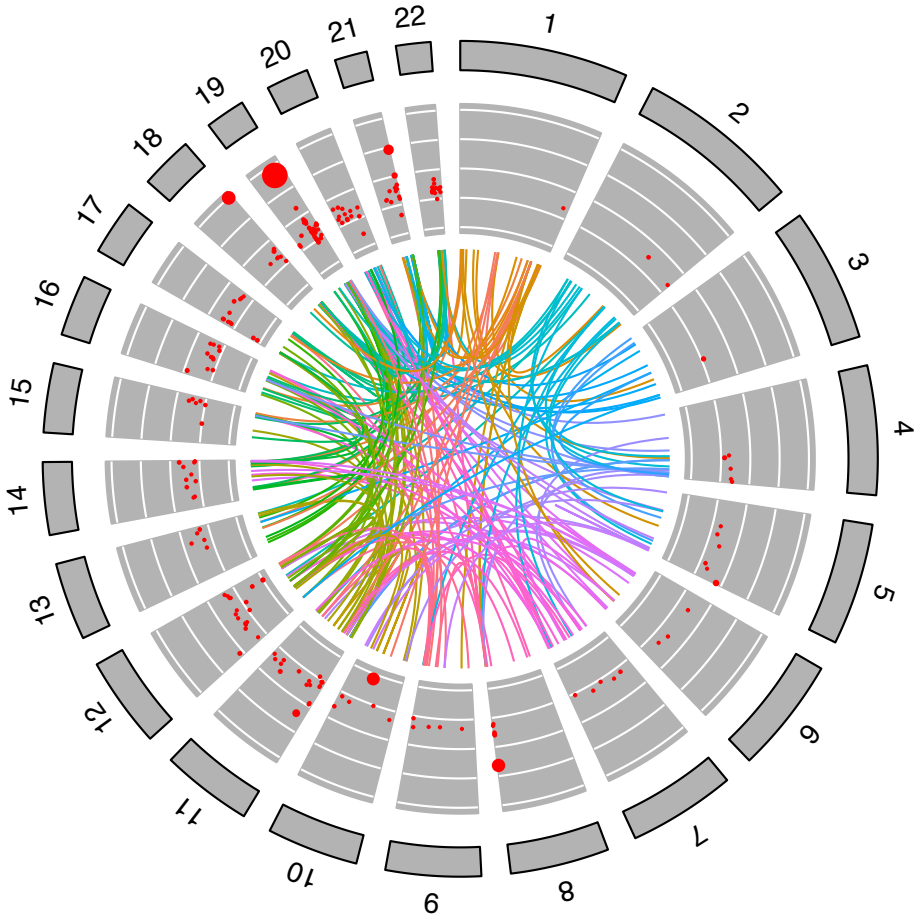
Technical University of Denmark

Master's Thesis in Computational Biology

Spatial and Genetic Interactions Influencing Human Gene Expression

Author:
Pascal Timshel

Supervisors:
Henrik Bjørn Nielsen, PhD
Tune H. Pers, PhD



Center for Biological Sequence Analysis
Department of Systems Biology
Technical University of Denmark

Hirschhorn Laboratory
Harvard Medical School
Broad Institute of MIT and Harvard

November 2015

DTU CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

BROAD INSTITUTE

HARVARD MEDICAL SCHOOL

Technical University of Denmark
Department of Systems Biology
Center for Biological Sequence Analysis
Kemitorvet, Building 208
2800 Kongens Lyngby, Denmark
Phone +45 45 25 24 77
www.cbs.dtu.dk
cbs@cbs.dtu.dk

Danish title:
Spatiale- og genetiske interaktioners indflydelse på human genekspression

Abstract

Genome-wide association studies (GWAS) successfully identified thousands of loci harbouring genetic variants associated with common human diseases and traits. GWAS has elucidated novel biological disease mechanisms, but have so far been limited to single-locus associations. Single-nucleotide polymorphisms (SNPs) are often modeled as having independent additive effects on the phenotype, neglecting any interaction effects between SNPs. Epistasis (that is, the statistical interaction between SNPs) has been demonstrated to prevail in model organisms such as yeast, but results have been limited in human studies.

Our knowledge of the mechanisms that can cause epistasis is still very incomplete. Arguably, large-scale detection of epistasis has previously been too technically challenging owing to statistical and computational issues - the combinatorial explosion of statistical tests becomes intractable when considering a genome scan of all two-loci interactions. Thus, efficient computational methods using *a priori* biological knowledge are needed to constrain the search space for epistasis.

Recent advances in chromosome conformation capture techniques, such as the Hi-C, have generated 3D maps of the human genome. With this a new paradigm in genome biology has emerged: genomes are organized around gene regulatory factors that govern cell identity. This has left a hitherto unexplored territory of relationships between spatial and genetic factors.

My thesis aims to uncover regulatory molecular mechanisms that can cause epistatic interactions influencing human gene expression. I hypothesize that physical proximity of interacting genomic regions provides a spatial scaffold to identify genetic interactions in human genotyping data. Specifically, I focus on the question whether spatial interacting genomic loci do enrich for epistasis. To the best of my knowledge, this is the first work that connects 3D genome maps and genetic data to guide the search for epistasis.

By leveraging data integration of genome-wide genetic, transcriptomic and Hi-C data, I developed a novel computational framework for performing enrichment analysis of "spatial epistasis" using *a priori* biological knowledge. I argue that the framework can be generalized to support more advanced statistical models of epistasis and, more importantly, alternative biological priors to drive the search for interacting SNPs.

The hypothesis was tested under different scenarios of spatial epistasis, including genome structures from multiple cell lines. For one of the cell lines tested, I identified enrichment of spatial epistasis. However, the signal was eliminated, after more thorough quality control of the data. I conclude that this study was underpowered to detect epistasis at this scale. Nevertheless, these results does not support the existence of prevailing epistasis influencing human gene expression. It remains inconclusive whether spatially interacting genomic regions are enriched for epistasis, and ultimately, there is no evidence to definitively reject or accept the spatial epistasis hypothesis.

Resumé

Såkaldte ”genome-wide” associations studier (GWAS) har med stor succes identificeret tusindvis af loci husende genetiske varianter associeret med udbredte humane sygedomme og egenskaber. GWAS har opdaget nye biologiske sygdomsmekanismer, men har indtil videre været begrænset til enkelt-locus associationer. Enkeltnukleotid-polymorfier, eller SNPs, modelleres ofte som uafhængige additive effekter på fænotypen, hvor man ser bort fra genetiske interaktioner. Genetiske modeller af epistasis (statistisk interaktion mellem SNPs) er ofte fremhævet som mere biologisk realistisk, og dog har denne type model haft begrænset succes med at detektere epistasis. En mulig forklaring på den udeblevne succes er, at omfattende søgning efter epistasis førhen har været for teknisk krævende på grund af de statistiske og beregningsmæssige udfordringer.

Dette speciale har til formål at opdage regulatoriske molekylære mekanismer, der kan forsage epistatiske interaktioner med indflydelse på human genekspression. Jeg opstiller hypotesen om “spatial epistasis”, hvor den fysiske afstand af interagerende genetiske loci udformer en spatial platform, der kan benyttes til at identificere genetiske interaktioners indflydelse på human genekspression. Jeg tester specifikt hypotesen om, at spatial interagerende genetiske loci er beriget for epistasis. Ud fra min viden er dette arbejde det første til at koble genomets arkitektur og genetisk data til at guide søgning efter epistasis.

Ved at benytte dataintegration af human genotype, genekspression og lokal 3D foldning af kromosomer, udvikles et værktøj til at udføre spatial epistasis berigelsesanalyse. Analyserne udføres i flere cellelinjer uden at finde robust tegn på epistasis. Det negative fund kan med stor sandsynlighed tilskrives manglende statistisk styrke til at detektere epistasis, grundet det beskedne antal individer inkluderet i dette studie. Dog antyder disse resultater at genetiske interaktioners indflydelse på human genekspression ikke er omfattende udbredt. I sidste ende forbliver det usikkert hvorvidt spatial interagerende genetiske regioner beriger for epistasis, eftersom der ikke er evidens til endegyldigt at forkaste eller acceptere spatial epistasis hypotesen.

Preface

"We used to have these boundaries of disciplines, the chemistry department in the chemistry building, the biology department in the biology building, the math department in the math building, the computer science department in the computer science building. The young students coming in today, they have no respect for these boundaries - and nor they should they. They just munge together... and that is the right way to be .... People are now exploring the fusion cuisine that comes out across all these different disciplines. There has never been a more exciting time to be a scientist."

Eric Lander, Founder of the Broad Institute

Having studied the interaction of biology, statistics and computer science for the preceding five years, it has been my privilege to work on this project. The project is a great example of why I chose to work in the field of computational biology. Soon after having started at the university as biotechnology student, I became interested in systems biology and bioinformatics. It seemed so natural to approach the complexity of biological systems holistically and try to comprehend and predict the way that the components of these systems interact*. I was amazed and allured by the fact that computational biologists could do experiments in front of their computers. I quickly found myself spending more and more time on statistics and computer science as a "Data Scientist", to be able to explore and integrate the flood of biological information ("Big Data") available today†.

My first work within bioinformatics was centered around "omics" analysis of microorganisms. Although this was learningful experience, I wanted to venture out in the complexity of analyzing human "omics" data, though I had no idea the genetics would play such great role in my coming work‡.

I feel that there could not be a better way to work within systems biology and genomics, than by studying epistasis influencing gene expression. The scale and complexity of the problem is vast and the project have enabled me to apply and extend my scientific skills in computational biology.

* I was particularly hooked on the field of systems biology after reading the intriguing paper by Lazebnik (2002) and the book by Alon (2006). † I apologize for using these two buzzwords in one sentence, but I could not resist after having read "Data Scientist: The Sexiest Job of the 21st Century" by Davenport and Patil (Harvard Business Review, 2012) ‡ My mom, a clinical doctor in genetics, was thrilled by my choice of this new field. I can definitely see the irony in this situation and relate to the saying "the acorn never falls far from the tree".

Acknowledgements

First and foremost I would like to thank Tune H. Pers for making my stay in Boston a truly unforgettable experience - both professionally and personally. He helped me make my dream of going to Boston come true and he made it a dream being there. I have learned a lot through his mentorship and friendship.

Secondly, I would like to thank Joel Hirschhorn for welcoming me in his lab. Without his sharp insights and helpful critique during this project, I would have lost contact with the important scientific aspects. Joel inspires all his lab members to do better science - something I hope to live up to in my future scientific career. I would also like to thank the rest of the Hirschhorn lab members for their kindness and assistance with scientific issues. Especially I owe great thanks to Tõnu Esko for his friendship and providing me with data for the project (and smokey scotch whisky); "Rigel" Yingleong Chan for our many humorous and stimulating scientific discussions, and helping me apprehend the basics of human genetics; Christina M. Astley for generating many of the good ideas for this project and providing biological insights to my analysis (and not to forget, where and how to get good coffee in Boston). Working at vibrant environment of Broad Institute of MIT and Harvard has been an unforgettable experience and immensely intellectually stimulating. It has been a great pleasure and honor to work alongside so many experienced researches and young talents.

I am grateful to people at Broad Information Technology Services for their tireless support and rapid responses to my countless number of late-hour emails. My code would still be running if they had not empowered me with super computing resources.

I would also like to thank my external collaborators that have provided assistance and expert knowledge in domains outside my expertise. Thierry Schüpbach (Swiss Institute of Technology) provided an improved and customized release of the most essential software used in my thesis (FastEpistasis), along with his technical insights of computer architectures to resolve bugs - both which have been instrumental to this project. Ferhat Ay (University of Washington) helped me navigate the fast-paced field of 3D genome architecture and provided pre-processed Hi-C.

Thanks to Henrik Bjørn Nielsen for mentoring me through my honors master programme at DTU. His encouraging comments and care for my well being have been important throughout the last two years.

I also acknowledge the long list of Danish grants that provided financial support to my stay in Boston, so I could concentrate on my research. Especially thanks to the Novo Nordisk for awarding me with their Scholarship Programme.

At last, and most importantly, a special thanks and appreciation to Simone, my loving girlfriend, and the rest of my family for their overbearingness and omnipresent support in all of my life's projects.

Abbreviations

Cohorts

EGCUT	Estonian Genome Center of the University of Tartu

Genetics

MAF	Minor Allele Frequency
QC	Quality Control
eQTL	expression Quantitative Trait Loci
GWAS	Genome-Wide Association Study

Contents

Introduction 1

Human geneticists in the post-genomic era have worked intensively on understanding the genetic code. The advent of high-throughput sequencing technology has resulted in the ability to measure millions of genetic variants from thousands of individuals. These high-dimensional data have paved the way for enhancing our understanding of disease etiology, but our current genetic models have far from embraced full complexity of the genome. In this chapter I motivate a more complete genetic model through the study of epistasis, which will be the main topic throughout this thesis. Finally, I describe the aims and approaches of this thesis.

1.1 Thesis Structure

My thesis consists of 9 chapters. The first two chapters familiarizes the reader with the background theory and literature for this thesis.Chapter 2 presents the definition of epistasis and a general genotype model. The chapter provides the framework for the statistical models used in this thesis for epistasis discovery. Chapter 3 introduces the spatial organization of the human genome along with the chromosome capture conformation experimental technologies. The aim is to give an overview of the fast moving field and most recent discoveries. The 3D map of the human genome will serve as the scaffold for bringing together genotypes and gene expression, into a broader perspective of regulatory biology. In Chapter 4 I explain and justify the "spatial epistasis" hypothesis - the main hypothesis of my thesis. Chapter 5 presents the main methods and data sets used for this work. The implementation of the methods and illustration of the main computational workflow is given in Chapter 6. I have deliberately chosen to keep the technical and algorithmic description of the computational infrastructure to a minimum, even though this constituted a major part of my project (as evident in from my GitHub repository, Appendix C). The main results of this thesis is presented in Chapter 7 and discussed in Chapter 8, followed by a brief outline the possible future directions. Chapter 9 concludes my thesis and summarizes its the main contributions.

My thesis contains a relatively comprehensive appendix. Appendix A contains the supplementary methods and describes selected pre-processing steps of the expression and Hi-C data. Supplementary analysis and results are given in Appendix B. Appendix C contains an overview of the code development in this thesis along with selected source code. A glossary can be found in Appendix D.

Words printed with this special font can be found in the glossary. Occasionally I squeeze in footnotes - for the most part you may just ignore them*.

* As you will learn, I mostly used footnotes to endure writing this thesis by keeping me entertained - I cannot guarantee that you will enjoy them in the same way.

1.2 Uncovering the Epistatic Needles in Genome-Wide Haystacks

Genome-wide association studies (GWAS) have identified a plethora of trait-associated polymorphisms. Genetic models often model single-nucleotide polymorphisms (SNPs) effect on the phenotype as additive and independent main effects. However, it is argued that this is not a realistic biological model and that epistasis (statistical interactions between SNPs) should be included. In this section I will highlight some of the motivations for including epistasis in genetic models and the implications of epistasis.

Firstly, searching for epistasis aligns well with the objectives of GWAS. The aims and efforts of GWAS can be described in two main categories. The first is to use knowledge of the causal variants to elucidate the biological and biochemical pathways that underpin disease - that is, to understand biology. The second is to improve prediction of phenotypic outcomes by leveraging estimated genetic effects. If epistasis contributes notably to the genetic architecture of a trait, then detection of epistasis will be valuable to both of the objectives of GWASs. Epistasis have also perspectives that extends beyond GWAS: the great promise of personal genomics and precision medicine calls for advancements of our current incomplete genetic models (Greene et al., 2010; Moore and Williams, 2009; Weigelt and Reis-Filho, 2014). One step forward is to enhance the detection of epistasis.

Secondly, epistasis may be a key factor in explaining heritability of traits. Most variants identified by GWAS are characterized by weak effects on the phenotype, and explain only a small proportion of trait heritability, leading to question how the remaining, "missing heritability" can be explained (Manolio et al., 2009; Eichler et al., 2010). The mystery of the missing heritability have haunted the field of genetics for more than a decade now (Maher, 2008), and have even been referenced to as the "dark matter of the genome" by NIH director Francis Collins (Collins, 2010) - an analogy to the immense quantities of invisible mass in the universe that astrophysicists have inferred but have struggled to find. The (broad sense) heritability of a trait is defined as the ratio (i) variation attributable to genetics (numerator), to (ii) total variance in a population for that particular trait (denominator) (Visscher, Hill, and Wray, 2008). The predominant view among geneticists has been that the explanation for missing heritability lies in the numerator - that is, in as-yet undiscovered variants (Gibson, 2012). However, recent theoretical work by Zuk et al. (2012) suggests that the absence of epistasis in genetic models can explain most of the missing heritability.

Thirdly, studying epistasis is an important factor for closing the gap between genetics and systems biology (Aylor and Zeng, 2008; Moore and Williams, 2005). Systems biologists consider cellular dynamics and development as controlled by interactions between biological components - a portion of these interactions are genetically determined. Human geneticists need better and more sophisticated genome-scale models of the network of genes and regulatory elements, and of how they interact to produce a phenotype. As an example of the complexity and interaction of genetic systems, consider the findings by Westra et al. (2013): expression of hundreds of human genes is controlled by several *trans*-acting genetic

variants.

On a personal note, I think of epistasis as one of the important unsolved problems in genetics. Throughout my involvement in the field of human genetics, I have witnessed dozens of debates around epistasis and its limited evidence. However, relatively little effort has been made to detect epistasis in humans* (Mackay and Moore, 2014). One of the motivating factors for my interest in epistasis was the luring challenge: on the imaginary tree with "genetic fruits", SNPs with linear main effects on the phenotype will be the low-hanging genetic fruits, while SNPs with nonlinear effects will be the high-hanging genetic fruits, hidden in the tree crown (along the other complex genotype-to-phenotype relationships, such as environmental interactions). My hope with thesis is to contribute to the debate and help settle the dispute "epistasis - undiscovered nuggets or fool's gold?".

Then Why Study Human Gene Expression? By now, I hope you have worked up an appetite for the coming Chapter 2 on page 5 about epistasis†. Since this thesis studies epistasis influencing human *gene expression*, I will briefly motivate why expression phenotypes are interesting.

Genetic associations with gene expression have been reported to exhibit very large effect sizes (Powell et al., 2013), making them good candidates to search for epistasis. This is in contrast to the small effect sizes observed for most complex diseases.

GWAS have been used to study a wide range of traits and diseases. The results of such efforts are directly interpretable, for example SNPs associated with BMI, may on average add 0.1 BMI units (kg per m^2)(Locke et al., 2015) per BMI-increasing allele or Crohn's disease risk alleles may confer odds ratios of up to 2.66 (Franke et al., 2010). The effect of a genetic variant on a trait or disease has to be mediated by cellular, tissue, and organ phenotypes. Many GWAS variants do not change coding sequences (Gusev et al., 2014), suggesting that SNPs exert phenotypic control by modifying gene expression. Hence, the genetics of gene expression is central to understanding of the genetic basis of complex traits.

1.3 Aims and Approaches

Before you delve deeper into this thesis, I want to clearly state the main contributions, problem definitions and hypothesis put forth in this thesis.

The overarching aim of my thesis is to uncover molecular mechanisms that can cause epistatic interactions. I will focus on how spatial and genetic interactions influence human gene expression. Specifically, my thesis sets out to answer the following questions:

* A simple search in PubMed for the terms "Epistasis" AND "Human" yielded just under 1,650 entries (as of November 2015), but many of these do not relate directly to results in humans but model systems. For comparison, the terms "GWAS" AND "Human" yielded twice the number of entries. † I promise you, it is going to be a royal feast of biological and statistical insights. But please leave room for the dessert on page 47. And maybe, if you are not completely stuffed, help your self at the Appendix A buffet.

1. What are the regulatory molecular mechanisms that cause epistasis? In particular, are the mechanisms responsible for folding of the genome involved in epistatic interactions?
2. Can regulatory molecular mechanisms facilitate identification of epistatic effects in human complex traits and in particular gene expression endophenotypes?
3. To which extent does epistasis influences human gene expression?

I propose a novel computational framework to address these questions, by searching for statistical epistasis while targeting a mechanistic biological hypothesis. To this end, I will use high-dimensional biological data to shed light on regulatory biology. Typically, genetic and expression data have been explored in a linear 1D context, but the advent of 3D maps of the human genome allows for new integrative approaches. I will leverage data integration of genomic 3D maps, genetic and expression data to detect epistasis, and ultimately test the main hypothesis of this thesis described in further into my thesis (Chapter 4).

Epistasis 2

Because epistasis plays such a big part of my thesis* a clear definition of epistasis is needed. This chapter will introduce biological and statistical epistasis, and highlight the difference between the two definitions. Next I will present two models of epistasis. This chapter is concluded with a discussion of the present evidence of epistasis affecting human complex traits.

2.1 Definition of Epistasis

Informally, epistasis means that something different take place when a particular set of alleles from different loci are found together than when they are apart. Although this general definition holds true for all variations of epistasis, there is a need for a more specific introduction. This is evident in the literature of genetics, where there has been confusion over definitions and interpretations of epistasis (Phillips, 2008; Cordell, 2002). This section provides a historical background and presents two different definitions of epistasis: biological and statistical.

2.1.1 Biological Epistasis

The term "epistatic", which translates as "standing upon" (Moore and Williams, 2005), was first used in 1909 by geneticist William Bateson (Bateson, 1909) to describe the phenomenon where an allele at one locus masks the effects of alleles at one or more other loci. Figure 2.1 shows the concept of biological epistasis in the classical sweet pea experiment conducted by Bateson and R. C. Punnett. It is possible to cross two white flowers and recover purple flowers in the offspring. It was during the studies of the non-Mendelian segregation ratios (9:7 ratio of purple to white flowers) resulting from dihybrid crosses of the pea plants, that Bateson coined the word "epistasis" (Miko, 2008). The two genes, C and P, are responsible for producing the purple plant pigment anthocyanin from a precursor molecule. The two gene products operates within a single biochemical pathway, hence mutations in gene C mask the effect of a mutation in gene P; gene C is said to be epistatic to gene P.

A more explicit definition of Bateson's definition of epistasis is shown in Figure 2.2. In this example, the two loci, A and B, influence the hair color trait in mice. Locus A and B have two possible alleles, A or a and B or b, respectively. The figure illustrates that independent of the genotype at locus A, mice with any copies of the B allele have grey coat color, that is allele B is *dominant* to allele b. Similarly, at locus A allele A, producing black coat, is dominant to allele a. More importantly,

* In a meticulous word count I found 164 instances of the word "epistasis". So I encourage you to familiarize yourself with the definition of epistasis before heading any deeper into my thesis.

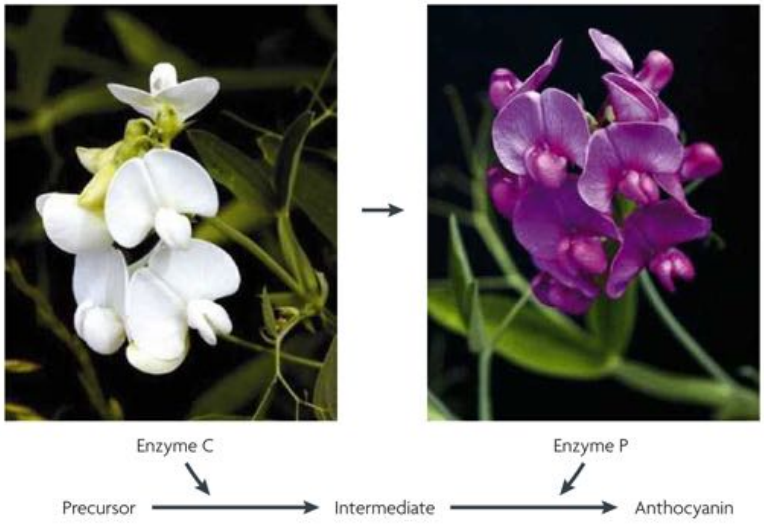


Figure 2.1: Bateson and Punnett classical sweet pea experiment. Illustration taken from Phillips (2008).

if the genotype at locus B is different from b/b, then the effect at locus A is not detectable, i.e. that effect at locus A is masked by that of locus B; locus B is said to be epistatic to locus A.

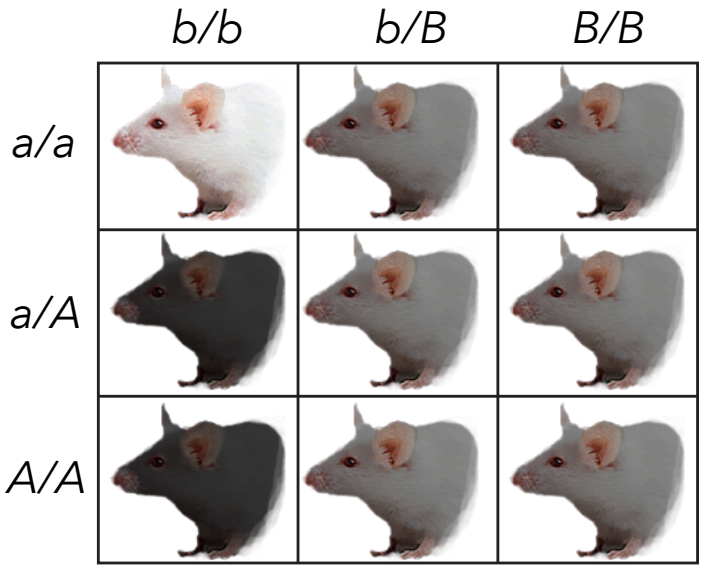


Figure 2.2: Genotype-phenotype map. Example of coat color of mice obtained from different genotypes at two loci. The two loci, A and B are acting epistatically under Bateson's (1909) definition of epistasis.

2.1.2 Statistical Epistasis

In the field of quantitative genetics, R.A. Fisher developed another definition of epistasis, or "epistacy", as Fisher referred to it. Fisher defined epistasis as a departure from the additive expectation of allelic effects (Fisher, 1919). This definition is closer to the conventional concept of statistical interaction: departure from a specific linear model describing the relationship between predictive variables. In Section 2.2 on the next page, the concept of epistasis and statistical models will be treated more in depth.

2.1.3 Biological vs. Statistical Epistasis

It is important to stress that Fisher's definition is not equivalent to Bateson's 1909 definition. The finding of epistasis in the biological context (e.g. from a segregation experiment of sweet pea cross) does not guarantee that there will be epistasis in the statistical sense. Even more troubling, lack of epistasis in the statistical sense does not mean that there are no relevant interactions between loci in the stricter biological sense (Phillips, 2008). Statistical epistasis is a pattern of genotype-phenotype relationships that results from genetic variation in a population, whereas biological epistasis happens at the cellular level in an individual. Bateson's definition is based on the criteria that one locus is "acting" while the other locus "modifies", suggesting asymmetric roles of two interacting loci. In Fisher's definition, the asymmetry between loci is absent.

In the chapters that follows, it should be clear that I adapt the statistical definition of epistasis to detect it (Chapter 6). However, the hypothesis proposed in Chapter 1

is clearly reaches beyond mere statistical associations and seeks to describe the molecular interactions that can cause epistatic interactions. As proposed by Phillips (2008), this meaning of epistasis is best defined as "functional epistasis". However, throughout this thesis, I will use "epistasis" in its broadest sense to refer to the dependence of the outcome of a mutation on the genetic background.

2.2 Models of Epistasis

In this section I introduce the most generic linear genotype model. I will focus on a special case of this model, called the multiplicative genotype model, which will be used for further analyses in this thesis. It is beyond the scope of this text to present vast amount genetic models* that model epistasis. For three excellent reviews on models and detection of epistasis in human complex traits see Wei, Hemani, and Haley (2014), Cordell (2009), and Upton et al. (2015).

2.2.1 Saturated Genotype Model

The saturated genotype model is perhaps the most commonly used parameterization of a two locus genetic model (Cordell, 2002). Table 2.1 illustrates the saturated two locus genotype model for biallelic loci A and B. It is known as a "saturated" model because it is fully parameterized: there are nine two-locus genotype classes, which are modeled by nine parameters. These parameters may be estimated to fit the observed nine two-locus phenotype values exactly. No other model exists that can improve the fit the observed phenotypes. All other models can be considered as sub- or nested models of this general model (Walters, Laurin, and Lubke, 2014, supplementary material). As such, recessive, dominant and allelic multiplicative models are contained within the general model and can be considered special cases of the model. The saturated model provides the best possible fit to the observed data, with the downside that it includes many parameters. In statistics, we are often interested in comparing the full model to its nested sub-models, that is determining whether a model with fewer parameters provide an equally good goodness of fit. For example, to test whether the interaction terms are required at all, we may use the 4 degree of freedom (df) test of interaction ($i_{11} = i_{12} = i_{21} = i_{22} = 0$). To reduce the df and thereby increase power, we may make parameter restrictions to the model (while retaining one or more interaction parameters)(Marchini, Donnelly, and Cardon, 2005). In Section 2.2.2 on the facing page we shall see one example of reducing the df and derive at the multiplicative "allelic" model.

One common issue with the saturated genotype model is that empty cells in the contingency table will invalidate the interaction test because of reduced df. This issues may arise when considering rare variants or small sample sizes, where no individuals have a combined genotype. The problem increases dramatically when modeling higher-order interactions. One solution is to increase the sample size, but

* For example, in the meticulous work by Li and Reich (2000), they arrive at 102 non-redundant biallelic two-locus, two-phenotype, fully penetrant disease models.

this may not be practically possible. Another option is to use assume additivity within a locus, as shown in Section 2.2.2.

Genotype		Locus B		
		b/b	b/B	B/B
Locus A	a/a	α	$\alpha + \gamma_1$	$\alpha + \gamma_2$
	a/A	$\alpha + \beta_1$	$\alpha + \beta_1 + \gamma_1 + i_{11}$	$\alpha + \beta_1 + \gamma_2 + i_{12}$
	A/A	$\alpha + \beta_2$	$\alpha + \beta_2 + \gamma_1 + i_{21}$	$\alpha + \beta_2 + \gamma_2 + i_{22}$

Table 2.1: Saturated two locus genotype model, with the biallelic loci A (alleles a and A) and B (alleles b and B). α represents the "baseline" or reference genotype effect, *aabb*; β_1 and β_2 represent the effects replacing one or both a alleles at locus A with A allele; γ_1 and γ_2 represent the effects replacing one or both b alleles at locus B with B allele; i_{11}, i_{12}, i_{21} and i_{22} are interaction effects.

2.2.2 Multiplicative "allelic" model

Let us now consider a restriction of the saturated genotype model that converts the nine-parameter "genotype" model into a four parameter multiplicative "allelic" model. We assume alleles act additively within a locus, which corresponds to assuming

$$\begin{aligned} \beta_2 &= 2\beta_1 \\ \gamma_2 &= 2\gamma_1 \\ i_{12} &= i_{21} = 2i_{11} \\ i_{22} &= 4i_{11} \end{aligned}$$

The resulting four parameter model in shown in Table 2.2. We may also express the model mathematically by the linear model

$$M1 : y = \alpha + \beta_1 \cdot SNP_A + \gamma_1 \cdot SNP_B + i_{11} \cdot SNP_A \cdot SNP_B \tag{2.1}$$

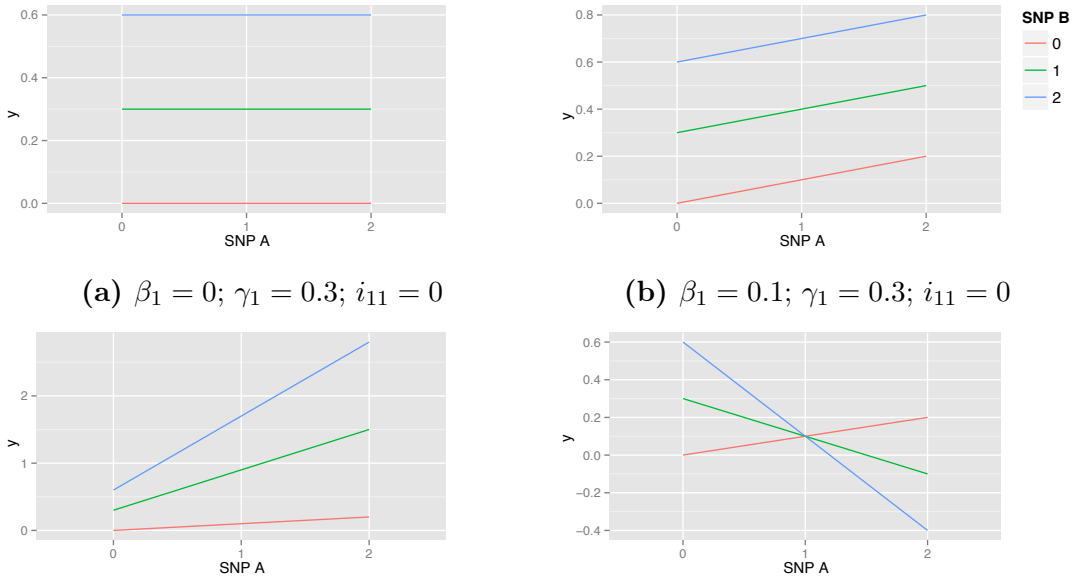
where y is a quantitative phenotype, SNP_A and SNP_B are variables taking values $(0, 1, 2)$ according to the number of copies of risk alleles at locus A and B respectively. Figure 2.3 shows different scenarios of epistasis under the multiplicative model. This model contains a single interaction parameter i_{11} that may be freely estimated. To test for epistasis, we set $i_{11} = 0$,

$$M2 : y = \alpha + \beta_1 \cdot SNP_A + \gamma_1 \cdot SNP_B \tag{2.2}$$

and compare the goodness of fit of $M1$ and $M2$ using a likelihood-ratio test. Depending on the software used for testing for epistasis, alternative implementations of the statistical test may be used (see Section 6.3.1 on page 38). An workflow for fitting the multiplicative model to genotype data using the likelihood-ratio test is presented in Listing C.2 on page 97 (code written in the statistical language `R`).

Genotype		Locus B		
		b/b	b/B	B/B
Locus A	a/a	α	$\alpha + \gamma_1$	$\alpha + 2\gamma_1$
	a/A	$\alpha + \beta_1$	$\alpha + \beta_1 + \gamma_1 + i_{11}$	$\alpha + \beta_1 + 2\gamma_1 + 2i_{11}$
	A/A	$\alpha + 2\beta_1$	$\alpha + 2\beta_1 + \gamma_1 + 2i_{11}$	$\alpha + 2\beta_1 + 2\gamma_1 + 4i_{11}$

Table 2.2: Multiplicative "allelic" model. The biallelic loci A and B have a multiplicative effect, where the presence of homozygous $AABB$ alleles results in a 4-unit effect on the phenotype compared to the heterozygous $AaBb$ alleles. This model has four parameters to fit, as compared to the nine parameter saturated model shown in Table 2.1 on page 9.



