenberg. Histone modification patterns associated with the human X chromosome. *EMBO reports*, 7(6):628–634, June 2006.
- C. J. Brown and H. F. Willard. The human X-inactivation centre is not required for maintenance of X-chromosome inactivation. *Nature*, 368(6467):154–156, Mar. 1994.
- C. J. Brown, A. Ballabio, J. L. Rupert, R. G. Lafreniere, M. Grompe, R. Tonlorenzi, and H. F. Willard. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, 349(6304):38–44, Jan. 1991.
- J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12):1213–1218, Oct. 2013.
- P. S. Burgoyne. Genetic homology and crossing over in the X and Y chromosomes of Mammals. *Human genetics*, 61(2):85–90, 1982.

- C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282(5396):2012–2018, Dec. 1998.
- C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC), F. Wensley, P. Gao, S. Burgess, S. Kaptoge, E. Di Angelantonio, T. Shah, J. C. Engert, R. Clarke, G. Davey Smith, B. G. Nordestgaard, D. Saleheen, N. J. Samani, M. Sandhu, S. Anand, M. B. Pepys, L. Smeeth, J. Whittaker, J. P. Casas, S. G. Thompson, A. D. Hingorani, and J. Danesh. Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ (Clinical research ed.)*, 342:d548, 2011.
- P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. M. Semple, M. S. Taylor, P. G. Engström, M. C. Frith, A. R. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, 38(6):626–635, June 2006.
- L. Carrel and H. F. Willard. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, 434(7031):400–404, Mar. 2005.
- L. H. Chadwick. The NIH Roadmap Epigenomics Program data resource. *Epigenomics*, 4(3):317–324, June 2012.
- W. Chao, K. D. Huynh, R. J. Spencer, L. S. Davidow, and J. T. Lee. CTCF, a candidate trans-acting factor for X-inactivation choice. *Science*, 295(5553):345–347, Jan. 2002.

- C. M. Clemson, J. A. McNeil, H. F. Willard, and J. B. Lawrence. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *The Journal of cell biology*, 132(3):259–275, Feb. 1996.
- G. M. Cooper, E. A. Stone, G. Asimenos, N. C. S. Program, E. D. Green, S. Batzoglou, and A. Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, 15(7):901–913, July 2005.
- G. E. Crawford, S. Davis, P. C. Scacheri, G. Renaud, M. J. Halawi, M. R. Erdos, R. Green, P. S. Meltzer, T. G. Wolfsberg, and F. S. Collins. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature methods*, 3(7):503–509, July 2006.
- S. Cuddapah, R. Jothi, D. E. Schones, T.-Y. Roh, K. Cui, and K. Zhao. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research*, 19(1):24–32, Jan. 2009.
- C. R. Darwin. *On the Origin of Species by Means of Natural Selection Or The Preservation of Favoured Races in the Struggle for Life*. 1861.
- G. Davey Smith and S. Ebrahim. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ (Clinical research ed.)*, 330(7499):1076–1079, May 2005.
- J. F. Degner, A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard. DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature*, Feb. 2012.

- J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, Feb. 2002.
- M. E. Donohoe, L.-F. Zhang, N. Xu, Y. Shi, and J. T. Lee. Identification of a Ctfc cofactor, Yy1, for the X chromosome binary switch. *Molecular cell*, 25(1):43–56, Jan. 2007.
- A. W. Drong, C. M. Lindgren, and M. I. McCarthy. The Genetic and Epigenetic Basis of Type 2 Diabetes and Obesity. *Clinical Pharmacology & Therapeutics*, 92(6):707–715, Oct. 2012.
- I. Dunham, N. Shimizu, B. A. Roe, S. Chisoe, A. R. Hunt, J. E. Collins, R. Bruskewich, D. M. Beare, M. Clamp, L. J. Smink, R. Ainscough, J. P. Almeida, A. Babbage, C. Bagguley, J. Bailey, K. Barlow, K. N. Bates, O. Beasley, C. P. Bird, S. Blakey, A. M. Bridgeman, D. Buck, J. Burgess, W. D. Burrill, and K. P. O'Brien. The DNA sequence of human chromosome 22. *Nature*, 402(6761):489–495, Dec. 1999.
- P. Elliott, J. C. Chambers, W. Zhang, R. Clarke, J. C. Hopewell, J. F. Peden, J. Erdmann, P. Braund, J. C. Engert, D. Bennett, L. Coin, D. Ashby, I. Tzoulaki, I. J. Brown, S. Mt-Isa, M. I. McCarthy, L. Peltonen, N. B. Freimer, M. Farrall, A. Ruokonen, A. Hamsten, N. Lim, P. Froguel, D. M. Waterworth, P. Vollenweider, G. Waeber, M.-R. Jarvelin, V. Mooser, J. Scott, A. S. Hall, H. Schunkert, S. S. Anand, R. Collins, N. J. Samani, H. Watkins, and J. S. Kooner. Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. *JAMA*, 302(1):37–48, July 2009.
- ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sept. 2012.
- ENCODE Project Consortium, E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S.

Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. J. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W.-K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. UCLA, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammanna, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C.-L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaöz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Löytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, N. C. S. Program, B. C. of Medicine

- Human Genome Sequencing Center, W. U. G. S. Center, B. Institute, C. H. O. R. Institute, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King, A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. H. Lee, P. Ng, A. Shahab, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, and A. E. Urban. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, June 2007.
- J. M. Engreitz, A. Pandya-Jones, P. McDonel, A. Shishkin, K. Sirokman, C. Surka, S. Kadri, J. Xing, A. Goren, E. S. Lander, K. Plath, and M. Guttman. The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science*, 341(6147):1237973–1237973, Aug. 2013.
- J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shores, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, Mar. 2011.
- D. M. Evans and S. Purcell. Power calculations in genetic studies. *Cold Spring Harbor Protocols*, 2012(6):664–674, June 2012.
- J. B. Fan, X. Chen, M. K. Halushka, A. Berno, X. Huang, T. Ryder, R. J. Lipshutz, D. J. Lockhart, and A. Chakravarti. Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome research*, 10(6):853–860, June 2000.
- R. S. N. Fehrmann, R. C. Jansen, J. H. Veldink, H.-J. Westra, D. Arends, M. J. Bonder, J. Fu, P. Deelen, H. J. M. Groen,

