The Genetics of Cellular Phenotypes

Zhihao Ding

Wellcome Trust Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Darwin College August 2014

To my grandmother, my parents and my wife

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This thesis does not exceed the length limit of 60,000 words specified by the Biology Degree Committee.

Zhihao Ding
August 2014

Acknowledgements

It has been a lengthy journey for me to reach my PhD. My interest and motivation in science is mostly influenced by my grandmother and my parents to whom I owe the most gratitude. I thank my friends and lecturers in Wuhan University. The undergraduate program in the College of Life Science led me into the door of biology. I thank Doctor Andrew Coulson and Professor Mark Blaxter in the University of Edinburgh, who gave me strong support and guidance during my master study, which eventually led me to the bioinformatics field that I thoroughly enjoy. Before my PhD, I worked in the lab of Professor Carlos Caldas, who offered me great opportunities to participate in cancer projects and supported me for my application for a PhD.

During my PhD, I owe most thanks to my supervisor Richard Durbin. His patience in guidance, precision as a mathematician, and outstanding creativity in research have profoundly helped me in my journey of research and influenced my views on science in general. I mostly enjoy the environment of the Durbin group, which has always been a rich resource for discussions and advices. Strong numerical skills in people around me have been tremendously helpful in my PhD education. I particularly thank Andrew Brown, Leopold Parts, Jared Simpson, Kees Albers, Kimmo Palin, Thomas Kean, Shane McCarthy, Aylwyn Scally, Stephan Schiffels, Vladimir Shchur and Andreas Leha for their help on many questions from me during my work. I thank the Wellcome Trust Graduate Program that funded this study. I thank Annabel Smith, Christina Hedberg-Delouka, Alex Bateman and Julian Rayner who supervised the PhD program and offered me great help both on my study and on my status as an international student. The program has been a wonderful training framework that offered precious training opportunities both on science and on personal skills that I

feel greatly benefited.

In the past four years, I worked with excellent collaborators on many aspects of my projects. I would like to thank John Winn in Microsoft Research for his insightful input on the work in chapter 3; Ewan Birney for his leadership in the CTCF project; and Oliver Stegle for his advice on statistical models. I also appreciate the time I worked as a rotation student in the laboratories of Mike Stratton and David Adams. It has been a great pleasure to work with these and many other people.

In the end, this thesis is dedicated to my family, who supported me the entire way.

Abstract

Waves of genome wide association studies (GWAS) have identified a large number of loci associated with disease predisposition and natural traits in the past decade. A number of identified variants have revealed potential causal mechanisms for the associated diseases. However, despite the early success, much of the phenotypic variation is not explained by the GWAS variants and the effect sizes tend to be very small. The real challenge in advancing our understanding, and subsequently making it relevant for clinical application, is deciphering the biological functions of these loci, which remain largely uncertain. Compared to the whole organism phenotypes that are distal to the genetic variants, cellular phenotypes are closer to genetic regulation, thus not only tend to offer effect size, as shown in expression QTL studies, but also are likely to mediate between genotypes and whole organism phenotypes, supporting biological functions.

In chapter 2, I describe a genetic association study on binding of a primary transcription factor CCCTC binding factor (CTCF) in human populations. We search for quantitative trait loci (QTL) for tens of thousands of CTCF binding sites in a group of 51 individuals, making this the first well powered QTL study on a major transcription factor in humans. We discovered a large number of QTLs and revealed a strong genetic component that contributes to binding variation. We found the associated variants are often located near to predicted binding sites, some perturbing the binding motif directly, and others affecting indirectly. We observed allele specific effect (intra-individual) consistent with QTL signals (inter-individuals), supporting a strong genetic component in CTCF binding variation.

In chapter 3, I address the problem of low power in associations between gene

expression levels and phenotypes. This is largely driven by the high degree of stochasticity in the measured gene expression levels. We showed that by applying factor analysis both to remove global confounding effects and to create summarizing factors for biological pathways, the heritability and association strength can be substantially elevated as a result. We applied this idea to a cohort with skin expression data with ageing phenotypes, and discovered heritable ageing pathways.

It is also of great interest to develop new methods for obtaining measurements of cellular phenotypes. In chapter 4 I describe a novel computational method to estimate telomere length from whole genome or exome sequencing data. Using data from the TwinsUK cohort that has both DNA sequencing data and experimental telomere length measurements available, I show that our method can effectively extract telomere length information. The method has been applied to a few cancer studies in collaboration and achieved early success in confirming experimental findings.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Life presents enormous diversity. From the white snow flower growing in the over 2000 meters high plateau in Tibet to the lionfish swimming in the Indo-Pacific sea, each life form has dramatically different appearance, structure, behavior, reproduction etc. Yet they fit in a global ecosystem finding their own ways of living, with their positions and roles shaped by the force of evolution. Fascinatingly, given how distinctive each life is, there are things that are shared and principles that they follow. Understanding how diverse life arises and how the characters transmit and spread between generations and species is key to reveal the basic principles that govern their biology.

Some of the patterns must have been realized by ancient humans. These include that children are more likely to have similar appearance to their parents in having similar eye colors, skin colors, height and so on. However, it was not at all straightforward to understand the reason for such similarity between parents and their offspring. To answer this question, a number of prerequisite questions need to be addressed first. How is information faithfully maintained in individuals, and the cells within them? How is this information transferred from one generation to the next and how does it control the characteristics of an individual?

Here I briefly review the progress in the genetics of traits that are at the level of individuals, and how we can use a similar approach to study the genetics of molecular traits at a cellular level, which must come between the genetic material and the organismal trait. I firstly introduce the history of mapping Mendelian traits and quantitative traits. I then discuss widely used methods that have been developed to date for genetic mapping. Finally, I describe the progress in studying cellular traits using the same principles.

1.1 Hunting for genetic determinants of phenotypes

1.1.1 Mendelian traits

The understanding of the basic principles of genetics has come a long way in the last 150 years. In early and mid nineteenth century, the use of hybridization in plants to obtain flowers with desired colors was already studied scientifically (Gärtner, 1849). But the principles that govern the formation of the traits were yet to be formulated. Gregor Mendel chose peas (*Leguminosae*), which has clearly distinguishable characteristics and good protection at flowering time from foreign pollen contamination, for his studies(Mendel, 1865). After eight years of counting the number of peas with different seed coat colors, shape etc, his milestone paper in 1865 illustrated the principles that were later referred to as Mendelian Laws, which became the corner stones of the genetic field.

During the same period, scientific progress on cytology discovered possible physical molecules or structures that can be linked to Mendelian factors. In 1866 Ernst Haeckel postulated that the nucleus is responsible for heredity from the observation that sperms largely contain nuclei. Deoxyribonucleic acid (DNA) was first isolated by a Swiss physician Friedrich Miescher in 1869 and later in 1875 Strasburger discovered chromosomes. Sutton and Boveri in 1903 proposed the “chromosomal theory of inheritance”, which suggested a direct link between the Mendelian factors and a physical cellular molecule. The inheritance material was confirmed later in 1952 in the famous bacteriophage experiment by Alfred Hershey and Martha Chase.

1.1.2 Quantitative traits

Mendelian factors can be observed in traits that are separated into clear categories, such as the color of seed coats. But there are also, perhaps more prevalently, traits that are continuous and not clearly separable into discrete classes, such as the weight of peas or the height of plants. Many of these traits are also highly heritable. Initially, there seemed to be little connection between Mendelian factors and continuous traits. Early scientists were unable to discover a simple rule of heredity in these traits. Breeding studies such as East 1916 suggest that absolute dominance is rare. Even a Mendelian trait such as the plant color, with careful examination, still shows some variation. This suggests that quantitative characteristics probably result from the action of the environment on the segregation of many Mendelian loci. Statistical developments at the same time were helpful in reconciling the disconnection. Fisher's paper (Fisher, 1919) first introduced variance decomposition, which mathematically illustrated that the variance of a trait can be separated into different components, including those driven by genetic factors as well as non-genetic factors. Many of the concepts and methods in Fisher's paper became the foundations of quantitative genetics that we still use today.

The first quantitative trait loci (QTL) mapping was done by Karl Sax in 1923. He found that the weight of beans (*Phaseolus vulgaris*) followed a similar distribution to that of the pigmentation colors. The beans homozygous for color are about twice as heavy as the beans heterozygous for color. This observation suggested that either the Mendelian factor for color also affects weight as a quantitative factor, or there exists two tightly linked Mendelian factors that control the color and the weight of the beans, and that the effect on the weight is additive.

1.1.3 Genetic variation and markers

Mendel's law of segregation applies directly to alleles on different chromosomes. However, alleles on the same chromosome can be transmitted together as linkage groups and it is difficult to distinguish their individual effects. Recombination is the primary mechanism that separates them. In sexually reproducing diploid genomes, a pair of homologous chromosomes synapse followed by individual chromatids exchange segments of DNA in meiosis. The frequency of two genes being separated by a recombination event can be used to define their genetic distance, i.e. $m = -\frac{1}{2}ln(1-2r)$ by Haldane (Haldane, 1919), where r is the recombination frequency. It has been noticed that genetic distances do not uniformly distribute along the chromosome as the nucleotide distance, but instead have hotspots and cold spots (Jeffreys et al., 2005; Myers et al., 2005), and also varies considerably between genders (Kong et al., 2002). Recombination provides an important source of genetic variation and allows for evaluating the marginal effects of genes. Genetic markers that are experimentally accessible for capturing such variation are thus critical for mapping traits.

For most of the 20th century QTL studies have been greatly constrained by the lack of markers that can be densely spaced in the genome to capture the genetic variation. The development of DNA restriction fragment length polymorphism (RFLP) was the first method that substantially increased the resolution to DNA-level polymorphism. Eric Lander and David Botstein (Lander and Botstein, 1989) proposed statistical methods to dissect Mendelian factors in quantitative traits, which became the main stream approach in the following years.