(a) $\beta_1 = 0$; $\gamma_1 = 0.3$; $i_{11} = 0$

(b) $\beta_1 = 0.1$; $\gamma_1 = 0.3$; $i_{11} = 0$

(c) $\beta_1 = 0.1$; $\gamma_1 = 0.3$; $i_{11} = 0.5$

(d) $\beta_1 = 0.1$; $\gamma_1 = 0.3$; $i_{11} = -0.3$

Figure 2.3: Multiplicative models of two biallelic loci A and B. The plots in the top panel (**(a)** and **(b)**) have no epistatic term, whereas the plots in the bottom panel (**(c)** and **(d)**) have a multiplicative epistatic effect. Geometrically, multiplicative epistasis corresponds to non-parallel lines. The figures were generated from the model in Equation (2.1) with $\alpha = 0$, β, γ and i_{11} as shown below the plots. See Listing C.1 for source code.

2.3 Evidence of Epistasis

The extent to which epistasis influences human complex traits remains highly debated. The skeptics of epistasis' importance argue that additive genetic effects often account for more than half of the total genetic variance, leaving little room for non-additive epistatic variance (Hill, Goddard, and Visscher, 2008). The skepticism is further fueled by the lack of empirical findings of epistasis. Advocates for epistasis' importance in the genetic architecture of complex traits, point to the evidence of pervasive epistasis in model organisms (Mackay, 2014; Carlborg and Haley, 2004). For example, a large-scale study in *Saccharomyces cerevisiae* examined 5.4 million gene-gene pairs and found that $\sim 3\%$ of the interactions have significant effects (Costanzo et al., 2010). Other studies have shown similar results (Brem et al., 2005; Tong et al., 2004).

One likely explanation for the absence of robust epistasis findings in humans, is that the statistical power to detect it is, in principle, much lower than that of detecting main effects. Wei, Hemani, and Haley (2014) identified three primary challenges related to this issue. Firstly, and of most importance, the "curse of dimensionality" implies a higher multiple testing penalty caused by the large search space for e.g. two-locus interactions. Secondly, dependence on high linkage disequilibrium* between the causal and observed variants is much higher for genetic interactions, compared to additive effects. Considering a pair of variants where none of the causal SNPs are genotyped, the use of tag-SNPs to detect the true interaction leads to a dramatic drop in statistical power. Thirdly, the increased model complexity of epistatic models requires more parameters to fit, and hence more degrees of freedom when assessing statistical significance of the model terms.

Despite these challenges, some attempts have been made to detect large-scale robust epistasis in humans. Hemani et al. (2014) provided the pioneering large-scale study of epistasis in humans influencing gene expression. Using specialized GPU software (Hemani et al., 2011) and large high performance computing facilities, the authors carried out an exhaustive pairwise epistasis scan, investigating more than 10^{15} SNP-pairs. The authors reported 30 epistatic SNP-pairs affecting 19 different gene transcripts. Since its publication, the study has been widely disputed. In a commentary, Wood et al. (2014) provided an alternative explanation of the "apparent epistasis", reporting that all of the epistatic variants could be explained by the presence of a single untyped variant in linkage disequilibrium with one of the epistatic SNPs. Almost all the non-additive genetic variation were eliminated when adjusting for the effect of this causal variant. In summary, at present there remains little hard evidence for prevailing epistasis affecting complex traits in humans.

* Linkage disequilibrium (LD) is the non-random association of alleles at different loci. A simple measure of LD is the correlation between alleles.

Spatial Organization of the Human Genome 3

The discovery of the DNA double helix (Watson and Crick, 1953) unlocked the basic structure of DNA. DNA need extremely tight packaging in order to fit within the nucleus, while at the same time DNA stay accessible to various proteins, to facilitate biological processes such as replication and transcription.
We often think of the genome as a linear object, but in reality chromosomes are folded in a highly complex fashion. For example, many cis-acting regulatory elements (e.g. enhancer elements) are located distal to their promoter targets, but need to come within close physical proximity via folding and looping.
It is becoming increasingly clear that apart from enabling compactness of the genetic material, the 3D structure of the genome is also important for regulating gene expression. Hence, understanding the spatial organization of the human genome is critical for understanding regulatory mechanisms within the cell. With the advent of high-throughput sequencing technology, we can now study genome-wide chromosome conformation using a technique called Hi-C*.

In this chapter I will familiarize the reader with the most basic concepts of chromatin organization and give a brief overview of experimental techniques used to explore the chromatin structure. I will discuss the folding of the genome and its role in regulating gene expression. My hope is you will learn some new surprising facts about the topology of the genome†. It is beyond scope of this text to give a comprehensive review of the plethora of exciting biological phenomenons the Hi-C technique has already unlocked. To this end, I highly recommend the excellent reviews by Gibcus and Dekker (2013), Dekker, Marti-Renom, and Mirny (2013), Fraser et al. (2015), and Sexton and Cavalli (2015).

3.1 Experimental methods

In the past decade, with the introduction and development of Chromosome Conformation Capture (3C) technologies, it has been possible to get insight into how the genome folds by interrogating physical interactions within the genome.

The original 3C technique described in Dekker et al. (2002), was the first molecular assay to study genome organization. In classical 3C experiments the contents of the nucleus are cross-linked with formaldehyde, forming covalent bonds between physically proximal parts of chromatin. The chromatin is then digested with a

* I learned about Hi-C during my stay at the Broad Institute where the technique was developed. It was a mind-boggling experience to learn about this new paradigm in genome biology; like the many other ingenious discoveries by "Broadies". † and maybe even inspire you - just as I was - to test new hypotheses related to the folding of the genome.

restriction enzyme, and the fragments ligated (hereafter referred to as hybrid DNA). The hybrid DNA represents pairwise interactions (physical 3D contacts). Finally the crosslinks are then reversed, proteins are degraded and the DNA is purified. The result is a 3C library consisting of stretches of DNA combining two fragments from two distinct genomic locations. In classical 3C experiments, these interactions are detected by PCR one at the time using locus-specific primers. This technique can be extremely laborious when studying interactions for even a small part of the genome.

More high-throughput 3C-based technologies have since been developed, including Chromosome Conformation Capture-on-Chip (4C) (Simonis et al., 2006), Chromosome Conformation Capture Carbon Copy (5C) (Dostie et al., 2006) and Hi-C (Lieberman-Aiden et al., 2009). These methods use 3C as the principal methodology by which they capture genomic interactions and differ only in the way they detect and quantify the hybrid DNA (Figure 3.1). Hi-C was the first method for unbiased sampling genome-wide physical interactions and hence provide a complete map of the folding genome. There are also noteworthy extensions of Hi-C, such as single-cell Hi-C Nagano et al. (2013), *in situ* Hi-C (Rao et al., 2014) and Capture Hi-C (Mifsud et al., 2015).

Because the data used in this thesis was generated using the Hi-C technique, the remainder of this chapter will focus on this particular flavor of the 3C technologies. The below description follows the original Hi-C protocol published by (Lieberman-Aiden et al., 2009; Berkum et al., 2010) (Figure 3.2).

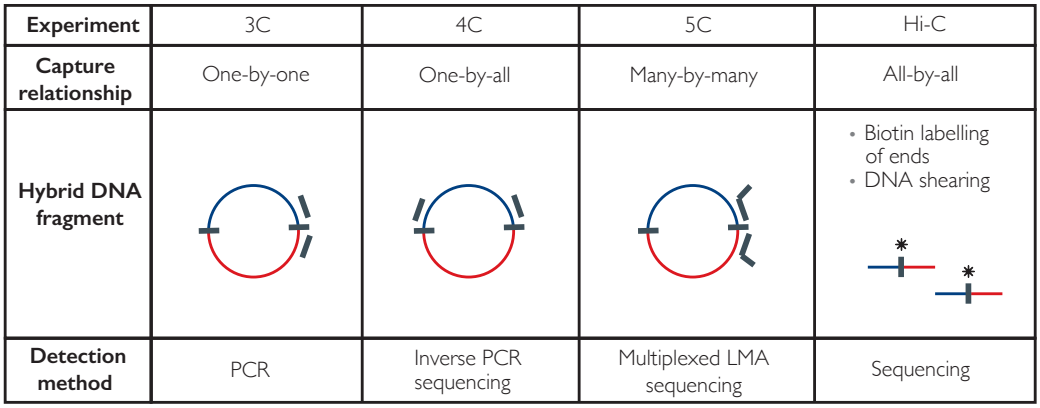


Figure 3.1: Overview of 3C-based techniques. The techniques differ in detection methodology (qPCR, DNA microarrays or DNA sequencing), and the selection criteria and amplification of the hybrid DNA (shown as blue and red circular DNA). Illustration modified from Dekker, Marti-Renom, and Mirny (2013).

Hi-C protocol In Hi-C experiments, the chromatin is treated with formaldehyde to create protein–DNA and protein–protein interactions. The DNA is subsequently digested with the HindIII restriction enzyme. The digested DNA (hereafter referred to as DNA segments) is ligated in the presence of biotin-labeled nucleotides in a diluted environment and then treated with exonuclease to digest linear DNAs but leave DNA loops protected. The DNA is sheared using sonication, and the

biotinylated hybrid DNA fragments (to be distinguished from DNA segments) are pulled down with Streptavidin beads to enrich for DNA fragments containing a ligation junction. The hybrid DNA fragments are PCR amplified using generic primers, and subjected to paired-end sequencing. Each hybrid fragment identifies a potential interaction between two loci based on where the two ends of each fragment are mapped to the genome.

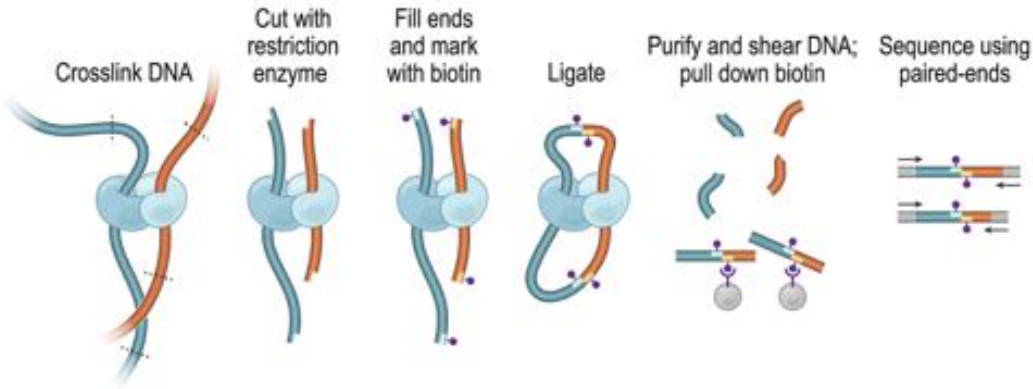


Figure 3.2: The Hi-C protocol published in Lieberman-Aiden et al. (2009). Illustration borrowed from Berkum et al. (2010).

3.2 Hi-C data analysis

Analyzing Hi-C data comes with a unique set of computational challenges that need to be overcome in order to uncover the 3D folding of the genome. In this section I will explain the major data analysis steps and challenges when going from raw sequence reads to interpretable maps of the genome. See Belton et al. (2012), Lajoie, Dekker, and Kaplan (2014), and Ay and Noble (2015) for a more technical and comprehensive review on this topic.

Step 1: Read Mapping
The output of the Hi-C experiment is paired-end sequences from the two ends of the Hi-C ligation products. However, the paired-end mode of standard read alignment algorithms (e.g. Bowtie (Langmead et al., 2009)) cannot be used because they assume that two ends lie relatively close to each other. Hence each side of the paired end read must be mapped separately to the reference genome.

Step 2: Fragment filtering
Read pairs for which both ends successfully pass through the initial mapping filters are further segregated into two categories: informative and non-informative pairs. The genomic alignment location of each mapped read is assigned to one of the restriction fragments. The restriction fragments can be calculated by *in silico* digestion of the reference genome with the restriction enzyme used in the Hi-C experiment. Mapped reads position should fall close to a restriction site. Any read that lies far away from a restriction site cannot be a true ligation product and is subsequently filtered out.

If the read pair maps to the same restriction fragment, it can either represent a self-circularized ligation product or an un-ligated "dangling end" product. Each of these two cases are considered non-informative, and are therefore filtered out. Finally the read pairs are filtered to remove any redundant PCR artifacts.

Step 3: Aggregating and Binning
The maximal resolution of a Hi-C dataset is determined by the restriction enzyme used. The space of all possible interactions, which is surveyed by Hi-C experiments for the human genome, is very large. Using a 6-bp cutting restriction enzyme (e.g. HindIII), there are $\approx 10^6$ restriction fragments*, leading to an interaction space of $\binom{10^6}{2} \approx 10^{12}$ possible pairwise interactions†. Thus, achieving sufficient coverage to support maximal resolution is a significant challenge.
At the cost reducing the resolution, we may increase the effective coverage by aggregating adjacent restriction fragments into fixed-size genomic intervals ("metafragments"). Binning data reduces the interaction space and smooth out noise (Lajoie, Dekker, and Kaplan, 2014) and thus increase the statistical power for calling significant interactions. Typically bins are in the range of 50kb-1Mb, depending on the size of the genome and sequencing depth of the Hi-C experiment (Sugar, 2014).

Step 4: Bias Correction
A challenge for any genome-scale study is experiment-induced biases. Hi-C data is affected by several confounding factors. Some of these biases are shared with other high-throughput sequencing assays, such as RNA-Seq, while some are unique to genome conformation-type assays. The primary technical biases for Hi-C data relate to genomic characteristics such as mappability of sequence reads (sequence uniqueness), fragment length (distance between restriction sites) and GC-content of the ligation products (Yaffe and Tanay, 2011).
Several normalization methods have been proposed to eliminate these biases. The most prominent method to date is ICE (iterative correction and eigenvector decomposition) algorithm (Imakaev et al., 2012). ICE is based on the assumption that all loci should have equal "visibility": since we are interrogating the entire interaction space in an unbiased manner, each fragment/bin should be observed approximately the same number of times in the experiment (coverage) (Imakaev et al., 2012, Supplementary Figure 3). That is, it assumes that all biases manifest in the difference of coverage. The main advantage of this approach is that it implicitly accounts for known as well as unknowns biases (it can be quite difficult to know each and every bias). Figure 3.3 shows the effect of normalizing a Hi-C data to make all loci equally visible. As most of the biases affect the two ends of the interactions *independently*, the expected coverage bias of the *interactions* can be computed as the *product* of the coverage biases for the two ends. The coverage biases is corrected iteratively, where at each iteration the interactions are normalized with the coverage of both ends. Because the normalization procedure changes the coverages at each iteration, it has to be repeated several times, until

* We expect to observe the restriction site in the genome at every $4^6 = 4,096$ bp, giving $3.2 \cdot 10^9 / 4,096 \approx 10^6$ restriction fragments in the human genome † Interestingly, this implies that we are currently only capturing $\approx 0.02\%$ of all possible interactions in a cell. (Of course, we do not expect all genomic loci to interact.) See Table 5.2 on page 34 for the number of pairwise interactions identified in the Hi-C used in this work.

the coverage of all bins converges to 1. The described iterative correction procedure is a matrix balancing problem which has been studied in several decades. The balancing algorithm is called Sinkhorn–Knopp balancing (Sinkhorn and Knopp, 1967) and is guaranteed to converge - although slowly. More efficient and scalable matrix balancing algorithms and decomposition methods have since been developed (Rao et al., 2014, Supplementary Note II; Li et al., 2015).

Step 5: Statistical Analysis
Hi-C data captures long-range interactions, either between loci on the same chromosome (intra-chromosomal) or on different chromosomes (inter-chromosomal). A typical goal of Hi-C data analysis is to identify significant interacting genomic regions. Once the Hi-C data is normalized and the biases are removed, identifying statistically significant inter-chromosomal interactions is relatively simple: without any prior knowledge on the pairwise distances between chromosomes, all possible pairs of inter-chromosomal loci are expected to interact equally under the null hypothesis. This means that the contact count (an integer value) between two loci may directly be used to select interactions (as described in more details in Appendix A.3.2 on page 75). However, there is currently no standard of what contact count threshold that constitutes high-confident inter-chromosomal interactions. In contrast, identifying high-confident inter-chromosomal interactions requires a well-defined statistical model. The genomic distance between the loci influences the number of contacts between two intra-chromosomal loci. This dependence can be attributed to random looping of the DNA and not biological signal from than formation of specific chromatin loops or domains. Hence, it is important to correct for this random polymer looping when assigning statistical significance to the observed contact counts. Several statistical methods have emerged to carry out this task, but here I will focus on the non-parametric method Fit-Hi-C Ay, Bailey, and Noble, 2014b used for processing the Hi-C data in this thesis.

Fit-Hi-C jointly models the random polymer looping effect (or 1D genomic distance bias) and technical biases (GC content, fragment length, mappability etc.). To do this, Fit-Hi-C uses non-parametric smoothing splines to find an initial fit and then iteratively refines the initial fit to account for genuine (non-random) contacts. The non-parametric approach learns almost any distance dependence of contact counts, whereas standard methods uses a fixed power-law decay function (Lieberman-Aiden et al., 2009)). The advantage of the non-parametric fit comes from the non-linear dependence between contact counts and resolution, genomic distance range, and sequencing depth. Fit-Hi-C then computes confidence estimates using the refined fit while incorporating biases computed by the matrix balancing normalization method ICE, as described earlier. The resulting P-values are subsequently subjected to multiple testing correction.

3.3 Genome Contact Maps

After this mapping, filtering and aggregation of the Hi-C data described in Section 3.2, one can construct the genome-wide interaction matrix - also referred to as a contact map. Figure 3.3 shows an example of a contact map from the

human genome. A contact map is a matrix with rows and columns representing non-overlapping regions or "bins" across the genome. Each entry in the matrix contains a count of read pairs (Figure 3.3 left side) or interaction intensity (Figure 3.3 right side) that connect the corresponding bin pair in a Hi-C experiment. The values in the contact map represents reflects the degree of interaction between two genomic bins. The values may be interpreted as the probability of contact between genomic loci, depending on the specific normalization method used to normalize the data.

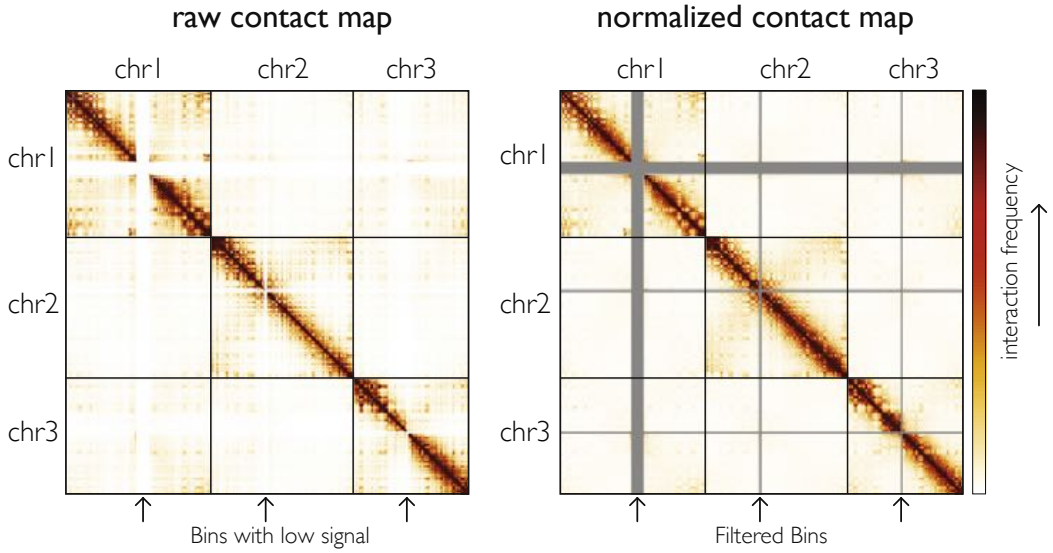


Figure 3.3: Genome contact map. The left contact map shows raw Hi-C data. The right contact map is filtered and balanced Hi-C data. After the balancing procedure, all loci are equally "visible" (sum of each row and column is equal). This results in an overall smoother heatmap. The coloring of the heatmap corresponds to the interaction frequency. The arrows below the maps mark genomic regions ("bins") that have been removed during quality control (filtering). Both contact maps show that chromosomes primarily interact in *cis*, confirming long-established concept of "chromosomes territories". Illustration modified from Lajoie, Dekker, and Kaplan (2014).

3.4 The hierarchy of the 3D genome

One of the major challenges of Hi-C data analysis is to differentiate biological signal from noise and to identify and interpret the interaction patterns. In the remainder of this chapter I will explain the outcomes of going from maps to biological knowledge. It is beyond the scope of this text to explain all the recent discoveries made from contact maps. For a comprehensive presentation of the genome organization see the recent reviews by Gibcus and Dekker (2013), Dekker, Marti-Renom, and Mirny (2013), Fraser et al. (2015), and Sexton and Cavalli (2015) I will focus on three types of patterns observed in mammalian genomes: chromosomal territories, chromosome compartments and TADs. For each pattern, I will explain 1) how it can be identified visually in the contact map (Figure 3.4);

2) how it can be interpreted biologically (Figure 3.5); and 3) how it is defined in terms of scale and hierarchy (Figure 3.7). But before we embark upon the journey into the 3D genome and peer into the architectural structures, let us consider an analogy to the well-known hierarchical organization of proteins.

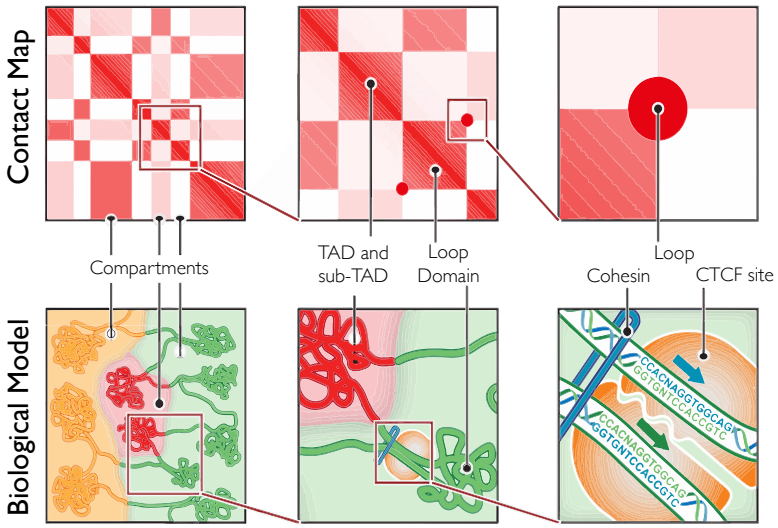


Figure 3.4: Going from contact map to biological model. This cartoon shows the on overview of the genome structures that can be identified from Hi-C contact maps. Left: major contacts patterns identifies large genome structures, such as chromosome compartments or subcompartments. Middle: quadrate elements of enriched contact frequency along the diagonal suggest the presence of domains of condensed chromatin or topological associated domains (TADs). Right: high-frequency adjacent contacts indicate the presence of loops. Loops are enriched at TAD boundaries and anchored at CTCF sites where the proteins cohesin and CTCF facilitate loop formation. Illustration modified from Rao et al. (2014)

Analogy to Protein Hierarchical Structure The genome, analogously to proteins, forms an ordered hierarchical structure comprising primary, secondary, tertiary and quaternary structures. This is a useful starting analogy for understanding the hierarchy of the 3D genome (Figure 3.7, analogy borrowed from Sexton and Cavalli (2015)). The primary structure, comprising sequence elements and nucleosomes; secondary structure, comprising interactions between nearby nucleosomes that shape local chromatin architecture (TADs); and tertiary structure comprising long-range 3D features (chromosome compartments); and quaternary structure comprising the entire genome (chromosomal territories). In the following paragraphs I will explain these structures in more details.

Chromosome Territories The long-established concept of chromosome territories (CTs) constitutes a basic feature of nuclear architecture of chromosomes in interphase (Figure 3.6). Interphase chromosomes adopt a form of territorial organization where *inter-chromosomal* contacts are minimized (Cremer et al., 2006). The first evidence showing that chromosomes occupy distinct, self-associated territories, were provided by light microscopy techniques such as fluorescence *in*

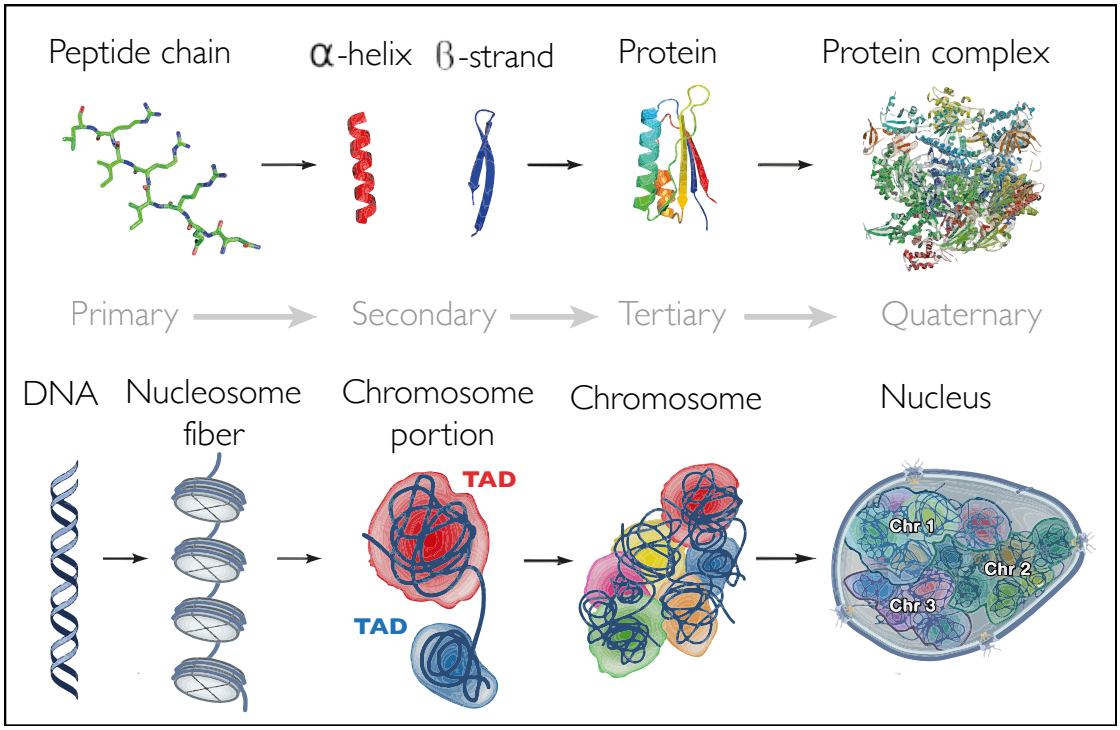


Figure 3.5: Analogous hierarchical organization of protein and genome structure. The genome structure is divided into four size scales, by analogy with protein structure. Primary structures comprising the linear sequence of polymers (peptide chains or nucleotide sequences packaged into a nucleosomal fiber in chromatin) form locally stabilized interactions to fold into secondary structures (α-helices/β-strands or chromatin TADs). These domains in turn hierarchically co-associate to form tertiary structures (proteins or chromosomes). The co-associations of multiple, separately encoded subunits forms the final quaternary structure (protein complexes or genomes). The analogy of protein and genome structure is taken from Sexton and Cavalli (2015). The illustration is inspired by Sexton and Cavalli (2015). Definitions of structure scales follow the sizes from Risca and Greenleaf (2015).

situ hybridization (FISH) (Cremer and Cremer, 2010). In the pioneering study, Lieberman-Aiden et al. (2009) confirmed the presence of CTs, by showing that the observed Hi-C data fits the contact probability distribution of the fractal globule rather than the equilibrium globule. Recently, Hi-C studies, accompanied by computational methods such as HaploSeq Selvaraj et al., 2013, have shown that intra-chromosomal contacts are more frequent than inter-chromosomes, even for loci hundreds of megabases apart on a given chromosome. CTs appear to be in multi-megabase scale ($\sim 1 - 10$ mb) (Dekker, Marti-Renom, and Mirny, 2013) The fact that chromosomes mostly keep to themselves, can be visualized on the contact maps shown in Figure 3.3 on page 18.

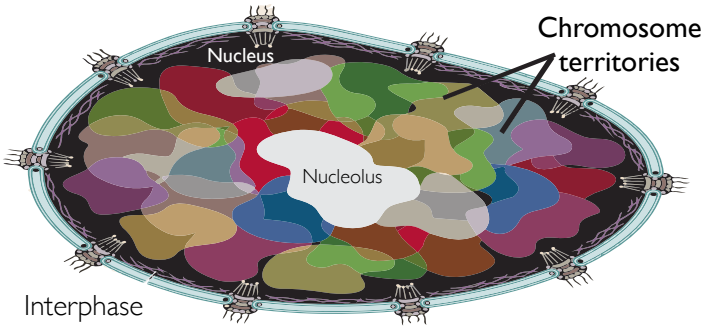


Figure 3.6: The territorial organization of chromosomes in interphase. A cartoon of the 3D nucleus where the human genome is organized into chromosome territories (CTs). Figure modified from Fraser et al. (2015).

Chromosome Compartments Hi-C identified another key architecture of the nucleus called chromosome compartments (Lieberman-Aiden et al., 2009). The two distinct types of chromosome compartments A and B, constitute a spatially segregation of the genome between transcriptionally permissive, euchromatic regions, and transcriptionally inert regions enriched for features of constitutive heterochromatin. The compartments were discovered in the work of Lieberman-Aiden et al. (2009) by using eigenvalue decomposition of the contact map (similar to the map shown in Figure 3.3 on page 18). The genome compartments appear as a "checkerboard-like" interaction pattern on contact maps (Figure 3.4 on page 19). Compartments appear to be in megabase scale (~ 5 mb) (Dekker, Marti-Renom, and Mirny, 2013). The A and B compartments showed enriched interaction frequencies between the same compartment type, and depleted interactions between different types.

Importantly, the genome compartments are closely linked to regulatory and functional partitions of the genome and correlates with indicators of transcriptional activity, such as DNA accessibility, gene density, replication timing, GC content and several histone marks (Dekker, Marti-Renom, and Mirny, 2013). Hence, the compartments are in in line with the notion of euchromatin and heterochromatin.

The compartment signal is not simply biphasic (representing just two states) but is continuous (Dekker, Marti-Renom, and Mirny, 2013). The regions that change the compartment A/B status typically correspond to a single or series of TADs (Figure 3.7). This finding suggests that TADs, which will be described next,

are the units of dynamic alterations in chromosome compartments (Dixon et al., 2015).

Topological Associated Domains Chromatin is organized into megabase-sized topologically associating domains (TADs) that represent the building blocks of genome organization. TADs were first reported by Dixon et al. (2012) and defined as *local chromatin interaction* domains. These TADs appear to be the fundamental modular unit of chromatin organization and describe the spatial neighborhoods of high-frequency chromatin interactions. The boundaries of TADs are largely *conserved* across mammalian species and multiple cell types in the same organism. TADs are among strongest signal in hierarchy of nuclear domains, and its presence of distinct domains can be identified as densely interacting squares on the diagonal of the contact map (Figure 3.4 on page 19). TADs appear to be in multi-kilobase scale (median size $\sim$ 500 kb) (Dekker, Marti-Renom, and Mirny, 2013). It should be noted that slightly different definitions of TADs exists (Rao et al., 2014). This discrepancy is likely caused by differences in the study designs, rather than the biological function of TADs. Variations in the resolution of the Hi-C data and different (hidden Markov) models used to identify TAD boundaries may result in a bias to detect either small- or large-scale domain features.

Several studies have provided evidence for TADs role in regulation genome function. For example, eQTLs and their target genes are enriched within TADs (Duggal, Wang, and Kingsford, 2014), and several epigenetic markers have been found to correlate with TAD boundaries (Rao et al., 2014). Another study observed extensive TAD reorganization during stem cell differentiation, which suggests TADs play an important role in cellular identity Phillips-Cremins et al., 2013; Dixon et al., 2015. Arguably, TADs are one of the most intensive research areas of genome architecture (Gibcus and Dekker, 2013) and important for understanding the physical wiring diagram of the genome.

Architectural Proteins of the Nucleus Although TADs play a central role in genome organization, their linear modularity does not explain alone their spatial organization. Indeed, TAD formations are mediated by both the CCCTC-binding factor protein (CTCF) and general transcription coactivator such as the mediator complex and cohesins (Figure 3.4 on page 19) (Bonora, Plath, and Denholtz, 2014). As such CTCF, cohesin, and mediator act as the "architectural" proteins of the nucleus. CTCF is required for proper cohesin localization to CTCF-enriched sites (Ciabrelli and Cavalli, 2015). A recent elegant experiment used CRISPR inversion of CTCF sites to show that the relative orientations of CTCF binding sites in regulatory elements, such as promoters and enhancers, determine the directionality of DNA looping and regulation of gene expression (Guo et al., 2015). Phillips-Cremins et al. (2013) recently provided evidence of that CTCF plays a critical role in the establishment and maintenance of cellular state through connecting long-range interactions via cohesin. Hence, it is with good reason that CTCF has been dubbed "master weaver of the genome" (Phillips and Corces, 2009).