- A. Smolonska, R. K. Weersma, R. M. W. Hofstra, W. A. Buurman, S. Rensen, M. G. M. Wolfs, M. Platteel, A. Zhernakova, C. C. Elbers, E. M. Festen, G. Trynka, M. H. Hofker, C. G. J. Saris, R. A. Ophoff, L. H. Van den Berg, D. A. van Heel, C. Wijmenga, G. J. te Meerman, and L. Franke. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS genetics*, 7(8):e1002197, Aug. 2011.
- S. Feng, S. E. Jacobsen, and W. Reik. Epigenetic reprogramming in plant and animal development. *Science*, 330(6004):622–627, Oct. 2010.
- R. A. Fisher. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02):399–433, Jan. 1919.
- R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, Jan. 1922a.
- R. A. Fisher. *On the Mathematical Foundations of Theoretical Statistics*. 1922b.
- R. A. Fisher. *Statistical Methods For Research Workers*. 1925.
- R. A. Fisher. *The Genetical Theory of Natural Selection*. 1930.
- R. A. Fisher. *The design of experiments*. 1935.
- R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512, July 1995.
- P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kähäri,

- D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suárez, J. Harrow, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. J. Searle. Ensembl 2012. *Nucleic acids research*, 40(Database issue):D84–90, Jan. 2012.
- J. Flint and T. F. C. Mackay. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome research*, 19(5): 723–733, May 2009.
- M. J. Fullwood, M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Y. Chew, P. Y. H. Huang, W.-J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. S. A. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. M. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W.-K. Sung, E. T. Liu, C.-L. Wei, E. Cheung, and Y. Ruan. An oestrogen-receptor- α -bound human chromatin interactome. *Nature*, 462(7269):58–64, Nov. 2009.
- T. S. Furey. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews Genetics*, 13(12):840–852, Dec. 2012.
- F. Galton. *Hereditary Genius*. 1869.
- F. Galton. The history of twins, as a criterion of the relative powers of nature and nurture. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 5:391–406, 1876.

- F. Galton. The geometric mean, in vital and social statistics. *Proceedings of the Royal Society of London*, 29(196-199):365–367, 1879.
- F. Galton. Eugenics: Its definition, scope, and aims. *The American Journal of Sociology*, 1904.
- F. Galton. *Inquiries Into Human Faculty and Its Development*. 1907.
- K. J. Gaulton, T. Nammo, L. Pasquali, J. M. Simon, P. G. Giresi, M. P. Fogarty, T. M. Panhuis, P. Mieczkowski, A. Secchi, D. Bosco, T. Berney, E. Montanya, K. L. Mohlke, J. D. Lieb, and J. Ferrer. A map of open chromatin in human pancreatic islets. *Nature genetics*, 42(3):255–259, Jan. 2010.
- D. Gentilini, D. Mari, D. Castaldi, D. Remondini, G. Ogliari, R. Ostan, L. Bucci, S. M. Sirchia, S. Tabano, F. Cavagnini, D. Monti, C. Franceschi, A. M. Di Blasio, and G. Vitale. Role of epigenetics in human aging and longevity: genome-wide DNA methylation profile in centenarians and centenarians' offspring. *Age (Dordrecht, Netherlands)*, 35(5):1961–1973, Oct. 2013.
- G. Gey, W. D. Coffman, and M. T. Kubicek. Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. *Cancer research*, 12(4):264–265, 1952.
- D. Gianola and D. Sorensen. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics*, 167(3):1407–1424, July 2004.
- N. Gilbert, S. Boyle, H. Fiegler, K. Woodfine, N. P. Carter, and W. A. Bickmore. Chromatin Architecture of the Human Genome. *Cell*, 118(5):555–566, Sept. 2004.
- W. Gilbert and A. Maxam. The nucleotide sequence of the lac operator. *Proceedings of the National Academy of Sciences of the United States of America*, 70(12):3581–3584, Dec. 1973.

- P. G. Giresi and J. D. Lieb. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods*, 48(3): 233–239, July 2009.
- P. G. Giresi, J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research*, 17(6):877–885, June 2007.
- P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, June 2012.
- H. H. H. Göring, J. E. Curran, M. P. Johnson, T. D. Dyer, J. Charlesworth, S. A. Cole, J. B. M. Jowett, L. J. Abraham, D. L. Rainwater, A. G. Comuzzie, M. C. Mahaney, L. Almasy, J. W. MacCluer, A. H. Kissebah, G. R. Collier, E. K. Moses, and J. Blangero. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature genetics*, 39(10):1208–1216, Oct. 2007.
- J. Graves, M. J. Wakefield, and R. Toder. The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Human molecular genetics*, 1998.
- J. A. M. Graves. Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5):901–914, Mar. 2006.
- C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P.

- Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, Mar. 2007.
- D. Gresham, M. J. Dunham, and D. Botstein. Comparing whole genomes using DNA microarrays. *Nature reviews Genetics*, 9(4):291–302, Apr. 2008.
- J. D. Groman, M. E. Meyer, R. W. Wilmott, P. L. Zeitlin, and G. R. Cutting. Variant cystic fibrosis phenotypes in the absence of CFTR mutations. *The New England journal of medicine*, 347(6):401–407, Aug. 2002.
- D. S. Gross and W. T. Garrard. Nuclease hypersensitive sites in chromatin. *Annual review of biochemistry*, 57:159–197, 1988.
- D. F. Gudbjartsson, G. B. Walters, G. Thorleifsson, H. Stefansson, B. V. Halldorsson, P. Zusmanovich, P. Sulem, S. Thorlacius, A. Gylfason, S. Steinberg, A. Helgadóttir, A. Ingason, V. Steinthorsdóttir, E. J. Olafsdóttir, G. H. Olafsdóttir, T. Jonsson, K. Borch-Johnsen, T. Hansen, G. Andersen, T. Jørgensen, O. Pedersen, K. K. Aben, J. A. Witjes, D. W. Swinkels, M. den Heijer, B. Franke, A. L. M. Verbeek, D. M. Becker, L. R. Yanek, L. C. Becker, L. Tryggvadóttir, T. Rafnar, J. Gulcher, L. A. Kiemeny, A. Kong, U. Thorsteinsdóttir, and K. Stefansson. Many sequence variants affecting diversity of adult human height. *Nature genetics*, 40(5):609–615, May 2008.
- T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 1943.
- C. H. Haering, A.-M. Farcas, P. Arumugam, J. Metson, and K. Nasmyth. The cohesin ring concatenates sister DNA molecules. *Nature*, 454(7202):297–301, July 2008.