In the last decade, technological advances made it possible to detect single nucleotide polymorphism (SNP) (see review Brookes, 1999), which is the most abundant form of genetic variation and offers a single base pair resolution. The International HapMap Project (The International HapMap 3 Consortium, 2010) is one of the key resources in defining a map of SNPs using nucleotide arrays. The project eventually genotyped 1.6 million SNPs in 1,184 individuals from eleven populations, focusing mostly on the common variants with allele frequency >5%. The 1000 Genomes Project was the first project to sequence a large number of individuals with a goal of cataloging genetic variation across populations. The pilot phase of the project (The 1000 Genomes Consortium, 2010) has identified 15 million SNPs, 1 million short insertion and deletions and 20,000 structural variants in 179 individuals from four populations. The most recent phase of the project (phase three) has identified 80 million SNPs in 2,523 individuals from 26 populations (unpublished). These projects have provided essential information for genetic mappings. Recently a number of projects aim to further improve cataloging genetic variation by sequencing a large number of individuals with particular focuses. This includes population wide sequencing project such as UK10K (http://www.uk10k.org/) for the British population, GoNL (http://www.nlgenome.nl/) for the Netherlands population, and SardiNIA (http://genome.sph.umich.edu/wiki/SardiNIA) for the Sardinian population, or disease focused projects such as the GoT2D for type 2 diabetes and the International Cancer Genome Project (International Cancer Genome Consortium, 2010, https://icgc.org/) for cancer.

1.2 Mapping quantitative traits

Heritable factors can be inferred from phenotypic distributions such as the frequencies of peas with different colors in Mendel's experiments. However, most traits do not have an intuitive phenotypic distribution as that of the pea color. The distributions can be very complex, particularly when a phenotype is controlled by multiple loci. In these cases, the marginal effect of individuals genes can hardly be detected or distinguished, and QTLs can not be discovered by modeling phenotype data only.

Using information provided by the molecular markers is an obvious way to resolve this puzzle. Although it is not possible to know the locations of the QTLs beforehand, with a dense marker map, some of the tested markers are likely to be in linkage disequilibrium with genuine QTL loci. These tagging markers can be mapped in a number of approaches, and quantitative methods have been developed to define the relationships between the markers and the traits.

1.2.1 Linkage analysis and its limitation

Linkage analysis aims to identify genetic factors influencing traits by analyzing the cosegregation of markers with the traits across generations in families. Based on this idea, linkage analysis has been tremendously successful in identifying Mendelian diseases. Some of the examples include the identification of multiple mutations in the CFTR gene causing cystic fibrosis (Tsui et al., 1985; Riordan et al., 1989), the disease haplotypes in Huntington's disease (Gusella et al., 1983; MacDonald et al., 1992) etc. The susceptible variants are often rare, possibly shaped by negative selection, but for the method to work they need to be highly penetrant.

Linkage analysis has also been applied to common diseases and quantitative traits. For example, variants have been identified associated with inflammatory bowel disease (IBD) (reviewed in Mathew and Lewis, 2004), type I diabetes (Luo et al., 1995; Mein et al., 1998) and schizophrenia (Williams et al., 1999; Ekelund et al., 2000). However, the heritability accounted for by the identified variants is typically very modest, even when the heritability of the disease is much higher, e.g. IBD and schizophrenia. Clearly, the reported loci are only a fraction of the full picture of the genetic architecture of these diseases. When the genetic genetic architecture is complex, where the phenotype is determined by a collection of variants with low penetrance, often a very large number of families is needed to discover and differentiate these effects. For example the association of type 2 diabetes with the Pro12Ala variant in the peroxisome proliferative activated receptor-γ gene (PPARG), which has an effect size of 1.25 fold, could only be detected using linkage studies of over one million sib pairs (Altshuler et al., 2000). It is impractical to recruit enough families with several affected generations to obtain a sufficient number of informative meioses, especially given that human families tend to be small. It becomes even more challenging if the study disease has late onset. These results suggest that in contrast to Mendelian disease, where a limited number of high penetrance loci are responsible for the disease phenotypes, complex diseases have much more complex genetic architecture that the linkage analysis approach is not well powered to discover.

1.2.2 Population association analysis

Instead of using a linkage study design, a simple statistical association can be used, which merely states the co-occurrence of genotypes and phenotypes in a population. Such an association may exist due to the fact that a DNA segment that contains a variant affecting disease susceptibility can be inherited by many individuals that share a common ancestor who carries the factor. This approach has been extremely powerful over the last 8 years, resulting in over 10,000 genotypes to phenotype associations (Wellcome Trust Case Control Consortium, 2007; The NHGRI GWAS Catalog, Welter et al., 2014).

1.2.2.1 Mapping disease variants with case control phenotypes

For many diseases, there is no clear quantitative measure indicating the disease status. The phenotype is then reduced to a binary form of whether an individual does or does not have the disease. In this scenario, a case control design is often used. Healthy and disease individuals, often on the order of thousands, are recruited to a study and genotyped for a large number of variants, on the order of hundreds of thousands to millions. The basic idea is to compare the genotype frequency of the markers between the cases and the controls. A highly divergent marker frequency would suggest a possible link between the marker and the disease status. For each variant, the number of individuals with AA, AB and BB genotypes can be counted for the healthy individuals (m_{0j}) and the disease individuals (m_{1j}) to form a contingency table as below.

Genotype	AA	AB	BB	Total
Case	m_{11}	m_{12}	m_{13}	$m_{1.}$
Control	m_{01}	m_{02}	m_{03}	$m_{0.}$
Total	$m_{.1}$	$m_{.2}$	$m_{.3}$	m

The association can be tested using a χ^2 test: $\chi^2 = \sum_{i=0}^{1} \sum_{j=1}^{3} \frac{(m_{ij}-E[m_{ij}])^2}{E[m_{ij}]}$ with two degrees of freedom, where $E[m_{ij}] = \frac{m_{i.}m_{.j}}{m}$. The effect of a genotype can then be estimated as an odds ratio $OR_{AA} = \frac{m_{13}/m_{11}}{m_{03}/m_{01}}$.

1.2.2.2 Mapping QTLs using a simple *t*-test

A variety of methods have been developed for QTL mapping (Leal, 1998; Balding et al., 2008). Here I introduce the widely used t–tests, the analysis of variance and more recently linear mixed models.

The marginal effect of substituting allele A with allele B can be evaluated by comparing the homozygous individuals AA and heterozygous individuals AB, assuming effects are normally distributed in each genotype group with same variance. Let μ_0 and μ_1 be the genuine means of the phenotypes. The test hypothesis can be formulated as $H_0 : \mu_0 = \mu_1$, and $H_1 : \mu_0 \neq \mu_1$. The test statistic is thus

$$t = \frac{m_1 - m_0}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_0})}}, \ s^2 = \frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}$$

where (m_0, m_1), (n_0, n_1), and (s_0, s_1) are the sample means, samples sizes and sample standard deviations of AA and AB respectively. H_0 is rejected if t exceeds a significant threshold, such as $\alpha = 5\%$ when $t > t_{(0.025)}$ for a two tailed test with n_0+n_1-2 degrees of freedom.

1.2.2.3 Mapping QTLs using linear regression models

More generally, population samples contain three genotypes (AA, AB, BB). The quantitative phenotype y can be modeled as resulting from the sum of genetic effects and environmental effects in a simple linear model

$$y_i = \mu + \beta x_i + \epsilon_i, \ i = 1...n,$$

where y_i is the phenotypic value of the ith individual; x_i is the genetic dosage of the ith individual, which is the allele count of one allele such as (0,1,2) for the number of B alleles in genotypes (AA, AB, BB). ϵ represents the random error in y that can not explained by x, which is assumed to be independently and identically distributed. To satisfy the assumption for the distribution of the error term, often phenotypic measurements need to be transformed, e.g. using log, square root or mapping to normal quantiles. This can also be extended to generalized linear models that allow for response variables that have a variety of error models.

A simple way to make inference about the parameters (β, σ^2) is to use the least square approach, which gives

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y},\ \hat{\sigma}^2 {=} \frac{1}{n-1}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)$$

where $\mathbf{y}^T = (y_1, y_2, ..., y_n)$ is a vector of phenotypes and $\mathbf{X}_j^T = (x_{j1}, x_{j2}, ..., x_{jn})$ is a vector of genotypes at variant j for each individual. β is often interpreted as the effect size, representing the contribution of a unit change in the genetic dosage encoded in x to the phenotype y.

This is equivalent to a single factor analysis of variance (ANOVA) of the genetic effect. The mean sum of squares within genotype groups SS_{within} reflects any other residual variation that is non-genetic. The difference between the total sum of squares (SS_{total}) and SS_{within}, $SS_{between}$, reflects the QTL genotypes effect on the phenotypes. The ratio between them $\frac{SS_{between}/(3-1)}{SS_{within}(n-1)}$ is an F value that can be used to test for the genetic effect. The statistical significance level can be computed by comparing to the F distribution with degrees of freedom 2 and $n-3$. If we let $\sigma_e$be the environmental variance and $\sigma_g$be the genetic variance, the heritability can be expressed as $h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$, where σ_e can be estimated by $SS_{within}/(n-1)$ and σ_g^2 can be estimated by $(SS_{between} - SS_{within})/k$ where k is a factor adjusting for group size of three genotype groups ($k = 3/(\frac{1}{n_0} + \frac{1}{n_1} + \frac{1}{n_2})$). In case of comparing two genotype groups, $F = t^2$.