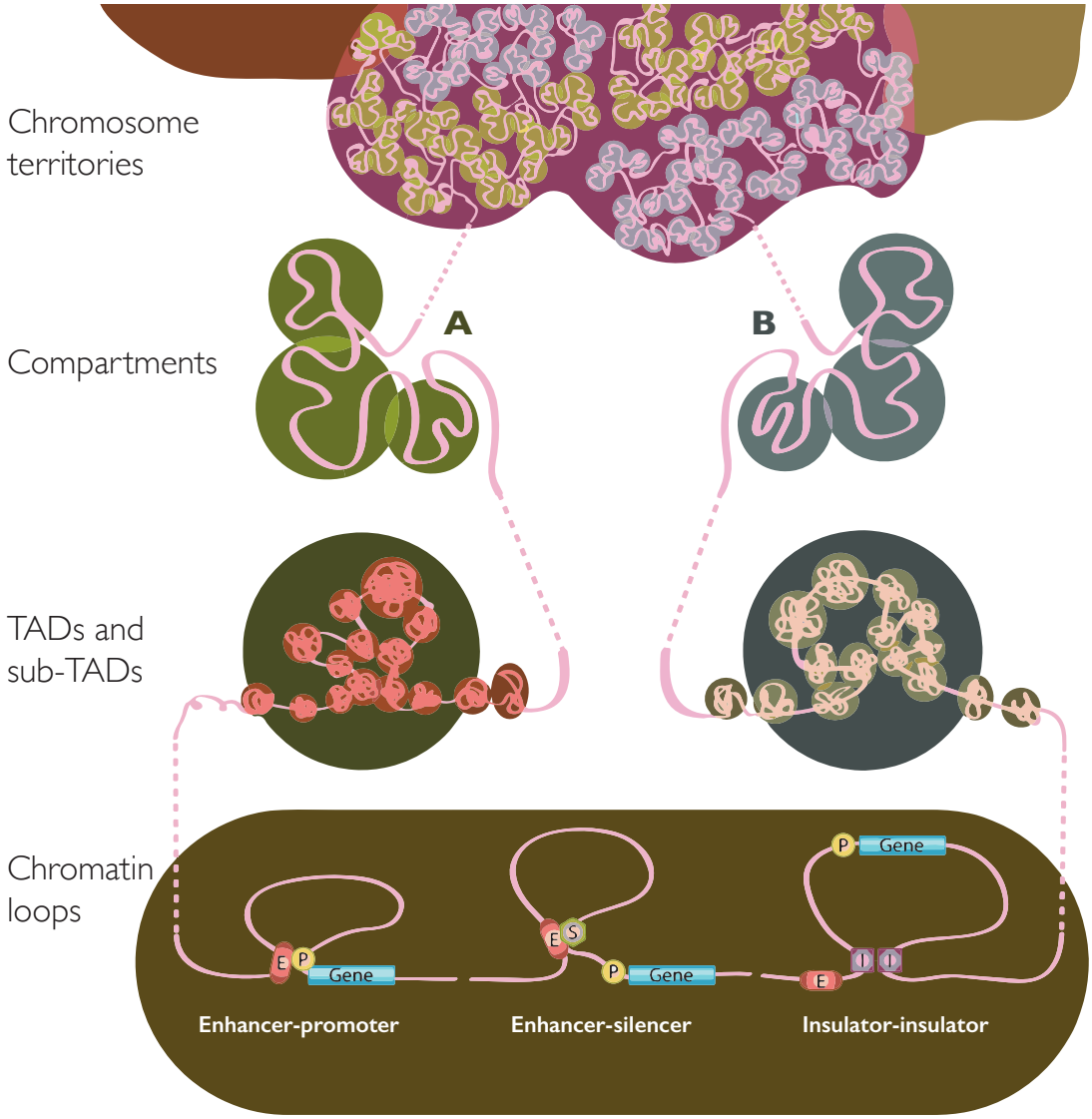


Figure 3.7: 3D organization of the genome. Chromatin conformations across genomic scales illustrate the four different organization levels outlined in the main text. The hierarchical organization of chromatin conformations ordered from low resolution (top) to high resolution (bottom). Chromosome territories, i.e. the domain of a nucleus occupied by a chromosome, are built up from a hierarchy of nucleosome fibers (pink colored band). Chromosomal compartments ($\sim 1 - 10$ mb scale) A and B are shown separately to illustrate their distinguished nature. A/B compartment status typically correspond to a single or series of TADs (mb scale). Three examples of gene regulation in the third dimension are shown: chromatin looping (sub-mb scale) regulate gene expression via long-range interactions amongst enhancers, promoters (left), repressors, silencers (middle) and insulators (right). E, enhancer; P, promoter; S, silencer; I, insulator. Figure modified from Fraser et al. (2015).

3.5 A New Representation of the Human Genome

It is becoming increasingly clear that TADs have a causal role regulating genome function (Dekker, 2008; Hebenstreit, 2013; Zirkel and Papantonis, 2014; Rao et al., 2014; Sexton and Yaffe, 2015). The regulatory and several epigenetic markers have been found to correlate with genome folding. Rao et al. (2014) provokingly proposed that epigenetics is *de facto* genome folding, backed up by the argument "folding drives functions". This raises the chicken and egg problem: is the genome conformation driving gene regulation or vice versa? Are specific genome conformations rather the cause or the effect of gene expression activity? Some evidence suggests the former: for example chromatin loops have been demonstrated to be a cause transcriptional activation rather than a consequence (Deng et al., 2012).

Whatever the answers to these questions are, the new genome conformation technologies have provided us with the tool to answer them. In this new representation of the human genome, we can image that epigenetic markers may no longer be viewed as stacked linear tracks on conventional genome browsers, but could instead be viewed in its local or global 3D genomic context of the spatial relationship of chromosomes or TADs. And even more importantly, to provide a *scaffold* for overlaying or integrating additional high-dimensional biological data. What may appear as a obscure pattern in a linear track genome browser, might form obvious relationships in a spatial representation of the data. I will end this chapter with a quote from a famous German philosopher:

> "And those who were seen dancing were thought to be insane by those who could not hear the music." (Friedrich Nietzsche)

Spatial Epistasis Hypothesis

4

Epistatic interactions are poorly understood at the molecular level. Consequently they are difficult to detect and almost impossible to predict *de novo*. In this chapter I first provide an overview of our current understanding of the molecular mechanisms that can cause epistasis and subsequently motivate new molecular causes of epistasis. I will then put forward the main premise of my thesis and propose the "spatial epistasis" hypothesis. Finally I reason over its biological plausibility. You should find a close connection between this chapter and research aims presented in Section 1.3. Hopefully you appreciated the concepts of epistasis and the spatial organization of the genome introduced in Chapters 2 and 3, because we are about to venture into a new domain brought together by these two concepts.

4.1 Background

There are few examples of replicated large-scale epistatic interactions. One explanation for the lack of empirical findings, is that detecting epistasis is too technically challenging owing to statistical and computational issues. The use of prior biological knowledge to guide the search for epistasis can facilitate discovery of epistatic SNPs (Ritchie, 2011). Layers of biological machinery exist between the genotype-to-phenotype relationships, and imposing this extra dimension of biological knowledge into statistical analyses may help to detect epistasis. Such approaches are beneficial for two reasons: they increase the efficiency when searching for epistasis, and facilitate the biological interpretation of the data. Several other plausible mechanisms for epistasis exist. For example, intra-gene epistasis may result from non-independent effects of mutations on RNA stability and enzyme activity or protein stability (Ortlund et al., 2007), However, detecting intra-gene epistasis (or local epistasis) is complicated by confounding effects of the haplotype and linkage disequilibrium structure (see Wood et al. (2014) for a description of this problem). Inter-gene epistasis may result from the redundancy (or 'buffering') between pathways, structure of metabolic networks or protein interactions (Lehner, 2011). However data from databases in the public domain does not readily support investigating two first mentioned putative drivers of epistasis. However, protein interaction databases have successfully been used to prioritize SNP pairs and increase power to detect statistical epistasis (Pattin and Moore, 2008; Emily et al., 2009). Given that the far majority of GWA-associated loci are located within non-protein-coding regions, and such regulatory variants explain up to 79% of the genetic heritability of diseases and traits (Gusev et al., 2014), a central challenge of association studies is to gain biological insights from regulatory variants. A major limitation of protein interaction-based approaches is that they are constrained to

investigate genetic interactions residing in protein coding regions. Hence these methods fail to investigate the large portion of regulatory variants in intergenic regions.

Göndör and Ohlsson (2009) was the first to suggest that it may be useful to integrate chromosome interactomes when exploring the genetic (and epigenetic) background of complex diseases. Recently, Hemani et al. (2014, Supplementary Figure 15) followed up on this idea by proposing interacting genomic loci as a mechanism underlying epistatic interactions. The authors performed an exhaustive search for epistasis and cross-referenced the epistatic SNPs with a map of chromosome interacting regions and found that 44 out of 501 epistatic interactions mapped within 5 mb (empirical P-value $P = 1.8 \cdot 10^{-10}$). However, the detected epistatic signal could be explained by confounding of unobserved data (Wood et al., 2014), hence the there is currently no support for genome interactions as a driving factor for epistasis.

4.2 Hypothesis

I suggest a mechanism by which biological function can lead to epistatic effects. It has been shown that different chromosomal regions spatially co-localize in the cell through chromatin interactions (Lieberman-Aiden et al., 2009). I hypothesize that physical proximity of interacting genomic regions provides a spatial scaffold to identify genetic interactions in human genotyping data - hereafter referred to as "spatial epistasis". I postulate that spatial epistasis influencing human gene expression is mediated through the spatial proximity of chromatin interactions. Specifically, I formulate the following hypothesis:

Hypothesis: Spatial genomic interactions are enriched for genetic interactions

4.3 Biological Mechanisms

I first consider a general biological model of chromatin crosstalk that may result in epistatic effects on the phenotype. Chromatin loop formation has been demonstrated to be sensitive to specific combinations of SNPs (Steidl et al., 2007). Particular sets of SNPs may influence interactions between different structures of the genome by reinforcing or diminishing loop formation (see Figure 4.1). The specific set of SNPs may have non-additive epistatic effects on the phenotype. Synergistic effects may result in a enhanced interaction network and increased expressivity, whereas antagonizing variants may perturb the network.

I consider two instances of the above mentioned chromatin crosstalk model, that may justify the proposed hypothesis (Figure 4.2. In Chapter 3, I reviewed the involvement of chromatin organization in regulating the genome function. Therefore I will limit this chapter to explaining the basic premise of my hypothesis and reason over its biological plausibility.

Firstly, I speculate that the genome folding and chromatin organization orchestrated by CTCF and other architectural proteins, may drive spatial epistasis (Figure 4.2a). Following the same logic from the general chromatin crosstalk pertubation model, SNPs in the (well-characterized) CTCF binding sites may disrupt CTCF binding, change the genome folding and ultimately change the genome expressivity.

Secondly, I speculate that transcriptional factories may be a causal factor for spatial epistasis (Figure 4.2b). It has been proposed that transcriptional activation is associated with subcompartments termed transcription factories. Co-regulated genes routinely co-express in these nuclear subdomains, which is thought to support simultaneous transcription of functionally related genes (Schoenfelder et al., 2010). The transcription factory is speculated to consists of a protein-rich core with several DNA and RNA binding proteins, such as transcription factors (TFs) and ribonucleoproteins (RNPs) (Eskiw et al., 2008). RNA polymerase II are attached to the surface of the protein-rich core. As illustrated in the cartoon, transcriptional factories provides an opportunity for both inter- and intra-chromosomal chromatin interactions.

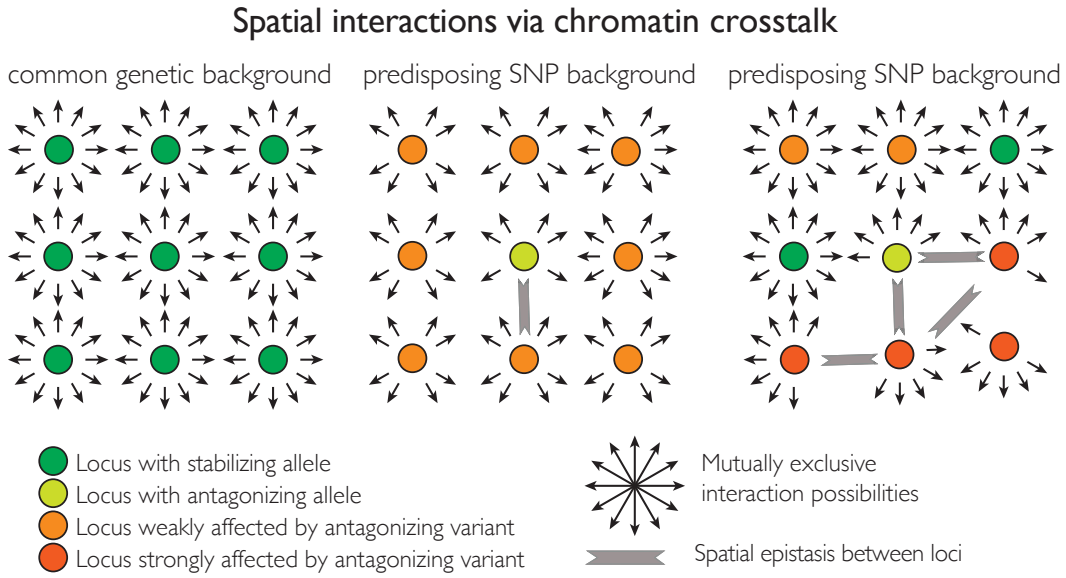
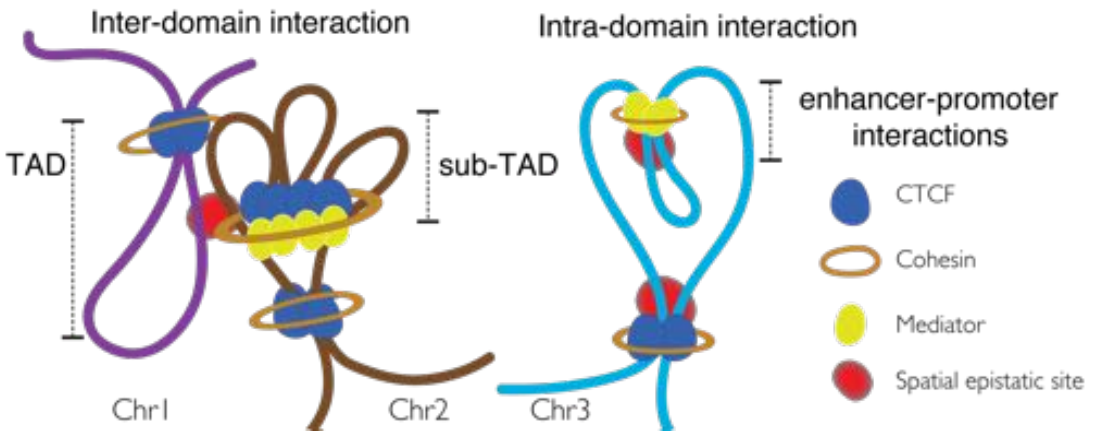
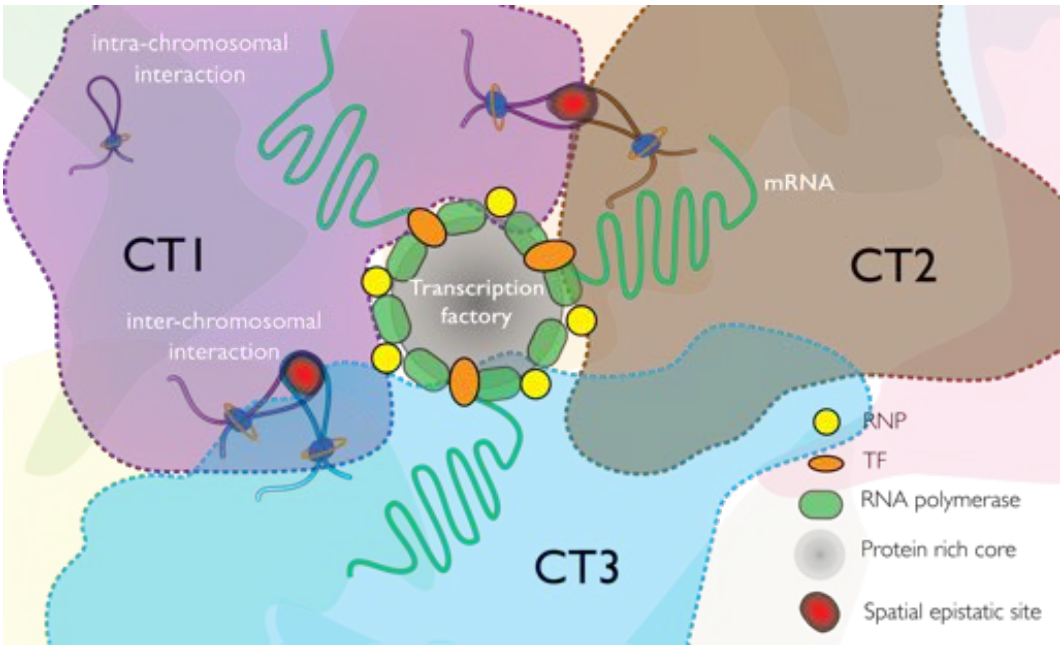


Figure 4.1: Spatial epistasis mediated by chromatin crosstalk. The genetic makeup may regulate the genome expressivity through chromatin interactions (e.g. intra-chromosomal loops or inter-chromosomal bridges). This cartoon shows the chromosomal wiring network of interacting loci (nodes). The network is postulated to tweak transcription through the interplay of genetic variants. On the left side of the illustration, common genetic variants facilitate a fine-tuned regulation of expressivity by allowing many possible chromatin interactions (edges). Changes in genome expressivity may result from genetic variants within loci, that either stabilize or antagonize the chromatin interactions of the network. The middle part of the figure depicts the scenario of antagonizing variants that partially perturb the chromatin crosstalk. If only one direction of the connection is lost between two nodes, the remaining oppositely directed interaction will have a compensatory effect, so no phenotypic changes will be observed. The presence of many antagonizing variants perturb the network and results in a severely altered phenotype (right side of the cartoon). Epistatic effects on the phenotype may arise when the connection is lost in both directions between two nodes. The same logic can be applied to stabilizing variants, that may result in synergistic effects. Illustration adopted from Göndör and Ohlsson (2009).



(a) Chromatin interactions at different scales mediating spatial epistasis. This cartoon shows several possible scenarios for spatial epistasis facilitated by architectural proteins act combinatorially to orchestrate the folding of the genome. The proteins are involved in interactions at different size scales (TADs, sub-TADs and loops). CTCF sites are enriched TAD boundaries (Phillips and Corces, 2009) and specific motif is a potential cause of spatial epistasis. Other factors, such as the cohesin and coactivator mediator complex, participates in the formation of smaller sub-TADs (Kagey et al., 2010) (left). TADs from different chromosomes may interact and form a putative spatial epistasis site (left, inter-chromosomal interaction chr1:2). CTCF and cohesin also act together within TAD boundaries to support enhancer-promoter interactions (right, intra-chromosomal interaction). Putative spatial epistasis sites for intra-chromosomal interaction are shown for both the enhancer-promoter and TAD boundary interactions. Illustration inspired by Bonora, Plath, and Denholtz (2014).



(b) Transcription factories and mediating spatial epistasis. This cartoon illustrates an instance of spatial epistasis where the "spatial scaffold" is provided by a transcription factory. Inter-chromosomal chromatin interactions, or so-called chromatin bridges, are shown between chromosomal territories (CTs) 1:2 and 1:3. CT1 also includes an intra-chromosomal, *cis* acting chromatin loop. All of these sites constitute putative epistatic sites. Notice that in order to interact with other chromosomes, one or both partners of chromatin interactions must reach beyond the confines of its CT. Illustration inspired by Rieder, Trajanoski, and McNally (2012).

Figure 4.2: The spatial epistasis hypothesis implicating spatial proximate chromatin interactions as a driver for epistasis. Putative spatial epistatic sites are shown as red rounded rectangles.

Datasets 5

Computational biology often relies on large quantities of high-dimensional biological data. Correction interpretation of downstream analyses, requires carefully pre-processing and cleaning of the input data. This work is no exception. I used several key data sets to test for enrichment of spatial proximate epistasis. This chapter describe the pre-processing of genotype, expression and Hi-C data. In the subsequent Chapter 6, I describe how these data sets were used in a computational pipeline.

5.1 Genotypes

The Estonian Genome Center of the University of Tartu (EGCUT) Gene Expression Cohort (Metspalu, 2004) is composed of unrelated individuals from the Estonian population. This study used $n = 832$ samples from individuals with both expression and genotype data available. All steps described below, except genotyping and imputation, was carried out by the author.

5.1.1 Genotyping and Imputation

Genotyping and imputation was carried out by researches at Tartu University. Genotyping was performed using Illumina Human370CNV arrays (Illumina Inc., San Diego, US). The data set consisted of 311,563 genotyped SNPs. Imputation was done using IMPUTE2 (Howie, Donnelly, and Marchini, 2009) with the HapMap CEU phase 2 (Frazer et al., 2007) as reference panel (release # 24). All SNP genome positions was relative to the reference human genome NCBI build 37 (UCSC hg19). The data set comprised 2,552,493 SNPs after imputation.

5.1.2 Data Pre-processing

A list of 10 sample mix-ups was corrected. The genotype data was converted from IMPUTE2's probabilistic dosage format (.gen) to hard-called, best guess genotype calls (.ped) using GTOOL (Freeman and Marchini, 2007) with a threshold of 0.9. The genotype data contained purely biallelic single nucleotide polymorphisms. That is, no structural or indel variants.

5.1.3 Quality Control

Quality control filters were applied as suggested by Anderson et al. (2010) and Turner et al. (2011) resulting in 1,779,693 high quality SNPs. QC was conducted

using PLINK 1.9 (Purcell and Chang, 2014) with the parameters shown in Table 5.1.

PLINK argument	threshold	description
`-maf`	0.05	Minor allele frequency
`-geno`	0.1	Per variant missingness
`-mind`	0.1	Per individual missingness
`-hwe`	1e-6	Hardy-Weinberg equilibrium test

Table 5.1: Stardard genotype quality control parameters were used. See Anderson et al. (2010) for details.

Finally, duplicated variants were identified and removed. Duplicate variants were defined as SNPs with identical variant identifier (rsID) or genomic coordinate. 1,779,665 SNPs remained after removal of duplicated variants.

5.2 Gene Expression

The EGCUT Gene Expression Cohort consists of whole peripheral blood samples of $n = 832$ individuals. The gene expression data described in the following sections comes from the same genotyped individuals described earlier (Section 5.1). All steps described below, except the array experiments, was carried out by the author.

5.2.1 Array Platform

Whole-genome gene-expression levels were obtained by Illumina Human HT12v3 arrays (Illumina Inc, San Diego, US) according manufacturers protocols. A subset of the expression dataset is available at GEO (Gene Expression Omnibus) public repository under the accession GSE48348.

Raw expression data was exported from Illumina GenomeStudio. The data was exported without any normalization, transformation, imputation, or background correction method that is offered by Illumina's GenomeStudio. A total of $n = 48,803$ probes where exported.

5.2.2 Data Transformation and PEER Analysis

PEER (Stegle et al., 2012) was used to remove confounding factors from the expression data set. Prior to running PEER gene, expression levels were variance stabilized using a $\log_2$ transformation. Using PEER, we inferred 50 latent factors; four EGCUT genetic MDS components (Figure A.1 on page 61) were included as covariates when factors were constructed. See Appendix A.2.4 on page 66 for a more detailed description of the analysis. The latent factors and covariates were subsequently regressed out using an additive linear model and the resulting gene expression residuals were used for epistasis discovery.

5.2.3 Probe Annotation

The Illumina manifest file was retrieved from Illumina Inc (2014, see download URL). To allow for comparison of results to Hemani et al. (2014), the default Illumina probe identifier `Array_Address_Id` were mapped to `Probe_Id`. Probes with the value `RefSeq` in the "Source" column were marked as RefSeq probes. The values of the column "Symbol" were used for as HGNC symbols. Finally, the genomic coordinates ("Chromosome" and "Probe_Coordinates") and strand orientation ("Probe_Chr_Orientation") for the probes were extracted.

5.3 Hi-C data

The Hi-C data used in this study was derived from two different data sets, spanning three different cell lines. The two data sets was of different resolution and hence processed differently. For both data sets, only inter-chromosomal interactions were retained in the final processing step. For a background description on Hi-C data analysis see Section 3.2 on page 15.

5.3.1 hESC and hIMR90 cell lines

The publicly available Hi-C data sets from human embryonic stem cells (hESC) and human lung fibroblast (hIMR90) published by Dixon et al. (2012) were processed with Fit-Hi-C (Ay, Bailey, and Noble, 2014b) as described in Libbrecht et al. (2015). The Hi-C data used hg19 as reference genome. Here follows a summary of the processing pipeline applied to the data by Libbrecht et al. (2015).

The raw paired-end libraries were processed with a pipeline that merges reads from two replicates per cell line, maps the reads to the reference genome, extracts the read pairs where both read ends map uniquely, and finally excludes PCR duplicates. The human genome was then partitioned into non-overlapping 10 kb windows, where each end of a read pair is assigned to the nearest 10 kb window mid-point, to yield a whole genome contact map. The contact map was sparse with $\approx 0.3\%$ non-zero entries. The contact map comprised both intra- and inter-chromosomal contacts. After eliminating all loci that are less than 50% uniquely mappable, the bias correction method ICE (Imakaev et al., 2012) was applied to the contact map to estimate the bias associated with each 10 kb locus (see Section 3.2 on page 15 for a description of the ICE method). Next, the computed biases and raw contact maps were used as input to Fit-Hi-C to assign statistical confidence to the contact counts (see Section 3.2 on page 15 for more detail on Fit-Hi-C). A P-value of interaction were estimated for each pair of 10 kb loci with non-zero contact counts (the P-value was set to 1 for pairs with zero contacts). Finally, the combined collection of P-values is corrected for multiple testing using the false discovery rate (FDR), which is estimated using the standard Benjamini & Hochberg method. In practice, Fit-Hi-C reports the statistical confidence associated with a given contact as a q-value, which is defined as the minimum FDR attained at or above a given interaction significance (Storey, 2002; Noble, 2009).

Table 5.2 lists the number of Hi-C interactions. This work used various q-value thresholds to define the set of inter-chromosomal interactions considered for epistasis discovery. Because Fit-Hi-C normalizes for 1D genomic distance, the majority of significant contacts were inter-chromosomal or long distance interactions. See Appendix A.3.2 on page 75 for an explorative analysis and visualization of the Hi-C data.

no. interactions	hESC	hIMR90
all*	91,234,032	157,113,166
inter-chromosomal	39,240,868	84,773,727
inter-chromosomal ($q = 10^{-3}$)	2,515,748	1,006,394

Table 5.2: Number of interactions for different subsets of the Dixon et al. (2012) data. *Intra and inter-chromosomal interactions with raw contact count > 0.

5.3.2 K562 cell line

The Hi-C data set described in this section was part of the original Hi-C publication by Lieberman-Aiden et al. (2009). Since its publication, the data set has been widely used and reanalyzed by various methods. I used the list of interacting loci derived from human erythroleukemia cell line K562, published by Lan et al. (2012). In short, Lan *et al.* re-analyzed the K562 Hi-C data from Lieberman-Aiden et al. (2009) using a Mixture Poisson Regression Model. The model identifies significant genomic interaction points using a power-law decay as background contact probability distribution. Focal point estimates of interacting genomic regions were obtained from Lan et al. (2012, Supplementary File 3), comprising 96137 interactions. Genomic coordinates were mapped to hg19 using LiftOver (Min Kang and Abecasis, 2014, Software); 3205 interactions could not be mapped to hg19. 3497 autosomal inter-chromosomal interactions were extracted. The following outlier removal procedure applied: outliers were defined as interactions containing at least one genomic region within the 99.9-th percentile interaction counts. Three genomic regions with interaction count > 96.23 were identified (all from chromosome 5), resulting in exclusion of 976 (14%) interactions. After outlier removal, 1258 (49.9%) duplicate interactions were removed. Duplicate interactions were defined as interactions with identical interacting genomic regions. The final set of genomic interactions comprised 1263 interactions composed of 2021 unique interaction loci. See Appendix A.3.1 on page 72 for more details on the issues of using the K562 cell line and an explorative analysis of the data.

Design Choices and Implementation

6

This chapter describes the main computational pipeline used in my thesis. I will introduce the rationale behind each of the steps in spatial proximate epistasis discovery. I will explain the general concepts and terminology*. I have deliberately chosen not to describe every technical detail but rather give an overview of the design choices. The technical details were fun and challenging to solve. However, my fear of loosing the reader on computer science issues that caused countless of "all-nighters", restraint me from including many tiresome technical and algorithmic sections†. A substantial amount of computational infrastructure underlie the pipeline outlined in this chapter. So when reading this chapter, bare in mind that a significant of part of it is "written" in `python`, but has been hidden away on GitHub (Appendix C on page 95 bear witness to more than 10,000 lines of `python` code).

6.1 Design Principles and Generalizability

Before we approach the implementation of the framework outlined in this chapter, it is important to understand the main design features. The key principle of the framework is to use information from the spatial organization of the human genome, to limit the search space of possible for interactions between SNPs. In Chapter 4, I argued why known chromatin interactions might comprise a reasonable biological prior for detecting genetic interactions. A part from using biological knowledge to drive the search for interacting SNPs, there are three other noteworthy design principles.

Firstly, generalizability is a core feature of the framework. There are two major adaptable components. The first component is the "statistical engine", that is the underlying model used to detect statistical epistasis. Because the framework is "agnostic" to the statistical model of epistasis, the default linear multiplicative 'allelic' model may be substituted with non-linear models or more advanced Bayesian approaches that explicitly models the biological prior (see Section 6.3 for alternative models to detect epistasis). The second component is the biological prior used to select SNP pair candidates. A substituting of the default "spatial proximity" prior need to meet two requirements, in addition to formulating a reasonable biological prior. Firstly, the relationship between the biological entities in the prior and the genetic variants needs a well-defined mapping. E.g. when postulating that

* Remember that definitions of terms printed in this special font can be found in the glossary (Appendix D on page 99) † Please bear with me if I, in a moment of weakness, let one or two nerdy technical sections slip into this chapter.

interacting proteins enrich for epistasis, the mapping is clear: genetic variants within genes can be mapped to the corresponding gene product (proteins, the biological entities of the model). Secondly, in order to assess whether the biological prior enrich for epistasis, a complement set of biological entities needs to be defined. This facilitates the generation of a null distribution, as explained in Section 6.4. Following the protein interaction example, the complement set of biological entities are non-interacting proteins. Clearly, for any biological prior, the most demanding step, which requires careful considerations, is to define the interacting and non-interaction biological entities.

Secondly, the enrichment test is based on an empirical null distribution, as opposed to an analytical definition of a null model, where the underlying assumptions would be difficult to justify. In Section 6.4, I will explain how the empirical null distribution directly controls for confounding factors. The framework is based on a summarized statistic to facilitate the enrichment test. It considers the cumulative effects of SNP-pairs in interacting genomic regions, to test the hypothesis of spatial epistasis enrichment. The second feature is reminiscent of rare variant burden tests (Morgenthaler and Thilly, 2007), based on collapsing or summarizing the rare variants within a region by a single value to increase statistical power. Specifically, the framework is based on summarizing count data within pairs of genomic regions. This principle will be describe further in Section 6.4.3.

Lastly, the framework includes an important feature to model different scenarios of spatial epistasis. Two biological variables are parameterized: the confidence of physical interaction between genomic loci (as discussed in Appendix A.3.2 on page 75), which ultimately represents the frequency or intensity of the interaction; and the range or width of the interaction (this parameter is introduced in Section 6.5). The first parameter enable the study of how strong genomic loci need to interact in order to drive spatial epistasis. The second parameter enables investigations of how much the physical interaction causing spatial epistasis extends from its focal contact point.

6.2 Computational Feasibility

A fundamental challenge for epistasis discovery is computational feasibility. To illustrate this, the single-phenotype exhaustive pairwise epistasis search in individuals with one million genotyped SNPs, would require $\binom{10^6}{2} = 5 \cdot 10^{11}$ tests. If we consider going beyond pairwise models, one million SNPs generate $1.7 \cdot 10^{17}$ three-SNP models. (Consider that this number is three order of magnitude larger than the number of cells in the human body (Bianconi et al., 2013).) This number of tests is computationally intractable without the use of specialized hardware, which is not readily accessible to most research groups*. Hence, the computational complexity needs to be addressed prior to the epistasis discovery to avoid undertaking endless calculations.

* And certainly not a pathetic master student unless... you are that the Broad Institute. The people at Broad IT Services was generous enough to let me use more than 12 years of computing time (see Table 7.1 on page 49). I hope I remembered to mention them in my acknowledgement.

The run time, T, of computing the tests for epistasis is dependent on the number of genetic variants (N_G) and the number of expression traits considered (N_P). T may be expressed using the big O notation as follows:

$$T(N_G, N_P) = \mathcal{O}\left(N_P \cdot N_G^2(l, q)\right) \tag{6.1}$$

where l is the interaction width defining the length of the genomic region considered for epistasis discovery (Figure 6.2 on page 45 (**a**) illustrates this principle of l). q is the q-value threshold used to control the statistical confidence of the Hi-C interactions considered for epistasis discovery (see Appendix A.3.2 on page 75 for a more detailed description of q). Note that N_G is squared because the all pairwise interactions are calculated for SNPs considered for epistasis*.

The next section explains, in practical terms, how each input data type was subsetted and its consequence for the computation time. The downside of subsetting the data will also be discussed.

6.2.1 Data Subsetting

SNP Subsetting Equation (6.1) illustrates the importance of limiting the squared term N_G for computational feasibility. This study does not have the statistical power to detect epistasis between rare variants, as discussed further in Appendix B.3. For this reason, rare variants (MAF $< 5\%$) was removed during quality control of the genotype data. This removed one third of the variants. The interaction width, l, was also used to limit N_G. Values of $l = [500 - 50,000]$bp was used. Interestingly, this parameter has a biological interpretation for the model of spatial epistasis. Small l (narrow interaction width) suggests that spatial epistasis exists in the close vicinity of interacting genomic regions, while large l (wide interaction width) imply a broad region of spatial epistasis.

Hi-C Interaction Subsetting The number of Hi-C interactions that are considered for spatial epistasis discovery ($n_{interactions}$) influences N_G. An important design choice in this framework is that only inter-chromosomal interactions are considered for spatial epistasis, hence the intra-chromosomal interactions are excluded. Firstly, this dramatically reduces the search epistasis for epistasis. Secondly, detecting epistasis between SNPs located on the same chromosome (local epistasis) is complicated by confounding effects of the haplotype and linkage disequilibrium structure†(see Wood et al. (2014) for a further description of this issue). Different q-value thresholds were used to control $n_{interactions}$ in the interval $\approx 1,000 - 40,000$. The q-value threshold can in this context be interpreted as parameter controlling the strength or tightness between spatial epistasis interactions. E.g. a stringent threshold, resulting in few $n_{interactions}$, imply that spatial epistasis exists only between the most strongly interacting genomic regions.

* the bound on all pairwise combinations can be calculated using the binomial coefficient as $\binom{N_G}{2} = \frac{(N_G-1)\cdot N_G}{2} = \mathcal{O}\left(N_G^2\right)$ † Unexplained haplotype and linkage disequilibrium effects were the main criticisms of the Hemani et al., 2014 paper. Although the controversy of their epistasis findings is not yet resolved, it is more convenient to simply avoid local epistasis.

Probe Subsetting A natural way of limiting the expression traits for epistasis detection, is to remove all probes that do not map to well annotated genomic sequences (e.g. the RefSeq database). Analysis containing probes mapping to uncharacterized genomic regions can make downstream biological analysis hard to interpret. More than one third of the probes were discarded because they did not map to the RefSeq database. Probes mapping to the sex chromosomes were discarded because this work was not concerned with any sex specific hypothesis. Most importantly, a variance and mean criteria was used to select probes exhibiting the highest variance and mean expression. High variance/mean probes increases the statistical power to detect epistasis (Westra et al., 2013).

It should be noted that more advanced schemes exists for subsetting and quality control of probes (Becker et al., 2012; Westra et al., 2013). Additional criteria include removing probes containing SNPs to avoid hybridization bias (Alberts et al., 2007); removing probes that do not map over the full length (50 bp) to a contiguous genomic location (i.e. no intron-spanning probes) or removing non-uniquely mapping probes (`nuID` uniqueness score). These schemes were not applied in this study because they are cumbersome to implement relative to their effect on decreasing the number of selected probes.