- L. Handoko, H. Xu, G. Li, C. Y. Ngan, E. Chew, M. Schnapp, C. W. H. Lee, C. Ye, J. L. H. Ping, F. Mulawadi, E. Wong, J. Sheng, Y. Zhang, T. Poh, C. S. Chan, G. Kunarso, A. Shahab, G. Bourque, V. Cacheux-Rataboul, W.-K. Sung, Y. Ruan, and C.-L. Wei. CTCF-mediated functional chromatin interactome in pluripotent cells : Nature Genetics : Nature Publishing Group. *Nature*, 43(7):630–638, June 2011.
- H. H. He, C. A. Meyer, S. S. Hu, M.-W. Chen, C. Zang, Y. Liu, P. K. Rao, T. Fei, H. Xu, H. Long, X. S. Liu, and M. Brown. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature methods*, 11(1):73–78, Jan. 2014.
- X. He, C. K. Fuller, Y. Song, Q. Meng, B. Zhang, X. Yang, and H. Li. Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. *The American Journal of Human Genetics*, 92(5):667–680, May 2013.
- L. B. Herzing, J. T. Romer, J. M. Horn, and A. Ashworth. Xist has properties of the X-chromosome inactivation centre. *Nature*, 386(6622):272–275, Mar. 1997.
- J. R. Hesselberth, G. E. Crawford, X. Chen, Z. Zhang, P. J. Sabo, L. Song, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods*, 6(4):283–289, Apr. 2009.
- A. G. Hinch, N. Altemose, N. Noor, P. Donnelly, and S. R. Myers. Recombination in the Human Pseudoautosomal Region PAR1. *PLoS genetics*, 2014.
- L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy*

- of Sciences of the United States of America*, 106(23):9362–9367, June 2009.
- Y. Hochberg and Y. Benjamini. More powerful procedures for multiple significance testing. *Statistics in medicine*, 9(7): 811–818, July 1990.
- M. M. Hoffman, J. Ernst, S. P. Wilder, A. Kundaje, R. S. Harris, M. Libbrecht, B. Giardine, P. M. Ellenbogen, J. A. Bilmes, E. Birney, R. C. Hardison, I. Dunham, M. Kellis, and W. S. Noble. Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, Dec. 2012.
- R. Holliday and J. E. Pugh. DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, Jan. 1975.
- E. E. Holohan, C. Kwong, B. Adryan, M. Bartkuhn, M. Herold, R. Renkawitz, S. Russell, and R. White. CTCF genomic binding sites in *Drosophila* and the organisation of the bithorax complex. *PLoS genetics*, 3(7):e112, July 2007.
- A. H. Horakova, S. C. Moseley, C. R. McLaughlin, D. C. Tremblay, and B. P. Chadwick. The macrosatellite DXZ4 mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome. *Human molecular genetics*, 21(20):4367–4377, Oct. 2012.
- S. Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):R115, 2013.
- D. Houle. Comparing evolvability and variability of quantitative traits. *Genetics*, 130(1):195–204, Jan. 1992.
- International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789–796, Nov. 2003.
- International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, Sept. 2005.

- International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164): 851–861, 2007.
- R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–264, Apr. 2003.
- V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409(6819):533–538, Jan. 2001.
- D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172): 776–779, Feb. 2014.
- L. J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. H. Staerfeldt, K. Rapacki, C. Workman, C. A. F. Andersen, S. Knudsen, A. Krogh, A. Valencia, and S. Brunak. Prediction of human protein function from post-translational modifications and localization features. *Journal of Molecular Biology*, 319(5):1257–1265, June 2002.
- Y. Jeon and J. T. Lee. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*, 146(1):119–133, July 2011.
- J. Jiang, Y. Jing, G. J. Cost, J.-C. Chiang, H. J. Kolpa, A. M. Cotton, D. M. Carone, B. R. Carone, D. A. Shivak, D. Y. Guschin, J. R. Pearl, E. J. Rebar, M. Byron, P. D. Gregory, C. J. Brown, F. D. Urnov, L. L. Hall, and J. B. Lawrence. Translating dosage compensation to trisomy 21. *Nature*, 500(7462):296–300, July 2013.

- A. D. Johnson and C. J. O'Donnell. An open access database of genome-wide association results. *BMC medical genetics*, 10:6, 2009.
- D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, June 2007.
- P. A. Jones and S. B. Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, Feb. 2007.
- P. A. Jones and P. W. Laird. Cancer epigenetics comes of age. *Nature genetics*, 21(2):163–167, Feb. 1999.
- L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J.-P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Büning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskis, S. Kugathasan, A. Lattiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. De Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeisig, B. Zhang, C. K. Zhang, H. Zhao, International IBD Genetics Consortium (IIBDGC), M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, E. E. Schadt, M. J. Daly, A. Franke, M. Parkes,

- S. Vermeire, J. C. Barrett, and J. H. Cho. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, Nov. 2012.
- R. Joubert-Caron, J.-P. Le Caër, F. Montandon, F. Poirier, M. Pontet, N. Imam, J. Feuillard, D. Bladier, J. Rossier, and M. Caron. Protein analysis by mass spectrometry and sequence database searching: A proteomic approach to identify human lymphoblastoid cell line proteins. *Electrophoresis*, 21(12):2566–2575, July 2000.
- M. H. Kagey, J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, Sept. 2010.
- P. R. Kamstrup, A. Tybjaerg-Hansen, R. Steffensen, and B. G. Nordestgaard. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *JAMA*, 301(22):2331–2339, June 2009.
- M. B. Katan. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet*, 1(8479):507–508, Mar. 1986.
- M. J. Kearsey and A. G. Farquhar. QTL analysis in plants; where are we now? *Heredity*, 80 (Pt 2):137–142, Feb. 1998.
- H. G. Khorana. Total synthesis of a gene. *Science*, 203(4381):614–625, Feb. 1979.
- H. Kilpinen, S. M. Waszak, A. R. Gschwind, S. K. Raghav, R. M. Witwicki, A. Orioli, E. Migliavacca, M. Wiederkehr, M. Gutierrez-Arcelus, N. I. Panousis, A. Yurovsky, T. Lapalainen, L. Romano-Palumbo, A. Planchon, D. Bielser, J. Bryois, I. Padioleau, G. Udin, S. Thurnheer, D. Hacker, L. J. Core, J. T. Lis, N. Hernandez, A. Reymond, B. Deplancke,

- and E. T. Dermitzakis. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, 342(6159):744–747, Nov. 2013.
- T. H. Kim, Z. K. Abdullaev, A. D. Smith, K. A. Ching, D. I. Loukinov, R. D. Green, M. Q. Zhang, V. V. Lobanenko, and B. Ren. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell*, 128(6):1231–1245, Mar. 2007.
- V. Kumar, M. Muratani, N. A. Rayan, P. Kraus, T. Lufkin, H. H. Ng, and S. Prabhakar. Uniform, optimal signal processing of mapped deep-sequencing data. *Nature Biotechnology*, 31(7):615–622, July 2013.
- M. H. Kuo and C. D. Allis. In vivo cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment. *Methods*, 19(3):425–433, Nov. 1999.
- W. P. Kuo, T.-K. Jenssen, A. J. Butte, L. Ohno-Machado, and I. S. Kohane. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, 18(3):405–412, Mar. 2002.
- B. T. Lahn and D. C. Page. Four evolutionary strata on the human X chromosome. *Science*, 286(5441):964–967, Oct. 1999.
- E. Lalonde, K. C. H. Ha, Z. Wang, A. Bemmo, C. L. Kleinman, T. Kwan, T. Pastinen, and J. Majewski. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome research*, 21(4):545–554, Apr. 2011.
- F. H. Lam, D. J. Steger, and E. K. O’Shea. Chromatin decouples promoter threshold from dynamic range. *Nature*, 453(7192):246–250, Apr. 2008.
- E. S. Lander and D. Botstein. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199, Jan. 1989.