Many studies also use maximum likelihood approaches to estimate genetic effects. With the same linear model, the likelihood function is

$$L(\mu, \beta, \sigma) = \prod_{i}^{n} z(y_i - (\mu + \beta x_i), \sigma^2)$$

where z is the standard normal density function $z(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. The inference of the parameters is often done using the Expectation-Maximization algorithm (Moon, 1996). The likelihood of the full model with estimates $(\hat{\mu}, \hat{\beta}, \hat{\sigma})$ can be compared against a null model $(\hat{\mu_0}, 0, \hat{\sigma_0})$ where the genetic effect is removed by setting $\beta = 0$

to compute likelihood ratio statistic

$$LOD = -2Log(\frac{L(\hat{\mu}, \hat{\beta}, \hat{\sigma})}{L(\hat{\mu_0}, 0, \hat{\sigma_0})})$$

LOD has an asymptotic χ^2 distribution with one degree of freedom, which can be used to determine statistical significance.

1.2.2.4 Mapping QTLs using linear mixed models

Spurious associations can arise when study samples in association analysis have variable genetic relationships, in which case the ϵ_i are not independent. Such confounding factors of relatedness may not be known to the researcher from phenotypic data collection. To adjust for it, the linear mixed model approach has become a popular method of choice recently. These models typically use an additional random variable with a specific covariance structure to capture the genome wide sample relatedness (Kang et al., 2008; Zhang et al., 2010; Listgarten et al., 2012; Zhou and Stephens, 2012):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \epsilon$$

where y is the phenotype vector, β is an unknown fixed effect for the candidate genetic marker, and u is an additional random effect reflecting the genetic effect due to relatedness. u is normally distributed with $u \sim N(0, K\sigma_g^2)$, where K is the kinship matrix, with each $k_{i,j}$ the correlation between the markers either genome wide (Kang et al., 2010) or a selected subset (Listgarten et al., 2012), of individual i and j. σ_g^2 is the unknown genetic variance. For statistical testing, similarly, a likelihood ratio statistic can be computed by comparing against a null model. This model successfully removes false positives due to sample structure. It can also help to refine genuine signals by controlling for the other genetic markers that are not the candidate locus being tested, such as using only markers on chromosomes except the one that the test mark locates in (Listgarten et al., 2012).

The main limitation of the linear mixed model approach is the computation cost, which in the full model is of the order of MN^3 (Kang et al., 2008). There have been improvements on reducing the cost by making approximations. One approach is based on the assumption that the genetic effects of total markers by u is approximately shared between individual markers, thus the relationship matrix only needs to be built once instead of each time per marker. The data is then rotated by the eigendecomposition of the relationship matrix to remove the structure (Listgarten et al., 2012). Another approach is based on the observation that a relatively small number of independent markers can be selected to capture the information about relatedness. A careful selection of markers could dramatically reduce the size of the relationship matrix and allows for exact analysis in each test (Listgarten et al., 2012).

1.2.3 Multiple testing correction

In an association scan, a collection of statistical tests is typically evaluated for a large number of markers. While there are good reasons for doing so, such as one wishes to allow as many genetic causes as possible, this leads to a major issue in the greatly increased probability of declaring false positives. Typically a nominal $p = 0.05$ is used to claim an effect is statistically significant. This means that the probability of rejecting null hypothesis is 5% by chance. However, in cases where a data set is used to test for many hypothesis, the probability of reaching $p = 0.05$ is substantially elevated by chance. For example, in 100 tests, the probability of observing at least one test significant at 5% level is

$$Pr(\min_i p_i \leqslant 0.05) = 1 - Pr(all\ p_i > 0.05) = 1 - (1 - 0.05)^{100} = 0.99$$

which means it is almost guaranteed to have at lease one nominally significant association.

Methods have been developed to resolve this problem by adjusting the threshold when multiple tests are performed. The minimum p value distribution, which is substantially skewed to low p values, is used as the p value distribution under the null hypothesis instead of the individual p value distribution, which is uniform. The corresponding error rate is often referred to as the family wise error rate (FWER).

The Bonferroni correction is a simple method to control for $FWER < \alpha$ by using a threshold of $p < \frac{\alpha}{m}$, where m is the number of tests.

Bonferroni correction can be too strict in many cases. A more liberal approach is to control for a false discovery rate, where the significance is declared while accepting a fraction of false positives. One popular method is the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995), which relies on the assumption that p values under the null model are uniformly distributed. A false discovery rate can thus be calculated by comparing the observed value against its percentile rank: $p < \frac{k}{m}q$ to declare k signals out of m tests at a false discovery rate q.

More recently, a q value approach (Storey and Tibshirani, 2003) was developed and frequently used in many studies (Degner et al., 2012; Maurano et al., 2012; McVicker et al., 2013). It further calibrates the balance between the fraction of the declared significances and the false positives in an automated way. It estimates the proportion of tests π_0 that are drawn from the null by fitting a cubic spline to the p value distribution and taking the frequency of the p values at the end of the distribution, which reflects the proportions of nulls when there is no association. This π_0 can then be used to calibrate a false discovery rate at any p value level.

The distribution of p values under the null hypothesis can also be empirically estimated using permutations. This is normally conducted by random assigning phenotypes or genotypes to individuals. The nominal p values from the original test can be compared to the p values from the permutations to establish an FDR level.

1.2.4 Statistical power in genetic associations

The statistical power to detect associations between genotypes and phenotypes depends on a number of factors. Situations where variants have small effects are particularly hard to map. The linkage strength between a marker and a genuine QTL also adds to the complexity. Below I discuss how these factors relate to each other using the simplest t-test model.

Assume that we want to seek for results controlling for a type I error rate α and a type II error rate β, then

$$1 - \beta = Prob(t_1 > z_{\alpha/2}) = 1 - \Phi(z_\alpha - t)$$

where z_α is the critical value for the confidence level $1 - \alpha$under the null hypothesis $t_1 = 0;\Phi$ is the standard normal cumulative density function. If we assume the ratio of AA:AB:BB is 1:2:1, a QTL is linked with the tested marker at a recombination rate r, and the genuine additive effect is a, then the difference between AA and BB is $m_2 - m_0 = (1 - 2r)2a$, where (m_0, m_2) are the phenotypic sample means of AA and BB. The t statistics is calculated as

$$t = \frac{m_2 - m_0}{\sqrt{s^2(\frac{4}{n} + \frac{4}{n})}} = \frac{(1 - 2r)2a}{\sqrt{8s^2/n}}$$

Replacing t with $z_{\alpha/2} + z_\beta$,

$$n = 8[\frac{z_{\alpha/2} + z_\beta}{(1 - 2r)2a/s}]^2$$

We can see that QTL can be detected with small n if the effect size a is large, the linkage r between the marker and the QTL is strong and the residual noise s is small. Note that the QTL effect is only mediated via the marker locus that is linked to the causative variant, thus the real effect can be under estimated, and it is not easily distinguishable between a strong effect via weak linkage and a weak effect via strong linkage.

So far the most reliable way to validate a discovery is to replicate the result in an independent sample cohort. The general principle of choosing the replicate setting is to repeat the initial experimental design as closely as possible, with samples drawn from the same population and phenotypically ascertained using the same procedure. The position of the associated loci in the replication cohort must be identical to the original position or in strong linkage disequilibrium, with an effect in similar order and in same direction. One caveat is that the effect in the initial association can be over estimated due to winner's curse (Zollner and Pritchard, 2007). On the other hand, fewer tests are conducted in replication, reducing the multiple testing burden. Estimates from multiple replicates can regress towards the genuine mean effect.

1.2.5 GWAS results and interpretation

The Wellcome Trust Case Control Consortium (Wellcome Trust Case Control Consortium, 2007) performed the first large association study by comparing disease individuals and healthy individuals (case-control design) for 7 diseases. Some of the early successes using the GWAS approach include the discovery of TNFSF15 as susceptibility gene to the Crohn's disease (Yamazaki et al., 2005) and TCF2 (or HNF1B) for type 2 diabetes and prostate cancer (Gudmundsson et al., 2007). Variants are also found in genes that can be targeted by drugs, such as PPARG and KCNJ11 associated with type 2 diabetes, and IL12B associated with psoriasis, targeted by thiazolidinediones, sulfonylureas and anti-p40 antibodies (Krueger et al., 2007; Manolio et al., 2008). These early results were followed by an explosive growth of studies with more study individuals and using more dense genetic markers. Recently, the International IBD Genetics Consortium (IIBDGC) (http://www.ibdgenetics.org/) reported the discovery of 163 loci associated with Crohn's disease using a very large study cohort consisting of over 75,000 individuals (Jostins et al., 2012). Many of the loci reported in this study are implicated in other immune-mediated disorders, e.g. ankylosing spondylitis and psoriasis.

In some cases, the causal relationship is plausible, such as the IFIH1 gene identified as associated with diabetes (Nejentsev et al., 2009). The gene is known to play a role in antiviral infection and there is strong link between type 1 diabetes and viral infection(Nejentsev et al., 2009). However, more often there are cases where the functional relevance is not obvious. Variants reported in different studies sometimes reveal unexpected connections, e.g. CDKN2A is reported to be associated with Coronary disease, type 2 diabetes, and invasive melanoma (Kamb et al., 1994; Helgadottir et al., 2007; Scott et al., 2007).

One important observation in genome wide association studies is that the odds ratio for associated variants is modest, typically between 1 to 1.5 (Hindorff et al., 2009).

For example, a recent large scale genome wide association study on type 2 diabetes in more than 150,000 individuals revealed more than 70 loci but only explain 11% of T2D heritability(Morris et al., 2012). Similarly, a large scale study on Crohn's disease showed that the heritability is only 23% (Franke et al., 2010). The reasons are two fold. First, it is possible that the variants identified by the genome wide association study are only a small subset of the variants that contribute to the disease etiology. Due to the statistical power and the winner's curse, the rest is not sufficiently powered to be discovered, particularly the ones with low allele frequency, e.g. MAF<1%. Second, the genetic effect of a DNA variant must propagate through multiple levels of cellular networks, regulated by other mechanisms such as epigenetics or by environments, which substantially reduces the initial effect, manifesting a weak effect at higher level that can result from initial strong effects at cellular level.