6.3 Epistasis Search and FastEpistasis

The software used to test for epistasis plays an important role in the design of the study, because large-scale epistasis calculations are highly computational intensive. A plethora of software packages for have been developed to detect epistasis (Wei, Hemani, and Haley, 2014; Moore and Williams, 2015; Upton et al., 2015). The methods represent a wide range of model classes, including simple linear frequentist models (Cordell, 2009), Bayesian models (Zhang and Liu, 2007; Albrechtsen et al., 2007), information theory (Seo, Kim, and Moon, 2003), neural networks (Beam, Motsinger-Reif, and Doyle, 2014) and artificial intelligence (Moore and Hill, 2015). Arguable, the most widely used software for epistasis test is PLINK (Purcell et al., 2007) which uses the multiplicative model described in Section 2.2.2 (Equation (2.1) on page 9). PLINK's implementation of the model poses issues related to scalability and speed; and ultimately prevents its usage in large-scale epistasis discovery.

6.3.1 Model Definition

I here describe the aforementioned "statistical engine" used as the underlying workhorse for computing two-locus tests for epistasis effects. The regression based framework is implemented in FastEpistasis (Schüpbach et al., 2010; Schuepbach, 2014). The software is an efficient solution extending the PLINK epistasis module. FastEpistasis uses a parallel algorithm that is efficient, scalable and optimized to handle multiples phenotypes. The two main reasons for choosing FastEpistasis over the algorithms mentioned in Section 6.3, was computational speed and model simplicity. The latter is also an important design choice for power considerations, because of modest sample size in this study.

Consider the multiplicative allelic model I introduced in Section 2.2.2 on page 9 (Equation (2.1)). With alleles A and a, the genotypes at a single locus are part of the set $\mathbb{G} = \{AA, Aa, aa, \text{missing}\}$. The linear model for the phenotype y is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \tag{6.2}$$

where $x_i \in \mathbb{R}$, $i = 1, 2$ is numerical counterpart assigned to each element of $\mathbb{G}$, chosen to the number of minor alleles (except when missing), hence $x_i \in \{0, 1, 2\}$. ϵ is a independent and identically distributed (i.i.d.) random Gaussian variable, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. As described earlier in Section 2.2.2, we may assess the significance of the interaction coefficient β_3 by comparing the fit of the interaction model to that of the marginal model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{6.3}$$

A significant interaction term carries the interpretation that the model Equation (6.2) explains a significantly larger proportion of phenotypic variance than model Equation (6.3).

This model can be written more compactly and more generally for multiple phenotypes using matrix notation. Let X^G be the $N \times M$ genotype matrix, defined by the genotypes of M SNPs for N individuals. Let Y be a corresponding $N \times K$ phenotype vector, where K is the number of phenotypes (expression traits).

$$Y = X^M B, \qquad Y \in \mathbb{R}^{N \times K}, \qquad X^M \in \mathbb{R}^{N \times 4}, \qquad B \in \mathbb{R}^{4 \times K} \tag{6.4}$$

where coefficient $B_{4,k}$ corresponds to the genetic interaction term of phenotype k. We have design the model matrix X^M, so for each $\binom{M}{2}$ combination of SNPs in X^G, we construct the $N \times 4$ matrix:

$$X^M = \begin{bmatrix} | & | & | & | \\ 1 & x_m & x_p & x_m \odot x_p \\ | & | & | & | \end{bmatrix}$$

where x_m is the m$^{\text{th}}$ column of X^G (i.e., the m$^{\text{th}}$ SNP over all individuals), and $x_m \odot x_p$ is the Hadamard product (element-wise product) between vector x_m and x_p.

As shown in Equation (6.4), we wish to discover the linear relationship between SNP pairs and the phenotypes. That is, we wish to estimate a length-four coefficient vector B, such that $X^M B \approx Y$. Given the independent Gaussian noise assumption, this leads to the well-known maximum likelihood estimate of B, arriving at the least square solution:

$$\hat{B} = \left(X^{M\top} X^M\right)^{-1} X^{M\top} Y \tag{6.5}$$

We may then calculate the estimated output phenotype vector $\hat{Y}^M = X^M \hat{B}$, with the residual, δ, sum of squared error:

$$\delta^M_{SSE} = \sum_{n=1}^{N} \left(Y_n - \hat{Y}^M_n\right)^2 \tag{6.6}$$

We can obtain the standard error of the estimated coefficients, $\text{Sd}\left(\hat{B}\right)$ using Mean Squared Error (MSE) to estimate the variance. We scale the δ^M_{SSE} by the number of parameters in the model ($p = 4$) to obtain the MSE:

$$\text{Var}\left(\hat{B}\right) = \text{MSE}\left(X^{M\top}X^M\right)^{-1} = \frac{\delta^M_{SSE}}{N-4}\left(X^{M\top}X^M\right)^{-1} \tag{6.7}$$

$$\text{Sd}\left(\hat{B}\right) = \sqrt{\frac{\delta^M_{SSE}}{N-4}\left(X^{M\top}X^M\right)^{-1}} \tag{6.8}$$

The significance of the genetic interaction term is assessed via the Wald test,

$$W = \left(\frac{\hat{B}_{4,k}}{\text{Sd}\left(\hat{B}_{4,k}\right)}\right)^2 \tag{6.9}$$

where W follows a χ^2 distribution with degree of freedom 1.

Readers interested in more statistical details and assumptions of the tests, may refer to Listing C.2 on page 97, that presents a workflow for fitting the multiplicative model to genotype data using the the statistical language `R`.

6.3.2 Implementation

To overcome the computational burden associated with performing millions or billions of statistical tests, FastEpistasis have implemented highly speed optimized solutions to the linear algebra operations described above. The computational pipeline described in this chapter, used an unoffical release `2.07` of FastEpistasis. This unofficial release contain several additional optimization and options customized to my project. The release was kindly provided by the main developer of FastEpistasis, Thierry Schüpbach. I will briefly outline some of the advantages and speed-up gains of FastEpistasis.

FastEpistasis optimizes the computations by splitting the analysis tasks into three separate applications: pre-computation, core-computation and post-computation. During pre-computation, set of SNP pairs, defined by user-specified SNP subsets A and B, consists of all pairs of SNPs with one SNP in A and the other in B. SNPs are ordered by those in $A \setminus B$, $A \cap B$ and $B \setminus A$, to efficiently iterate through the set of SNP pairs during the core computations. As the name implies, the core-computation is were the statistical tests are carried out using the vectorized (SSE) data from the pre-computation as input. In the post-computation phase, data gathered during the computational phase is restructured to limit storage throughput and summary output files are generated.

One important speed gain of FastEpistasis, is the algebraic solution to the normal equation given in Equation (6.5). In contrast to PLINK, which estimates coefficients using singular value decomposition (SVD) factorization and estimates their standard errors by applying a separate SVD to obtain the Moore-Penrose pseudo-inverse, FastEpistasis uses a single QR decomposition to estimate the interaction coefficient

and standard error. Generally, the safest method in terms of numerical stability is SVD. However, SVD comes with a much larger numerical cost, relative to the QR decomposition. FastEpistasis supports the advanced vectorized SSE and AVX instruction sets for additional speed-up.

To speed computations of multiple phenotypes, FastEpistasis uses an entirely embarrassingly parallel framework to calculate epistasis tests over the dataset of SNPs. And because the framework is embarrassingly parallel, FastEpistasis can statically divide the work load among nodes and/or threads. Both Symmetric MultiProcessing (SMP) and Message Passing Interface (MPI) architectures are supported.

6.4 Empirical Enrichment of Spatial Epistasis

I used a empirical null distribution to compute the enrichment P-values for spatial epistasis enrichment. This approach is conservative because it directly controls for confounding variables and corrects for the data distributions. Empirical null distributions does not rely on an analytical definition of a null model, where the underlying assumptions would be difficult to justify and could easily result in statistical significance where a resampling scheme would not. Yang et al. (2011) showed that the typical null models for genetic interaction discovery is problematic and recommended using empirical distributions. The downside of using empirical distributions is that they are computationally intensive to generate because the rely on high numbers of resampling.

6.4.1 Empirical P-value

The concept of an empirical P-value is easy to grasp because it is closely related to traditional hypothesis testing. First, a sufficient number of samples are drawn from the null model to generate a null distribution with high enough resolution to determine statistical significance. More samples leads to better approximation of the analytical null distribution, corresponding to the blue histogram approximates the black curve in Figure 6.1. Often 100 or 1000 samples are sufficient to form the empirical null distribution. Next, the number null samples at least as extreme as the observed score is determined and normalized by the number of null samples generated to give the P-value (striped blue area in Figure 6.1) (Noble, 2009).

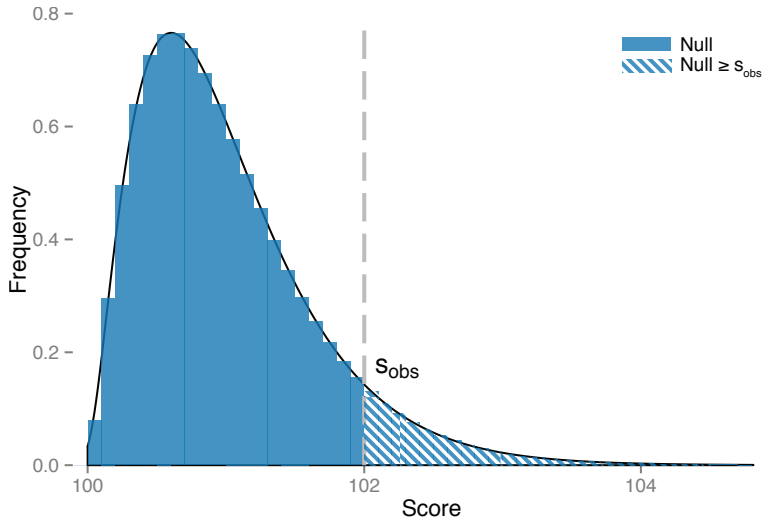


Figure 6.1: The concept of the empirical P-value calculation. The blue histogram illustrates the empirical null distribution of "scores" observed from shuffled interaction pairs. The solid black curve is the true analytical null function (here modeled as a $S \sim F(5, 100) + c$, where $c = 100$). The P-value associated with an observed score of $S_{obs} = 102$ is equal to the area under the curve to the right of 102 (striped blue). The P-value is *estimated* by counting null scores ≥ 102, leading to an estimate of 7.6%.

6.4.2 Null Distribution

A key feature of this framework is the generation of the null distribution. The following paragraph explains the assumptions and implementation of the null distribution.

I assume that the *observed* interaction pairs from the Hi-C data is a realization of sampling pairs of loci from a distribution of true physically interacting genomic loci. Following this line of thought, the observed Hi-C data represents *one* such "spatial proximate" sample (see the legend of Figure 6.2). In turn, the null distribution is assumed to be a realization of sampling pairs of loci from a distribution of true non-interacting genomic loci. In practical terms, the null distribution is generated by shuffling interaction pairs to form non-interacting pairs. In this pipeline, 1,000 permutations were used to generate 1,000 null samples. We ensure that all pairs in the null are not observed in the interaction table. Internally in the pipeline, the null distribution is represented as a mapping of interactions fragments to null samples* Importantly, by shuffling the interaction pairs, we preserve the "genomic properties" of the interaction table and avoid potential confounding factors in the null distribution: e.g. bias might be introduced if the null distribution was generated by sampling pairs of random loci from the entire genome.

* In gory details, the generation of the null distribution is facilitated by a nested dictionary that maps interaction fragments to the interaction table. The data structure contains Experiment Identifier (EID) (a unique identifier for each sample from the null distribution) at the top-level, and Experiment Interaction Identifier (EIID) (a unique identifier for each interaction for each sample from the null distribution) at the outer level.

6.4.3 Mathematical Formulation

Here follows a more formal and mathematical formulation of the empirical P-value. Suppose a vector, S, contains random values from the null distribution of non-interacting genomic regions. Let S_i be the number of genetic interactions from the i-th null sample and let S_{obs} be the value of the observed number of genetic interactions from interacting genomic regions. Then a right-tailed empirical P-value for S_{obs} can be computed as follows*:

$$P_{empirical} = \frac{1 + \sum_{i=1}^{N} 1 \cdot [S_i \geq S_{obs}]}{N + 1} \tag{6.10}$$

Note that the above formula includes a commonly used pseudocount (North, Curtis, and Sham, 2002; Knijnenburg et al., 2009) to avoid P-values of zero. Equation (6.10) reveals two properties of the empirical P-value. First, the resolution of obtainable P-values is $1/N$. Secondly, the smallest achievable P-value is $1/N$.

6.5 Computational Pipeline

With the basic components of the the pipeline explained earlier in this chapter, we are now ready to put it all together. This section explains the individual steps in the pipeline for the spatial epistasis discovery shown in Figure 6.2. This subject is inherently technical and I recognize that the density of new terminology might require an extra look at the glossary (Appendix D on page 99). Also, do not worry about the footnotes - they are only intended for the most technical interested.

In step **(1)** the Hi-C data is transformed into an interaction table. The interaction table defines the interaction pairs (pairs of genomic coordinates in close proximity). Each interaction pair consists of two interactions fragments from different chromosomes (inter-chromosomal). The size of the interaction fragments is determined from the resolution of the Hi-C data and the fragment binning (see Section 3.2 on page 15 for more details). The pipeline currently uses a fixed resolution fragment length of 10 kb.

In step **(2)** we define the local search space for spatial epistasis with the interaction width (l) parameter. The search space is bounded by the symmetric distance $l/2$ bp upstream and downstream from the midpoint of the interaction fragment considered. We extract SNPs within a distance $l/2$ from the center of each interaction fragment. The extracted SNPs are then partitioned into two sets†, of size n and m, where all pairwise combinations are analyzed for epistasis using FastEpistasis. By now we can appreciate the *most* important feature of this pipeline: when considering all pairwise interactions between the two SNP sets, we test for epistasis between *both* spatial proximate SNP-pairs and null SNP-pairs. (The later would in other

* readers may want look up the Iverson bracket notation † Partitioning of SNPs is important to avoid calculation of redundant SNP-pairs. I developed a simple algorithm that sorts and partitions SNPs based on the genomic coordinates to obtain minimum redundancy and maximize speed. See also Section 6.3.2 for a description of the SNP sets A and B.

cases be considered a "waste product".) In other words, we obtain the the null distribution "for free" using this approach. This principle is illustrated in Figure 6.2 step **(2)**, where the red colored pairwise interactions represent SNP pairs in spatial proximity and blue lines represent SNP pairs from the null distribution.

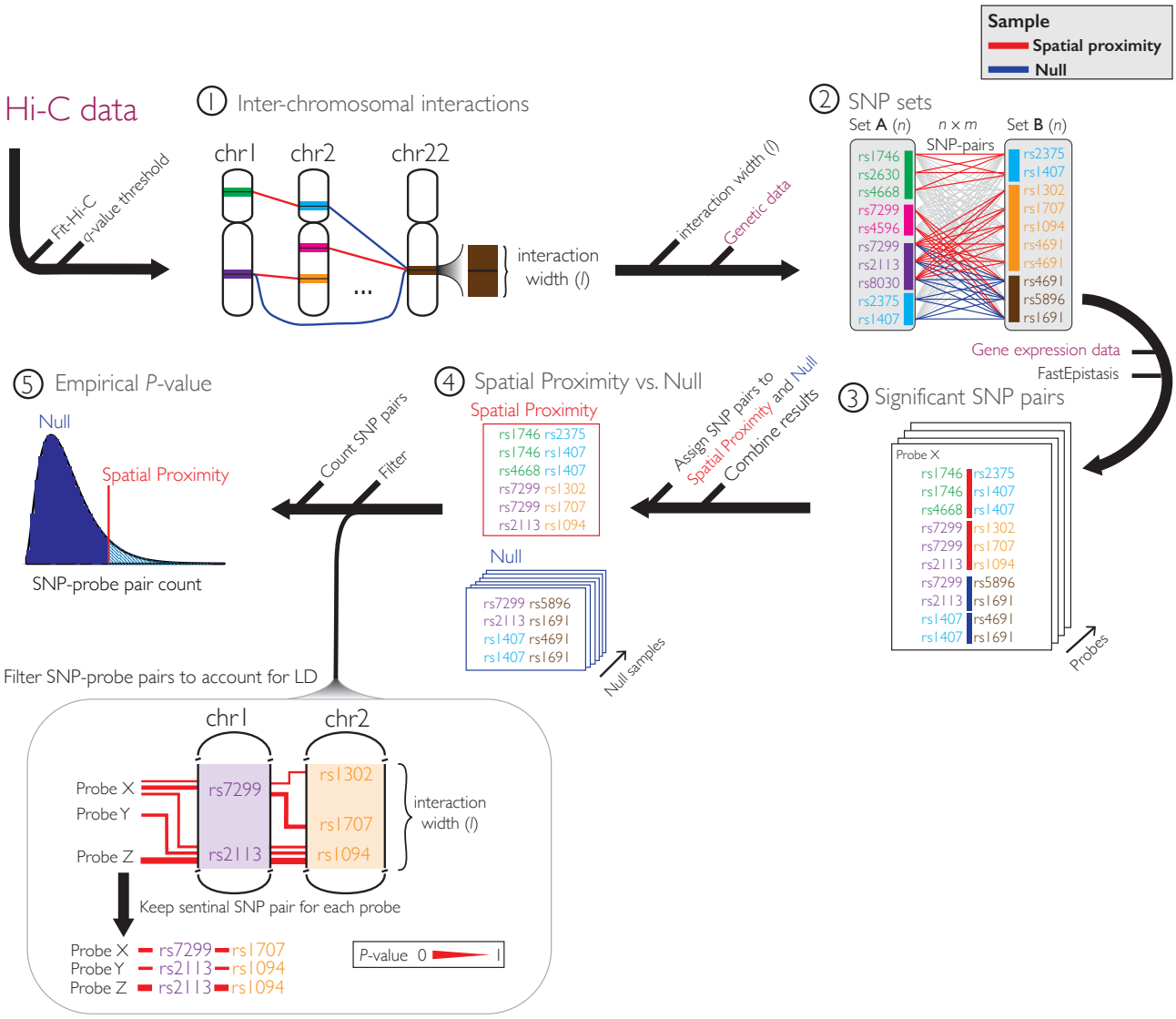
In this pipeline, Bonferroni adjusted P-values are used to correct for multiple hypothesis testing. FastEpistasis only outputs putative epistatic SNP pairs if they reach a user specified significance threshold. Hence the Bonferroni significance threshold needs to be precisely estimated prior to running FastEpistasis: if the threshold is set too high, we might overlook epistatic SNP pairs, leading to an increase in the false negative rate; conversely, if the threshold is set too low, the computation process becomes I/O bound and thereby extremely slow*. Fortunately, we are able to calculate the exact number of statistical tests as the product of the set size n and $m^{\dagger}$, prior to running FastEpistasis.

The output of FastEpistasis is compiled in step **(3)**. For each probe, we extract the Bonferroni significant SNP pairs (hereafter referred to as SNP-probe pairs).

In step **(4)** the SNP-probe pairs are mapped to either the spatial proximate sample or the null samples. The SNP-probe pairs are subsequently filtered using a "interaction pair" filter. This filter retains independent SNP-probe pairs by indirectly accounting for the LD structure in the genotype data. For a given interaction pair, the filter keeps only the sentinal SNP pair for each probe. In step **(5)**, the remaining SNP-probe pairs are counted for the spatial proximate sample and each of the 1,000 null samples. This generates a distribution over SNP-probe pair counts. We then identify the extremity of the observed spatial proximate sample compared to the null distribution. Finally, we use Equation (6.10) on page 43 to compute the empirical P-value.

* I/O bound refers the condition in which a process progress is limited by the speed of the I/O system. In this case the bottleneck will be writing terabytes to the hard drive. † On a technical note, the set size of n and m are slightly different between each of the null samples and the observed sample (spatial proximate). The difference is in set size is due to variations in the density of genotyped SNPs throughout the genome. However, the Bonferroni correction is uniformly distributed across the null samples.

Figure 6.2 ***(facing page)*:** Spatial epistasis discovery pipeline.
The realization of the spatial proximate and null samples are coded in objects colored in red and blue, respectively. The purple text color represents data integration steps. **(1)** Pairs of inter-chromosomal spatial proximate interactions (interaction pairs, red lines) are formed from the Hi-C interaction fragments. The null distribution (blue lines) is generated by shuffling interaction pairs to form non-interacting pairs. **(2)** For a given interaction width (l), SNPs within a distance $l/2$ from the center of each fragment are extracted and partitioned into two sets, and then analyzed for epistasis. The red colored lines represent SNP pairs in spatial proximity; blue lines are SNP pairs from the null distribution; gray lines are SNP pairs that does not map to any interactions. **(3)** FastEpistasis outputs the significant SNP pairs for each probe. **(4)** The SNP-probe pairs pairs are mapped to either the spatial proximate or null distribution. **(5)** The SNP-probe pairs are filtered using a "interaction pair" filter. For a given interaction pair (here illustrated in light purple and orange color), this filter keeps only the sentinal SNP pair for each probe. Finally the empirical P-value can be calculated by counting the remaining SNP-probe pairs and comparing the spatial proximate count to the null distribution.



Results 7

This chapter contains the main results of my thesis. I will first show a "proof of concept" by replicating the epistatic SNP-probe pairs reported by Hemani *et al.* Next I will present the enrichment results of spatial proximate epistasis. Selected enrichment results are followed up by more detailed analyses.

Appendices are not always being read - I myself skip them most of the time. However, I spent quite some time on the analysis in the appendix, so bare in mind that they are good to read for additional insights.

7.1 Replication of Hemani *et al.*

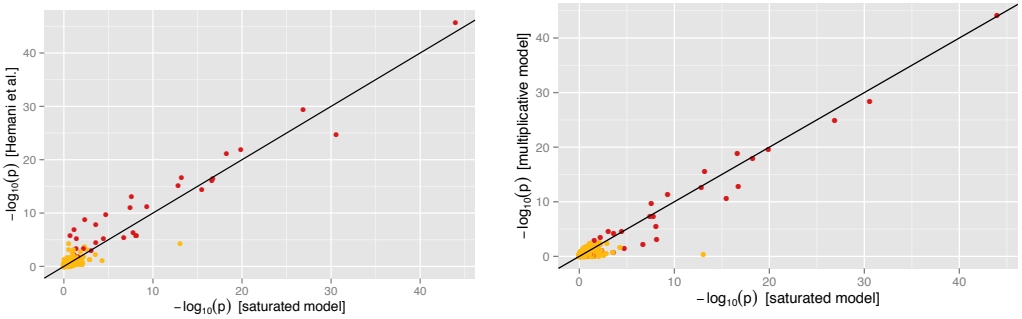
This study sought to replicate the work of Hemani et al. (2014) as it provides a way to validate the data processing pipeline used in this work. The replication of Hemani *et al.* was made possible because the authors had published the significance of the genetic interactions in the replication cohorts, including the EGCUT cohort. The *P*-values reported by Hemani *et al.* for the EGCUT cohort were derived from rank transformed expression levels normalized using the "eQTL Mapping Pipeline" (Deelen, Westra, and Franke, 2014).

434 out of the 501 SNP-probe pairs published by Hemani *et al.*, could be recovered in the EGCUT cohort after quality control performed in this study. The saturated genotype model with nine degrees of freedom used by Hemani *et al.* was fitted to the genetic data. The expression residuals derived from the PEER analysis was used as response variable. The *P*-values for the genetic interactions were found to agree well with data published by Hemani *et al.* (Figure 7.1 **a**; Pearson correlation of *P*-values $r = 0.66$; log-transformed *P*-values $r_{log} = 0.96$;). Not surprisingly, the 30 SNP-probe pairs replicated in the independent cohorts, showed the highest correlation. Examples of replicated and non-replicated SNP-probe pairs can be found in Appendix B.1 on page 81.

Next, the *P*-values were tested for sensitivity to the choice of genotype model. The full saturated model was compared to the much simpler multiplicative model, because the latter model was used for spatial epistasis discovery (Section 7.2 on the next page). The *P*-values derived from the two models were similar ($r = 0.46$, $r_{log} = 0.96$).

Together, this analysis showed that the difference in gene expression data normalization method used by Hemani *et al.* and this work, does not play a critical role for the epistatic signals with large magnitude - they still remain significant. Likewise, the multiplicative and saturated genotype model did not affect whether a SNP-probe pair with strong signal would be deemed significant or not. However, the choice of normalization method and genotype model may influence the interpretation of genetic interactions with less signal. Unfortunately the effect estimates

of epistatic SNPs were not made public, so it was not possible to compare the concordance of direction of effects to the work of Hemani *et al.*



(a) Replication of Hemani *et al.*

(b) Comparison of saturated and multiplicative model.

Figure 7.1: *P*-value comparison for 434 genetic interactions. Points colored in red represents the 30 SNP-probe pairs replicated across cohorts. The straight line represents $x = y$.

7.2 Spatial Proximate Epistasis Enrichment

This section contains the primary results of my thesis. I describe the results in two tiers of analysis depth. First I give an overview of cell type-specific spatial epistasis enrichment. Next, I scrutinize the initial results, and provide a more detailed analysis of the validity of the results.

7.2.1 Cell Type-specific Spatial Epistasis Enrichment

Hi-C data from three different cell lines were used to define sets of proximate inter-chromosomal genomic interactions. Spatial epistasis was calculated between these regions for 9,269 probes using FastEpistasis. The resulting SNP-probe pairs were subsequently filtered to account for the genotype correlations (LD structure). Spatial epistasis was assessed by counting epistatic SNP-probe pairs after pair level filtering to account for LD structure (referred to as "interaction pair" filtering, see Figure 6.2 on page 45). A null distribution was constructed from genomic regions that were not spatially proximate. Different scenarios or "parametrizations" of the 3D genomic proximity were used, i.e. different thresholds for defining spatial proximate regions (interaction width and q-value). A set of negative control parametrizations was constructed for the hIMR90 and hESC cell lines by random sampling of interactions with contact count $= 1$.

In total, more than 12 years of single threaded computation time were used to carry out the statistical test for epistasis shown in Table 7.1. A total of $1.33 \cdot 10^{14}$ linear regression models models were fitted and tested for significance of the interaction term. Although the largest of the of the calculations (hESC$[l = 2,500\text{bp}; q = 10^{-13}]$, ~ 5 years computation single threaded time) included more than 500 trillion tests

for epistasis, it only covered 0.35% of the possible two-locus epistasis tests for the genotyping data. For the same parametrization, the genomic loci searched for epistasis covered 6.62% of the genome.

Table 7.1 shows the enrichment P-values resulting from different parametrizations of spatial epistasis. Histograms underlying the P-values can be found in Appendix B.2 on page 83. Two parametrizations of spatial proximity for the hIMR90 cell line were enriched for spatial epistasis (one-sided test, $\alpha = 0.05$): hIMR90$[l = 1,000\text{bp}; q = 10^{-6}]$ and hIMR90$[l = 2,500\text{bp}; q = 10^{-7}]$. These parametrizations used medium range interaction width and a non-conservative q-value threshold. hIMR90$[l = 1,000\text{bp}; q = 10^{-6}]$ showed a two-fold enrichment in the number of epistatic SNP-probe pairs compared to the mean of the null (Figure B.2 on page 83, **b**). Using a more narrow interaction width (hIMR90$[l = 500\text{bp}; q = 10^{-6}]$) resulted in marginal significance while a very wide interaction width (hIMR90$[l = 50,000\text{bp}; q = 10^{-9}]$) showed no enrichment (Figure B.2 on page 83, **a** and **d**). No hESC or K562 parametrizations exhibited enrichment for spatial epistasis. The negative control parametrizations did not enrich for spatial epistasis as expected. Two additional filtering strategies of the SNP-probe pairs were applied to test the robustness of the enrichment P-values. Using no filter and the "chromosome pair" filter used by Hemani *et al.* did not change the significance of the P-values in Table 7.1.

Cell line	**Interaction width, *l***	***q*-value** ($n_{interactions}$)	**No. epistasis tests *per probe*†**	***P*-value‡**
hIMR90	500 bp	10^{-6} (26,325)	260,273,820	0.078
	1,000 bp	10^{-6} (26,325)	1,042,826,407	0.001 *
	2,500 bp	10^{-7} (8,114)	733,442,624	0.020 *
	50,000 bp	10^{-9} (1,021)	4,284,422,018	0.927
hESC	500 bp	10^{-16} (2,665)	3,222,025	0.257
	500 bp	10^{-14} (11,919)	57,787,904	0.620
	1,000 bp	10^{-12} (39,966)	2,236,957,368	0.643
	2,500 bp	10^{-13} (24,084)	5,560,830,560	0.143
K562	1,000 bp	NA (1,263)	1,485,960	0.540
	5,000 bp	NA (1,263)	28,045,983	0.780
$\text{hIMR90}_{\text{control}}$	1,000 bp	NA (5,000)	45,668,788	0.169
$\text{hESC}_{\text{control}}$	1,000 bp	NA (5,000)	45,714,894	0.317

Table 7.1: Initial spatial epistasis enrichment. P-values were calculated after filtering the SNP-probe pairs using the "interaction pair" filter. An asterisk marks significant enrichment ($\alpha = 0.05$). The q-values threshold is not applicable for the K562 cell line because the model used to analyze the Hi-C data does not output FDR values. The control cell lines were constructed by sampling 5,000 interactions with contact count = 1.

†total number of statistical tests between interacting and non-interacting loci *per probe*.

‡ P-value derived from filtered SNP-probe pairs.

7.2.2 Dubious Epistasis underlying hIMR90 Enrichment

The initial results shown in Table 7.1 indicated presence of spatial epistasis enrichment for two parametrizations of 3D genomic proximity for the hIMR90 interactions: hIMR90$[l = 1,000\text{bp}; q = 10^{-6}]$ and hIMR90$[l = 2,500\text{bp}; q = 10^{-7}]$. This section focuses solely on these two parametrizations and studies the validity of enrichment results in more details. Specifically, I made two observations indicating that the spatial epistasis enrichment is of likely dubious nature.

SNP-probe Pairs do not Co-localize One interesting biological aspect of epistasis influencing gene expression, is how the epistatic SNPs and probes are spatially organized. To answer this question I investigated the chromosomal co-localization between SNP pairs and probes, i.e how often at least one of the epistatic SNPs was located on the same chromosome as the probe.

Both hIMR90$[l = 1,000\text{bp}; q = 10^{-6}]$ and hIMR90$[l = 2,500\text{bp}; q = 10^{-7}]$ showed low SNP-probe co-localization: 8.99% and 7.38% of the spatial proximate epistasis pairs, respectively. The same tendency was apparent when also considering the non-spatial pairs (Table B.2 on page 91). Interestingly, the fraction of co-localizing SNP-probe was equally low for the set of SNP-probe pair with higher confidence of epistasis (minimum GCC ≥ 3, see Appendix B.3 on page 85 for a description of minimum GCC filtering). These data suggest that both epistatic SNPs act *in trans*. This result does not agree with published findings of marginal effects (eQTLs) in human blood (Westra et al., 2013), where the majority of the associations act *in cis* (*cis*-eQTLs).

SNP-probe Pairs Overfit the Data A major issue with the spatial epistasis enrichment signal shown in Table 7.1 is that the underlying epistatic effect estimates may not be valid. To avoid any bias in results, I considered the genotype class count (GCC) of putative epistatic loci as a measure of confidence for the genetic interaction. Specifically, I defined the minimum GCC, as the minimum GCC for any combination of genotypes. Hence minimum GCC measures how well the two-locus genotype count matrix is populated and is in units of "data points". I demonstrated that low GCCs may cause spurious epistasis results (Appendix B.3 on page 85). The multiplicative model fitted to SNP pairs with low GCCs may give misleading estimates of the effect size and P-value for the interaction term because of *overfitting* and the presence of high leverage points (see Figure B.6 for examples of this issue).

To investigate whether minimum GCC biased the observed enrichment signal, SNP pairs were filtered with a second filter based on their minimum GCC. In this filter only SNP pairs with at least three data points in each genotype class were kept (a threshold of five data points was used in Hemani et al. (2014)). See Appendix B.4 on page 88 for figures and tables related to the GCC filtered results. The SNP-probe pair counts decreased dramatically when the mimimum GCC filter was applied (Table B.1). For both hIMR90$[l = 1,000\text{bp}; q = 10^{-6}]$ and hIMR90$[l = 2,500\text{bp}; q = 10^{-7}]$ only two SNP-probe pairs remained after filtering, resulting in removal of spatial epistasis enrichment ($P_{GCC\ filter} = 0.28$ and 0.26,

respectively). Figure B.8 on page 89 shows the enrichment histograms when applying the minimum GCC filter. Reversing the order of the "interaction pair" filter and minimum GCC did not change the significance of the results (Figure B.9 on page 90).

In summary, scrutinizing the epistatic signal underlying the enrichment signal showed that the majority of SNP-probe pairs were a result of the underpowered design of this study, resulting in model overfitting for variants with rare alleles. These data suggest that the spatial epistasis enrichment was false.

Discussion 8

Several contributions of this project have already been discussed in their respective chapters, so I order to minimize redundancy, I will not repeat the treatment of these subjects. In particular, the proposed spatial epistasis hypothesis was discussed in Chapter 4 and a discussion of the design features of the associated computational framework was covered in Chapter 6.

I have presented a new framework, which uses information of the spatial organization of the human genome, to limit the search space when searching for interactions between SNPs in large-scale genotyping data. I proposed the "spatial epistasis" hypothesis and reasoned over its biological plausibility. I argued why known chromatin interactions might provide a reasonable biological prior for genetic interactions. The developed framework is distinguished from existing methods that propose either to exhaustively scanning all SNP pairs or to only test marginally significant loci. My method provides an alternative to these approaches by facilitating biological interpretation of the findings, something lacking in many association studies. If the biological prior of spatial epistasis is truly informative, the framework is able to select potentially good SNP pair candidates, and thereby increase the statistical power; and furthermore, detect true epistastic SNP pairs missed by methods testing all pairs exhaustively. Few other methods have been proposed that use biological knowledge to guide the search for epistasis. These methods have used protein–protein interaction networks to drive the search for interacting SNPs. However, protein interaction-based methods are restricted to select SNP pairs in gene regions that encode the corresponding proteins. Hence these methods fail to investigate the large portion of regulatory variants in intergenic regions. I argued that my framework is more suited to detect and provide biological insights of associated regulatory variants. My framework integrates 3D maps of the genome folding, which have been shown to inform about multiple regulatory features of the genome. As such, the developed framework contains biological knowledge critical for gaining identifying and interpretation of regulatory variants. This quality represents an important contribution of my work, since most genetic associations discovered today are located in gene deserts with a regulatory function. Moreover, I argued that the method constitutes a conceptual framework, which can be generalized to use other biological priors or statistical models, and ultimately test future hypothesis around the molecular mechanisms driving epistasis. I provided evidence that this approach is computationally feasible for large-scale search of epistasis, although it was difficult to quantify its statistical powerful. I found that the study was underpowered to detect large-scale epistasis. Thus my work did not provide compelling evidence of epistasis influencing human gene expression.