- E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczy, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chisoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou,

- R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kasprzyk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowski, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001.
- J. J. M. Landry, P. T. Pyl, T. Rausch, T. Zichner, M. M. Tekkedil, A. M. Stütz, A. Jauch, R. S. Aiyar, G. Pau, N. Delhomme, J. Gagneur, J. O. Korb, W. Huber, and L. M. Steinmetz. The Genomic and Transcriptomic Landscape of a HeLa Cell Line. *G3*, 3(8):1213–1224, 2013.
- T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. C. 't Hoen, J. Monlong, M. A. Rivas, M. González-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padiou, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom,

- Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, Á. Carracedo, S. E. Antonarakis, R. Häsler, A.-C. Syvänen, G.-J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, E. T. Dermitzakis, A. Palotie, J. F. Deleuze, U. Gyllenstein, H. Brunner, J. Veltman, A. Cambon-Thomsen, J. Mangion, D. Bentley, and A. Hamosh. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, Sept. 2013.
- D. A. Lashkari, J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13057–13062, Nov. 1997.
- D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, Apr. 2008.
- J. T. Lee, L. S. Davidow, and D. Warshawsky. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nature genetics*, 21(4):400–404, Apr. 1999.
- G. Li, M. J. Fullwood, H. Xu, F. H. Mulawadi, S. Velkov, V. Vega, P. N. Ariyaratne, Y. B. Mohamed, H. S. Ooi, C. Tennakoon, C.-L. Wei, Y. Ruan, and W.-K. Sung. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biology*, 11(2):R22, 2010.
- C. Libert, L. Dejager, and I. Pinheiro. The X chromosome in immune functions: when a chromosome makes the difference. *Nature Reviews Immunology*, 10(8):594–604, Aug. 2010.
- E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny,

- E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct. 2009.
- J. Q. Ling, T. Li, J. F. Hu, T. H. Vu, H. L. Chen, X. W. Qiu, A. M. Cherry, and A. R. Hoffman. CTCF mediates interchromosomal colocalization between *Igf2/H19* and *Wsb1/Nf1*. *Science*, 312(5771):269–272, Apr. 2006.
- V. V. Lobanenkov, N. RH, V. V. Adler, H. Paterson, E. M. Klenova, A. V. Polotskaja, and G. H. Goodwin. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5′-flanking sequence of the chicken c-myc gene. *Oncogene*, 5(12):1743–1753, Dec. 1990.
- K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature genetics*, 33(2):177–182, Feb. 2003.
- C. Lorthongpanich, L. F. Cheow, S. Balu, S. R. Quake, B. B. Knowles, W. F. Burkholder, D. Solter, and D. M. Messerschmidt. Single-cell DNA-methylation analysis reveals epigenetic chimerism in preimplantation embryos. *Science*, 341(6150):1110–1112, Sept. 2013.
- C. B. Lozzio and B. B. Lozzio. Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood*, 45(3):321–334, Mar. 1975.
- K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, Sept. 1997.
- M. F. Lyon. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature*, 190:372–373, Apr. 1961.

- T. F. C. Mackay, E. A. Stone, and J. F. Ayroles. The genetics of quantitative traits: challenges and prospects. *Nature reviews Genetics*, 10(8):565–577, Aug. 2009.
- J. H. Malone and B. Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, 9:34, 2011.
- M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sept. 2005.
- H. Marks, J. C. Chow, S. Denissov, K.-J. François, N. Brockdorff, E. Heard, and H. G. Stunnenberg. High-resolution analysis of epigenetic changes associated with X inactivation. *Genome research*, 19(8):1361–1373, Aug. 2009.
- J. Martens and H. G. Stunnenberg. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*, 2013.
- U. Maskos and E. M. Southern. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic acids research*, 20(7):1679–1684, Apr. 1992.
- J. R. Masters. HeLa cells 50 years on: the good, the bad and the ugly. *Nature Reviews Cancer*, 2(4):315–319, 2002.

- M. T. Maurano, H. Wang, T. Kuttyavin, and J. A. Stamatoyannopoulos. Widespread Site-Dependent Buffering of Human Regulatory Polymorphism. *PLoS genetics*, 8(3):e1002599, Mar. 2012.
- A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564, Feb. 1977.
- M. I. McCarthy. Genomics, type 2 diabetes, and obesity. *The New England journal of medicine*, 363(24):2339–2350, Dec. 2010.
- R. McDaniell, B.-K. Lee, L. Song, Z. Liu, A. P. Boyle, M. R. Erdos, L. J. Scott, M. A. Morken, K. S. Kucera, A. Battenhouse, D. Keefe, F. S. Collins, H. F. Willard, J. D. Lieb, T. S. Furey, G. E. Crawford, V. R. Iyer, and E. Birney. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, 328(5975):235–239, Apr. 2010.
- G. McVicker, B. van de Geijn, J. F. Degner, C. E. Cain, N. E. Banovich, A. Raj, N. Lewellen, M. Myrthil, Y. Gilad, and J. K. Pritchard. Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science*, 342(6159):747–749, Nov. 2013.
- D. Melzer, J. R. B. Perry, D. Hernandez, A.-M. Corsi, K. Stevens, I. Rafferty, F. Lauretani, A. Murray, J. R. Gibbs, G. Paolisso, S. Rafiq, J. Simon-Sanchez, H. Lango, S. Scholz, M. N. Weedon, S. Arepalli, N. Rice, N. Washecka, A. Hurst, A. Britton, W. Henley, J. van de Leemput, R. Li, A. B. Newman, G. Tranah, T. Harris, V. Panicker, C. Dayan, A. Bennett, M. I. McCarthy, A. Ruokonen, M.-R. Jarvelin, J. Guralnik, S. Bandinelli, T. M. Frayling, A. Singleton, and L. Ferrucci. A Genome-Wide Association Study Identifies Protein Quantitative Trait Loci (pQTLs). *PLoS genetics*, 4(5):e1000072, May 2008.
- G. Mendel. Versuche uber Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereins Brünn*, 4:3–47, 1866.