Albeit the challenges and limitations, the GWAS results nevertheless highlight informative clues on the underlying biology. To seek further understanding, one has to investigate deeper into tissues and cells, these reasons have motivated studies into mapping molecular phenotypes measured in a cell, where most statistical methods are also applicable.

1.3 The promise of cellular phenotypes

1.3.1 Moving towards cellular phenotypes

Cellular processes are more directly subject to genetic regulation. For that reason, the effect size, defined as the magnitude of change in the downstream measurement by a change in the genetic allele, could be much higher than individual level traits.

Another important advantage is that cellular phenotypes can be linked to interpretable cellular products. The regulation process can be seen as a generative process, starting from decoding the information stored in DNA to transcribing into RNA and then to translating into proteins. The measurement of the product abundance at each step could reveal the mechanistic process with direct relational context. In practice, it is not yet technically possible to capture all these types of information simultaneously, but it is already feasible to measure each step separately using various molecular assays and integrate the measurements later in computational analysis.

1.3.2 The measurement of cellular phenotypes

The first major leap in large scale measurement of cellular phenotypes is perhaps the microarray. It is based on the same idea as Southern Blot (Southern, 1992) that DNA fragment can be hybridized to known complementary DNA sequences, called probes. This can be used to measure gene expression levels, where mRNA is reverse transcribed into cDNA, which can then be hybridized to a microarray. The method allows simultaneous quantitation of a large number of probes, designed to target a number of genes. The first study using microarray profiling gene expression was published in 1995 (Schena et al., 1995).

In 2005, the birth of next generation sequencing started a new era for genomic assays. It has further revolutionized the sequencing of DNA using an idea of sequencing by synthesis for a large amount of short DNA fragments in a massively parallel way(Bentley et al., 2008). The price has dropped exponentially as a result, from $10M in 2005 to $4000 in 2014 per human genome at 30x coverage (NHGRI, www.genome.gov/sequencingcosts). It is gradually making investigating genetics at genome wide scale for a large group of individuals practically feasible.

Next generation sequencing technology also gives rise to a large variety of assays that are designed to measure other molecules. The basic idea is to transform the desired molecular information into a collection of DNA sequences, which can be sequenced. During the past few years, a rich collection of methods have been developed. For example, for RNA transcription related information, RNA-seq (transcript abundance, Chu and Corey, 2012) and GRO-Seq (binding sites of active Pol II, Core et al., 2008); For translation, Ribo-Seq (ribosome profiling, Ingolia, 2014) etc have been developed. For DNA Methylation, Bisulfite Sequencing (BS-Seq, Krueger et al., 2012) and MeDIP-Seq (Taiwo et al., 2012) are widely used. For DNA-Protein interactions, there are DNase-Seq, FAIRE-Seq and ChIP-Seq (See review Furey, 2012). A recent refinement of ChIP-seq, ChIP-exo, is able to identify the exact bases that are bound by a factor (Rhee and Pugh, 2011). Chromosome conformation can be measured by assays such as Hi-C/3C-Seq and more recently 5C, which relies on the cross linked DNA generated due to interactions between two factors (Dostie and Dekker, 2007; Simonis et al., 2009; Lieberman-Aiden et al., 2009). There are also assays designed to measure special sequence elements, such as Tn-Seq for transposon sequencing.

The resulting sequences from these assays can be aligned to a reference sequence to reveal the information about where the event has occurred and how much of target products exist in the starting material. Chapter 2 of this thesis uses ChIP-seq technology in particular for measuring bindings of the CCCTC (CTCF) binding factor. In detail, it works by extracting segments involved in protein-DNA interactions using Chromatin Immunoprecipitation(ChIP) followed by sequencing. When protein-DNA interaction occurs in a cell, binding proteins and DNA segments are temporarily bonded as a complex. Such structure can be chemically strengthened using cross linking agent, after which the long DNA molecules are then shared into ~500bp fragments by sonication. This produces a mixture of DNA fragments, within which some are bonded by proteins. The ones of interest are then selectively immunoprecipitated from cell debris using specific antibodies, such as anti-CTCF in the case of chapter 2. The target molecule is thus enriched and purified. Once this material is obtained, the associated DNA fragments can be extracted out and sent for sequencing. The initial locations of the protein-DNA interactions can then be determined by aligning these sequences back to the reference genome. The quantity of the fragments corresponds to the number of the molecules in the starting material, representing the intensity of the binding. One caveat is that there is a number of sources of technical variation involved in the data production (Taub et al., 2010). For example, non specific fragments may remain in the purified material, which then become background for the real binding sites. Computational methods have been developed to differentiate signals from background.

The fast development of cellular assays has opened the door to obtain cellular information in an economical, genome wide, and simultaneous way. This has allowed to investigate the genetic landscape of molecular traits such as gene expression, transcription factor binding, histone modification etc. The relations or dependencies between these molecular events can be examined in a scale that has never been reached before. QTL approaches can be applied to discover genetic loci that play a role in regulation in various levels and aspects of the molecular processes in a cell.

1.3.3 Latent variables in analyzing high dimensional genotypes and cellular phenotypes

In association mapping with disease traits, tens of thousands to tens of millions genotypes are assayed. In QTL mapping of cellular phenotypes, in addition to a large number of genotypes, a high dimensional phenotype is also measured, e.g. by microarray or by next generation sequencing assays. The measured phenotypic variation can come from sources such as cellular fluctuations (Liebermeister, 2002; Dueck et al., 2005; Gibson, 2008), regulation of gene expression (Sanguinetti et al., 2006; Pournara and Wernisch, 2007), and environmental conditions (Hastie et al., 2000), many of which are confounding factors that need to be accounted for to prevent loss of power in discovering true signals and also false discovery of spurious signals (Leek and Storey, 2007; Hyun et al., 2008).

One way of disentangling the mixture in a high dimensional dataset is to use dimension reduction techniques to identify key components that reflect the data structure. On the one hand, these components can be used to reflect the relationships between samples learned from the measured dataset, thus become useful indicators when independent sampling is assumed. On the other hand, they can help identify sources that are influential to a large number of traits, which often come from a non-interesting source, such as technical batches. Data for some of the factors that affect transcript levels may have been collected by researchers, such as age, experimental batch, etc., the inclusion of which as covariates in association models have shown to improve QTL discoveries (Emilsson et al., 2008). However, perhaps more prevalently, the confounding factors are hidden to researchers. In this case computational methods can help identify them, which can be considered alongside known covariates in association mappings (Leek and Storey, 2007; Hyun et al., 2008; Stegle et al., 2010; Nica et al., 2011; Fusi et al., 2012). At a smaller scale, such as genes in a pathway, such component can themselves become phenotypes that reflect the commonality of the traits that are functionally linked. This thesis explores this in Chapter 3 in the context of associating gene expression to ageing phenotypes.

Among the dimension reduction techniques, principal component analysis (PCA) is perhaps the most widely used method. In PCA the original data is converted by an orthogonal projection onto a lower dimensional linear space known as principal space. The corresponding dimensions, known as principal components, are learned from the data by either maximizing the variance of the projected data (Hotelling, 1933) or equivalently minimizing the averaged projection costs defined as the mean square error between the projections and the initial data points (Pearson, 1901). Consider a dataset $\{\mathbf{x}_n\},\ n = 1, ..., N$ where $\mathbf{x}_n = \{x_{nm}\},$ m={1,...M} with N observations each of M dimensions. PCA attempts to project the data onto a space with dimension $D < M$. The dimension of the new linear space can be defined by a unit vector $\mathbf{u}$ with constraint $\mathbf{u}^T\mathbf{u} = 1$. The variance of the projected data is thus given by $\mathbf{V} = \mathbf{u}^T\mathbf{S}\mathbf{u}$ where $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$. Maximize $\mathbf{V}$ with the constraint on $\mathbf{u}$ gives a quantity λ that satisfies $\mathbf{S}\mathbf{u} = \lambda\mathbf{u}$, where $\mathbf{u}$ and λ are the eigenvector and eigenvalue of $\mathbf{S}$. This process can be repeated to obtain D principal components $\{\mathbf{u}_1, ..., \mathbf{u}_d\}$ with corresponding eigenvalues $\{\lambda_1, ..., \lambda_m\}$ with $\lambda_1 > \lambda_2 > ... > \lambda_d$.

As a non parametric method, PCA has an advantage of not requiring model assumptions. Other advantages also include fast computational speed, and very easy visualization for separating samples based on their high dimensional measurements, e.g. Novembre et al. (2008) showed the first two principal components learned from one million genotypes correctly separated European populations into geographic groups. The PC projection can also be directly linked to the genealogical history of samples (McVean, 2009), although this may not be unique as multiple processes such as isolation, migration and admixture can give similar projections. In disease mappings, because of this property, PCs are useful to control for population stratification, where they can be included as fixed covariates in association models (Price et al., 2006; Novembre and Stephens, 2008).

Alternative to the linear projection, PCA can also be expressed as a probabilistic solution for latent variables using maximum likelihood, known as the probabilistic PCA

(Tipping and Bishop, 1997). Consider $\mathbf{z}$ latent variables corresponding to the principal components with a prior probability $p(\mathbf{z}) = N(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix, the conditional probability of the observations is $p(\mathbf{x} \mid \mathbf{z}) = N(x \mid \mathbf{W}\mathbf{z} + \mu, \sigma^{\mathbf{2}}\mathbf{I})$, where the columns of $\mathbf{W}$ define a D dimensional linear subspace. The observations in $\mathbf{x}$ is thus reconstructed from a mapping between a space spanned by $\mathbf{W}$ to the original data space by $\mathbf{x}$, with additional Gaussian noise with a variance of σ^2. Different from the conventional PCA, the numerical solutions of $\mathbf{W}$ such as via the EM algorithm do not guarantee that the columns of $\mathbf{W}$ are orthogonal to each other.