Despite the lack of statistical power to detect epistasis at this scale, my method allows to test biological hypothesis of the molecular mechanisms causing epistasis. To summarize, I regard this framework useful for the following reasons concerning its *potential* to: ascertain of the role of epistasis in the genetic architecture of

human complex traits, thus contributing to the controversy around epistasis and missing heritability; identify (regulatory) biological mechanisms driving epistasis; uncover new genetic interactions in human disease data.

Replication of Previous Epistatic Effects As a proof-of-concept, I rediscovered the epistatic effects reported by Hemani et al. (2014). The effect size estimates of the genetic interaction term was found to be highly similar ($r_{log} = 0.96$). Firstly, this result suggests that the three parameter multiplicative allelic model performs at least equally well in detecting epistasis, compared to the nine parameter saturated genotype model used by Hemani *et al.* However, there is no basis for real power comparison of the two models. Secondly, the result indicated that normalizing gene expression data using PEER is comparable to using a mixed effect model with known covariates. Additional analysis provided further evidence that PEER is effectively removing batch effects and other confounding factors from gene expression data (Figure A.7). Hence for future studies, I recommend using PEER for a fast, robust and effective way of removing confounding effects.

Cell type discrepancy The EGCUT expression data was derived from whole peripheral blood samples. To define spatial interactions between genomic regions, I used publicly available Hi-C data derived from two studies spanning three different cell lines: K562 from Lan et al. (2012), hESC and hIMR90 from Dixon et al. (2012). K562 leukemia cells share many properties to hematopoietic cells (Young and Hwang-Chen, 1981), that give rise to all human blood cells. Hence, the cell type of the K562 Hi-C data are, to some degree, consistent with the cell type(s) of the expression levels. Unfortunately the K562 cell line poses another problem: a hallmark of leukemia cells is the presence of the Philadelphia chromosome, a specific chromosomal abnormality that is the result of a reciprocal translocation between chromosome 9 and 22 (Drexler, MacLeod, and Uphoff, 1999). A explorative analysis showed that the interactions observed in K562 cells between 9q34 and 22q11 are labeled as inter-chromosomal, although they are actually intra-chrosomomal in the K562 genome. In total more than 20% of the inter-chromosomal interactions in the K562 cells are between chromosome 9 and 22. Interestingly, this suggests that Hi-C data can be used to study chromosomal translocation in cancer cell lines, as they will exhibit higher than expected inter-chromosomal Hi-C contact intensities. Such application of Hi-C have previously been suggested by others (Engreitz, Agarwala, and Mirny, 2012). A careful inspection of the data, revealed that the Lan *et al.* data quality is questionable and contains several "genomic hotspot" artifacts (Figure A.11).
The two remaining Hi-C cell lines were derived from human embryonic stem cells and human lung fibroblast. Hence there is a mismatch between the cell types from the Hi-C and expression samples. This discrepancy between cell types might influence downstream results. The Hi-C and expression data should be matched by tissue or cell type to ensure that the genome folding and expression levels reflect the same cellular developmental state and biology. However, increasing evidence suggests substantial conservation of 3D genome structure across the mammals (Dixon et al., 2012; Rao et al., 2014). In particular certain features, like TADs, seem to be cell-type invariant (Nora et al., 2012; Dixon et al., 2012; Rao et al.,

2014). If this holds true, data from different cell types will be of less concern. However, single-cell Hi-C experiments have consistently revealed that long-range structures of the genome organization (e.g. interdomain and trans-chromosomal contact structure) are highly variable between individual cells (Nagano et al., 2013). Hence that exact effect this cell type discrepancy remains unclear. It is likely that technological advancements will render this concern important: the promise of affordable single-cell RNA-seq (Macosko et al., 2015) and single-cell Hi-C (Nagano et al., 2013) will allow researches to generate high-resolution profiles of individual cells, hence facilitating the study cell-to-cell variability of spatial epistasis.

Spurious Epistatic Signals The initial spatial epistasis signal from the hIMR90 cell line were examined in two ways. The co-localization analysis revealed that only a small fraction of the SNP-probe pairs co-localized on chromosomes, suggesting a *trans* acting epistatic network. Although this result does not agree with the eQTLs discovered in blood (Westra et al., 2013) and the (apparent) epistatic networks reported by Hemani et al. (2014), it is consistent with the proposed involvement of transcriptional factories in mediating epistasis (Göndör and Ohlsson, 2009). This multi-interaction scenario is tentatively supported by the identification of up to five chromosomes simultaneously in contact with each other Zhao et al., 2006. However, this would require that a least two of the three chromosomes involved in the interactions, would loop out of confines of their CTs.

Although there is convincing evidence that spatial enrichment signal was false, it is interesting to note that three of the four parametrizations involving hIMR90 chromosomal contacts exhibited marginal significant enrichment. This raises an interesting question: if spatial epistasis exist, is it cell type specific? If the answer is yes, then the decrease in statistical power when using whole peripheral blood samples, containing mixtures of cell types may hamper, or simply prohibit, detecting of spatial epistasis.

The strength of the projects is, to my best judgment, not its results but rather the frameworks developed and proposed hypothesis. These contributions have opened up for new research questions and future analysis involving new (three-dimensional) perspectives of genetics and genome biology.

Conclusion 9

This thesis aimed to uncover regulatory molecular mechanisms that can cause epistatic interactions influencing human gene expression, thus providing a more complete understanding of epistasis. In Chapter 1, I motivated the need for more complete and biological realistic genetic models that include epistatic effects. Clarifying the definition of biological and statistical epistasis was a requirement for understanding genotype-phenotype relationships in the presence of epistatic effects. The linear genetic models described in Chapter 2 provided the fundamental basis for modeling epistasis. In Chapter 3, I described a new paradigm for genome biology wherein genomes are organized around gene regulatory factors that govern cell identity. I introduced the 3D genome maps generated by the experimental technique Hi-C, that have paved the way for this new paradigm. I focused on the hierarchical structure of the genome and the basic components of chromatin organizations, such as TADs.

Together Chapters 2 and 3, set the stage for putting forward the main premise of my thesis in Chapter 4. I proposed the "spatial epistasis" hypothesis and reasoned over its biological plausibility. I hypothesized that physical proximity of interacting genomic regions provides a spatial scaffold to identify genetic interactions in human genotyping data.

I developed a computational framework for testing the spatial epistasis hypothesis by performing enrichment analysis of spatial epistasis. As presented in Chapter 6, the model is designed to test for spatial epistasis in a fast and efficient way, hence overcoming some of the major statistical and computational challenges related to searching for epistasis. The framework was designed with the ability to model different scenarios of spatial epistasis by parameterizing the confidence of physical interaction between genomic loci and their interaction width. I explained that the framework employed an empirical null distribution that directly controls for confounding factors, when assessing enrichment significance. Finally, I argued that the framework can easily be generalized to support more advanced statistical models of epistasis and, more importantly, alternative biological priors driving the search for interacting SNPs. Ultimately, this framework provides a powerful way of testing a broad spectrum of hypothesis beyond spatial epistasis.

As a proof of concept, the framework successfully replicated previously reported epistatic effects. This also served as quality assurance of the genetic data and normalized expression data, and paved the way for testing the spatial epistasis hypothesis.

In Chapter 7, the hypothesis was tested under different scenarios of spatial epistasis, including genome structures from multiple cell lines. For human lung fibroblast cells, I identified a strong enrichment signal of spatial epistasis ($P_{emperical} < 0.001$). After scrutinizing the putative epistatic SNPs, it was established that more than 99% was spurious epistatic effects attributed to the underpowered statistics.

I conclude that although this study was underpowered to detect epistasis at this scale, my results do not support the existence of prevailing epistasis influencing human gene expression. Ultimately, there is no evidence to definitively reject or accept the spatial epistasis hypothesis. Nevertheless, I expect the conceptual framework to be valuable in future genetic association studies, because it has the potential to provide the much needed biological interpretation of regulatory variants.

Appendix and Epilogue

A major part of my work could not be fitted into regular chapters of my thesis. Instead, to avoid distractions from the main story, I have chosen to place the additional material in the following appendices*. Appendix A describes PEER analyses of the expression data and explorative analyses of the Hi-C data. Supplementary epistasis results are given in Appendix B.

* Interestingly, there is evidence that suggests avoiding appendices. Kahneman (2011) proposed the peak-end rule, which states that people largely judge an experience based on how they felt at its peak (the most intense point) and at the end of the experience, rather than based on the total sum or average of whole the experience. (This fascinating rule was discovered by studying colonoscopy examinations: one group of patients had a colonoscopy which was short but intensely painful throughout. In another group of patients, the colonoscopy was far longer and just as painful at its peak - thus involving more pain in total - with one important modification: the pain was *reduced at the end* of the examination. The second group of patients reported a *better* experience and less discomfort, even though they had experienced more pain overall.) Hence, the critical point is how the story ends. Knowing that most appendices are not as exciting as the main text, and this is no exception, I risk ruining the reader's whole experience of my thesis. I will at least let this be the excuse if you ended up not liking my thesis...

Supplementary Methods A

A.1 Population Structure

A.1.1 Population Structure

PLINK 1.9 (Chang et al., 2015) was used to calculate genome-wide IBS pairwise distance measures using the `-genome` flag. Four MDS components was calculated using `-cluster -mds-plot 4`.

Population structure can be identified by inspection of scatter plots of MDS components. Appearance of clusters on such plots, indicates population structure. Figure A.1 shows that there is little population structure in the EGCUT cohort - the samples are centrally homogeneous distributed. This result agrees with the reported little admixture of the Estonian population (Nelis et al., 2009).

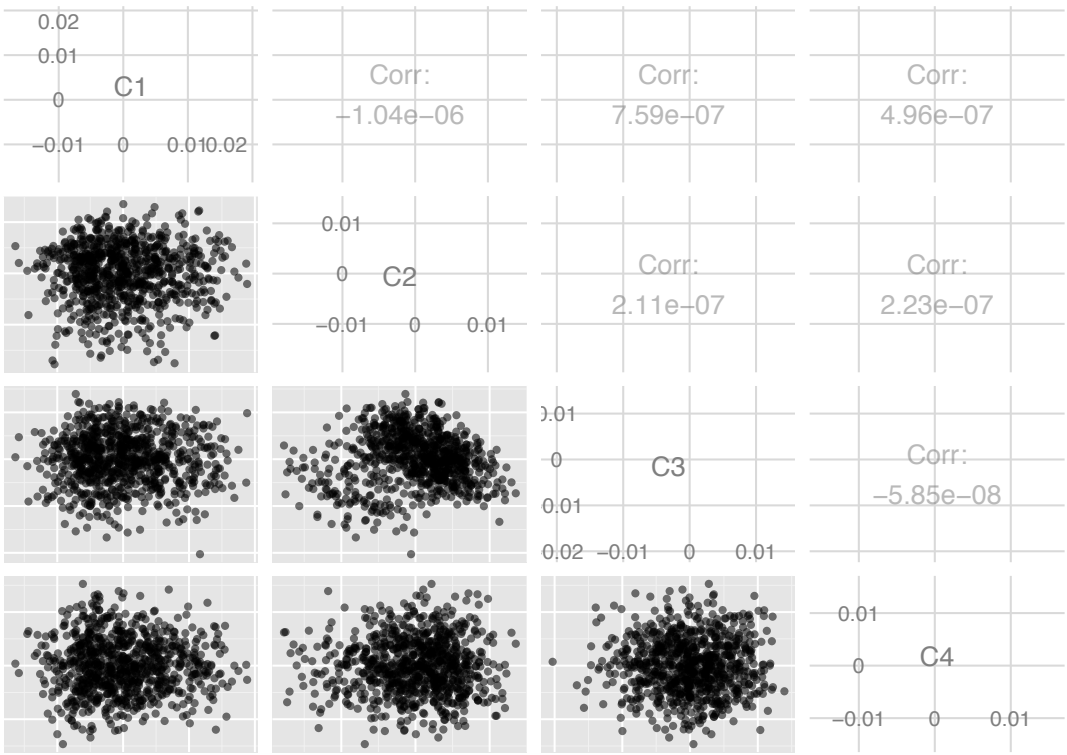


Figure A.1: EGCUT population structure. The matrix elements are the four MDS components ($C1 - C4$). The lower and upper diagonal show scatterplots and Pearson correlation, respectively. Each sample ($n = 832$) is represented by a single point.

A.2 PEER Analysis

Gene expression levels measured by any high-throughput technology are inherently noisy and contain biases needed to be controlled or removed. Leek and Storey (2007) showed that batch effects and other confounding effects in gene expression data reduce power in the analysis and may introduce spurious signal to many genes. PEER (Stegle et al., 2012) is a software package for learning confounding effects in gene expression data. Several other groups have developed statistical approaches to account for latent variables of expression data. Leek and Storey (2007) developed surrogate variable analysis (SVA) - a factor analysis method for identifying expression heterogeneity components. Another tool, "eQTL mapper" (Deelen, Westra, and Franke, 2014) uses principal component analysis (PCA) to identify confounding non-genetic components. PEER's main advantage compared to other methods, is the flexibility and complexity control of a probabilistic Bayesian model.
PEER is implemented in C++ with a user-friendly interface to R (R Core Team, 2015).

A.2.1 Bayesian Model

The model underlying this PEER assumes that gene expression levels can be modeled as additive effects from independent sources. Specifically, these contributions can be modeled as known and hidden factors. Figure A.2 shows different levels of accounting for additive effects of gene expression levels for eQTL discovery. Most studies include known factors that are easily obtained, such as age and gender. However, fewer studies account for hidden factors, such as technical and batch effects.

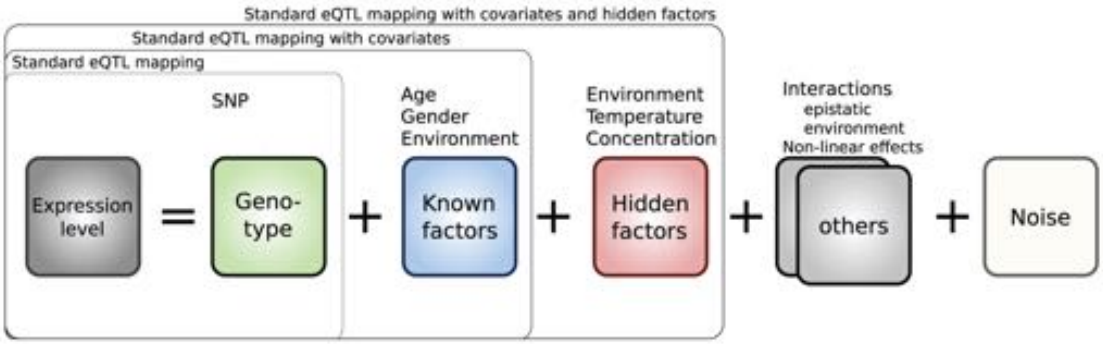


Figure A.2: General additive model for gene expression variability. Each factor in the model contributes to the gene expression variability. Typical eQTL studies model only genotypes and known factors. Illustration adopted from Stegle et al. (2010).

PEER learns the effects of the known and hidden factors jointly across individuals and gene probes. A graphical representation of the PEER model is shown in Figure A.3. Simply put, PEER models the observed gene expression levels as the sum of contributions from the known and hidden factor models. Once these factors have

been learned, they can be regressed out using linear regression, and the expression residuals can be used for downstream analysis, e.g. epistasis discovery.

An important feature of PEER is the gamma prior on the inverse variance α_c and β_k for weights of observed and hidden factors, respectively. This prior introduces automatic relevance detection (ARD), forcing the weights of unused factors to zero and thereby switching them off. In this way, ARD provides model complexity control by automatically learning the the effective number of covariates. Hence, the user needs only set K to a sufficiently large value. This is in contrast to non-probabilistic factor analysis methods (e.g. PCA-type models), where there is no automatic way of estimating optimal number of unobserved factors.

PEER learns the effects of the factors using variational Bayesian inference. The main purpose of variational Bayes is to approximate the posterior distribution of variables in the model. In each iteration, the parameters for the posterior estimates are updated to maximize the lower bound. The lower bound is the marginal likelihood ("evidence") of the observed data. The algorithm can be considered an extension of the EM (expectation-maximization) algorithm and provides locally-optimal solutions. A more detailed introduction of variational Bayesian inference is beyond the scope of this section; see Bishop (2007) and Jordan (1999) for an introduction to this framework.

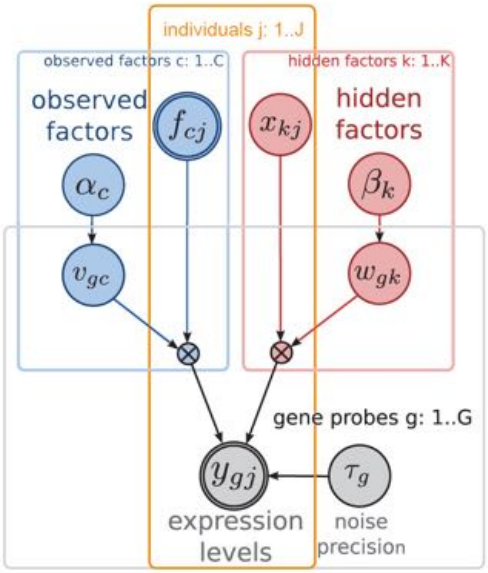


Figure A.3: Bayesian network of the PEER model for gene expression variation. The model combines observed factors (blue) and hidden factors (red) to explain the gene expression levels (gray). The solid rectangles indicate that contained variables are duplicated for each individual (j), gene probe (g) observed factor (c) and hidden factor (k). f_{cj} is the measured covariate for each individual j; v_{gc} is the weight for each gene g, with α_c as the inverse variance parameter of the gamma prior. x_{kj} is the latent variables for each individual; w_{gk} is the weight for each gene g, with β_k as the inverse variance parameter of the gamma prior. τ_g is the precision parameter for the global Gaussian noise. Illustration adopted from Stegle et al. (2010).

A.2.2 Running PEER

The *required* input for PEER is a gene expression matrix and specification of the maximal number of hidden factors (K) to infer from the data. Optionally, the user can specify input matrices shown in Table A.1 for more advanced modeling.

Optional settings	Description
Covariates matrix	Matrix specifying the known factors, e.g. age and sex.
Uncertainty matrix	Matrix of uncertainty estimates (i.e. variance) specific to each gene probe.
Prior connectivity matrix	Matrix specifying *a priori* probabilities of the effect of hidden factors on specific genes.

Table A.1: Optional PEER settings.

A.2.3 PEER in Practice - Lessons Learned

To become comfortable with PEER's model in practice, several properties of PEER were investigated. This section summaries the most important findings. Both the EGCUT expression data and a smaller synthetic example data set was used to perform simulations.

PEER Run Time

PEER's run time scales linearly with number of individuals and number of genes, and quadratically with the number of learned factors.

Sensitivity to Variance Stabilization Method

This work found that PEER is sensitive to the variance stabilization method. Three types of variance stabilization were compared: $\log_2$, inverse normal transformation (INT) and no variance stabilization. The INT was performed using `rntransform` from the GenABEL R-package (Aulchenko et al., 2007). PEER did not converge when using non-variance stabilized data. The algorithm converged very quickly when using the INT data. This is likely because of normalization method enforces a strict distribution of the data. $\log_2$ transformation of the data provided a good compromise between convergence time and simplicity in the transformation function. Hence this work recommends using this method for variance stabilization.

PEER Models Gene Probes Jointly - More Data is Better

It is of interest to know whether PEER models genes independently or jointly. To investigate this, a five nested data sets, D_n were created from the EGCUT

cohort:

$$D_1 \supseteq D_2 \supseteq \cdots \supseteq D_5$$

Where D_1 is the full data set with all gene probes and D_n contains $1/n$ part of the gene probes. For each data set, the model PEER was fitted and the residuals extracted. The Pearson correlation, $r_{1,n}$, between the residuals of the full data set and the nested subset were calculated. The correlation decreased with decrease in size nested data set, i.e. $r_{1,n+1} < r_{1,n}$.
PEER learns the confounding effects best with more data - at the expense of increasing computation time (correspondence with the PEER team; Leopold Parts, 11/29/14). For this reason, it is advantageous to subset on probes (QC, removing unmapped probes) *after* running PEER. The intuition is that all probes, including low quality probes, will have the same confounding pattern, which jointly improves PEER's ability to learn the confounding factors.

Choosing k and the Effect of Automatic Relevance Detection

Running PEER with different number of hidden factors, k, gives comparable PEER expression residuals for the EGCUT cohort. Mean correlation between PEER expression residuals for $k = 10$ and $k = 50$ was $r = 0.956$, indicating that there is little effect of using a very large k, except for the computational cost. (One may contrast this result to the one shown in Figure A.9 on page 71). This work found that k needs only to be set to a sufficiently large value, so that $k \geq k_{true}$. This is due to the ARD feature of PEER, which switches off unimportant factors. k_{true} can be estimated from inspecting the posterior variance of the factor weights (see Figure A.5 on page 68).

PEER as a Simple Linear Model

This work sought to find PEER's relation to a simple regression model. First, PEER was run on a synthetic data set with $k = 0$ and including four covariates. Next, a linear regression model was fit to the data with the four covariates as explanatory variables. The residuals derived from the two models were compared and found to correlate perfectly. This result shows that using PEER with $k = 0$ is equivalent to using a regression model. We conclude that PEER includes the covariates as fixed covariates and regress them out using a linear model.

A.2.4 PEER Analysis on EGCUT

PEER Settings

We compared several settings using PEER to remove confounding effects from the EGCUT expression data. We considered convergence of the algorithm and computation time as the main criteria for selecting the final settings shown in Table A.2.

Primary settings	**Value**
Prior variance stabilization method	$\log_2$ transformation
Number of hidden factors	50
Secondary settings	**Value**
Know factors (covariates)	4 MDS genetic components
Max iterations	1000
Global mean	No
Uncertainty matrix	No
Prior connectivity matrix	No

Table A.2: Final PEER settings for EGCUT cohort.

Diagnostics

PEER performs Baysian variational inference where, in each iteration, the parameters for the posterior estimates are updated to maximize the lower bound. The lower bound is the log of the evidence $P(data)$. Thus, monitoring the lower bound is useful for checking that the inference performs correctly (i.e. the bound is increasing) and monitoring convergence (i.e. the bound is not increasing a lot). This improvement in the bound between two iterations is tested against the cutoff to determine when to stop the algorithm.

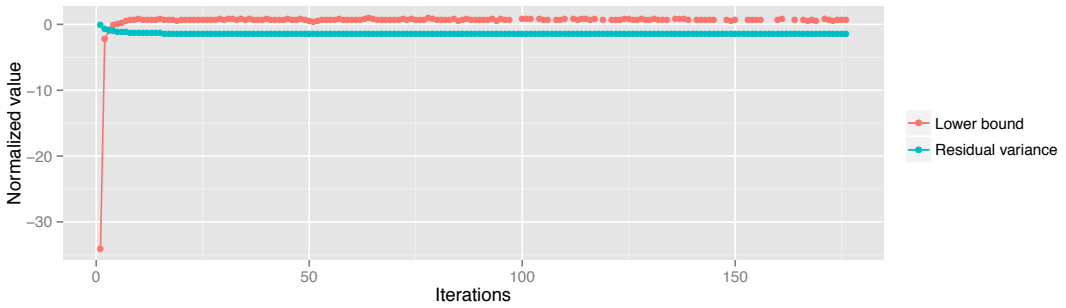


Figure A.4: PEER convergence diagnostics. The variational Bayesian inference algorithm converged after $i = 176$ iterations. The lower bound (red line; see also Appendix A.2.1 on page 62) describes the fit of the model. The residual variance (blue line) is used to monitor the magnitude of the change between the iterations. Both lines show an initial sharp change before flatten out, indicating that the model converges. For the purpose of clarity in the trends, the lower bound and residual variance are max-scaled; the residual variance is furthermore $\log_2$ transformed.

Interpretation of Inferred Factors

The importance of the inferred factors were assessed by the posterior variance of their weights. The factors were grouped into three categories based on their importance (Figure A.5). Five factors with a positive posterior variance ($\alpha > 0$) were identified as "important". The four covariates (genetic MDS components) were

found to be unimportant, indicating that the population structure of the EGCUT is not relevant as a covariate. Figure A.6 shows the inter-correlation structure of the five most important inferred factors.

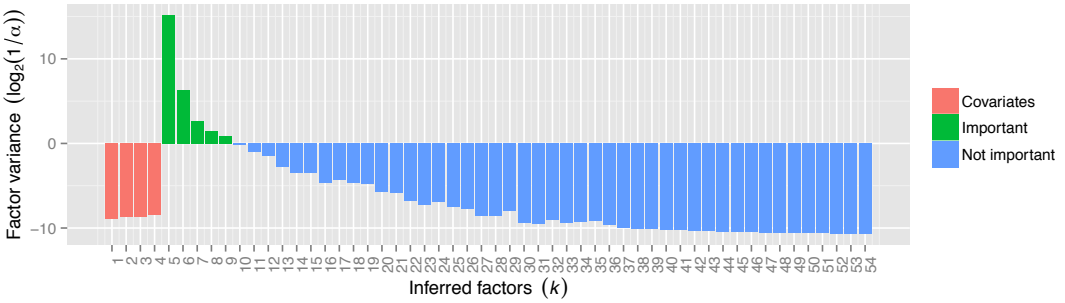


Figure A.5: PEER inferred factor importance. The variance of the inferred PEER factors can be interpreted as the "factor importance". Factors $k = 5...9$ (green) with a positive score on the y-axis are considered "important" factors. Factors $k = 1..4$ (red) are the four genetic MDS components included as covariates. This plot further illustrates the ARD feature of PEER (see Appendix A.2.1 on page 62, indicated by the unimportant group of factors, $k = 10...54$ (blue).

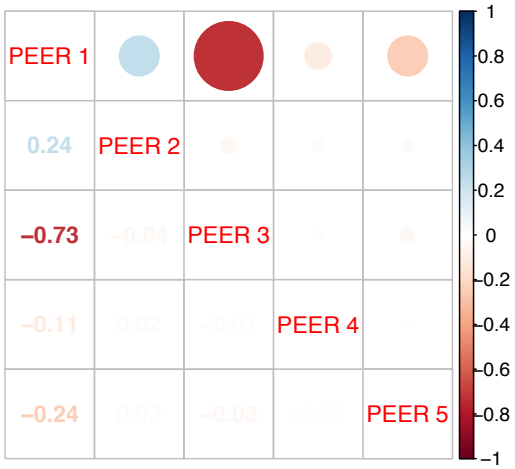


Figure A.6: Factor inter-correlation matrix. Inter-correlation of PEER factors. `PEER 1` and `PEER 3` are highly correlated

To validate that PEER learns confounding factors in the expression data, the inferred factors were correlated with measured covariates (Figure A.7). The measured covariates used for this analysis are typical covariates such as gender, age, batch, RIN and blood count numbers (see Table A.3 on page 70 for a full list of the covariates). These covariates were not included when the PEER factors were constructed, for the purpose of interpretation of the inferred PEER factors. The strongest correlation was observed between the most important PEER factor, `PEER 1`, and the `Batch` covariate ($r = -0.45$). Additionally `PEER 2` showed correlation to both RNA quality (RIN) and neutrophil count. None of the inferred factors

correlated with the height - a variable not expected to confound the gene expression levels. Together these results verify that PEER is identifying confounding factors in gene expression data.
It may be natural to question the decision of not including all measured covariates when removing confounding factors from the expression data. In defense of this work, there are no golden standard for which covariates to include. It is challenging to assess which approach performs best when cleaning expression data. However, PEER is a general framework, that does not need covariates as input - it will learn the confounding patterns from the data. For this reason, this work recommends running PEER without inputting covariates.

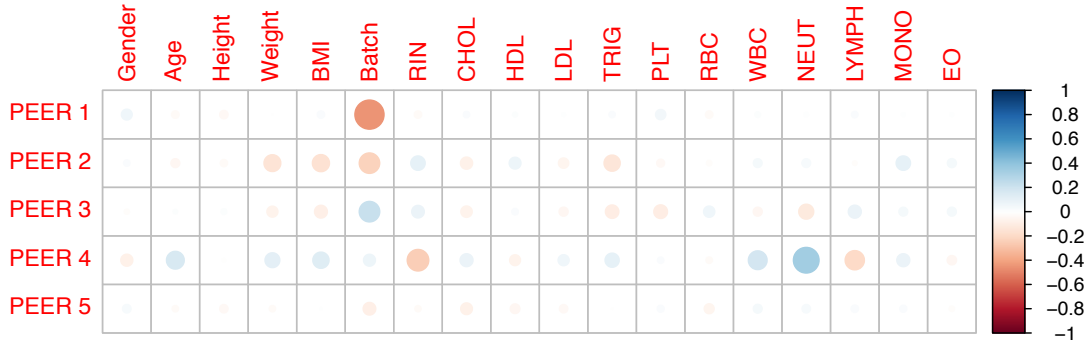


Figure A.7: Top 5 most important PEER factors correlations to measured covariates. `PEER k` represents the kth most important inferred PEER factors based on the factor variance, α (see Figure A.5, green colored factors). The size and color of the dot in the matrix represent the magnitude and direction of the Pearson correlation (r), respectively. See Table A.3 for a description of the variables.

Variable Name	Description
Gender	Individual gender
Age	Individual age
Height	Individual height
Weight	Individual weight
BMI	Body Mass Index
Batch	Sample Batch
RIN	RNA Integrity Number
CHOL	Cholesterol
HDL	High-Density Lipoprotein
LDL	Ligh-Density Lipoprotein
TRIG	Triglycerides
PLT	Platelet Count
RBC	Red Blood Cell count
WBC	White Blood Cell count
NEUT	Neutrophil count
LYMPH	Lymphocyte count
MONO	Monocyte count
EO	Eosinophil count

Table A.3: EGCUT measured covariates used for correlation analysis of the PEER inferred factors.

Residual Expression

An important step in any data modeling is learning the distribution of the data - before and after any transformation is applied to the data. The effect of variance stabilization and PEER factor removal on the sample expression value distribution is shown in Figure A.8. The distributions of samples exhibited higher variation and a right tail skew prior to PEER analysis. After PEER analysis the distributions of samples show a centralized normal distribution.

To investigate the magnitude of effect of using PEER, the original expression values were compared with the expression residuals after removal of confounding factors. For each individual we calculated the Pearson correlation of the expression values across all probes. The expression values showed low correlation ($r = 0.125 \pm 0.024$; Figure A.9), indicating that removing the inferred factors from expression data has large effects.

Together, these analyses illustrate the importance of removing confounding effects from gene expression data prior to downstream analysis.

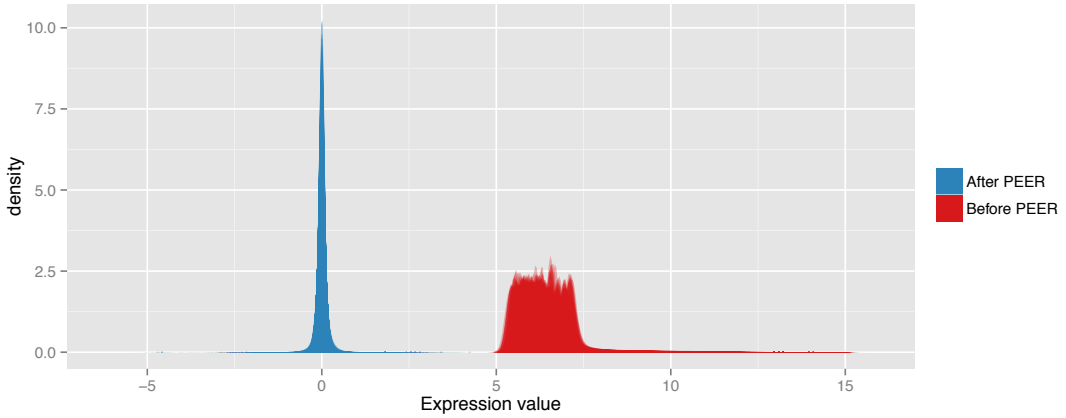


Figure A.8: Distribution of expression values across individuals. Values before PEER analysis are shown in red; values after $\log_2$ transformation and PEER analysis are shown in blue. The distribution of expression values for each individual across all gene probes is smoothed using a $\mathcal{N}(0, 1)$ kernel.

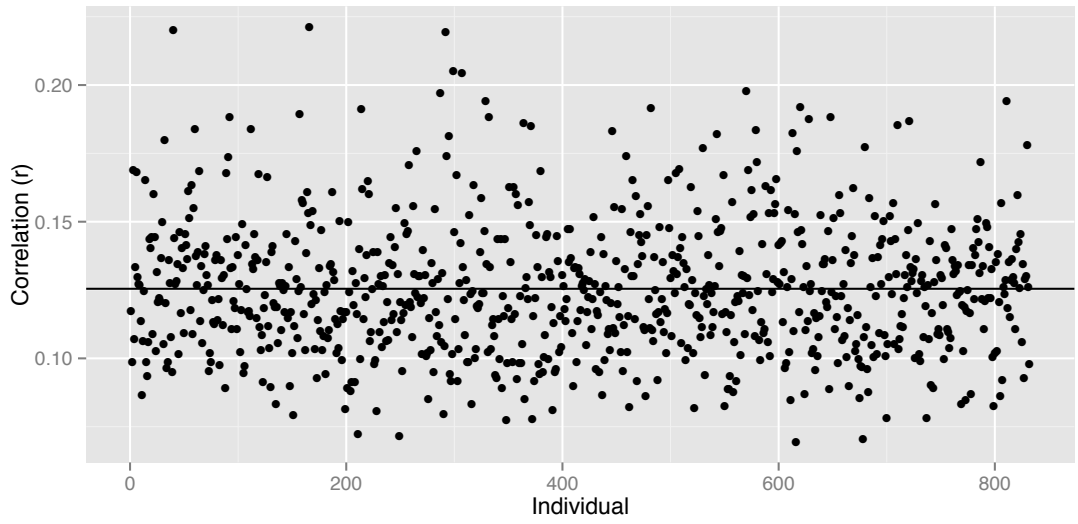


Figure A.9: Correlation between original expression values and PEER residuals. For each individual in the EGCUT cohort ($n = 832$) the Pearson correlation was computed across all probes. The horizontal line indicates the mean correlation ($r = 0.125 \pm 0.024$).

A.3 Hi-C Data Exploration

Learning the main characteristics of data from a novel high-throughput technology is a challenge to every researcher. All high-throughput technologies comes with certain biases that needs to be addressed. The data processing step often requires deeper insights into the technology and knowledge about how the data was generated. This section aims to describe the exploitative steps to characterize the Hi-C data and learn the structure and proper representation of the data.

I will also illustrate how important explorative analysis for quality control of the data. To this end, I will communicate the biological story of the K562 Hi-C data. The story is a good example of *data driven* computational biology: you learn the biological story from the data - without any hypothesis or domain specific knowledge. As you will see, without any prior knowledge of cancer biology, I learned about cancer chromosomal aberrations - just from a curious explorative analysis.