- M. Merckenschlager and D. T. Odom. CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets. *Cell*, 152(6): 1285–1297, Mar. 2013.
- M. L. Metzker. Sequencing technologies - the next generation. *Nature reviews Genetics*, 11(1):31–46, Jan. 2010.
- T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, Aug. 2007.
- G. Miller. Human lymphoblastoid cell lines and Epstein-Barr virus: a review of their interrelationships and their relevance to the etiology of leukoproliferative states in man. *The Yale journal of biology and medicine*, 43(6):358–384, June 1971.
- S. A. Monks, A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, S. Edwards, J. W. Phillips, A. Sachs, and E. E. Schadt. Genetic inheritance of gene expression in human cell lines. *American journal of human genetics*, 75(6):1094–1105, Dec. 2004.
- S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigó, and E. T. Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777, Apr. 2010.
- T. H. Morgan. *Evolution and Adaptation*. 1903.
- T. H. Morgan. The origin of five mutations in eye color in *Drosophila* and their modes of inheritance. *Science*, 1911.
- T. H. Morgan. *A Critique of the Theory of Evolution*. 1916.
- S. C. Moseley, R. Rizkallah, D. C. Tremblay, B. R. Anderson, M. M. Hurt, and B. P. Chadwick. YY1 associates with the

- macrosatellite DXZ4 on the inactive X chromosome and binds with CTCF to a hypomethylated form in some male carcinomas. *Nucleic acids research*, 40(4):1596–1608, Feb. 2012.
- N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, and M. Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, Apr. 2011.
- C. Ober, D. A. Loisel, and Y. Gilad. Sex-specific genetic architecture of human disease. *Nature reviews Genetics*, 9(12):911–922, Dec. 2008.
- R. Ohlsson, R. Renkawitz, and V. Lobanenko. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trend in Genetics*, 17(9):520–527, Sept. 2001.
- S. Ohno, W. D. Kaplan, and R. Kinosita. Formation of the sex chromatin by a single X-chromosome in liver cells of *Rattus norvegicus*. *Experimental cell research*, 18:415–418, Oct. 1959.
- Oxford Nanopore Technologies. Oxford Nanopore introduces DNA ‘strand sequencing’ on the high-throughput GridION platform and presents MinION, a sequencer the size of a USB memory stick, Feb. 2012. URL <https://www.nanoporetech.com/news/press-releases/view/39>.
- I. Paran and D. Zamir. Quantitative traits in plants: beyond the QTL. *Trend in Genetics*, 19(6):303–306, June 2003.
- B. Payer and J. T. Lee. X chromosome dosage compensation: how mammals keep the balance. *Annual review of genetics*, 42: 733–772, 2008.
- K. Pearson. *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*. The London, 1900.

- K. Pearson. On lines and planes of closest fit to systems of points in space. *The London*, 1901.
- K. Pearson. JSTOR: Biometrika, Vol. 13, No. 1 (Oct., 1920), pp. 25-45. *Biometrika*, 1920.
- G. D. Penny, G. F. Kay, S. A. Sheardown, S. Rastan, and N. Brockdorff. Requirement for Xist in X chromosome inactivation. *Nature*, 379(6561):131-137, Jan. 1996.
- J. E. Phillips and V. G. Corces. CTCF: master weaver of the genome. *Cell*, 137(7):1194-1211, June 2009.
- J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768-772, Mar. 2010.
- D. Pinkel, R. Seagraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, 20(2):207-211, Oct. 1998.
- T. Pollex and E. Heard. Recent advances in X-chromosome inactivation research. *Current opinion in cell biology*, 24(6):825-832, Dec. 2012.
- A. Portela and M. Esteller. Epigenetic modifications and human disease. *Nat Biotechnol*, 28(10):1057-1068, Oct. 2010.
- R. J. Pruim, R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines, T. P. Gliedt, M. Boehnke, G. R. Abecasis, and C. J. Willer. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336-2337, Sept. 2010.
- E. M. Pugacheva, V. K. Tiwari, Z. Abdullaev, A. A. Vostrov, P. T. Flanagan, W. W. Quitschke, D. I. Loukinov, R. Ohlsson, and

- V. V. Lobanenkov. Familial cases of point mutations in the XIST promoter reveal a correlation between CTCF binding and pre-emptive choices of X chromosome inactivation. *Human molecular genetics*, 14(7):953–965, Apr. 2005.
- D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, and R. Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol*, 30(8):777–782, Aug. 2012.
- T. E. Reddy, J. Gertz, F. Pauli, K. S. Kucera, K. E. Varley, K. M. Newberry, G. K. Marinov, A. Mortazavi, B. A. Williams, L. Song, G. E. Crawford, B. Wold, H. F. Willard, and R. M. Myers. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research*, 22(5):860–869, May 2012.
- A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375, Feb. 2003.
- B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309, Dec. 2000.
- A. D. Riggs. X inactivation, differentiation, and DNA methylation. *Cytogenetics and cell genetics*, 14(1):9–25, 1975.
- M. T. Ross, D. V. Grafham, A. J. Coffey, S. Scherer, K. McLay, D. Muzny, M. Platzer, G. R. Howell, C. Burrows, C. P. Bird, A. Frankish, F. L. Lovell, K. L. Howe, J. L. Ashurst, R. S. Fulton, R. Sudbrak, G. Wen, M. C. Jones, M. E. Hurles, T. D. Andrews, C. E. Scott, S. Searle, J. Ramser, A. Whittaker, R. Deadman, N. P. Carter, S. E. Hunt, R. Chen, A. Cree, P. Gunaratne, P. Havlak, A. Hodgson, M. L. Metzker, S. Richards,