Probabilistic PCA has some advantages over the conventional PCA. This includes more efficient inference using the EM algorithm, having a likelihood function that can be readily used for comparison with other models, automatic identification of the dimension of the subspace due to the Bayesian treatment etc (Bishop, 2006). A closely related method to the probabilistic PCA is factor analysis (Basilevsky, 1994; Tipping and Bishop, 1997). The only difference is that its covariance structure is defined as $\mathbf{\Psi}$, an $M \times M$ matrix, instead of an isotropic $\sigma^2\mathbf{I}$, where $\mathbf{\Psi}$ captures the independent variance associated with each coordinate. This feature has shown to be particularly useful in capturing natural correlation between variables, such as expression levels of genes that are functionally linked. The PEER package (Stegle et al., 2010) provides software for both conventional PCA and factor analysis for high dimensional genomic data.

1.3.4 The genetics of gene expression

As an important product of DNA coding, gene expression is technically feasible to measure and has thus attracted a large amount of research focus. Many studies have looked for expression QTLs (eQTLs), aiming to link genetic variation to expression levels of gene products. Genetic regulation of gene expression levels has been found to underlie phenotypes from human diseases (see reviews Kleinjan and van Heyningen, 2005; Wray and Wray, 2007) to the morphology of Darwin's finches (Abzhanov et al., 2004).

Genetics of gene expression variation Gene expression variation can result from a number of factors, including environmental effects, epigenetic effects, biological random fluctuations, and genetic effects. How much of phenotypic variation can be explained by genetics is one of the core questions that need to be addressed. Studies have shown that a high proportion of gene expression levels (over 40%) are heritable across individuals (Petretto et al., 2006; Stranger et al., 2007; Dixon et al., 2007; Göring et al., 2007; Price et al., 2011; Grundberg et al., 2012). The heritability varies but is mostly greater than 10%, which is much larger than that from a typical GWAS study for whole organism phenotypes (see review Skelly et al., 2009).

A recent study on over four hundred twins found 8,329 out of the 13,970 genes in the investigation have shown one or several QTLs (Lappalainen et al., 2013). This suggests that the expression levels of a large majority of the genes are under genetic control. An additional level of evidence supporting this can be seen when comparing the expression levels of the two alleles of a gene within heterozygous individuals, or the allele specific expression (Morley et al., 2004). This study has found in humans around 6.5% of sites per individual show allele specific expression, largely consistent with the eQTLs. As the sample size and the accuracy of measurement increases, more eQTLs are likely to be discovered, including ones with relative weak effects (Cheung et al., 2010; Grundberg et al., 2012). The current data suggest that variation in the expression of genes is substantially linked to the genetic background.

The eQTLs discovered so far are enriched in regions close to the transcription start site (TSS) and the transcription end site (TES), areas known to play a role in the regulation of gene expression, mostly via transcription factor binding or methylation modifications (Veyrieras et al., 2008; Dimas et al., 2009; Stranger et al., 2012). Many eQTLs are found within the promoter binding motif, directly perturbing the binding in the interface, making them very likely to be causal. These effects are presumably mediated via changing the binding affinity of the promoter complex, which subsequently affects the efficiency of gene transcription. This may also cause differential usage of promoters, which is known to be an important source of variation in gene expression (Forrest et al., 2014). Notably, 16% of disease GWAS variants are eQTLs (Lappalainen et al., 2013), evidence supporting an effect route from DNA to gene expression then to disease traits.

Genetic regulation in *cis* and *trans* Genetic effects on gene expression and other molecular phenotypes within a genomic location are often categorized into in *cis* and in *trans*. These term were initially introduced by Haldane (Needham, 1942) to describe different allele configurations in heterozygous individuals, with a meaning actually more similar to linkage disequilibrium. The terminology was later used by Lewis (Lewis, 1945) to describe whether two mutations are in the same gene.

In the eQTL literature, *cis* and *trans* are typically defined more based on the distance between the associated variants to the target genes. Genetic regions proximal to the target genes are referred to as *cis* regions while the ones at a different chromosome or far from the target genes are referred to as *trans* regions. This may be a reasonable classification as a large proportion of eQTLs are indeed close (<100kb) to the transcription start sites (TSS) of genes (**?**), representing an important type of *cis* elements. It however can also be problematic, for example the distance thresholds used for differentiating *cis* and *trans* is arbitrary in different studies, from a few hundred base pairs to one or two megabases.

Notably, the differentiation between the *cis* and *trans* effect can also be based on the principle that *cis* elements are allele specific, while *trans* element can act on both of the target alleles. One study design is to compare the ratio of transcription between the two alleles in a hybrid offspring and the gene expression levels between the two parents (Wittkopp et al., 2004). A consistent ratio would suggest a *cis* effect driven by the target gene while otherwise it suggests a *trans* effect driven by other factors somewhere else in the genome or epigenetics effects.

Localizing *cis* and *trans* effect elements involves hugely different levels of technical challenges. The search space for a *trans* association is the product of the number of genetic markers and the number of expression traits, which is several orders of magnitudes greater than that for a *cis* scan. As a result, *a trans* effect with a similar effect size as a *cis* effect is much harder to detect because of the multiple testing penalty. A *trans* scan also involves correcting for more confounding factors that further weakens the signal. Indeed, *trans* eQTLs discovered so far are only a small minority

(Stranger et al., 2007; Small et al., 2011) in human studies.

1.3.5 The genetics of transcription factor binding

Transcription factor binding variation is one of the primary mechanisms by which gene expression is modulated. Key questions include what is the variability of transcription factor binding, what drives it, and how does it affect variation of gene expression levels. Studies have largely taken one of two approaches: 1) investigate specific regions at which regulatory events occur to build transcription factor binding maps, and associate genetic sequence variation within the binding sites to the binding variation; 2) consider binding variation as a quantitative trait and apply QTL mapping.

Technically, transcription factor binding can be measured using ChIP-seq genome wide, which does not require prior knowledge of the binding sequence. The sequence reads from a ChIP-seq experiment can be aligned to the reference genome to recover where the binding events have occurred and how strong they are. This normally involves a computational analysis called peak calling, which essentially identifies regions with a higher density of reads compared to that in the background based on estimations using various models (see reviews Laajala et al., 2009; Park, 2009). Using the reads mapped at the identified binding peaks, algorithms have been developed to infer short sequence patterns, called motifs, with a length normally less than 20bp, predicted to be the binding interface between the transcription factor and the DNA nucleotide (Tompa et al., 2005; Elnitski et al., 2006).

Transcription factor binding variation Studies on binding variation between species suggest that many binding events are species specific, with large divergence between species. For example, Boyer et al. (2005) and Kunarso et al. (2010) showed that the binding of two key regulatory proteins (OCT4 and NANOG) in human and mouse embryonic stem cells show dramatic divergence. Such divergence was also seen in hepatocytes when comparing transcription factor binding profiles between human and mouse (FOXA2, HNF1A, HNF4A and HNF6, Odom et al., 2007). The binding profile are substantially diverged in closely related yeast (Ste12 and Tec1, Borneman et al., 2007) and fungi (MCM1, Tuch et al., 2008). A recent study comparing two transcription factors in five vertebrates reconfirmed the pattern shown in the previous studies (Schmidt et al., 2010), revealing that individual binding events are gained and lost rapidly during evolutionary time, although the conservation level varies largely between different transcription factors, suggesting different evolutionary constraint.

There is also substantial binding variation between individuals within species. Zheng et al. (2010) found 30% of sites of STE12 show binding variation in a group of yeast segregants from two divergent parents. Kasowski et al. (2010) profiled NFκb and Pol II in a small group of ten humans and showed that 25% and 7.5% respectively of sites vary between individuals. A subset of the binding variation correlates with downstream gene expression. This suggests that many differences in individuals and species are at the level of transcription factor binding, which plays a strong role in species diversity and gene regulation.

Genetics of transcription factor binding variation It is of great interest to understand what gives rises to the variation in transcription factor binding. Genetic factors and environmental effects can both play a role. Recent studies have increasingly shown that heritable genetic effects are responsible for a large component of the transcription binding variation. A clever experiment by Wilson et al. (2008) provided convincing evidence. The study used an aneuploid mouse strain carrying a human chromosome 21, and asked whether transcription factor binding on chromosome 21 is driven by human sequence or by the mouse nuclear environment. The results showed that transcription factor binding on the human chromosome is largely recapitulated, supporting the hypothesis that transcription factor binding is mostly governed by genetic sequence.

Associating genetic variation with binding variation in yeast has showed that *cis* regulation plays the primary role (Zheng et al., 2010). The linked genetic variants tend to reside within the binding motif of the target protein or related cofactors. This suggests that genetics affects transcription factor binding by affecting the binding affinity at the protein-DNA interface. The variants that affect sequence motif that subsequently affect binding affinity correlate with the binding signals (Kasowski et al., 2010; McDaniell et al., 2010). It is also common that some binding events of transcription factors are correlated with mutations near the binding motif (Kasowski et al., 2010; McDaniell et al., 2010). Notably, binding variation also depends on the accessibility of the DNA in chromatin configuration, supported by the findings that sequence variation that affects DNase I sensitivity sites, nucleosome positioning and DNA methylation also affect transcription factor binding (Segal and Widom, 2009).