A.3.1 K562 cell line

This section presents an explorative analysis of the Hi-C data from the K562 cell line. The data originated from the first Hi-C publication (Lieberman-Aiden et al., 2009) and was reanalyzed by Lan et al. (2012) (see Section 5.3.2 on page 34 for details) Due to the early stage of the Hi-C protocol when the paper was published, the data has several limitations compared to later Hi-C publications. This section will highlight the most important features of the data set.

Since this study focuses on inter-chromosomal interactions, a natural first step for data exploration is to consider the distribution of inter-chromosomal interactions over chromosomes. Given that the model used to derive the set of genomic interactions is accurate, the expected distribution of interactions over chromosomes is uniform. However, the distribution for the K562 cell line was found to be highly non-uniform (Figure A.10). Chromosome 5 (17%), 9 (17%) and 22 (13%) together constitute almost half of the total inter-chromosomal interactions.

To investigate the cause of this non-uniform distribution, the genomic interactions were visualized using a circos plot (Yin, Cook, and Lawrence, 2012, `ggbio` R package). The circos plot shows interaction hotspots between chromosomes 9 and 22 (Figure A.11).

Further analysis revealed that the interaction hotspot between chromosome 9 and 22 is located on 9q and 22q. This interesting finding can be explained by the nature of the cell line. K562 cells are derived from the bone marrow of a patient who had chronic myelogenous leukemia (CML) (Lozzio and Lozzio, 1975). A hallmark of CML is the presence of the Philadelphia chromosome, a specific chromosomal abnormality that is the result of a reciprocal translocation between chromosome 9 and 22 (Drexler, MacLeod, and Uphoff, 1999). This gives rise to a fusion gene, bcr-abl, that juxtaposes the *Abl*1 gene on chromosome 9 (region q34) to a part of the *BCR* ("breakpoint cluster region") gene on chromosome 22 (region q11). This

means that the interactions observed in K562 cells between 9q34 and 22q11 are labeled as inter-chromosomal, although they are actually intra-chrosomomal in the K562 genome. This work found that 20.6% of the inter-chromosomal interactions are between chromosome 9 and 22.

These findings are interesting for two reasons: firstly, it indicates that sizable fractions of the inter-chromosomal interactions identified in cancer cells may be intra-chromosomal for that particular genome - or vice versa; secondly, it highlights that Hi-C data can be used to study chromosomal translocations. Translocations in cancer cell lines can be detected since they exhibit higher than expected inter-chromosomal Hi-C contact intensities. Hi-C data have also been used to study the causality of translocations. Engreitz, Agarwala, and Mirny (2012) found that the three-dimensional genome architecture shapes the landscape of rearrangements and further demonstrated that the oncogenic *BCR-ABL* gene fusion results from elevated Hi-C contact frequencies.

Figure A.11 also shows that the largest hotspot is located on chromosome 5. A set of three genomic regions on chromosome 5 was found to be involved in 28.1% interactions (Table A.4). These hotspots are likely an artifact of the Poisson model used by Lan et al. (2012) and for this reason they were removed during QC (see Section 5.3.2 on page 34).

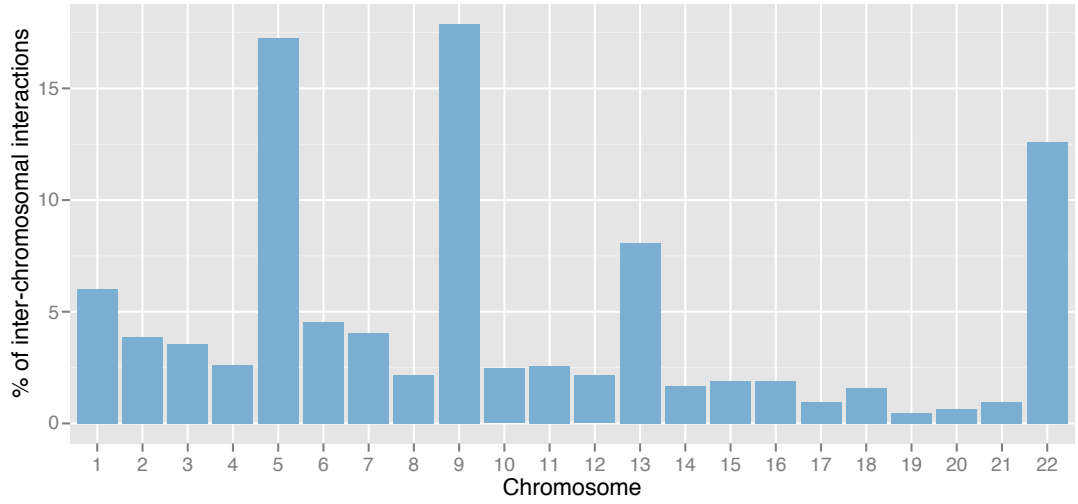


Figure A.10: Distribution of percentage of inter-chromosomal interactions over autosomal chromosomes. Note that interactions are not normalized by the chromosome length.

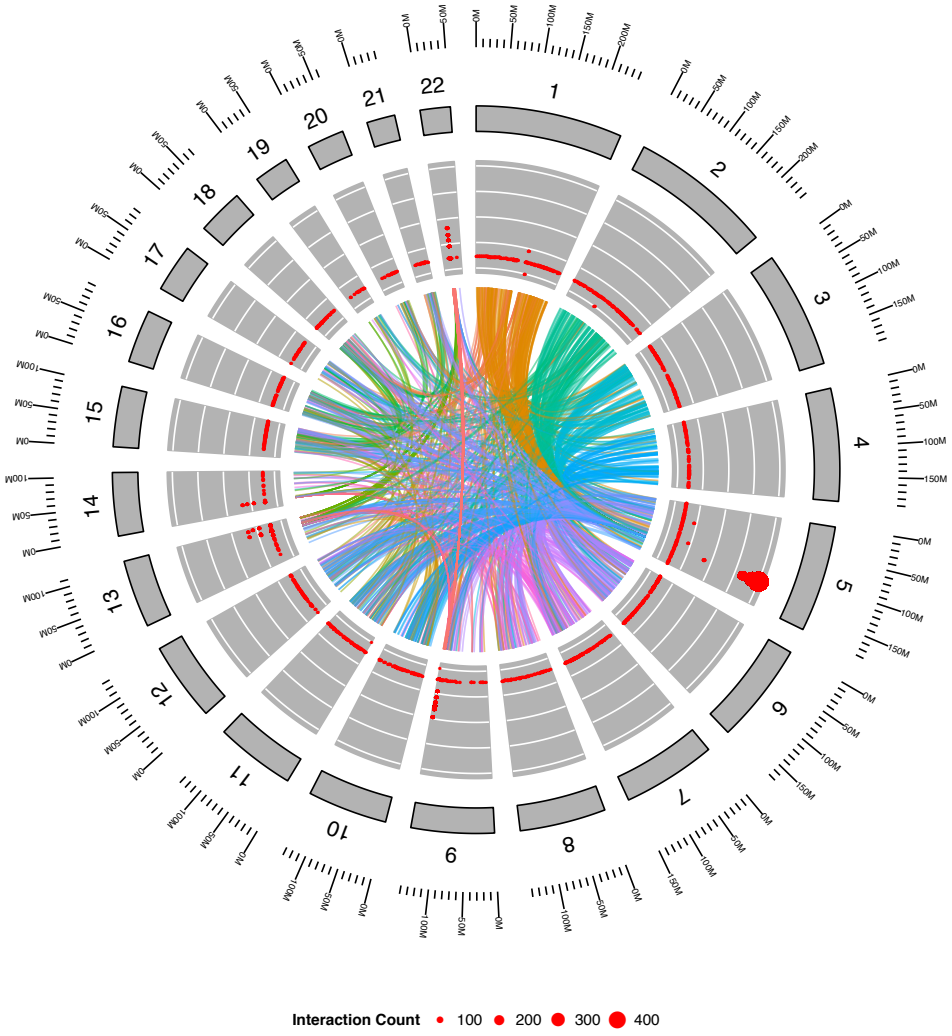


Figure A.11: Circos plot of autosomal inter-chromosomal interactions for Lan *et al.* data ($n_{interactions} = 3497$). Edges represents interactions between genomic regions; edges are colored according to their chromosome pair. $\log_2$(interaction count) for a given genomic position is shown as the radii of the red points. The "hotspots" between chromosome 9 and 22 can be explained by the presence of Philadelphia chromosome in the K562 cell line. Hotspots on chromosome 5 remains unexplained. Three outliers (top hotspots) were later removed, see Table A.4 on the next page and Section 5.3.2 on page 34.

Rank	Chromosome	Position	Interaction Count
1*	chr5	133628145	478
2*	chr5	133628590	324
3*	chr5	133625948	174
4	chr9	135862266	20
5	chr5	133621686	14
6	chr22	22244206	12
7	chr9	135873437	12
8	chr9	135919455	12
9	chr9	135481499	12
10	chr9	135660195	12

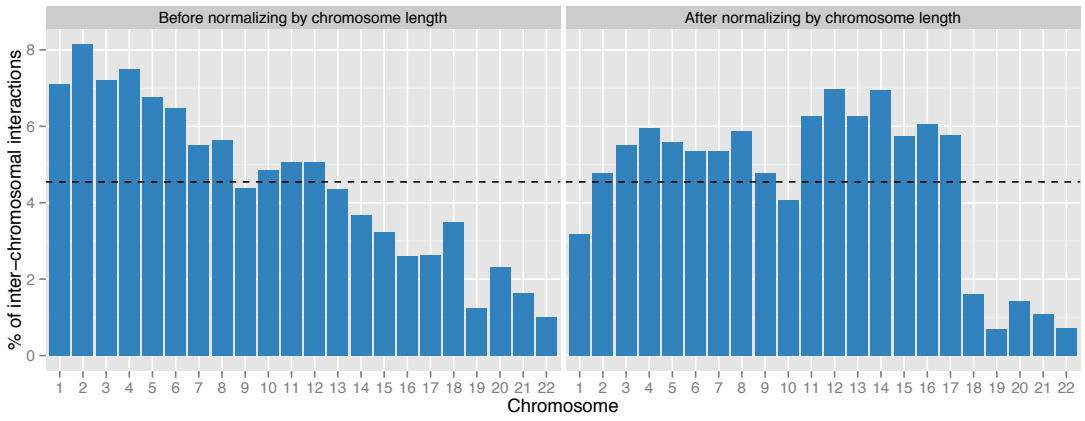
Table A.4: Top 10 interacting genomic regions in the K562 Hi-C data.
*Genomics regions were later removed during outlier detecting.

A.3.2 Dixon et al. data

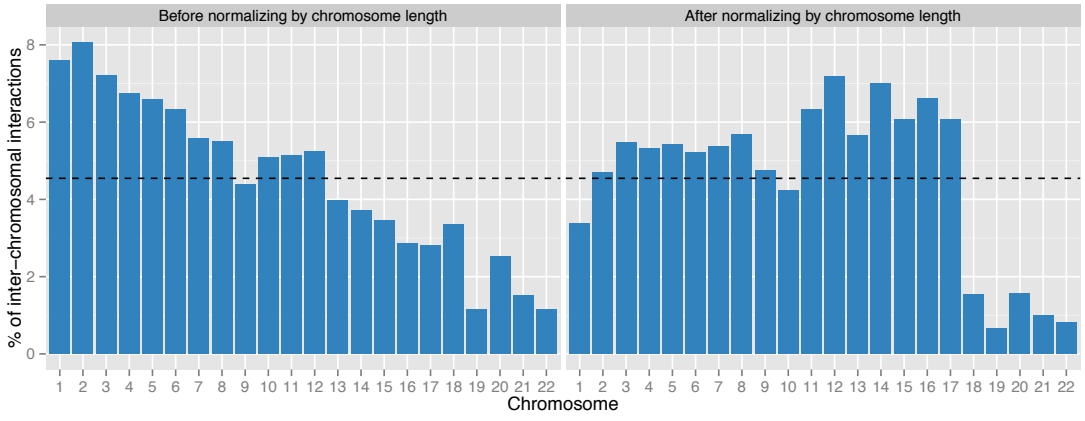
This section gives an overview of the Hi-C data produced by Dixon et al. (2012). See Section 5.3.1 on page 33 for a description of the methods applied to the data. The authors performed deep sequenced Hi-C experiments in human embryonic stem cells (hESC) and human lung fibroblast (hIMR90). The deep sequencing allows for a much higher resolution compared to previous studies (e.g. the K562 data presented in Appendix A.3.1 on page 72). For the hESC and hIMR90 cells, significant genomic interactions was assessed by the software Fit-Hi-C (Ay, Bailey, and Noble, 2014a). Briefly, Fit-Hi-C assesses significance of interactions by correcting for the 1D genomic distance affecting the Hi-C frequency counts (see Section 3.2 on page 15 for details). Fit-Hi-C uses the Benjamini-Hochberg procedure to compute the FDR, and deriving a q-value, defined as the minimum FDR that can be attained when calling that interaction significant(Storey, 2002). In this work, the q-value was used to control the number of interactions to include in the analysis*. Firstly, the q-value cut-off can be used to define a set of high confidence interactions. Secondly, it provides a practical way of limiting the computation time - it is unfeasible to include all interactions in the analysis and illustrations.

Similar to the analysis of the K562 cell line, the distribution of inter-chromosomal interactions over chromosomes were investigated. The analysis did not indicate any clear outlier chromosome (Figure A.12), although the distribution was not uniform - even after accounting for chromosome length. Chromosome $19 - 22$ contributed little to the inter-chromosomal interactions. The small chromosomes $(17 - 22)$ are generally not mappable to begin with, so there will be little Hi-C data mapping to them. It is possible that normalizing by the effective chromosome length may result in a more uniform distribution. The effective chromosome length accounts for black listed regions, repeats and centromeres as described by Rosenbloom et al. (2013) and UCSC Genome Browser (2015).

* It should be noted that this study later discovered that the q-value is not the most intuitive way of filtering the data for *inter-chromosomal* interactions. Contact count provides a more intuitive measure for doing stratified analysis based on e.g. interactions with contact count ≥ 30.



(a) hIMR90



(b) hESC

Figure A.12: Distribution of percentage of inter-chromosomal interactions over autosomal chromosomes. The dashed line represents a uniform distribution.

The Hi-C data from Dixon et al. (2012) contains information about the raw contact count of the genomic interaction. The contact count is the number sequence reads for the interaction and represents the strength of the interaction. For both cell lines, the majority of interactions have low contact counts. Such interactions will not be considered as significant when analyzed by Fit-Hi-C using a q-value threshold of 10^{-3}. Both cell lines have more intra- than inter-chromosomal interactions. The difference between the total interaction count of the two cell lines can be explained by the difference in sequencing depth (Dixon et al., 2012, Supplementary Table 1). Because the difference in total interaction count affects the hypothesis space for assessing statistical significance, the stringency of the q-value cut-off cannot be directly compared between the two cell lines. The technical details of this problem is beyond the scope of this text.

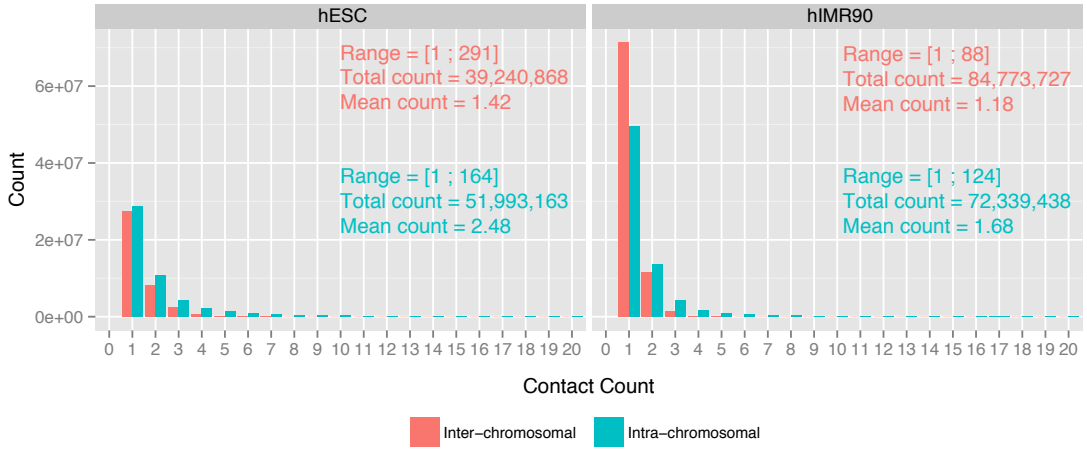


Figure A.13: Distribution of contact counts for two cell lines from Dixon *et al.*. Both intra- and inter-chromosomal interactions are included. Only the contact count range $[0; 20]$ is shown for simplicity of the figure. Summary data of contact counts for each cell line and interaction type is shown in colored text.

Similar to the K562 cell line, the properties of Dixon *et al.* Hi-C data was explored using a circos plot. Figures A.14 and A.15 show no sign of outliers or interaction "hotspots". That is, no single interaction fragment can account for the interaction count observed from its respective chromosome. The interactions are distributed across the whole chromosome. This is in contrast to the data from the K562 cell line shown in Figure A.11 on page 74. The circos plots only serve to visualize the general structure and patterns of the data and do not allow to directly compare the difference in the Hi-C data between the cell lines. This is because the scale of the visualized attributed is different for the two cell lines.

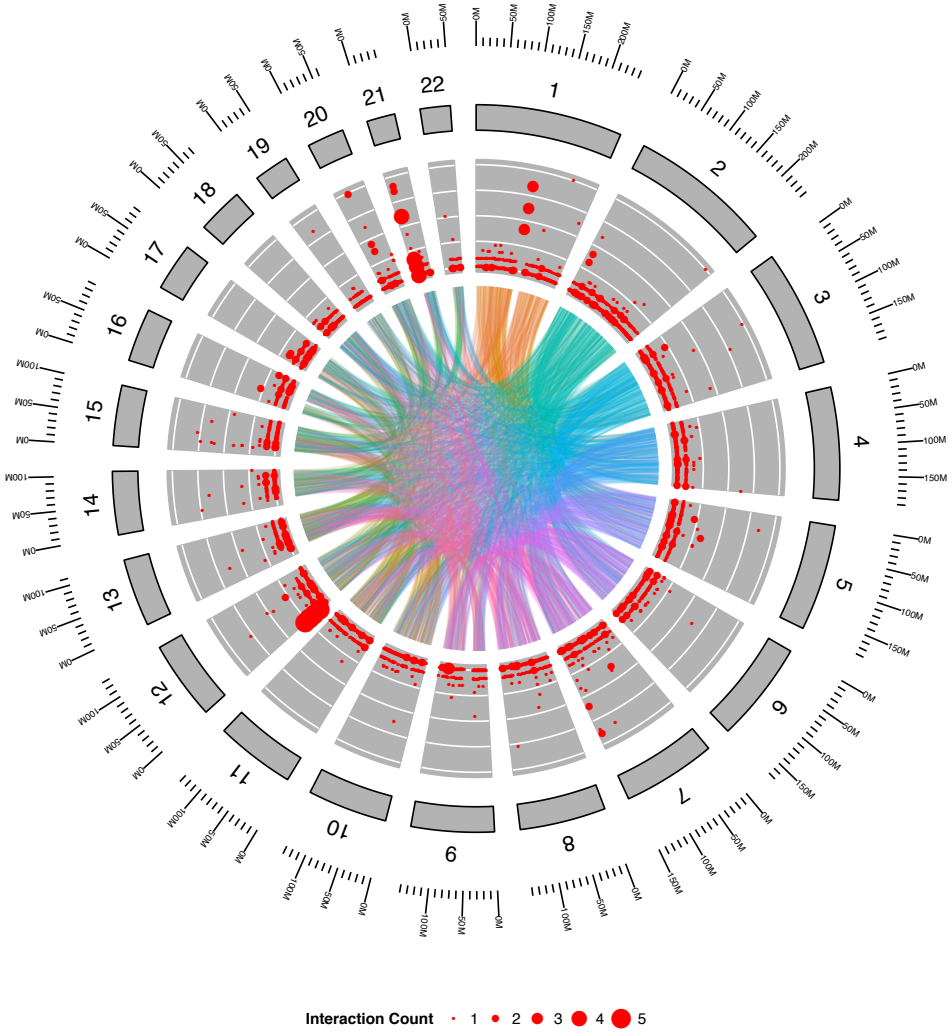


Figure A.14: Circos plot of autosomal inter-chromosomal interactions for hIMR90 cell line. $n_{interactions} = 3468$ were included in the visualization ($q = 10^{-8}$). Edges represent interactions between genomic regions; edges are colored according to their chromosome pair. The contact count for a given genomic position is shown as the radii of the red points; the size of the red points represents the interaction count.

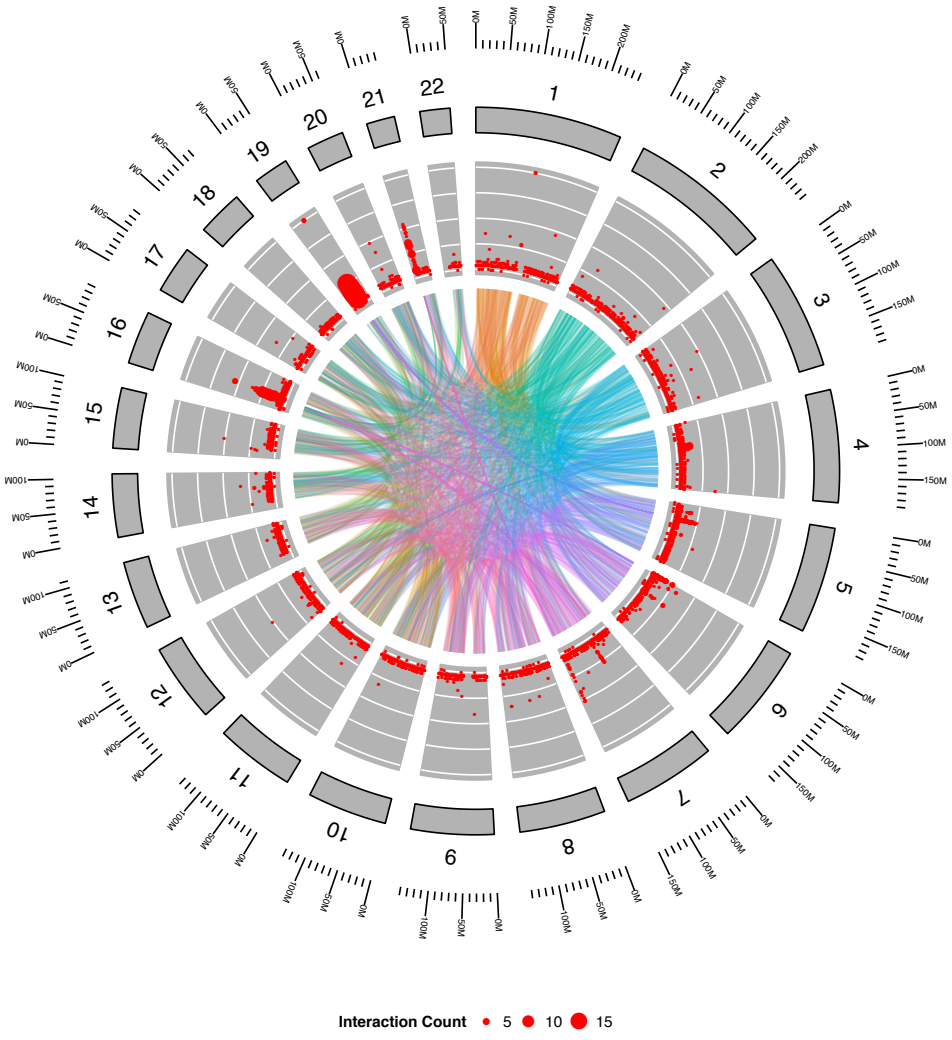


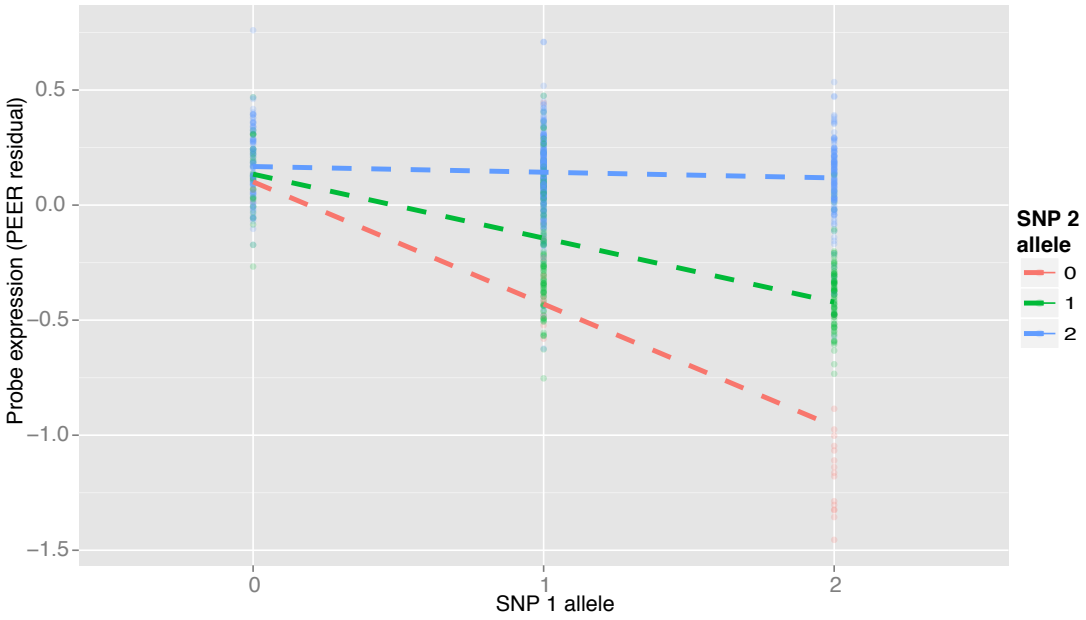
Figure A.15: Circos plot of autosomal inter-chromosomal interactions for hESC cell line. $n_{interactions} = 2665$ were included in the visualization ($q = 10^{-16}$). See Figure A.14 on page 78 for further descriptions.

Supplementary Analysis and Results B

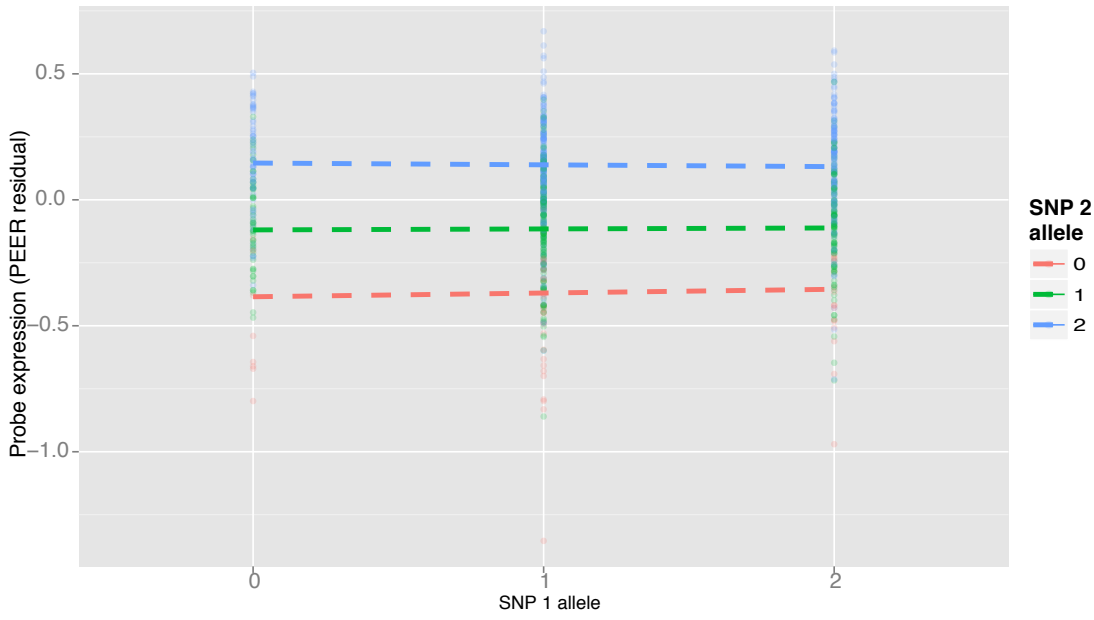
B.1 Examples of Hemani *et al.* Epistatic SNP-probe Pairs

This section provides examples of epistatic SNP-probe pairs published by Hemani et al. (2014, Supplementary Table 1). Plotting the data for all published SNP-probe pairs is uninteresting and beyond the scope of this section. The plots serve as a window into the "raw" data underlying the epistasis discovery. The intention is to familiarize the reader with what epistasis look like with "real data".

Two examples were chosen to illustrate the extreme cases of two different outcomes when replicating the work of Hemani *et al.*. The first example, Figure B.1 (**a**), shows a very strong epistatic SNP-probe pair that could be replicated in this work. The second example, Figure B.1 (**b**), shows a SNP-probe pair that, in contrast to what Hemani *et al.* found, is clearly is not epistatic.



(a) rs807491:rs7254601 is epistatic to the gene TMEM149 (Illumina ProbeID ILMN1786426). This is the most significant epistatic pair of all the SNP-probe pairs published by Hemani *et al.* $\beta_{interaction} = 0.25$, P-value $= 6.85 \cdot 10^{-45}$.



(b) rs10435352:rs1883613 is clearly not epistatic to the gene VNN2 (Illumina ProbeID ILMN1678939). $\beta_{interaction} = -0.01$, P-value $= 0.53$.

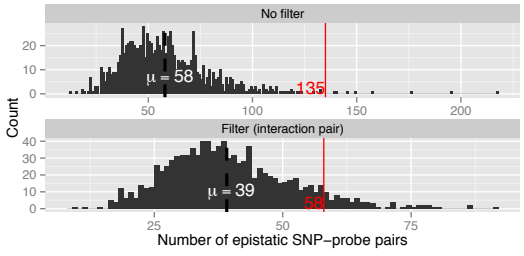
Figure B.1: Examples of epistatic SNP-probe pairs published by Hemani *et al.* The colored dashed line represents the linear model fitted to the data. Individual observations are represented by colored points. SNP 1 and 2 refer to the identifiers with the format SNP1:SNP2 shown in the subcaptions.

B.2 Spatial Epistasis Enrichment Histograms

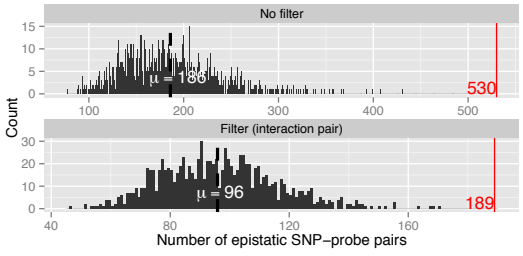
This section visually shows the evidence of presence or absence of spatial proximate epistasis for the cell lines analyzed in this work. This section presents the epistasis enrichment histograms that underlie the empirical P-values shown in Table 7.1 on page 49. These P-values describe the extremity of the spatial proximate epistasis observation, but they do not capture information about the scale of the distance to the null. The histograms shown in this section provide a meaningful way of assessing the degree of extremity in units of number of SNP-probe pairs.

The shape of the null distribution is on its own interesting to study. All null distributions analyzed in this work were characterized by a bell-shape with long right tail (right-skewed).

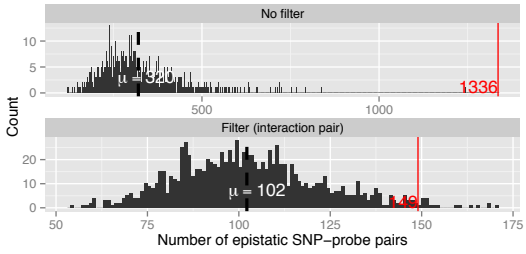
Surprisingly, the P-values for the unfiltered and filtered SNP-probe pairs were found to be similar (the filtering of SNP-probe pairs is described in Section 6.5 on page 43). This is illustrated in Figures B.2 to B.5 by considering the position of horizontal red line relative to the null distribution. The chromosome pair filtering strategy used in Hemani et al. (2014) was found to give close to identical results as the interaction pair filtering approach (data not shown).



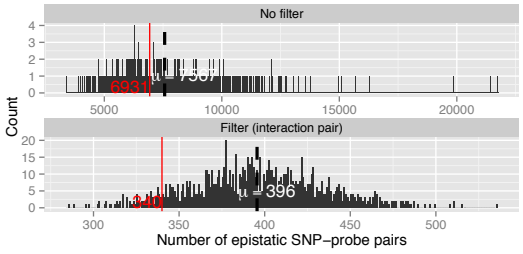
(a) $l = 500$ bp, q-value $= 10^{-6}$



(b) $l = 1,000$ bp, q-value $= 10^{-6}$

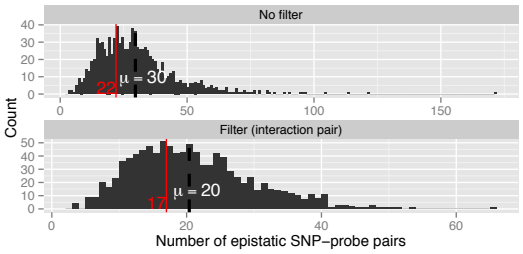


(c) $l = 2,500$ bp, q-value $= 10^{-7}$



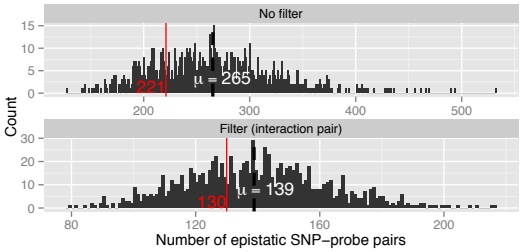
(d) $l = 50,000$ bp, q-value $= 10^{-9}$

Figure B.2: Histogram plots for epistasis enrichment for hIMR90 cell line. The top panel shows the unfiltered epistatic SNP-probe pair counts. The bottom panel shows the counts after filtering. The black colored histogram shows the null distribution of number SNP-probe pairs using $n = 1,000$ null samples. The red horizontal line shows the observed number of epistatic SNP-probe pairs for the spatial proximate genomic regions. The results are shown for different interaction width (l) and q-value thresholds in (**a**)-(**d**). Each of the four panels represents different parametrizations and contain two components. The first component shows the results without applying any filter. The second component shows the results after applying the "interaction pair" filter.