G. Scott, D. Steffen, E. Sodergren, D. A. Wheeler, K. C. Worley, R. Ainscough, K. D. Ambrose, M. A. Ansari-Lari, S. Aradhya, R. I. S. Ashwell, A. K. Babbage, C. L. Bagguley, A. Ballabio, R. Banerjee, G. E. Barker, K. F. Barlow, I. P. Barrett, K. N. Bates, D. M. Beare, H. Beasley, O. Beasley, A. Beck, G. Bethel, K. Blechschmidt, N. Brady, S. Bray-Allen, A. M. Bridgeman, A. J. Brown, M. J. Brown, D. Bonnin, E. A. Bruford, C. Buhay, P. Burch, D. Burford, J. Burgess, W. Burrill, J. Burton, J. M. Bye, C. Carder, L. Carrel, J. Chako, J. C. Chapman, D. Chavez, E. Chen, G. Chen, Y. Chen, Z. Chen, C. Chinault, A. Ciccodicola, S. Y. Clark, G. Clarke, C. M. Clee, S. Clegg, K. Clerc-Blankenburg, K. Clifford, V. Cobley, C. G. Cole, J. S. Conquer, N. Corby, R. E. Connor, R. David, J. Davies, C. Davis, J. Davis, O. Delgado, D. Deshazo, P. Dhami, Y. Ding, H. Dinh, S. Dodsworth, H. Draper, S. Dugan-Rocha, A. Dunham, M. Dunn, K. J. Durbin, I. Dutta, T. Eades, M. Ellwood, A. Emery-Cohen, H. Errington, K. L. Evans, L. Faulkner, F. Francis, J. Frankland, A. E. Fraser, P. Galgoczy, J. Gilbert, R. Gill, G. Glöckner, S. G. Gregory, S. Gribble, C. Griffiths, R. Grocock, Y. Gu, R. Gwilliam, C. Hamilton, E. A. Hart, A. Hawes, P. D. Heath, K. Heitmann, S. Hennig, J. Hernandez, B. Hinzmann, S. Ho, M. Hoffs, P. J. Howden, E. J. Huckle, J. Hume, P. J. Hunt, A. R. Hunt, J. Isherwood, L. Jacob, D. Johnson, S. Jones, P. J. de Jong, S. S. Joseph, S. Keenan, S. Kelly, J. K. Kershaw, Z. Khan, P. Kioschis, S. Klages, A. J. Knights, A. Kosiura, C. Kovar-Smith, G. K. Laird, C. Langford, S. Lawlor, M. Leversha, L. Lewis, W. Liu, C. Lloyd, D. M. Lloyd, H. Loulseged, J. E. Loveland, J. D. Lovell, R. Lozado, J. Lu, R. Lyne, J. Ma, M. Maheshwari, L. H. Matthews, J. McDowall, S. McLaren, A. McMurray, P. Meidl, T. Meitinger, S. Milne, G. Miner, S. L. Mistry, M. Morgan, S. Morris, I. Müller, J. C. Mullikin, N. Nguyen, G. Nordsiek, G. Nyakatura, C. N. O'Dell, G. Okwuonu, S. Palmer, R. Pandian, D. Parker, J. Parrish, S. Pasternak, D. Patel, A. V. Pearce, D. M. Pearson, S. E. Pelan, L. Perez, K. M. Porter, Y. Ramsey, K. Reichwald,

- S. Rhodes, K. A. Ridler, D. Schlessinger, M. G. Schueler, H. K. Sehra, C. Shaw-Smith, H. Shen, E. M. Sheridan, R. Shownkeen, C. D. Skuce, M. L. Smith, E. C. Sotheran, H. E. Steingruber, C. A. Steward, R. Storey, R. M. Swann, D. Swarbreck, P. E. Tabor, S. Taudien, T. Taylor, B. Teague, K. Thomas, A. Thorpe, K. Timms, A. Tracey, S. Trevanion, A. C. Tromans, M. d'Urso, D. Verduzco, D. Villasana, L. Waldron, M. Wall, and Q. Wang. The DNA sequence of the human X chromosome. *Nature*, 434 (7031):325–337, Mar. 2005.
- J. Rung, S. Cauchi, A. Albrechtsen, L. Shen, G. Rocheleau, C. Cavalcanti-Proença, F. Bacot, B. Balkau, A. Belisle, K. Borch-Johnsen, G. Charpentier, C. Dina, E. Durand, P. Elliott, S. Hadjadj, M.-R. Jarvelin, J. Laitinen, T. Lauritzen, M. Marre, A. Mazur, D. Meyre, A. Montpetit, C. Pisinger, B. Posner, P. Poulsen, A. Pouta, M. Prentki, R. Ribel-Madsen, A. Ruokonen, A. Sandbaek, D. Serre, J. Tichet, M. Vaxillaire, J. F. P. Wojtaszewski, A. Vaag, T. Hansen, C. Polychronakos, O. Pedersen, P. Froguel, and R. Sladek. Genetic variant near *IRS1* is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nature genetics*, 41(10):1110–1115, Oct. 2009.
- P. J. Sabo, M. S. Kuehn, R. Thurman, B. E. Johnson, E. M. Johnson, H. Cao, M. Yu, E. Rosenzweig, J. Goldy, A. Haydock, M. Weaver, A. Shafer, K. Lee, F. Neri, R. Humbert, M. A. Singer, T. A. Richmond, M. O. Dorschner, M. McArthur, M. Hawrylycz, R. D. Green, P. A. Navas, W. S. Noble, and J. A. Stamatoyannopoulos. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature methods*, 3(7):511–518, July 2006.
- F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, Dec. 1977.

- A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414): 109–113, Sept. 2012.
- K. Sax. The Association of Size Differences with Seed-Coat Pattern and Pigmentation in PHASEOLUS VULGARIS. *Genetics*, 8(6):552, Nov. 1923.
- E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, Mar. 2003.
- E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, July 2005.
- M. A. Schaub, A. P. Boyle, A. Kundaje, S. Batzoglou, and M. Snyder. Linking disease associations with regulatory information in the human genome. *Genome research*, 22(9): 1748–1759, Sept. 2012.
- M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct. 1995.
- W. F. Scherer, J. T. Syverton, and G. O. Gey. Studies on the propagation in vitro of poliomyelitis viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. *The Journal of experimental medicine*, 97(5):695–710, May 1953.

- D. Schmidt, P. C. Schwalie, C. S. Ross-Innes, A. Hurtado, G. D. Brown, J. S. Carroll, P. Flicek, and D. T. Odom. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome research*, 20(5):578–588, May 2010.
- D. Schmidt, P. C. Schwalie, M. D. Wilson, B. Ballester, A. Goncalves, C. Kutter, G. D. Brown, A. Marshall, P. Flicek, and D. T. Odom. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, 148(1-2):335–348, Jan. 2012.
- D. E. Schones and K. Zhao. Genome-wide approaches to studying chromatin modifications. *Nature reviews Genetics*, 9(3):179–191, Mar. 2008.
- E. G. Schulz and E. Heard. Role and control of X chromosome dosage in mammalian development. *Current opinion in genetics & development*, 23(2):109–115, Apr. 2013.
- A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaub-lomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park, and A. Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, June 2013.
- L. Sie, S. Loong, and E. K. Tan. Utility of lymphoblastoid cell lines. *Journal of Neuroscience Research*, 87(9):1953–1959, July 2009.
- S. K. Sieberts and E. E. Schadt. Moving toward a system genetics view of disease. *Mammalian Genome*, 18(6-7):389–401, July 2007.
- H. Skaletsky, T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S.-F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash,

- C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.-P. Yang, R. H. Waterston, R. K. Wilson, S. Rozen, and D. C. Page. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942):825–837, June 2003.
- S. A. Smallwood, H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8): 817–820, Aug. 2014.
- Z. D. Smith and A. Meissner. DNA methylation: roles in mammalian development. *Nature reviews Genetics*, 14(3):204–220, Feb. 2013.
- L. Song and G. E. Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb.prot5384, Feb. 2010.
- L. Song, Z. Zhang, L. L. Graseder, A. P. Boyle, P. G. Giresi, B.-K. Lee, N. C. Sheffield, S. Gräf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb, and T. S. Furey. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research*, 21(10):1757–1767, Oct. 2011.
- H. D. Soule, J. Vazquez, A. Long, S. Albert, and M. Brennan. A human cell line from a pleural effusion derived from a breast carcinoma. *Journal of the National Cancer Institute*, 51(5): 1409–1416, Nov. 1973.
- C. C. A. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing Genome-Wide Association Studies: Sample Size, Power, Im-