Validation of the function of transcription factor binding A variety of computational methods have been developed to infer the functions of QTL variants, mostly by summing evidence from published functional data sets as well as sequencing conservation (McLaren et al., 2010; Kircher et al., 2014). Eventually, an experimental validation such as by gene knock-down or nuclei base editing (Cong et al., 2013; Hwang et al., 2013) will be required to confirm the predictions. For example, Cusanovich et al. (2014) investigated differential transcription factor binding by knocking down 59 transcription factors in one HapMap lymphoblastoid cell line. The results show that most transcription factor changes only exert weak impact on the expression levels of genes within a 10kb window, and the ones that cause large changes tend to be located at transcription factor binding clusters, or at sites with high binding affinity or at enhancer regions. In a related study, the FANTOM consortium (The Fantom Consortium, 2014) knocked down 52 transcription factors in an acute monocytic leukemia-derived cell line (THP-1) throughout a time course of growth arrest and differentiation (Suzuki et al., 2009), revealing complex roles of transcription factors in the regulatory network, with no single transcription factor driving the differentiation process. These studies have identified a small number of functional transcription factors or core regulators, a perturbation of which cause immediate downstream gene expression changes.

In general, the connection between the transcription factor network and gene expression appears to be complex with individual effects being relatively weak. It is possible that the perturbation of a single transcription factor can be compensated by other factors in the same biological process. Sophisticated system biology approaches may help reveal the network relationships and their impact on the gene expression.

It is also noted that, although there has been a large volume of studies on transcription factor binding, primarily driven by technological advances such as ChIP-seq, it is still not possible to profile all transcription factors in a cell. High quality antibodies are still not available for all factors due to technical limitations. The scope of current studies is largely affected by this technical limitation to focus on a small number of transcription factors whose measurement is technically robust. A much bigger picture of the binding landscape is yet to be revealed.

1.3.6 The genetics of other epigenetic variation

The DNA molecule is physically organized into a three dimensional structure of chromatin. The scaffold of the structure that DNA coils around is made by protein complexes called nucleosomes that are composed of histone protein octamers. Regulatory information is conveyed by the positions of nucleosomes and the modification of histone proteins, the tails of which can be covalently modified by methylation or acetylation (Campos and Reinberg, 2009; Segal and Widom, 2009). Such modifications have been shown to correlate with downstream functions. Covalent modification can also occur on nucleotide with methyl groups added to cytosine. These modifications, which interact closely with DNA nucleotide itself and play important roles in the readout of DNA information, are generally termed as epigenetics.

Epigenetic elements involved in organizing chromatin structure The architecture of chromatin is not completely understood. Studies have shown that there exist areas that are attached to the nuclear lamina forming a particular spatial organization. These lamina associated domains are surrounded by CpG islands and insulators such as the CCCTC binding factor, and are associated with low gene expression (Guelen et al., 2008). These results suggest a functional impact of chromatin structure by delineating broad active or recessive environments for the readout of DNA information by transcription. A related study applied Hi-C technology to identify higher order chromatin interactions genome wide in human and mouse embryonic stem cells. It identified “topological domains” that are particularly involved in the interactions, and these domains are correlated with insulator binding protein CTCF, housekeeping genes, tRNA genes and short interspersed element (SINE) retrotransposons (Dixon et al., 2012). The active chromatin areas also correlate with open chromatin structure, with DNA in linear form and not wrapped around nucleosomes. These areas can be identified by using restriction enzyme DNase I, as only the open chromatin areas are exposed to excision sites that can be digested.

Genetic factors in epigenetic variation Epigenetic variation can result from genetic or non-genetic reasons. It is known that epigenetic modification helps to store the memory of the environmental exposures. A key question is to what extent epigenetic variation between individuals are due to genetic reasons? A related study (McDaniell et al., 2010) found that DNaseI foot print is highly heritable using six samples from a family of European ancestry. Another more recent study performed DNase I hypersensitivity site mapping in 70 HapMap cell lines of Yoruba ancestry, and identified a large number of genetic variants that are associated with the level of chromatin accessibility (dsQTLs). It estimates that over 50% of eQTLs are dsQTLs, with their effects mediated through chromatin accessibility. Based on the same set of cell lines, Bell et al. (2011) discovered methylation levels of 180 CpG-sites in 173 genes associated with *cis* QTL variants (10% FDR). Another study (Zhang et al., 2014) discovered that cytosine modifications at CpG sites are primarily driven by *cis* QTLs using over one hundred HapMap cell lines of European and African origin. A subset of these modifications colocalize with transcription factors to enhance or repress gene expression, often associated with changes in chromosome accessibility. These studies have established the regulatory connections between genetic variation and epigenetic variation (similar results are seen in McVicker et al., 2013).

Non-genetic factors in epigenetic variation It has also been seen that there exist substantial non-genetic causes for epigenetic variation. Environmental effects can also cause methylation variation thus in general the direction of causality is unknown. A change in methylation can either be the result of a genetic effect changing it to the current status, or an environmental exposure that is stored in a form of epigenetic modifications. The monozygotic twins with identical genetic background can help resolve this, as MZ twins have identical genetic background (Bell and Spector, 2012). In the aging context, Bell and Spector (2012) showed that differential methylation is associated with age and age related phenotypes in a twins cohort, but highlighted that a subset of these can be mediated by genetic reasons. Another study (Rakyan et al., 2011) investigated monozygotic twins pairs that are discordant for childhood-onset type 1 diabetes (T1D) identified methylation sites that are associated with the disease.

1.3.7 Tissue and environment effect in QTL mapping

One important caveat in QTL mapping is cell type and cell state. Studies on multiple tissues have shown that although there is a modest degree of sharing, quite often the regulatory effect of an eQTL is private to a specific tissue. It is therefore crucial to search for QTL in the correct tissue, i.e. in a tissue that is relevant to the disease of interest. This may not always be straightforward, as in some cases the regulatory effect is not in the tissue where a trait is manifested. Using a wrong tissue could be misleading. Cells with similar differentiation lineages have increased eQTL sharing relative to developmental distant tissues. However, a significant fraction of *cis*-eQTLs are cell type specific. This argues that variation that primarily affects late developmental processes may achieve sufficient power to be discovered by a *cis* scan (see review Gaffney et al., 2012). A number of projects have started to look into the landscape of gene expression as well as regulatory element signals in multiple cell types (**?**; The Fantom Consortium, 2014). This is still very challenging, as many tissues are not experimentally accessible, or are financially expensive when studied on a large scale, which can be necessary to detect weak effects or intra individual effects. The Human Induced Pluripotent Stem Cells Initiative (HipSci, http://www.hipsci.org/) is one of the first projects to systematically investigate genetics, epigenetic, proteomics and cell biology in induced stem cells and the differentiated daughter cells from them. With induced stem cells, HipSci is able to access tissues that are normally not very accessible from normal sample biopsy procedure, such as neurons, by differentiating stem cells into the target cells. Genetic mappings for a variety of tissues obtained in a number of differentiation stages may be powerful in revealing some of the key biological insights, such as how genetics regulates tissue differentiation. This could be the first step towards understanding this important process, and ideally one should measure such process *in vivo*.

Additional to cell type, cell state also contributes to the variation of molecular phenotypes. It is known that the transcription profile changes dramatically when a cell is at different stages of its life cycle (Marguerat et al., 2012). Most current studies are conducted in cells in quiescent state, which may not be the state relevant to the trait. Phenotypes are not necessarily present in the quiescent state, and the ones of interest can be hidden in this system. Applying assays that are targeting the correct cell state will be important in reveal the genuine regulatory architecture.

Environmental exposure is also an important source of variation. In the absence of accurate measurement, environmental factors will cause loss of power due to increased stochasticity. Most environmental exposures are hard to measure. The number of study cohorts with well annotated environmental measures, usually obtained from questionnaires, is very limited, and even if there is, it is not certain that the relevant quantities are measured.

1.3.8 Resolving the causative relationship

Cellular molecules work together in a system to achieve a biological function. Genes responsible for a particular biological function can be grouped together as a pathway, e.g. Kyoto Encyclopedia of Genes and Genomes (KEGG) catalogs some common pathways (Kanehisa et al., 2004). It is of interest to know the causative relationships among the molecules, which gives the knowledge of how a function is achieved. It is possible to do this from experiments by perturbing different combinations of various molecule levels. The relationship can however also be estimated numerically, for example, looking at conditional probabilities in different regulation configurations (e.g. Schadt et al., 2005). In the genomic context, one important characteristic is that genetic variation is generally fixed in an individual's life time, with the exception of somatic mutations. Thus it provides an important anchor to resolve such relationships (Lawlor et al., 2008), which is tremendously interesting in understanding its role in the network.

1.4 Overview of the remainder of this thesis

This thesis tackles a number of problems of mapping cellular traits as well as developing new methods for measuring cellular phenotypes. The remaining of the thesis is organized in four chapters

- Chapter 2 describes a first well powered systematic QTL study of a primary transcription factor CCCTC binding factor. This work is published in PLoS Genetics (Ding et al., 2014).

- Chapter 3 describes a method to substantially increase power in gene expression association to ageing. This work is accepted in G3 subject to minor revisions (joint first author).

- Chapter 4 describes a novel approach to measure telomere length using existing genome or exome sequencing data in a large scale. This work is published in Nucleic Acid Research (Ding et al., 2014, first author). The method has been applied to a melanoma study, which discovers several mutations in the Protection of Telomeres 1 (POT1) gene that are disease susceptible (Robles-Espinoza et al., 2014).

- Chapter 5 provides a conclusion, drawing together materials from the previous chapters and discusses future directions.