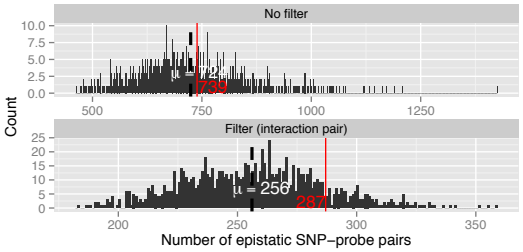


(a) $l = 500$ bp, q-value $= 10^{-14}$

(b) $l = 500$ bp, q-value $= 10^{-16}$

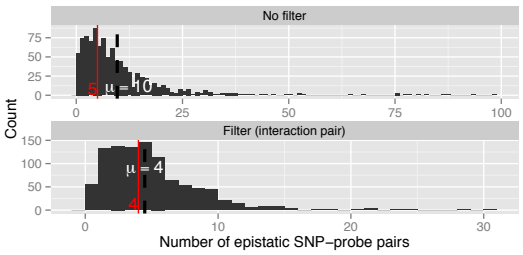


(c) $l = 1,000$ bp, q-value $= 10^{-12}$

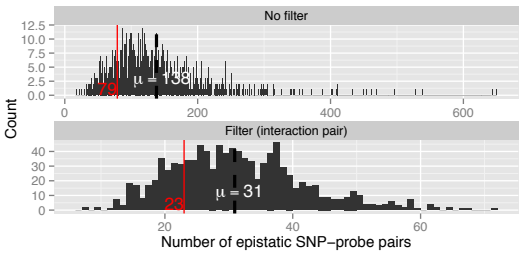


(d) $l = 2,500$ bp, q-value $= 10^{-13}$

Figure B.3: Epistasis enrichment for hESC cell line. See Figure B.2 for further description of the figure.

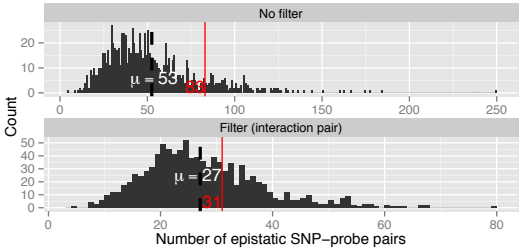


(a) $l = 1,000$ bp

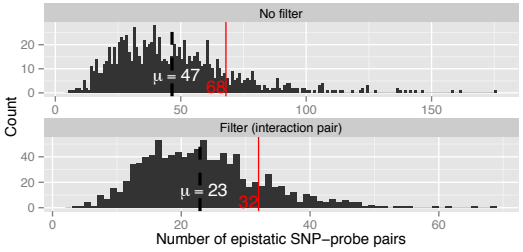


(b) $l = 5,000$ bp

Figure B.4: Epistasis enrichment for K562 cell line. No q-value threshold was used, because the analyzed K562 Hi-C data does not support it. See Figure B.2 for further description of the figure.



(a) hESC cell line, $l = 1,000$ bp



(b) hIMR90 cell line, $l = 1,000$ bp

Figure B.5: Epistasis enrichment for the negative controls with contactCount $= 1$. See Figure B.2 for further description of the figure.

B.3 Minimum GCC Filter

This section illustrates the importance of using a minimum genotype class count (GCC) filter to ensure that the model estimates can be interpreted correctly.

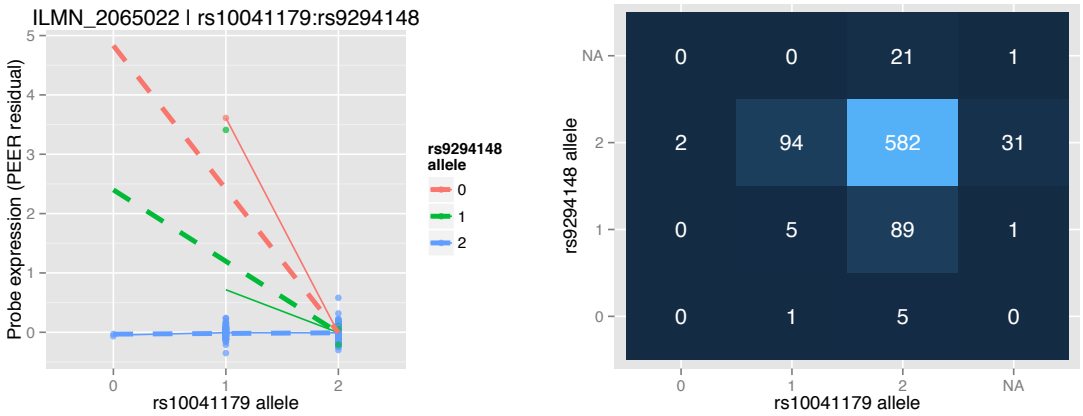
GCC and MAF MAF is the most widely used measure to exclude low frequency variants from genetic studies (e.g. GWAS lacking statistical power to detect rare variant associations). However, when modeling epistasis it is more useful to think in terms of minimum GCC and not MAF for setting a threshold of the SNP frequencies. The reason should become apparent in the next two paragraphs. Minimum GCC measures how well the two-locus genotype count matrix is populated and is in units of "data points". It is worth noting that the minimum GCC is tightly coupled with the minimum MAF of the SNP pair (Pearson correlation $r = 0.65$ in the EGCUT cohort).

Multiplicative Model and Minimum GCC This work found that the multiplicative model leads to spurious epistasis results for SNP pairs with low GCC. (Readers are encouraged to brush up their memory of epistatic models in Section 2.2.1 on page 8 before continuing studying this section.) Low GCC causes the model to *overfit*[*] the data. In the extreme case, when the minimum GCC is zero (i.e. missing genotype classes), the model is *extrapolating*, (Figure B.6 (**a**)). This leads to misleading model estimates, i.e. highly significant P-values and very large effect sizes. A detailed description is given in caption of Figure B.6.

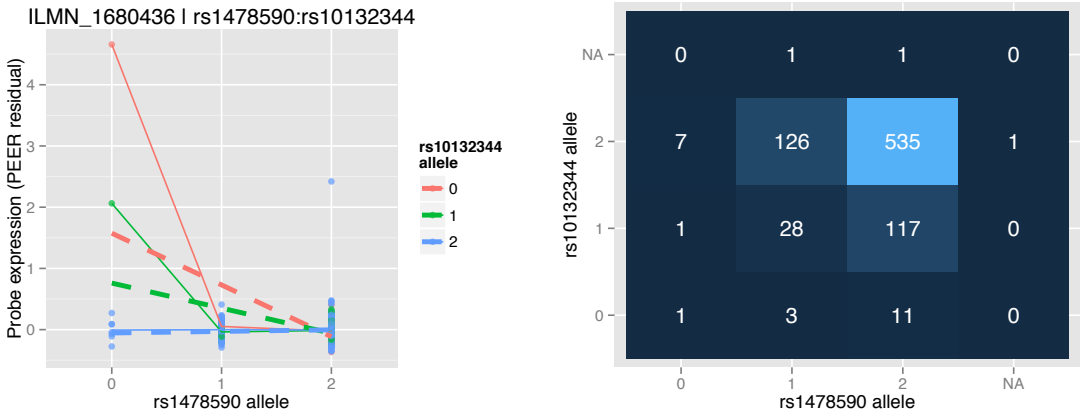
Minimum Genotype Class Count and *P*-value Bias To investigate the scale of the problem with low GCC SNP pairs confounding the analysis, a random sample of 100,000 significant SNP-probe pairs was drawn from the output of the spatial epistasis discovery of hIMR90$[l = 2,500\text{bp}; q = 10^{-7}]$. All SNP-probe pairs were significant after Bonferroni correction.

SNP pairs with missing genotype classes were found to be the most prevalent class, constituting $> 60\%$ of the "epistatic" SNP-probe pairs (Figure B.7 (**a**)). The P-values for epistasis was biased by the minimum genotype class counts: SNP pairs with low minimum genotype class count showed the most significant P-values (Figure B.7 (**b**)).

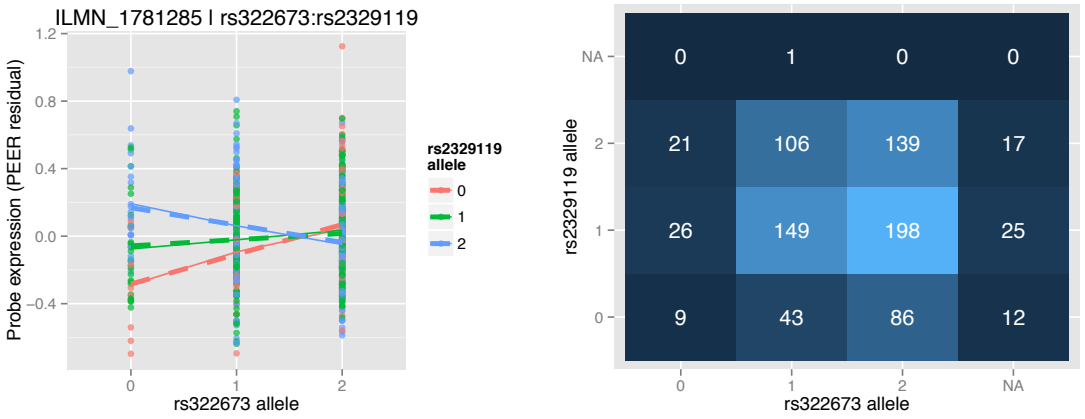
[*] The multiplicative model describes noise instead of the underlying expression-genotype relationship. Overfitting generally occurs when a model has too many parameters relative to the number of observations. Such a model will generally have poor predictive performance.



(a) Minimum GCC = 0. This example illustrates the problem of overfitting with the multiplicative model in case of *missing* GCC. This is an extreme example of a SNP-probe pair where the fitted multiplicative model suggest a strong epistatic signal ($\beta_{interaction} = 1.21$, P-value $= 2.00 \cdot 10^{-110}$), but in fact there is no evidence of epistasis.

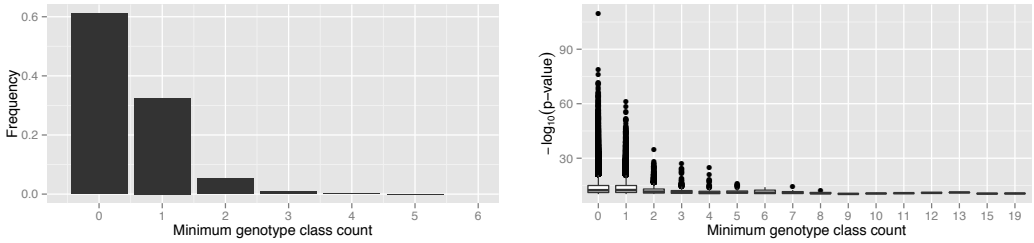


(b) Minimum GCC = 1. The model reports epistasis ($\beta_{interaction} = -0.44$, P-value $= 1.29 \cdot 10^{-39}$), although there little evidence of epistasis. The one minor allele homozygote individual $(0, 0)$ is a high leverage point and greatly effects the estimate of interaction regression coefficient.



(c) Minimum GCC 9. An example of statistical epistasis using a multiplicative model ($\beta_{interaction} = -0.14$, P-value $= 4.07 \cdot 10^{-11}$). No individuals are high leverage points, in part because there are no low GCC. The fitted model (thick dashed line) fits well to the mean of the observations (full line).

Figure B.6: Examples of the effect of minimum GCC when using a multiplicative model. The **left component** of each panel shows an interaction plot. The thick dashed line represents the fitted linear model; the full line represents the mean of the observations for combinations of alleles. A good model fit, without presence of overfitting and high leverage points, is characterized by the two lines being similar. The **right component** of each panel shows the matrix representation of two-locus genotype class counts. The minimum genotype class count is the lowest value in the matrix, when excluding missing data points (`NA`). The sum of all cells in matrix is 832, corresponding to the number of individuals in the EGCUT cohort.
See Table B.3 on page 93 for additional data on the SNP-probe pairs shown in this figure.



(a) Prevalence of minimum genotype class counts.

(b) P-value bias for low minimum genotype class counts.

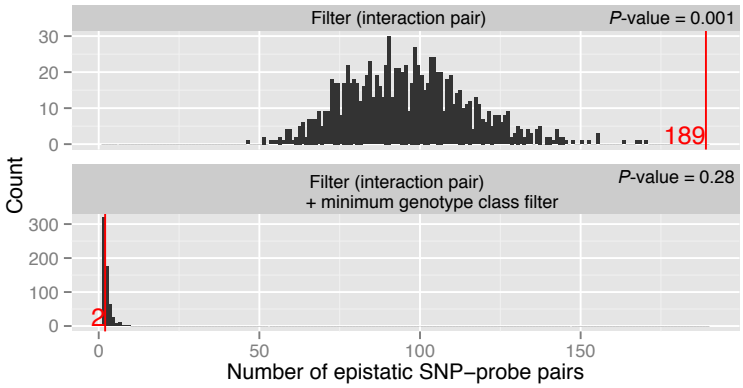
Figure B.7: Analysis of 100,000 random significant SNP-probe pairs. The pairs were drawn from the output of the spatial epistasis discovery of hIMR90$[l = 2,500\text{bp}; q = 10^{-7}]$. Notice that the x-limit of the left plot is restricted to $[1 \ldots 6]$ because class counts greater than this interval constitute $< 0.03\%$ of all SNP-probe pairs analyzed.

B.4 Deeper Insights into the hIMR90 Spatial Epistasis Enrichment

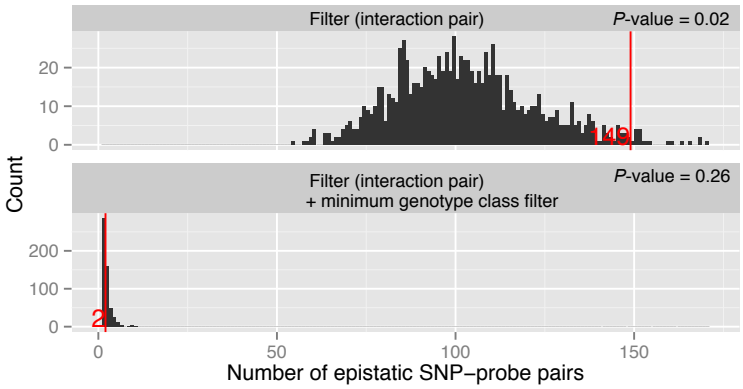
This section follows up on the seemingly spatial epistasis enrichment signals for two parametrizations of 3D genomic proximity for the hIMR90 cell line: hIMR90$[l = 1,000\text{bp}; q = 10^{-6}]$ and hIMR90$[l = 2,500\text{bp}; q = 10^{-7}]$ (see Table 7.1 on page 49). This section complements Section 7.2.2 on page 50 with plots and more detailed descriptions of the analysis procedure.

Minimum Genotype Class Count (GCC) Filtering As illustrated in the computational pipeline overview in Figure 6.2 on page 45, the SNP-probe pairs are filtered using an "interaction pair" filter. This filter keeps independent SNP-probe pairs by accounting indirectly for the LD structure in the genotype data. However, this filter does not ensure that the epistasis model fits the data and that the model estimates are reasonable. For this reason a second "GCC based" filter was applied, retaining only SNP pairs with minimum GCC ≥ 3.

To evaluate the effect of the minimum GCC filter, the number of significant SNP-probe pairs before and after applying the filter was calculated. Table B.1 shows that $\approx 1\%$ of the SNP-probe pairs remains after applying the GCC filter. The dramatic decrease in SNP-probe pairs indicates that the significance of the majority of the SNP pairs is due to *overfitting*. Furthermore, only two spatial proximate SNP-probe pairs remain for both hIMR90 parametrizations. Figure B.8 shows that after applying the GCC filter, the spatial epistasis enrichment is removed. Together these results reveal that the spatial epistasis enrichment shown in Table 7.1 on page 49 is caused by low minimum GCC SNP-probe pairs with little evidence of epistasis.



(a) $l = 1,000$ bp, q-value $= 10^{-6}$



(b) $l = 2,500$ bp, q-value $= 10^{-7}$

Figure B.8: Effect of minimum genotype class filter on enrichment histograms for the hIMR90 cell line. The panels (**a**) and (**b**) show the two parametrizations of hIMR90 found the be enriched for spatial epistasis. The top component in each panel shows the histogram of SNP-probe pair counts after applying the "interaction pair" filter (the same histogram is shown in Figure B.2 on page 83). The bottom component in each panel shows the histogram after applying the minimum genotype class filter with a minimum class count of three. After applying the minimum genotype class filter, the spatial epistasis enrichment is clearly removed. See Figure B.2 on page 83 for further description of the figure elements.

SNP-probe pair counts

Parametrization	Interaction pair filter		Interaction pair filter + minimum GCC filter	
	All pairs	Spatial pairs	All pairs	Spatial pairs
hIMR90$[l = 1,000\text{bp}; q = 10^{-6}]$	96,083	189	1,087	2
hIMR90$[l = 2,500\text{bp}; q = 10^{-7}]$	102,339	149	984	2

Table B.1: Effect of minimum GCC filtering on SNP-probe pair counts. The table shows counts of significant SNP-probe pair using the interaction pair filtering and in combination with minimum GCC filtering. A drastic decrease in SNP-probe pairs is observed when the minimum GCC filter is applied. "All pairs" counts include SNP pairs from the null and spatial proximate pairs. "Spatial pairs" counts include only the spatial proximate SNP pairs. Chromosomal positions, p-values, effect size estimates and more data appear in Table B.4 on page 94 for the two "spatial pairs" on the right hand side of the table are.
See Figure 6.2 on page 45 for an description of the interaction pair filter.

Reverse Order Filtering It is possible that the order of the filters influences the enrichment results. The interaction pair filtering retains only the most significant SNP-probe pair for a given Hi-C interaction pair. This work previously showed that the most most significant SNP-probe pair is biased towards low genotype class counts (see Appendix B.3 on page 85, Figure B.7). Figure B.9 shows that the distribution of SNP-probe pair counts is different when first applying the minimum genotype class filter - it is not bell shaped as previously seen. However, the end result, i.e. enrichment *P*-value after applying the interaction pair filter, remains nearly unchanged for the reversed order filters.

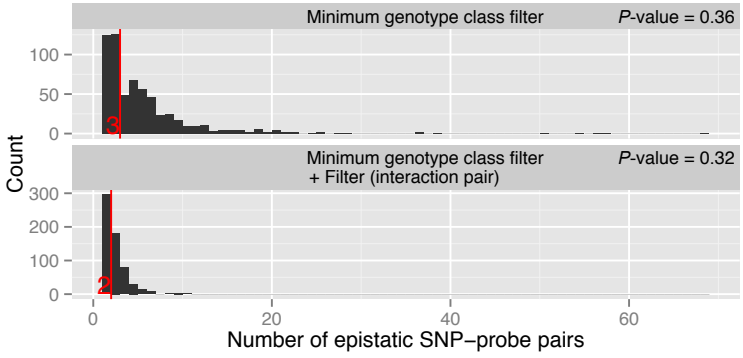


Figure B.9: Effect of reversing the order of the SNP-probe pair filters for the hIMR90$[l = 2,500\text{bp}; q = 10^{-7}]$ parametrization. The filters were applied in the reverse order compared to Figure B.8 (**b**). See Figure B.2 on page 83 for further description of the figure elements.

SNP-probe Pair Co-localization This section presents the SNP-probe pair co-localization analysis. Co-localizing is defined as a SNP-probe pair where at least

SNP-probe co-localization

Parametrization	Interaction pair filter		Interaction pair filter + minimum GCC filter	
	All pairs	Spatial pairs	All pairs	Spatial pairs
hIMR90[$l = 1,000$bp; $q = 10^{-6}$]	9.87% (9,482/96,083)	8.99% (17/189)	9.07% (106/1169)	50.00% (1/2)
hIMR90[$l = 2,500$bp; $q = 10^{-7}$]	9.68% (9,902/102,339)	7.38% (11/149)	9.04% (95/1,051)	50.00% (1/2)

Table B.2: SNP-probe pair co-localization. The table shows the percentage of SNP-probe pairs co-localizing. The numbers in parenthesis indicates the underlying counts. "All pairs" counts include SNP pairs from the null and spatial proximate pairs. "Spatial pairs" counts include only the spatial proximate SNP pairs. The **right side** of the table lists the co-localization after minimum GCC filtering.

one of the epistatic SNPs is localized on the same chromosome as the probe.

B.5 Epistasis Tables

I here present selected data for the (putative) epistatic interactions discovered in my thesis. I will only present epistasis data discovered when searching for spatial epistasis in the hIMR90 cell lines. I selected three categories of epistatic effects, grouped into three tables, that summarizes some of the knowledge obtained during this study. Importantly, the tables serve to illustrate the level of genomic and genetic annotations used in this thesis.

Table B.3 lists SNP-pairs with different confidence of epistasis (as measured by the minimum GCC). The SNP-pairs given in the table are also plotted in Figure B.6 on page 87. The table exemplifies the spurious epistatic SNPs underlying the hIMR90 spatial epistasis enrichment. Lastly, the table serves as an example of the importance of scrutinizing your results.

Table B.4 lists the two putative epistatic interactions that map to interacting genomic regions (spatially epistasis). Arguably, there is little evidence to support the spatial epistasis hypothesis.

Table B.5 lists the six SNP-probe pairs with the strongest evidence of epistasis found in this thesis. It is noteworthy that this short list of epistatic SNPs is the only confident epistatic signal detected in this work*. Consider that this list of epistatic effects are the result of testing trillions of SNP-pair interactions. Hence, the introduction of my thesis was appropriate when proclaiming "Uncovering the Epistatic Needles in Genome- Wide Haystacks" (Section 1.2 on page 2). Indeed, detecting epistasis remains technically challenging. Finally, it is worth noting that

* This is strictly not true. It is possible that the hESC or K562 cell line contained additional strong epistatic effects, that did not make it into this analysis. However, the overall tendency is unlikely to chance.

these epistatic SNPs have, to the best of my knowledge, not previous been reported. As such, they can be considered new findings credited to this work.

SNP1	rsID	MAF1	SNP2	rsID	MAF2	P-value	Beta	minGCC	Gene	Probe	EIID	EID
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	2.63E-23	-0.5636	0	ITK	5:156614448	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.46E-26	0.5306	0	CREB3L2	7:137210715	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	8.02E-18	0.5514	0	FAIM3	1:205144550	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.67E-31	0.5051	0	EGR2	10:64242090	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	7.56E-25	0.5474	0	LRMP	12:25152225	null_357_4752	null_357
5:152896308	**rs10041179**	**0.07**	**6:80265632**	**rs9294148**	**0.07**	**2.00E-110**	**1.2144**	**0**	**KIAA0672**	**17:12835495**	**null_357_4752**	**null_357**
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	8.45E-17	0.3753	0	KIAA1407	3:115166080	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	6.25E-13	0.4408	0	VIL2	6:159106859	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.63E-11	0.6132	0	FCRL2	1:155982443	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	2.35E-13	0.3674	0	SIAH1	16:46952618	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	4.57E-18	0.4974	0	COL9A2	1:40538963	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	7.09E-36	0.5992	0	FMOD	1:201576664	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.77E-30	0.5593	0	GPT2	16:45522597	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	2.60E-18	0.4203	0	STARD7	2:96214684	null_357_4752	null_357
5:152896308	rs10041179	0.07	6:80265632	rs9294148	0.07	1.01E-25	0.6053	0	FCRL5	1:155750161	null_357_4752	null_357
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	9.17E-52	0.2780	1	ASS1	9:132310134	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	1.23E-14	0.1545	1	RHOBTB1	10:62373868	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	8.18E-21	0.1546	1	DZIP3	3:109896160	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	4.13E-15	0.1247	1	FAM90A2P	8:7104545	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	1.73E-24	0.2147	1	GLP1R	6:39163079	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	1.40E-30	0.1971	1	TXNRD2	22:18299943	null_339_6338	null_339
2:218046715	**rs1478590**	**0.11**	**14:69666638**	**rs10132344**	**0.11**	**1.29E-39**	**0.4352**	**1**	**CSHL1**	**17:59341967**	**null_339_6338**	**null_339**
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	8.18E-17	0.1293	1	NCOR2	12:123377955	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	4.96E-15	0.1320	1	NUDT4P1	1:143848976	null_339_6338	null_339
2:218046715	rs1478590	0.11	14:69666638	rs10132344	0.11	1.52E-27	0.1737	1	C1orf200	1:9637055	null_339_6338	null_339
3:25307427	**rs322673**	**0.26**	**13:79204571**	**rs2329119**	**0.42**	**4.07E-11**	**-0.1396**	**9**	**DUSP1**	**5:172128082**	**null_974_3900**	**null_974**

Table B.3: Examples of epistastic effects with different levels of confidence (measured by minimum GCC). The SNP-probe pairs are sorted by their minimum GCC. The table presents data for three exemplary SNP-pairs. SNP-probe pairs marked in bold are plotted in Figure B.6 on page 87. This table serves to illustrate some of the spurious epistatic effects (minGCC < 3) underlying the hIMR90 spatial epistasis enrichment. `P-value`: 1-*df* test for the significance of the genetic interaction term; `Beta`: effect size estimate of the genetic interaction term; `Gene`: the gene-probe mapping listed in the Illumina manifest file; `Probe`: chromosomal coordinates of the probe; `EID`: Experiment Identifier; `EIID`: Experiment Interaction Identifier.

SNP1	rsID	MAF1	rsID	MAF2	SNP2	P-value	Beta	minGCC	Gene	Probe	EIID	EID
8:18885955	rs10104492	0.28	rs12451549	0.10	17:46875212	1.72E-11	0.1731	3	CSHL1	17:59341967	hic_1_4837	hic_1
2:30967126	rs1862959	0.20	rs10500890	0.18	11:21075311	2.24E-13	0.18653	3	APOD	3:196777220	hic_1_1090	hic_1

Table B.4: Top confident **spatial epistastic SNP-probe pairs** identified in this study. That is, these SNP-pairs map to psychical interacting genomic loci (in hIMR90 cells). The SNP-probe pairs were selected using the criteria min GCC ≥ 3 (these two SNP-probe pairs appear on the right hand side of Table B.1). See Table B.3 for a description of the table headers.

SNP1	rsID	MAF1	SNP2	rsID	MAF2	P-value	Beta	minGCC	Gene	Probe	EIID	EID
4:104066247	rs3974608	0.46	3:163183258	rs4299495	0.24	2.57E-11	0.066	10	LOC727752	19:51651	null_86_7175	null_86
3:57295044	rs9840663	0.32	22:47915370	rs80567	0.35	1.13E-11	0.086	12	ATPIF1	1:28436788	null_150_7035	null_150
9:81975039	rs7045654	0.32	16:12633764	rs7194011	0.42	2.64E-11	-0.078	19	SLC46A3	13:28172532	null_198_4075	null_198
1:176775553	rs6700266	0.35	7:82563554	rs6947662	0.4	3.10E-11	0.067	15	PIGR	1:205168794	null_420_1555	null_420
5:60895903	rs10036399	0.29	15:48836816	rs12911143	0.37	1.87E-11	0.064	11	KIAA1147	7:141003222	null_435_7721	null_435
2:135143088	rs4954158	0.39	16:63565417	rs35226	0.31	7.17E-12	-0.050	13	ABCC11	16:46823387	null_928_6011	null_928

Table B.5: Top confident **epistastic SNP-probe pairs** identified in this study. The SNP-probe pairs were selected using the criteria min GCC ≥ 10. These SNP-probe pairs are the strongest evidence of statistical epistasis found in this thesis. None of these SNP-probe pairs were reported by (Hemani et al., 2014). Notice that only two of the SNP-pairs co-localize with the probe. All the SNP-pairs map to non-interacting genomic loci ("null samples", see Figure 6.2). See Table B.3 for a description of the table headers.

Source Code

C

C.1 GitHub Meta-analysis

Git and GitHub were used for version control of the code written for this thesis*. I used `GitStats` for this small "meta-analysis" of my code. Table C.1 shows that `Python` and `R` were the most predominant programming languages used in this thesis. Figure C.1 reveals that the code was not contributed equally throughout the project period. This may be because I was too undisciplined to commit my code as soon as it was done. Figure C.2 shows **(a)** my very unproductive weekends, and **(b)** my habits of coding late in the night.

Extension	**Files**	**Lines**	**Lines/file**
R	33 (44.00%)	4115 (7.01%)	124
ipynb	9 (12.00%)	44484 (75.74%)	4942
pl	3 (4.00%)	162 (0.28%)	54
py	29 (38.67%)	8135 (13.85%)	280
txt	1 (1.33%)	1835 (3.12%)	1835
total	75	58731	

Table C.1: GitHub repository meta-analysis of programming language used for this thesis. Only selected file extensions are shown (`.R, .ipynb, .pl, .py, .txt`).

* Unfortunately a substantial part of my code was not version controlled, and hence not included in this "analysis". This includes the `R` code used to experiment with epistatic models, replicate the work of Hemani et al. (2014), perform the PEER analysis and generate circos plots

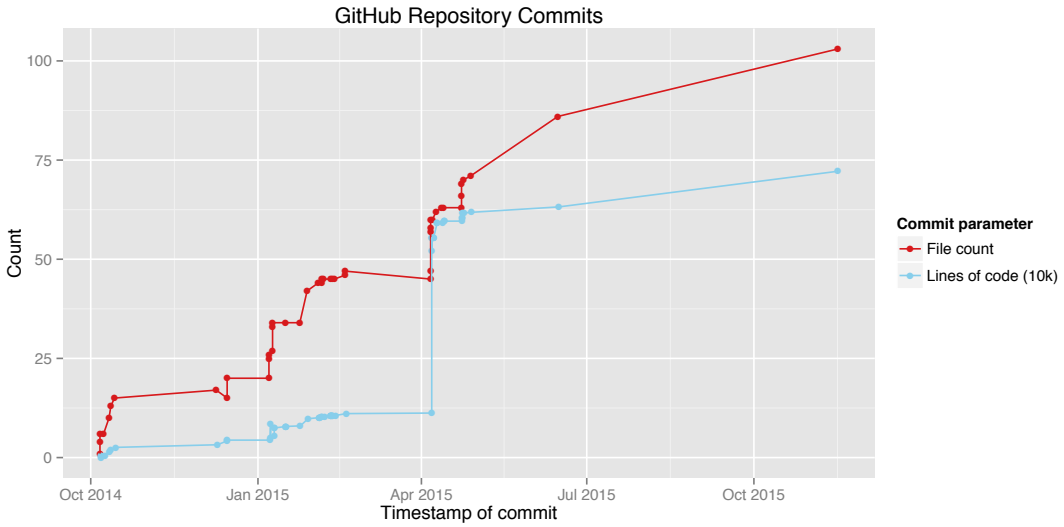


Figure C.1: File count and the number of lines of code (in values of 10,000) plotted against date.

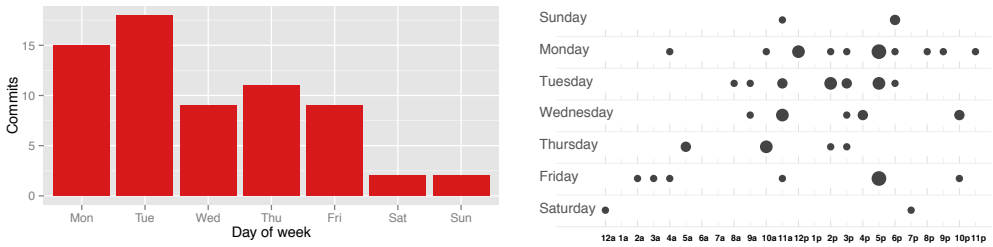


Figure C.2: Weekly distribution of commits.

C.2 Source Code

Listing C.1: Code for generating epistatic models shown in Figure 2.3 on page 10.

```
library(ggplot2)

# Generate Allele Codes
allele.codes <- rep(c(0,1,2), 3)
df.alleles <- unique(expand.grid(snp1=allele.codes, snp2=allele.codes))

# Set beta coefficients
beta.1 <- 0 # main effect SNP 1
beta.2 <- 0.3 # main effect SNP 2
beta.12 <- 0 # no epistasis

# Set response variable (phenotype)
y1 <- with(df.alleles, beta.1*snp1 + beta.2*snp2 + beta.12*snp1*snp2)
# Combine variables in data.frame, with SNPs as factors (allele codes)
df.epi <- with(df.alleles, data.frame(snp1=as.factor(snp1), snp2=as.factor(snp2), y=y1))
# Plot using ggplot
p <- ggplot(df.epi, aes(x=snp1, group=snp2, color=snp2, y=y)) + geom_point() + geom_line()
```

Listing C.2: . The code shows a workflow for testing for epistasis in R using the multiplicative allelic model. First the genotype data is loaded from is binary PLINK format via the `SnpStats` package. Next the two linear models are fitted to an example SNP-probe pair using the `lm()` function, and subsequently tested for model reduction using `anova()`.

```
library(snpStats) # used for PLINK data I/O functions

### Genotypes (from PLINK files)
files.plink <- "data/EGCUT_PLINK_GENOTYPES"
gwas.fn <- lapply(c(bed='bed',bim='bim',fam='fam'),
                  function(n) sprintf("%s.%s", files.plink, n))
geno <- read.plink(gwas.fn$bed, gwas.fn$bim, gwas.fn$fam) # SNPstats
genoBim <- geno$map # Obtain the SNP information from geno list
colnames(genoBim) <- c("chr", "SNP", "gen.dist", "position", "A1", "A2")
geno <- geno$genotypes # genotype matrix [individuals X SNPs] |
geno <- as(geno, "numeric") # Converting SnpMatrix to Numeric Matrix | # 0,1,2. Missing data will be NA.
dim(geno) # 832 individuals, ~2e6 SNPs

### Phenotypes (EGCUT expression matrix)
df.probes <- read.table("expression_matrix_samples_all.RData", h=T) # expression matrix [individuals X probes]
dim(df.probes)# 832 individuals, 9269 probes

#### Run Epistasis Tests ####
df.epistasis <- EpistasisTests(
  name.snp1="rs7989895",
  name.snp2="rs4846085",
  name.probe="ILMN_1651385",
  geno=geno,
  expression=df.probes)

#### Main Function ####
EpistasisTests <- function(name.snp1, name.snp2, name.probe, geno, expression) {
  # Extract data
  snp1 <- geno[, colnames(geno) == name.snp1]
  snp2 <- geno[, colnames(geno) == name.snp2]
  probe <- expression[, colnames(expression) == name.probe]

  if ( (length(snp1)<1) | (length(snp2)<1) | (length(probe)<1) ) {
    print("snp1, snp2 or probe not found")
    return(data.frame()) # returning empty data frame
  }

  SNP_LD <- cor(snp1, snp2, use="pairwise.complete.obs") # compute LD using Pearson Correlation
  n_usable_data_points <- sum(!is.na(snp1) & !is.na(snp2)) # total number of non-NA data points
  n_missing_data_points <- sum(is.na(snp1)) + sum(is.na(snp2)) # total number of missing data points
```

```
41
42    ### MAF calculation
43    # MAF Formula: number_of_SNP[x]_alleles/total_number_of_alleles
44    tmp.snp1.maf <- (sum(snp1==1,na.rm=T)*1 + sum(snp1==2, na.rm=T)*2)/(2*sum(!is.na(snp1)))
45    tmp.snp2.maf <- (sum(snp2==1,na.rm=T)*1 + sum(snp2==2, na.rm=T)*2)/(2*sum(!is.na(snp2)))
46    snp1.maf <- ifelse(tmp.snp1.maf <= 0.5, tmp.snp1.maf, 1-tmp.snp1.maf)
47    snp2.maf <- ifelse(tmp.snp2.maf <= 0.5, tmp.snp2.maf, 1-tmp.snp2.maf)
48
49    ### Factorize numeric encoded alleles
50    snp1.allele <- factor(snp1,levels=c("0", "1", "2", NA), exclude=NULL)
51    snp2.allele <- factor(snp2,levels=c("0", "1", "2", NA), exclude=NULL)
52
53    ### Construct two-loci genotype table for calculation
54    # of minimum Genotype Class Count (mGCC).
55    tab.allele <- table(data.frame(snp1.allele, snp2.allele)) # crosstabulation
56    tab.allele
57    df.twolocus <- as.data.frame(tab.allele)
58    df.twolocus.narm <- na.omit(df.twolocus) # this is the 3x3 matrix/table
59    min_genotype_class_count <- min(df.twolocus.narm$Freq) # minimum value in the contingency table
60
61
62    ##### Statistical tests - Multiplicative Allelic model #####
63    ## Full model
64    fullmod.multiplicative <- lm(probe ~ snp1 + snp2 + snp1:snp2)
65    fullmod.multiplicative
66    summary(fullmod.multiplicative)
67    beta_int_multiplicative <- fullmod.multiplicative$coefficients[4] # coefficient for interaction
          term (snp1:snp2)
68    ## Marginal model
69    margmod.multiplicative <- lm(probe ~ snp1 + snp2)
70    ## Test for model reduction | Maximum likelihood framework
71    # Note that the traditional approach of using the anova(lm(...)) call results in a "sequential
          test" for dropping model terms. This code uses a more reliable method by testing a specific
          hypotheses, by using the call "anova(fit.H0, fit.HA)". Alternatively, the call "anova(lm(...
          ), ssType=3)" or "drop1(aov())" uses a Type III test (partial), which is also appropriate
72    inttest_multiplicative <- anova(margmod.multiplicative, fullmod.multiplicative) # anova(fit.H0,
          fit.HA)
73    # Because we are assuming independent Gaussian noise (general linear model), the likelihood
          ratio test statistic we know the analytical distribution of the test: F-test because both
          the nominator and denominator are chi-sq distributed. Hence we do not have to rely on
          assymptotic results from the Wilk's Likelihood Ratio test theorem.
74
75    ### Extracting Model Estimates
76    F_statistic <- inttest_multiplicative$F[2] # F-test value
77      # Large values of F reflects a conflict between the data and H0, and hence lead to rejection
            of H0
78    F_pval <- inttest_multiplicative$P[2] # or more convenient: -log10(inttest_multiplicative$P[2])
79    # The P-value is calculated as F(F_statistic, df_denominator, df_numerator, lower.tail=F)
80
81    ### Degrees of freedom
82    F_df_denominator = margmod.multiplicative$df.residual - fullmod.multiplicative$df.residual
83      # Formula to calculate F_df_denominator: m_full - m_red = N_param_full - N_param_red = (3+1)
            -(2+1)=1
84      # The F_df_denominator will remain the same for all SNP-probe pairs
85    F_df_numerator = fullmod.multiplicative$df.residual
86      # Formula to calculate F_df_numerator: n - m_full = N_observations - N_param_full = 6-(3+1)=2
87      # The F_df_numerator is dependent on N_observations. Hence it is sensitive to missing data (
            unobserved combinations of alleles)
88
89    ### Return data frame results ###
90    df.res.tmp <- data.frame(
91      snp1=name.snp1, snp2=name.snp2, probename=name.probe,
92      beta=beta_int_multiplicative,
93      F_pval,
94      F_statistic,
95      F_df_denominator,
96      F_df_numerator,
97      SNP_LD=SNP_LD,
98      snp1_maf=snp1.maf,
99      snp2_maf=snp2.maf,
100     n_usable_data_points,
101     n_missing_data_points
102   )
103   return(df.res.tmp)
104 }
```

Glossary D

The terms listed in this glossary serve to provide precise definitions of terms most readers will not be familiar with. They are, for the most part, made up by the author of this thesis and are not yet* part of the genetics jargon.