- putation, and the Choice of Genotyping Chip. *PLoS genetics*, 5(5):e1000477, May 2009.
- R. J. Spencer, B. C. del Rosario, S. F. Pinter, D. Lessing, R. I. Sadreyev, and J. T. Lee. A boundary element between Tsix and Xist binds the chromatin insulator Ctf and contributes to initiation of X-chromosome inactivation. *Genetics*, 189(2): 441–454, Oct. 2011.
- M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Schöler, E. van Nimwegen, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schübeler. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378):490–495, Dec. 2011.
- J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–9445, Aug. 2003.
- B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, Jan. 2000.
- B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, Feb. 2007a.
- B. E. Stranger, A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, S. Montgomery, S. Tavaré, P. Deloukas, and E. T. Dermitzakis. Population genomics of human gene expression. *Nature genetics*, 39(10):1217–1224, Oct. 2007b.
- B. E. Stranger, E. A. Stahl, and T. Raj. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. *Genetics*, 187(2):367–383, Feb. 2011.

- A. H. Sturtevant. The linear arrangement of six sexlinked factors in *Drosophila*, as shown by their mode of association. *Journal of experimental zoology*, 14(1):43–59, Jan. 1913.
- A. H. Sturtevant, H. J. Muller, and T. H. Morgan. *The Mechanism of Mendelian Heredity*. 1915.
- K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, Aug. 2006.
- Team, R Development Core. *R: A language and environment for statistical computing*. Vienna, Austria, 2005.
- J. L. Thorvaldsen, J. R. Weaver, and M. S. Bartolomei. A YY1 Bridge for X Inactivation. *Cell*, 146(1):11–13, July 2011.
- R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutayavin, B. Lajoie, B.-K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, Sept. 2012.
- L. Tsui, M. Buchwald, D. Barker, J. Braman, R. Knowlton, J. Schumm, H. Eiberg, J. Mohr, D. Kennedy, N. Plavsic, and a. et. Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science*, 230(4729):1054–1057, Nov. 1985.

- A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S. M. Johnson. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome research*, 18(7):1051–1063, July 2008.
- J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin,

- J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Rendon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazarova, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291 (5507):1304–1351, Feb. 2001.
- J.-B. Veyrieras, S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, and J. K. Pritchard. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS genetics*, 4(10):e1000214, Oct. 2008.
- J. Vierstra, H. Wang, S. John, R. Sandstrom, and J. A. Stamatoyannopoulos. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nature methods*, 11(1):66–72, Jan. 2014.

- P. M. Visscher. Sizing up human height variation. *Nature genetics*, 40(5):489–490, May 2008.
- C. H. Waddington. *The epigenetics of birds*. 1952.
- C. H. Waddington. *The strategy of the genes*. 1957.
- J. A. Wallace and G. Felsenfeld. We gather together: insulators and genome organization. *Current opinion in genetics & development*, 17(5):400–407, Oct. 2007.
- J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22(9):1798–1812, Sept. 2012.
- K. Wang, W.-D. Li, C. K. Zhang, Z. Wang, J. T. Glessner, S. F. A. Grant, H. Zhao, H. Hakonarson, and R. A. Price. A Genome-Wide Association Study on Obesity and Obesity-Related Traits. *PloS one*, 6(4):e18939, Apr. 2011.
- Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics*, 10(1):57–63, Jan. 2009.
- J. D. Watson and F. H. Crick. The structure of DNA. *Cold Spring Harbor symposia on quantitative biology*, 18:123–131, 1953.
- Wellcome Trust Case Control Consortium, Australo-Anglo-American Spondylitis Consortium (TASC), P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, J. A. Todd, P. Donnelly, J. C. Barrett, D. Davison, D. Easton, D. M. Evans, H.-T. Leung, J. L. Marchini, A. P. Morris, C. C. A. Spencer, M. D. Tobin, A. P. Attwood, J. P. Boorman, B. Cant, U. Everson, J. M. Hussey, J. D. Jolley, A. S. Knight, K. Koch, E. Meech, S. Nutland, C. V. Prowse,

H. E. Stevens, N. C. Taylor, G. R. Walters, N. M. Walker, N. A. Watkins, T. Winzer, R. W. Jones, W. L. McArdle, S. M. Ring, D. P. Strachan, M. Pembrey, G. Breen, D. St Clair, S. Caesar, K. Gordon-Smith, L. Jones, C. Fraser, E. K. Green, D. Grozeva, M. L. Hamshire, P. A. Holmans, I. R. Jones, G. Kirov, V. Moskvina, I. Nikolov, M. C. O'Donovan, M. J. Owen, D. A. Collier, A. Elkin, A. Farmer, R. Williamson, P. McGuffin, A. H. Young, I. N. Ferrier, S. G. Ball, A. J. Balmforth, J. H. Barrett, T. D. Bishop, M. M. Iles, A. Maqbool, N. Yuldasheva, A. S. Hall, P. S. Braund, R. J. Dixon, M. Mangino, S. Stevens, J. R. Thompson, F. Bredin, M. Tremelling, M. Parkes, H. Drummond, C. W. Lees, E. R. Nimmo, J. Satsangi, S. A. Fisher, A. Forbes, C. M. Lewis, C. M. Onnie, N. J. Prescott, J. Sanderson, C. G. Matthew, J. Barbour, M. K. Mohiuddin, C. E. Todhunter, J. C. Mansfield, T. Ahmad, F. R. Cummings, D. P. Jewell, J. Webster, M. J. Brown, M. G. Lathrop, J. Connell, A. Dominiczak, C. A. B. Marcano, B. Burke, R. Dobson, J. Gungadoo, K. L. Lee, P. B. Munroe, S. J. Newhouse, A. Onipinla, C. Wallace, M. Xue, M. Caulfield, M. Farrall, A. Barton, Biologics in RA Genetics and Genomics Study Syndicate (BRAGGS) Steering Committee, I. N. Bruce, H. Donovan, S. Eyre, P. D. Gilbert, S. L. Hilder, A. M. Hinks, S. L. John, C. Potter, A. J. Silman, D. P. M. Symmons, W. Thomson, J. Worthington, D. B. Dunger, B. Widmer, T. M. Frayling, R. M. Freathy, H. Lango, J. R. B. Perry, B. M. Shields, M. N. Weedon, A. T. Hattersley, G. A. Hitman, M. Walker, K. S. Elliott, C. J. Groves, C. M. Lindgren, N. W. Rayner, N. J. Timpson, E. Zeggini, M. Newport, G. Sirugo, E. Lyons, F. Vannberg, A. V. S. Hill, L. A. Bradbury, C. Farrar, J. J. Pointon, P. Wordsworth, M. A. Brown, J. A. Franklyn, J. M. Heward, M. J. Simmonds, S. C. L. Gough, S. Seal, Breast Cancer Susceptibility Collaboration (UK), M. R. Stratton, N. Rahman, M. Ban, A. Goris, S. J. Sawcer, A. Compston, D. Conway, M. Jallow, M. Newport, G. Sirugo, K. A. Rockett, S. J. Bumpstead, A. Chaney, K. Downes, M. J. R. Ghorri, R. Gwilliam, S. E. Hunt, M. Inouye, A. Keniry, E. King, R. McGinnis, S. Potter,