Chapter 2

The genetics of CCCTC binding factor

Collaboration note. *This chapter contains work in collaboration with Yunyun Ni, Sander W. Timmer and others in the research groups of Gregory E. Crawford, Jason D. Lieb, Vishwanath R. Iyer and Ewan Birney. This work is published in PLoS Genetics (Ding et al., 2014). I am the lead author alongside the other two joint first authors Yunyun Ni and Sander W. Timmer. My contribution in this work includes data production and quality control, quantifying CTCF binding regions, genotype production, association mapping, and jointly with Yunyun Ni allele specific analysis. The manuscript also contains a novel discovery of three distinct CTCF binding modes on X chromosome, which was primarily conducted by Sander W. Timmer, and is not presented here.*

2.1 Overview

In the past decade a large number of variants have been discovered associated with traits or disease. Although they provide important hints, it is not at all straightforward to understand the underlying biological mechanisms. The majority of the loci that have been found are in non-protein coding DNA sequences, suggesting regulatory roles often responsible for the phenotypic effect (The 1000 Genomes Consortium, 2010). Sequence conservation based approaches can identify the regulatory regions that are under selection pressure, possibly due to binding of a protein factor or other regulators (Lindblad-Toh et al., 2011). Recent sequencing based technologies can give a more direct measure for regulatory events, such as the binding of CCCTC factor (Kunarso et al., 2010; Schmidt et al., 2010) and a number of other factors, e.g. Noonan and McCallion, 2010 and McVicker et al., 2013, revealing the landscape of the regulatory elements in human genome.

Studying the effect of genetic variants on gene regulation has become an important approach to find intermediates between genotype and whole organism phenotype. Using DNase I hypersensitivity and binding assays for the CTCF transcription factor on two family trios with known genome sequences, McDaniell et al. (2010) showed that allele-specific binding patterns consistent with strong genetic effects could be readily measured at heterozygous sites. Other studies have shown allele specific binding of RNA polymerase and NF-κB binding measured across a small number of individuals (Kasowski et al., 2010), or of a wider range of transcription factors in a single cell line (Reddy et al., 2012). Similarly, differences between mouse strains in binding of PU-1 and CEBPa at enhancer regions correlate with sequence differences and adjacent gene expression (Heinz et al., 2013). Intriguingly, some sites with prominent SNPs in the binding motifs of CTCF did not show a genetic effect in a study of its binding across an extended family (Maurano et al., 2012). Reciprocally, differences in transcriptor factor binding were seen between closely related species even where there was no sequence difference in the binding region (Stefflova et al., 2013).

In order to examine these phenomena further, and infer potential causative connections to disease GWAS results, we need to identify specific cases where a genetic variant affects binding. To do this we can use genetic association mapping. When applied to transcript expression levels as the measurements on 60 or more samples, this approach has identified thousands of expression quantitative trait loci (eQTLs) (Spielman et al., 2007; Stranger et al., 2007; Pickrell et al., 2010). A QTL study of human open chromatin (Degner et al., 2012) found 8,902 DNase I hypersensitivity sites that were correlated with genetic variants. However, there are currently no systematic association studies of how genetic variation in human populations affects the binding pattern of a specific transcription factor. Here we carry out such a study.

To identify transcription factor binding QTLs, we measured the binding of CTCF across a panel of lymphoblastoid cell lines (LCL). Previous studies have shown that there is resemblance between LCLs and the parent lymphocytes at a variety of molecular levels including transcription factor binding according to accumulated observations (See review Sie et al., 2009). Despite of some inherit limitations, such as aneuploidy, gene mutations and reprogramming, often associated with telomerase activity, which can be controlled experimentally to a certain level, LCLs has still been instrumental in general as a resource for functional screening that offers acceptable fidelity and is scalable compared to clinical trails or *in vivo* systems.

CTCF is a highly conserved multifunctional protein that serves both as a transcription factor as well an insulator binding protein, preventing interactions between enhancers and promoters and demarcating chromatin domains. Working with cohesin, CTCF can also mediate chromosomal looping interactions, and is involved in imprinting as well as X-inactivation (see Lee et al., 2012; Merkenschlager and Odom, 2013 for reviews). There have been extensive locus specific studies (Bell et al., 1999; Bell and Felsenfeld, 2000; Yusufzai et al., 2004; Splinter et al., 2006; Stedman et al., 2008; van de Nobelen et al., 2010; Sopher et al., 2011) and specific genome wide screens (Cuddapah et al., 2009; Phillips and Corces, 2009).

Schmidt et al. (2010) showed in breast cancer cell lines and hepatocellular carcinoma cell lines CTCF appears to work independently to cohesin. Schmidt et al. (2010) compared CTCF binding patterns across five species and showed that its binding variation correlates with the evolution distances between species. Previous studies have shown the extent of genetic effects on CTCF binding in families (McDaniell et al., 2010; Maurano et al., 2012), although specific loci underlying these effects have not been identified.

We used ChIP-seq to measure CTCF binding in 51 lymphoblastoid cell lines (LCLs) from the HapMap CEU population, each of which had already been sequenced as part of the 1000 Genomes Project (The 1000 Genomes Consortium, 2010) and had been subjected to RNA-seq analysis (Montgomery et al., 2010). Our data and analysis identified thousands of CTCF binding QTLs across the human genome. These data, together with the available full genome sequence of the cell lines, allowed us to explore parameters of genetic effects on protein-DNA binding. For example, we defined the relationship of the QTL location to the TF binding motif, estimated the relative impact of substitutions and insertions/deletions (INDELs), and measured whether allele-specific differences are indicative of population-wide variation.

2.2 Measuring CTCF binding in HapMap cell lines

ChIP-seq Chromatin immunoprecipitation was done at the University of Texas Austin and sequenced at the Wellcome Trust Sanger Institute. Cells were cross-linked with 1% formaldehyde for 7 min at room temperature. Formaldehyde was deactivated by adding glycine. Chromatin from harvested cells was sonicated with a Bioruptor to an average size of 500 bp DNA. Immunoprecipitation was performed using sonicated chromatin by adding anti-CTCF antibody (Millipore 07-729). For a subset of eight samples, including day replicates GM12891 and GM12892, the same procedure was applied but without using the anti-CTCF antibody, which gives information for estimating the input background. ChIP DNA was used to generate a ChIP-seq library according to the standard Illumina protocol. The library was then sequenced using the Illumina HiSeq platform in 50bp paired end reads. On average ~85.5M reads were produced per sample. Data have been submitted to the European Nucleotide Archive, available with accession number ERP002168. They are also deposited in ArrayExpress with accession number E-ERAD-141. Sequence lanes were assessed for multiple quality metrics including total yield, read quality, mapping quality, GC content distribution and duplication rate. All sequencing reads were aligned to the human reference sequence (GRCh37) using BWA v0.5.9-r16 (Li and Durbin, 2009) using default parameter settings. Duplicate reads were marked by the “MarkDuplicates” function of the software Picard (v1.47 http://picard.sourceforge.net/) and removed. We reason that as the binding interface is much smaller than the fragment size and we used paired-end sequencing, duplicates are more likely to be technical than biological. We applied a stringent filter by removing all the reads with mapping quality score below

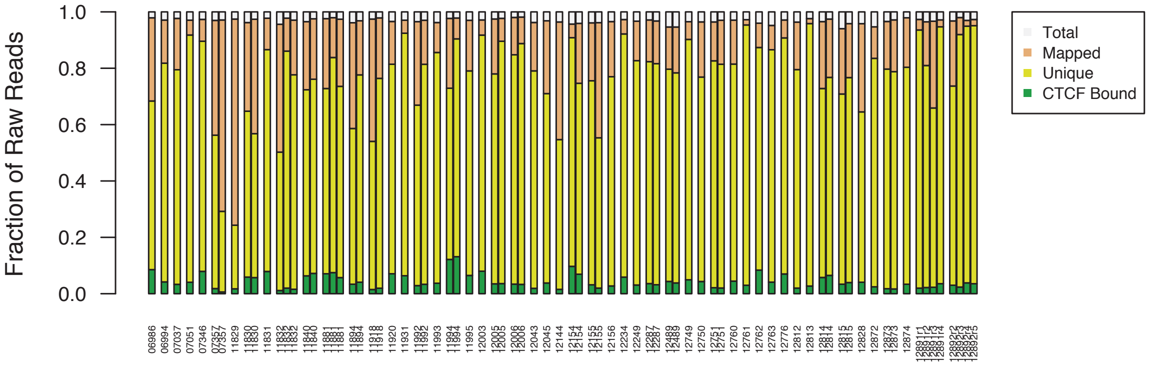

Figure 2.1: ChIP-seq production. The proportions of the mapped fragments, unique fragments, and CTCF bound fragments are plotted for each samples.

30, improperly paired (with 0x2 flag set in the BAM format), or with mate pairs more than 1kb apart (Figure 2.1). For allele specific analysis, we further performed local realignment using a variant-aware aligner glia (https://github.com/ekg/glia), which aligns reads against paths in a variant graph built by combining the reference sequence and known variants.

Binding region calling We performed binding region identification using a Parzen kernel density window algorithm that we applied in previous studies and achieved good performance (Shivaswamy et al., 2008; Lee et al., 2012). This procedure was applied to both experimental and input datasets after combining lanes and replicates into cell-line sample sets. Local maxima of these Parzen scores were used to define binding peak positions, and the interquartile range of the kernel density profile was used to determine the corresponding binding site of highest read density. The resulting set of candidate CTCF binding sites was then subjected to input correction, filtering for copy number artifacts, and determination of statistical significance. A input profile was built using data from a subset of eight samples that went through the exact same production except that the antibody was not used. First, in order to normalize for background represented by the input control, each binding site was paired with the corresponding input site with the highest read count within 200 bp. A binomial P-value was computed for each binding site under the null hypothesis that ChIP and input reads were equally likely. The ratio of total ChIP to input reads for each sample was used to normalize for differences in sequencing depth before calculating the binomial P-value, with the library having higher sequencing depth always scaled downward. Binding sites falling in previously defined genomic regions with aberrantly high signal due to copy number differences were discarded (Boyle, Davis et al. 2008, http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=334775099&c=chrX&g=wgEncodeMapability). Binding sites dominated by input were also discarded, retaining only sites where the ChIP read count scaled by sequencing depth exceeded input.