Genetics

- genotype class count (GCC): the number of observations for a particular combination of genotypes. This is the same as the two-locus genotype counts. A two-locus biallelic genotype model consists of nine ($3 \cdot 3 = 9$) different genotype classes.

Hi-C data

- interaction fragment: the genomic region making one of the interacting partners. This is the lowest level building parts of the Hi-C data set. The interaction fragment size determines the resolution of the Hi-C data. The fragment size can be either based on restriction fragments or fixed-sized (e.g. 40 kb) by aggregating restriction fragments into bins.
- interaction table: the complete set of interactions considered at a given FDR threshold. Each entry consists of of two interaction fragments, constituting a Hi-C interaction pair.
- Hi-C interaction pair: two interaction fragments interacting. That is, the fragments are in close spatial proximity.
- interaction count: the number of interactions an interaction fragment takes part of.
- contact count: the number of sequence reads for the interaction. The number is derived directly after mapping, filtering, pairing and duplicate removal of the paired-end Hi-C reads.
- normalized contact count: contact count normalized for the Hi-C bias at each locus (e.g. bias derived from the ICE method (Imakaev et al., 2012)). Normalized counts can be obtained by using the formula: $\text{contact\_count}_{norm} = \frac{\text{contact\_count}_{raw}}{b_1 \cdot b_2}$, where b_1 and b_2 is the bias for first and second locus, respectively.

Spatial Epistasis Enrichment

* Hopefully, the international scientific community will one day realize the potential of this work and adopt the many wonderful terms for spatial epistasis.

- spatial proximate epistasis (or spatial epistasis): statistical epistasis between genetic variants in close spatial proximity. Spatial proximity is defined as highly interacting genomic regions. That is, regions with high contact count measured by chromosome conformation capture methods (e.g. Hi-C data).
- interaction width (l): the length in bp of the genomic region that is considered for epistasis discovery. Thus the interaction width, l, defines the search space for "spatial genetic interactions". The region is bounded by the symmetric distance $l/2$ bp upstream and downstream from the midpoint of the interaction fragment.
- Experiment Identifier (EID): a unique identifier for each sample from the null distribution. That is, samples from non-interacting genomic regions.
  Format is [`experiment_type`]_[`experiment_number`], e.g. `null_1`, `null_2`, ..., `null_1000`.
- Experiment Interaction Identifier (EIID): a unique identifier for each interaction for each sample from the null distribution.
  Format is [`experiment_type`]_[`experiment_number`]_[`interaction_number`], e.g. null_1_1923 for interaction number 1923 from null sample 1.

References

Literature

Alberts, Rudi et al. (2007). "Sequence polymorphisms cause many false cis eQTLs". In: *PLoS ONE* 2.7, pp. 1–5. ISSN: 19326203. DOI: 10.1371/journal.pone.0000622.

Albrechtsen, Anders et al. (2007). "A bayesian multilocus association method: Allowing for higher-order interaction in association studies". In: *Genetics* 176.2, pp. 1197–1208. ISSN: 00166731. DOI: 10.1534/genetics.107.071696.

Alon, Uri (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC Mathematical and Computational Biology. CRC Press, p. 320. ISBN: 9781420011432.

Anderson, Carl a et al. (2010). "Data quality control in genetic case-control association studies." In: *Nature protocols* 5.9, pp. 1564–73. ISSN: 1750-2799. DOI: 10.1038/nprot.2010.116.

Aulchenko, Yurii S. et al. (2007). "GenABEL: An R library for genome-wide association analysis". In: *Bioinformatics* 23.10, pp. 1294–1296. ISSN: 13674803. DOI: 10.1093/bioinformatics/btm108.

Ay, Ferhat, Timothy L Bailey, and William Stafford Noble (2014b). "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts." In: *Genome research* 24.6, pp. 999–1011. ISSN: 1549-5469. DOI: 10.1101/gr.160374.113.

Ay, Ferhat and William S. Noble (2015). "Analysis methods for studying the 3D architecture of the genome". In: *Genome Biology* 16.1, p. 183. ISSN: 1474-760X.

Aylor, DL and ZB Zeng (2008). "From classical genetics to quantitative genetics to systems biology: modeling epistasis". In: *PLoS genetics* 4.3. DOI: 10.1371/journal.pgen.1000029.

Bateson, William (1909). *Mendel's Principles of Heredity*. Cambridge: Cambridge University Press.

Beam, Andrew L, Alison Motsinger-Reif, and Jon Doyle (2014). "Bayesian neural networks for detecting epistasis in genetic association studies." In: *BMC bioinformatics* 15.1, p. 368. ISSN: 1471-2105. DOI: 10.1186/s12859-014-0368-0.

Becker, Jessica et al. (2012). "A systematic eQTL study of cis-trans epistasis in 210 HapMap individuals." In: *European journal of human genetics : EJHG* 20.1, pp. 97–101. ISSN: 1476-5438. DOI: 10.1038/ejhg.2011.156.

Belton, Jon Matthew et al. (2012). "Hi-C: A comprehensive technique to capture the conformation of genomes". In: *Methods* 58.3, pp. 268–276. ISSN: 10462023. DOI: 10.1016/j.ymeth.2012.05.001.

Berkum, Nynke L. van et al. (2010). "Hi-C: A Method to Study the Three-dimensional Architecture of Genomes." In: *Journal of Visualized Experiments* 39, pp. 1–7. ISSN: 1940-087X. DOI: 10.3791/1869.

Bianconi, Eva et al. (2013). "An estimation of the number of cells in the human body". en. In: *Annals of Human Biology*.

Bishop, Christopher (2007). *Pattern Recognition and Machine Learning*. 1st ed. Springer. ISBN: 0387310738.

Bonora, Giancarlo, Kathrin Plath, and Matthew Denholtz (2014). "A mechanistic link between gene regulation and genome architecture in mammalian development". In: *Current Opinion in Genetics & Development* 27, pp. 92–101. ISSN: 0959437X. DOI: 10.1016/j.gde.2014.05.002.

Brem, Rachel B. et al. (2005). "Genetic interactions between polymorphisms that affect gene expression in yeast". In: *Nature* 436.7051, pp. 701–703. ISSN: 0028-0836. DOI: 10.1038/nature03865.

Carlborg, Orjan and Chris S Haley (2004). "Epistasis: too often neglected in complex trait studies?" In: *Nature reviews. Genetics* 5.8, pp. 618–25. ISSN: 1471-0056. DOI: 10.1038/nrg1407.

Chang, Christopher C et al. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* 4, pp. 1–16. ISSN: 2047-217X. DOI: 10.1186/s13742-015-0047-8.

Ciabrelli, Filippo and Giacomo Cavalli (2015). "Chromatin-Driven Behavior of Topologically Associating Domains". In: *Journal of Molecular Biology* 427.3, pp. 608–625. ISSN: 00222836. DOI: 10.1016/j.jmb.2014.09.013.

Collins, Francis (2010). "Has the revolution arrived?" In: *Nature* 464.7289, pp. 674–675. ISSN: 0028-0836. DOI: 10.1038/464674a.

Cordell, Heather J (2002). "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans." In: *Human molecular genetics* 11.20, pp. 2463–8. ISSN: 0964-6906.

– (2009). "Detecting gene-gene interactions that underlie human diseases." In: *Nature reviews. Genetics* 10.6, pp. 392–404. ISSN: 1471-0064. DOI: 10.1038/nrg2579.

Costanzo, Michael et al. (2010). "The genetic landscape of a cell." In: *Science (New York, N.Y.)* 327.5964, pp. 425–31. ISSN: 1095-9203. DOI: 10.1126/science.1180823.

Cremer, Thomas and Marion Cremer (2010). "Chromosome territories". In: *Cold Spring Harbor perspectives in biology* 2.3, a003889. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a003889.

Cremer, Thomas et al. (2006). "Chromosome territories - a functional nuclear landscape". In: *Current Opinion in Cell Biology* 18.3, pp. 307–316. ISSN: 09550674. DOI: 10.1016/j.ceb.2006.04.007.

Dekker, Job (2008). "Gene regulation in the third dimension." In: *Science (New York, N.Y.)* 319.5871, pp. 1793–4. ISSN: 1095-9203. DOI: 10.1126/science.1152850.

Dekker, Job, Marc a Marti-Renom, and Leonid a Mirny (2013). "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data." In: *Nature reviews. Genetics* 14.6, pp. 390–403. ISSN: 1471-0064. DOI: 10.1038/nrg3454.

Dekker, Job et al. (2002). "Capturing chromosome conformation." In: *Science* 295.5558, pp. 1306–1311. ISSN: 00368075. DOI: 10.1126/science.1067799.

Deng, Wulan et al. (2012). "Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor". In: *Cell* 149.6, pp. 1233–1244. ISSN: 00928674. DOI: 10.1016/j.cell.2012.03.051.

Dixon, Jesse R et al. (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." In: *Nature* 485.7398, pp. 376–80. ISSN: 1476-4687. DOI: 10.1038/nature11082.

Dixon, Jesse R. et al. (2015). "Chromatin architecture reorganization during stem cell differentiation". In: *Nature* 518.7539, pp. 331–336. ISSN: 0028-0836. DOI: 10.1038/nature14222.

Dostie, Josée et al. (2006). "Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements". In: *Genome Research* 16.10, pp. 1299–1309. ISSN: 10889051. DOI: 10.1101/gr.5571506.

Drexler, H G, R A MacLeod, and C C Uphoff (1999). "Leukemia cell lines: in vitro models for the study of Philadelphia chromosome-positive leukemia." In: *Leukemia Research* 23.3, pp. 207–215. ISSN: 01452126. DOI: 10.1016/S0145-2126(98)00171-4.

Duggal, Geet, Hao Wang, and Carl Kingsford (2014). "Higher-order chromatin domains link eQTLs with the expression of far-away genes". In: *Nucleic Acids Research* 42.1, pp. 87–96. ISSN: 03051048. DOI: 10.1093/nar/gkt857.

Eichler, Evan E et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." In: *Nature reviews. Genetics* 11.6, pp. 446–50. ISSN: 1471-0064. DOI: 10.1038/nrg2809.

Emily, Mathieu et al. (2009). "Using biological networks to search for interacting loci in genome-wide association studies." In: *European journal of human genetics : EJHG* 17.10, pp. 1231–40. ISSN: 1476-5438. DOI: 10.1038/ejhg.2009.15.

Engreitz, Jesse M., Vineeta Agarwala, and Leonid a. Mirny (2012). "Three-Dimensional Genome Architecture Influences Partner Selection for Chromosomal Translocations in Human Disease". In: *PLoS ONE* 7.9, pp. 1–9. ISSN: 19326203. DOI: 10.1371/journal.pone.0044196.

Eskiw, Christopher H et al. (2008). "RNA polymerase II activity is located on the surface of protein-rich transcription factories." In: *Journal of cell science* 121.Pt 12, pp. 1999–2007. ISSN: 0021-9533. DOI: 10.1242/jcs.027250.

Fisher, R. A. (1919). "The Correlation between Relatives on the Supposition of Mendelian Inheritance." English. In: *Transactions of the Royal Society of Edinburgh* 52.02, pp. 399–433. ISSN: 0080-4568. DOI: 10.1017/S0080456800012163.

Franke, Andre et al. (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci." In: *Nature genetics* 42.12, pp. 1118–25. ISSN: 1546-1718. DOI: 10.1038/ng.717.

Fraser, James et al. (2015). "An Overview of Genome Organization and How We Got There: from FISH to Hi-C". In: *Microbiology and Molecular Biology Reviews* 79.3, pp. 347–372. ISSN: 1092-2172. DOI: 10.1128/MMBR.00006-15.

Frazer, Kelly A et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." In: *Nature* 449.7164, pp. 851–861. ISSN: 0028-0836. DOI: 10.1038/nature06258.

Gibcus, Johan H. and Job Dekker (2013). "The Hierarchy of the 3D Genome". In: *Molecular Cell* 49.5, pp. 773–782. ISSN: 10972765. DOI: 10.1016/j.molcel.2013.02.011.

Gibson, Greg (2012). "Rare and common variants: twenty arguments". In: *Nature Reviews Genetics* 13.2, pp. 135–145. ISSN: 1471-0056. DOI: 10.1038/nrg3118.

Göndör, Anita and Rolf Ohlsson (2009). "Chromosome crosstalk in three dimensions." In: *Nature* 461.7261, pp. 212–7. ISSN: 1476-4687. DOI: 10.1038/nature08453.

Greene, Casey S et al. (2010). "Enabling personal genomics with an explicit test of epistasis." In: *Pacific Symposium on Biocomputing*. Pp. 327–36. ISSN: 2335-6936.

Guo, Ya et al. (2015). "Article CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer / Promoter Function Article CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer / Promoter Function". In: *Cell* 162.4, pp. 900–910. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.07.038.

Gusev, Alexander et al. (2014). "Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases". In: *The American Journal of Human Genetics* 95.5, pp. 535–552. ISSN: 00029297. DOI: 10.1016/j.ajhg.2014.10.004.

Hebenstreit, Daniel (2013). "Are gene loops the cause of transcriptional noise?" In: *Trends in Genetics* 29.6, pp. 333–338. ISSN: 01689525. DOI: 10.1016/j.tig.2013.04.001.

Hemani, Gibran et al. (2011). "EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards." In: *Bioinformatics (Oxford, England)* 27.11, pp. 1462–5. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr172.

Hemani, Gibran et al. (2014). "Detection and replication of epistasis influencing transcription in humans." In: *Nature* 508.7495, pp. 249–53. ISSN: 1476-4687.

Hill, William G, Michael E Goddard, and Peter M Visscher (2008). "Data and theory point to mainly additive genetic variance for complex traits." In: *PLoS genetics* 4.2, e1000008. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000008.

Howie, Bryan N, Peter Donnelly, and Jonathan Marchini (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies". In: *PLoS Genetics* 5.6. ISSN: 15537390. DOI: 10.1371/journal.pgen.1000529.

Imakaev, Maxim et al. (2012). "Iterative correction of Hi-C data reveals hallmarks of chromosome organization." In: *Nature methods* 9.10, pp. 999–1003. ISSN: 1548-7105. DOI: 10.1038/nmeth.2148.

Jordan, Michael I (1999). "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37, pp. 183–233.

Kagey, Michael H et al. (2010). "Mediator and cohesin connect gene expression and chromatin architecture." In: *Nature* 467.7314, pp. 430–435. ISSN: 1476-4687. DOI: 10.1038/nature09380.

Kahneman, Daniel (2011). *Thinking, fast and slow*. Macmillan.

Knijnenburg, Theo A et al. (2009). "Fewer permutations, more accurate P-values." In: *Bioinformatics (Oxford, England)* 25.12, pp. i161–8. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp211.

Lajoie, Bryan R, Job Dekker, and Noam Kaplan (2014). "The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines". In: *Methods* 72, pp. 65–75. ISSN: 10462023. DOI: 10.1016/j.ymeth.2014.10.031.

Lan, Xun et al. (2012). "Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages." In: *Nucleic acids research* 40.16, pp. 7690–704. ISSN: 1362-4962. DOI: 10.1093/nar/gks501.

Langmead, B et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome biology* 10.3, R25. ISSN: 1465-6914; 1465-6906. DOI: gb-2009-10-3-r25[pii]$\backslash$n10.1186/gb-2009-10-3-r25.

Lazebnik, Y (2002). "Can a biologist fix a radio?—Or, what I learned while studying apoptosis". In: *Cancer Cell* 2.September, pp. 179–182.

Leek, Jeffrey T and John D Storey (2007). "Capturing heterogeneity in gene expression studies by surrogate variable analysis." In: *PLoS genetics* 3.9, pp. 1724–35. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.0030161.

Lehner, Ben (2011). "Molecular mechanisms of epistasis within and between genes". In: *Trends in Genetics* 27.8, pp. 323–331. ISSN: 01689525. DOI: 10.1016/j.tig.2011.05.007.

Li, Wentian and Jens Reich (2000). "A complete enumeration and classification of two-locus disease models." In: *Human heredity* 50.6, pp. 334–49. ISSN: 0001-5652. DOI: 22939.

Li, Wenyuan et al. (2015). "Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data." In: *Bioinformatics (Oxford, England)* 31.November, pp. 1–3. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu747.

Libbrecht, Maxwell W et al. (2015). "Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell type-specific expression." In: *Genome research*. ISSN: 1549-5469. DOI: 10.1101/gr.184341.114.

Lieberman-Aiden, Erez et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." In: *Science* 326.5950, pp. 289–93. ISSN: 1095-9203. DOI: 10.1126/science.1181369.

Locke, Adam E. et al. (2015). "Genetic studies of body mass index yield new insights for obesity biology". In: *Nature* 518, pp. 197–206. ISSN: 0028-0836. DOI: 10.1038/nature14177.

Lozzio, C B and B B Lozzio (1975). "Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome." In: *Blood* 45.3, pp. 321–34. ISSN: 0006-4971.

Mackay, Trudy F C (2014). "Epistasis and quantitative traits: using model organisms to study gene-gene interactions." In: *Nature reviews. Genetics* 15.1, pp. 22–33. ISSN: 1471-0064. DOI: 10.1038/nrg3627.

Mackay, Trudy Fc and Jason H Moore (2014). "Why epistasis is important for tackling complex human disease genetics." En. In: *Genome medicine* 6.6, p. 124. ISSN: 1756-994X. DOI: 10.1186/gm561.

Macosko, Evan Z. et al. (2015). "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5, pp. 1202–1214. ISSN: 00928674. DOI: 10.1016/j.cell.2015.05.002.

Maher, Brendan (2008). "Personal genomes: The case of the missing heritability." en. In: *Nature* 456.7218, pp. 18–21. ISSN: 1476-4687. DOI: 10.1038/456018a.

Manolio, Teri a et al. (2009). "Finding the missing heritability of complex diseases." In: *Nature* 461.7265, pp. 747–53. ISSN: 1476-4687. DOI: 10.1038/nature08494.

Marchini, Jonathan, Peter Donnelly, and Lon R Cardon (2005). "Genome-wide strategies for detecting multiple loci that influence complex diseases." In: *Nature genetics* 37.4, pp. 413–7. ISSN: 1061-4036. DOI: 10.1038/ng1537.

Metspalu, Andres (2004). "The Estonian Genome Project". In: *Drug Development Research* 62.2, pp. 97–101. ISSN: 0272-4391. DOI: 10.1002/ddr.10371.

Mifsud, Borbala et al. (2015). "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C". In: *Nature Genetics*, pp. 1–12. ISSN: 1061-4036. DOI: 10.1038/ng.3286.

Miko, Ilona (2008). "Epistasis: gene interaction and phenotype effects". In: *Nature Education* 1.1, p. 197.

Moore, Jason H and Doug P Hill (2015). "Epistasis Analysis Using Artificial Intelligence". In: 1253. DOI: 10.1007/978-1-4939-2155-3.

Moore, Jason H and Scott M Williams (2005). "Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis". In: *BioEssays* 27.6, pp. 637–646. ISSN: 02659247. DOI: 10.1002/bies.20236.

– (2009). "Epistasis and its implications for personal genetics." In: *American journal of human genetics* 85.3, pp. 309–20. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2009.08.006.

Moore, Jason H. and Scott M. Williams, eds. (2015). *Epistasis*. Vol. 1253. Methods in Molecular Biology. New York, NY: Springer New York. ISBN: 978-1-4939-2154-6. DOI: 10.1007/978-1-4939-2155-3.

Morgenthaler, Stephan and William G. Thilly (2007). "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST)". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 615.1-2, pp. 28–56. ISSN: 00275107. DOI: 10.1016/j.mrfmmm.2006.09.003.

Nagano, Takashi et al. (2013). "Single-cell Hi-C reveals cell-to-cell variability in chromosome structure." In: *Nature* 502, pp. 59–64. ISSN: 1476-4687. DOI: 10.1038/nature12593.

Nelis, Mari et al. (2009). "Genetic structure of Europeans: a view from the North-East." In: *PloS one* 4.5, e5472. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0005472.

Noble, William S (2009). "How does multiple testing correction work?" In: *Nature biotechnology* 27.12, pp. 1135–7. ISSN: 1546-1696. DOI: 10.1038/nbt1209-1135.

Nora, Elphège P. et al. (2012). "Spatial partitioning of the regulatory landscape of the X-inactivation centre". In: *Nature* 485.7398, pp. 381–385. ISSN: 0028-0836. DOI: 10.1038/nature11049.

North, B V, D Curtis, and P C Sham (2002). "A note on the calculation of empirical P values from Monte Carlo procedures." In: *American journal of human genetics* 71.2, pp. 439–41. ISSN: 0002-9297. DOI: 10.1086/341527.

Ortlund, Eric a et al. (2007). "Crystal structure of an ancient protein: evolution by conformational epistasis." In: *Science (New York, N.Y.)* 317.5844, pp. 1544–8. ISSN: 1095-9203. DOI: 10.1126/science.1142819.

Pattin, Kristine A and Jason H Moore (2008). "Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases." In: *Human genetics* 124.1, pp. 19–29. ISSN: 1432-1203. DOI: 10.1007/s00439-008-0522-8.

Phillips, Jennifer E. and Victor G. Corces (2009). "CTCF: Master Weaver of the Genome". In: *Cell* 137.7, pp. 1194–1211. ISSN: 00928674. DOI: 10.1016/j.cell.2009.06.001. arXiv: NIHMS150003.

Phillips, Patrick C (2008). "Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems." In: *Nature reviews. Genetics* 9.11, pp. 855–67. ISSN: 1471-0064. DOI: 10.1038/nrg2452.

Phillips-Cremins, Jennifer E. et al. (2013). "Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment". In: *Cell* 153.6, pp. 1281–1295. ISSN: 00928674. DOI: 10.1016/j.cell.2013.04.053.

Powell, Joseph E et al. (2013). "Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data." In: *PLoS genetics* 9.5, e1003502. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003502.

Purcell, Shaun et al. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *The American Journal of Human Genetics* 81.3, pp. 559–575. ISSN: 00029297. DOI: 10.1086/519795.

Rao, Suhas S.P. et al. (2014). "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping". In: *Cell* 159.7, pp. 1665–1680. ISSN: 00928674. DOI: 10.1016/j.cell.2014.11.021.

Rieder, Dietmar, Zlatko Trajanoski, and James G. McNally (2012). "Transcription factories". In: *Frontiers in Genetics* 3.OCT, pp. 1–12. ISSN: 16648021. DOI: 10.3389/fgene.2012.00221.

Risca, Viviana I. and William J. Greenleaf (2015). "Unraveling the 3D genome: genomics tools for multiscale exploration". In: *Trends in Genetics* 31.7, pp. 357–372. ISSN: 01689525. DOI: 10.1016/j.tig.2015.03.010.

Ritchie, Marylyn D (2011). "Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies." In: *Annals of human genetics* 75.1, pp. 172–82. ISSN: 1469-1809. DOI: 10.1111/j.1469-1809.2010.00630.x.

Rosenbloom, Kate R et al. (2013). "ENCODE data in the UCSC Genome Browser: year 5 update." In: *Nucleic acids research* 41.Database issue, pp. D56–63. ISSN: 1362-4962. DOI: 10.1093/nar/gks1172.

Schoenfelder, Stefan et al. (2010). "Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells". In: *Nature Genetics* 42.1, pp. 53–61. ISSN: 1061-4036. DOI: 10.1038/ng.496.

Schüpbach, Thierry et al. (2010). "FastEpistasis: a high performance computing solution for quantitative trait epistasis." In: *Bioinformatics (Oxford, England)* 26.11, pp. 1468–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq147.

Selvaraj, Siddarth et al. (2013). "Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing". In: *Nature Biotechnology* 31.12, pp. 1111–1118. ISSN: 1087-0156. DOI: 10.1038/nbt.2728.

Seo, DI, YH Kim, and BR Moon (2003). "New entropy-based measures of gene significance and epistasis". In: *Genetic and Evolutionary Computation— ...* Pp. 1345–1356.

Sexton, Tom and Giacomo Cavalli (2015). "The Role of Chromosome Domains in Shaping the Functional Genome". In: *Cell* 160.6, pp. 1049–1059. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.02.040.

Sexton, Tom and Eitan Yaffe (2015). "Chromosome folding: driver or passenger of epigenetic state?" In: *Cold Spring Harbor perspectives in biology* 7.2, a018721. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a018721.

Simonis, Marieke et al. (2006). "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)." In: *Nature genetics* 38.11, pp. 1348–1354. ISSN: 1061-4036. DOI: 10.1038/ng1896.

Sinkhorn, Richard and Paul Knopp (1967). "Concerning nonnegative matrices and doubly stochastic matrices". In: *Pacific Journal of Mathematics* 21.2, pp. 343–348.

Stegle, Oliver et al. (2010). "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies." In: *PLoS computational biology* 6.5, e1000770. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000770.

Stegle, Oliver et al. (2012). "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses." In: *Nature protocols* 7.3, pp. 500–7. ISSN: 1750-2799. DOI: 10.1038/nprot.2011.457.

Steidl, Ulrich et al. (2007). "A distal single nucleotide polymorphism alters long-range regulation of the PU . 1 gene in acute myeloid leukemia". In: *Journal of Clinical Investigation* 117.9, pp. 2611–2620. ISSN: 00219738. DOI: 10.1172/JCI30525.suppressor.

Storey, John D (2002). "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3, pp. 479–498.

Sugar, Robert (2014). "Genome Analysis in Three Dimensions: Functional Analysis of Hi-C Derived Datasets". PhD thesis. Churchill College. University of Cambridge. URL: https://www.ebi.ac.uk/sites/ebi.ac.uk/files/shared/documents/phdtheses/RobertThesis{_}2014-12-12{_}v03{_}CORRECTED.pdf.

Tong, Amy Hin Yan et al. (2004). "Global mapping of the yeast genetic interaction network." In: *Science (New York, N.Y.)* 303.5659, pp. 808–13. ISSN: 1095-9203. DOI: 10.1126/science.1091317.

Turner, Stephen et al. (2011). "Quality control procedures for genome-wide association studies". In: *Curr Protoc Hum Genet* Chapter 1, Unit1 19. ISSN: 1934-8258. DOI: 10.1002/0471142905.hg0119s68.

Upton, Alex et al. (2015). "Review: High-performance computing to detect epistasis in genome scale data sets". In: *Briefings in Bioinformatics* March, bbv058. ISSN: 1467-5463. DOI: 10.1093/bib/bbv058.

Visscher, Peter M, William G Hill, and Naomi R Wray (2008). "Heritability in the genomics era - concepts and misconceptions". In: *Nature Reviews Genetics* 9.april, pp. 255–266. DOI: 10.1038/nrg2322.

Walters, Raymond K, Charles Laurin, and Gitta H Lubke (2014). "Epi2Loc: An R Package to Investigate Two-Locus Epistatic Models." In: *Twin research and human genetics : the official journal of the International Society for Twin Studies* 17.4, pp. 272–8. ISSN: 1832-4274. DOI: 10.1017/thg.2014.38.

Watson, James D and Francis H C Crick (1953). "Molecular structure of nucleic acids". In: *Nature* 171.4356, pp. 737–738.

Wei, Wen-Hua, Gibran Hemani, and Chris S. Haley (2014). "Detecting epistasis in human complex traits". In: *Nature Reviews Genetics* September. ISSN: 1471-0056. DOI: 10.1038/nrg3747.

Weigelt, Britta and Jorge S Reis-Filho (2014). "Epistatic interactions and drug response." In: *The Journal of pathology* 232.2, pp. 255–63. ISSN: 1096-9896. DOI: 10.1002/path.4265.

Westra, Harm-Jan et al. (2013). "Systematic identification of trans eQTLs as putative drivers of known disease associations". In: *Nature genetics* 45.10, pp. 1238–43. ISSN: 1546-1718. DOI: 10.1038/ng.2756.

Wood, Andrew R. et al. (2014). "Another explanation for apparent epistasis". In: *Nature* 514.7520, E3–E5. ISSN: 0028-0836. DOI: 10.1038/nature13691.

Yaffe, Eitan and Amos Tanay (2011). "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture." In: *Nature genetics* 43.11, pp. 1059–65. ISSN: 1546-1718. DOI: 10.1038/ng.947.

Yang, Can et al. (2011). "The choice of null distributions for detecting gene-gene interactions in genome-wide association studies." In: *BMC bioinformatics* 12 Suppl 1.Suppl 1, S26. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-S1-S26.

Yin, Tengfei, Dianne Cook, and Michael Lawrence (2012). "ggbio: an R package for extending the grammar of graphics for genomic data." In: *Genome biology* 13.8, R77. ISSN: 1465-6914. DOI: 10.1186/gb-2012-13-8-r77.

Young, N S and S P Hwang-Chen (1981). "Anti-K562 cell monoclonal antibodies recognize hematopoietic progenitors." In: *Proceedings of the National Academy of Sciences of the United States of America* 78.11, pp. 7073–7. ISSN: 0027-8424.

Zhang, Yu and Jun S Liu (2007). "Bayesian inference of epistatic interactions in case-control studies." In: *Nature genetics* 39.9, pp. 1167–73. ISSN: 1061-4036. DOI: 10.1038/ng2110.

Zhao, Zhihu et al. (2006). "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions". In: *Nature Genetics* 38.11, pp. 1341–1347. ISSN: 1061-4036. DOI: 10.1038/ng1891.

Zirkel, Anne and Argyris Papantonis (2014). "Transcription as a force partitioning the eukaryotic genome." In: *Biological chemistry* 395.11, pp. 1301–5. ISSN: 1437-4315. DOI: 10.1515/hsz-2014-0196.

Zuk, Or et al. (2012). "The mystery of missing heritability: Genetic interactions create phantom heritability." In: *Proceedings of the National Academy of Sciences of the United States of America* 109.4, pp. 1193–8. ISSN: 1091-6490. DOI: 10.1073/pnas.1119675109.

Software

Ay, Ferhat, Timothy Bailey, and William Noble (2014a). *Fit-Hi-C*. URL: https://noble.gs.washington.edu/proj/fit-hi-c/.

Deelen, Patrick, Harm-Jan Westra, and Lude Franke (2014). *eQTL Mapping Pipeline*. URL: https://github.com/molgenis/systemsgenetics/wiki/eQTL-mapping-analysis-cookbook.

Freeman, Colin and Jonathan Marchini (2007). *GTOOL*. URL: http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html.

Min Kang, Hyun and Goncalo Abecasis (2014). *LiftOver*. URL: http://genome.sph.umich.edu/wiki/LiftOver.

Purcell, Shaun and Christopher Chang (2014). *PLINK 1.9*. URL: https://www.cog-genomics.org/plink2.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.r-project.org.

Schuepbach, Thierry (2014). *FastEpistasis*. Version 2.07. URL: http://www.vital-it.ch/software/FastEpistasis/.

Online Resources

Illumina Inc (2014). *Illumina manifest file*. Accessed: 2015-02-01. URL: http://support.illumina.com/downloads/humanht-12_v3_product_files.html (visited on 02/01/2015).

UCSC Genome Browser (2015). *Mappability or Uniqueness of Reference Genome from ENCODE*. Accessed: 2015-03-30. URL: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19\&g=wgEncodeMapability (visited on 03/30/2015).