- R. Ravindrarajah, P. Whittaker, C. Widdén, D. Withers, N. J. Cardin, D. Davison, T. Ferreira, J. Pereira-Gale, I. B. Hallgrímsson, B. N. Howie, Z. Su, Y. Y. Teo, D. Vukcevic, D. Bentley, M. A. Brown, A. Compston, M. Farrall, A. S. Hall, A. T. Hattersley, A. V. S. Hill, M. Parkes, M. Pembrey, M. R. Stratton, S. L. Mitchell, P. R. Newby, O. J. Brand, J. Carr-Smith, S. H. S. Pearce, R. McGinnis, A. Keniry, P. Deloukas, J. D. Reville, X. Zhou, and A.-M. a. Sims. Association scan of 14,500 non-synonymous SNPs in four diseases identifies autoimmunity variants. *Nature genetics*, 39(11):1329–1337, Nov. 2007.
- H.-J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, A. Zhernakova, D. V. Zhernakova, J. H. Veldink, L. H. Van den Berg, J. Karjalainen, S. Withoff, A. G. Uitterlinden, A. Hofman, F. Rivadeneira, P. A. C. t. Hoen, E. Reinmaa, K. Fischer, M. Nelis, L. Milani, D. Melzer, L. Ferrucci, A. B. Singleton, D. G. Hernandez, M. A. Nalls, G. Homuth, M. Nauck, D. Radke, U. Völker, M. Perola, V. Salomaa, J. Brody, A. Suchy-Dicey, S. A. Gharib, D. A. Enquobahrie, T. Lumley, G. W. Montgomery, S. Makino, H. Prokisch, C. Herder, M. Roden, H. Grallert, T. Meitinger, K. Strauch, Y. Li, R. C. Jansen, P. M. Visscher, J. C. Knight, B. M. Psaty, S. Ripatti, A. Teumer, T. M. Frayling, A. Metspalu, J. B. J. van Meurs, and L. Franke. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics*, 45(10):1238–1243, Oct. 2013.
- K. Wetterstrand. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP), 2013. URL <http://www.genome.gov/sequencingcosts>.
- K. Wheatley and R. Gray. Commentary: Mendelian randomization—an update on its use to evaluate allogeneic stem cell transplantation in leukaemia. *International journal of epidemiology*, 33(1):15–17, Feb. 2004.

- H. Wickham. *ggplot2. elegant graphics for data analysis*. Springer-Verlag New York Inc, Aug. 2009.
- C. L. Woodcock and R. P. Ghosh. Chromatin higher-order structure and dynamics. *Cold Spring Harbor perspectives in biology*, 2(5):a000596, May 2010.
- S. Wright. Correlation and causation. *Journal of agricultural research*, 1921.
- L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131–1139, Sept. 2013.
- N. D. Young. QTL mapping and quantitative disease resistance in plants. *Annual review of phytopathology*, 34:479–501, 1996.
- D. Yusuf, S. L. Butland, M. I. Swanson, E. Bolotin, A. Ticoll, W. A. Cheung, X. Y. C. Zhang, C. T. D. Dickman, D. L. Fulton, J. S. Lim, J. M. Schnabl, O. H. P. Ramos, M. Vasseur-Cognet, C. N. de Leeuw, E. M. Simpson, G. U. Ryffel, E. W.-F. Lam, R. Kist, M. S. C. Wilson, R. Marco-Ferreres, J. J. Brosens, L. L. Beccari, P. Bovolenta, B. A. Benayoun, L. J. Monteiro, H. D. C. Schwenen, L. Grontved, E. Wederell, S. Mandrup, R. A. Veitia, H. Chakravarthy, P. A. Hoodless, M. M. Mancarelli, B. E. Torbett, A. H. Banham, S. P. Reddy, R. L. Cullum, M. Liedtke, M. P. Tschan, M. Vaz, A. Rizzino, M. Zannini, S. Fietze, P. J. Farnham, A. Eijkelenboom, P. J. Brown, D. Laperrière, D. Leprince, T. de Cristofaro, K. L. Prince, M. Putker, L. del Peso, G. Camenisch, R. H. Wenger, M. Mikula, M. Rozendaal, S. Mader, J. Ostrowski, S. J. Rhodes, C. Van Rechem, G. Boulay, S. W. Z. Olechnowicz, M. B. Breslin, M. S. Lan, K. K. Nanan, M. Wegner, J. Hou, R. D. Mullen, S. C. Colvin, P. J. Noy, C. F. Webb, M. E. Witek, S. Ferrell, J. M. Daniel, J. Park, S. A. Waldman, D. J. Peet, M. Taggart, P.-S. Jayaraman, J. J. Karrich, B. Blom, F. Vesuna, H. O’Geen, Y. Sun, R. M. Gronostajski,

- M. W. Woodcroft, M. R. Hough, E. Chen, G. N. Europe-Finner, M. Karolczak-Bayatti, J. Bailey, O. Hankinson, V. Raman, D. P. LeBrun, S. Biswal, C. J. Harvey, J. P. DeBruyne, J. B. Hogenesch, R. F. Hevner, C. Héliçon, X. M. Luo, M. C. Blank, K. J. Millen, D. S. Sharlin, D. Forrest, K. Dahlman-Wright, C. Zhao, Y. Mishima, S. Sinha, R. Chakrabarti, E. Portales-Casamar, F. M. Sladek, P. H. Bradley, and W. W. Wasserman. The transcription factor encyclopedia. *Genome Biology*, 13(3):R24, 2012.
- H. Zhang, L. Gao, J. Anandhakumar, and D. S. Gross. Uncoupling Transcription from Covalent Histone Modification. *PLoS genetics*, 10(4):e1004202–e1004202, Apr. 2014.
- K. Zhang, J. B. Li, Y. Gao, D. Egli, B. Xie, J. Deng, Z. Li, J.-H. Lee, J. Aach, E. M. Leproust, K. Eggan, and G. M. Church. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nature methods*, 6(8):613–618, July 2009.