The resulting set of filtered peak P-values was subjected to multiple hypothesis testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Next, binding regions for the cell lines at various significance levels were merged using bedtools v2.17.0 (Quinlan and Hall 2010) in such a way as to preserve the set of calling cell lines (bedtools merge -nms -scores collapse -n). We employed several metrics in order to determine an appropriate significance cutoff, including the relationship between binding region count and P-value (Figure 2.2) and the number of calling cell lines for each binding region (Figure 2.3). Raw P-values were used to define significant sites once the P-value threshold was determined. Binding regions with BH-adjusted P-value $\leq$ 1E-5 were initially retained as significant (n=127,351), as that value appeared to be the inflection point in the binding region versus P-value curve and had the largest reduction in binding regions called in just one sample.

Finally, in order to assess the quality of binding regions called in only one cell line, we used bedtools (bedtools intersect –c) to identify binding regions containing the extended CTCF motif (Figure 2.4). Binding regions called in only one cell line showed a significantly lower occurrence of the CTCF motif as compared to binding regions called by two or more cell lines. Therefore, we discarded binding regions called in only one cell line and retained the 63,753 merged binding regions at adjusted P-value

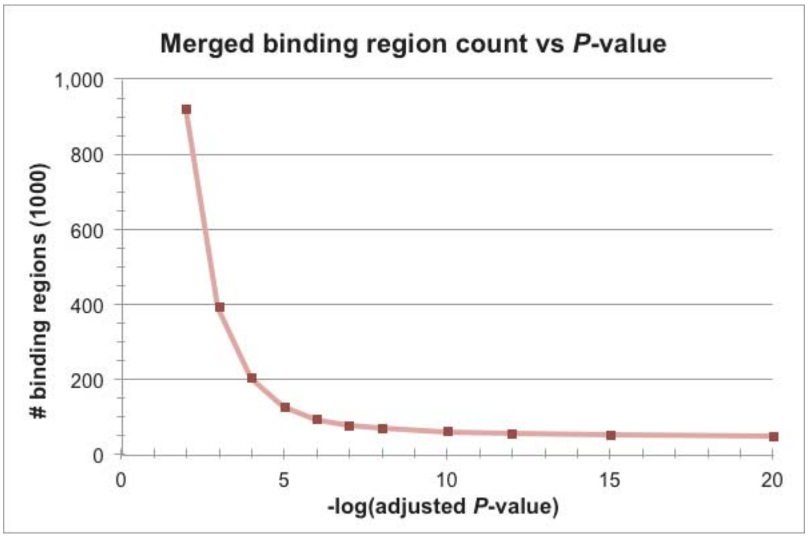

Figure 2.2: Number of merged binding regions plotted as a function of –log(BH-adjusted binomial P-value).

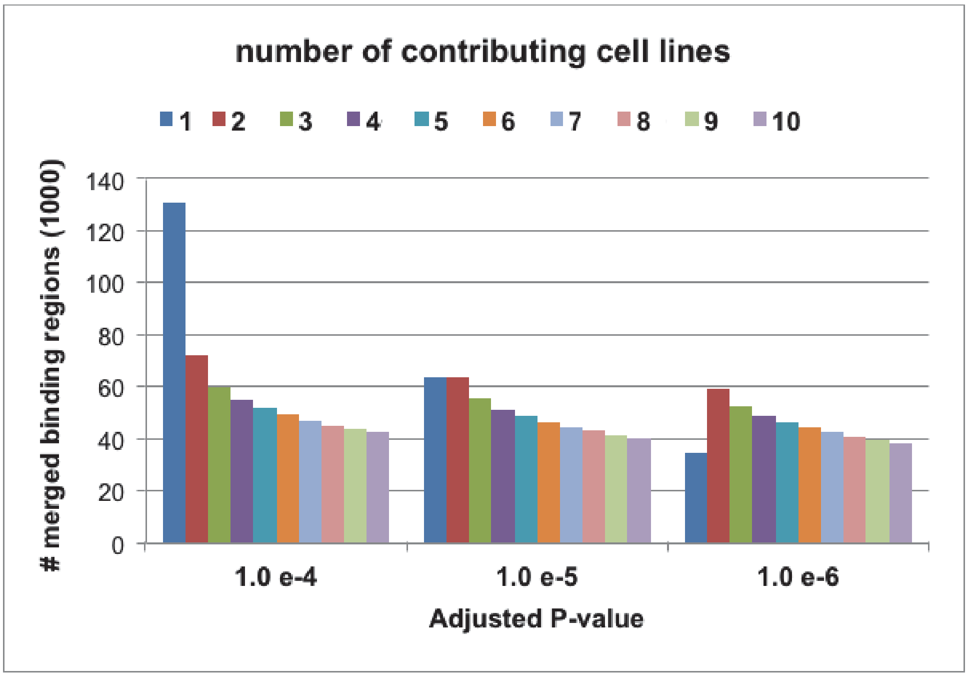

Figure 2.3: Number of merged binding regions as a function of number of calling cell lines, at three adjusted P-values.

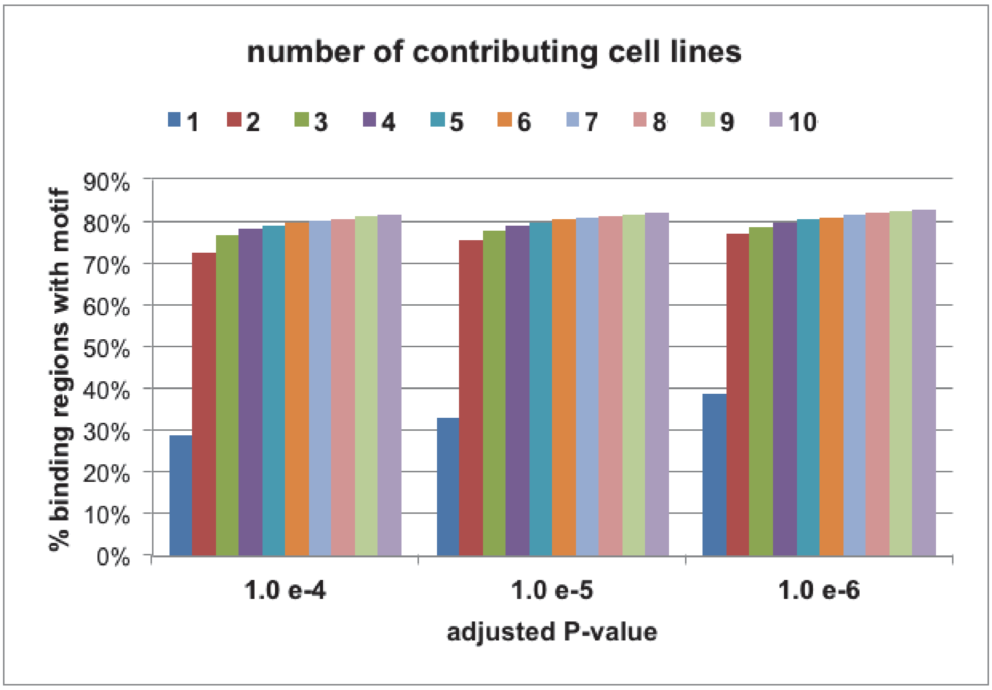

Figure 2.4: Proportion of merged binding regions as a function of number of calling cell lines, at three adjusted P-values.

1E-5 with two or more cell lines.

Blacklisting regions Out of 63,753 binding regions identified, we removed 2,898 binding regions falling in repeat sequences or in the Immunoglobulin heavy chain locus or major histocompatibility complex (MHC). In detail, 2,578 binding regions lie completely within repeat sequences marked by a merged set consisting of "Repeat Masker", "Segmental Dup" or "Simple Repeat" from the table browser of the UCSC Genome Browser, 35 binding regions lie within the Immunoglobulin heavy chain locus (chr14:106053226-106330470) and 285 fall in the MHC region (chr6:28477797-33448354).

Motif word identification We searched for instances of CTCF motif in the discovered binding regions using the CTCF canonical 19bp position weight matrix downloaded from the JASPAR database (Sandelin et al., 2004, http://jaspar.binf.ku.dk/). We extracted DNA sequences at the identified binding regions from human genome reference GRCh37 to construct a sequence database. The search was then performed using the software FIMO(Grant et al., 2011) of the MEME tool suite (Bailey et al., 2009) using parameter "–threshold 1E-4". This process identified at least one motif instance in 45,867 of our 57,428 binding regions. For the ones with multiple motif instances, we selected the motif with highest matching score as the nominal binding motif for the region for some analysis.

2.3 Quantification of CTCF binding

With the peak profile identified above, we quantified the signal for each binding region by counting the number of sequencing fragments (read pairs) when alignment overlapping the region. We applied stringent criteria by only counting the properly aligned read pairs with quality score at least 30 and excluding all the duplicated reads (samtools view -f 0x42 -F0x604 –q 30). We used Bedtools (v2.16.2) (Quinlan and Hall, 2010) to count the intersection between fragments and identified binding regions. This produced an $N \times M$ matrix, where N is the number of samples and M is the number of binding regions. To evaluate the variation in the ChIP experiments, for two samples we collected replicated data on four consecutive days. Using binding sites defined previously, we compared the correlation between replicates grown on consecutive days and the correlation between all other samples. We found a mean pairwise correlation coefficient of 0.83 and 0.82 for the replicate sets for NA12891 and NA12892, respectively, while the mean pairwise correlation coefficient between samples was 0.17. This suggests a good signal to noise ratio in the experiment. This could be considered as covariates in linear model. However, in our data, we do not see much deviation from uniform in the test results from our random control (results shown in 2.11), for simplicity, we do not add additional variables to our tests.

For the subsequent genetic analysis, we are interested in the binding regions that have good signal and also vary between individuals. The mean and variance of binding intensities are correlated by the nature of the Poisson process for the sequencing. We found a group of 4,516 binding regions (7% of the total binding regions identified) with little signal or variation - defined as binding regions mapped with fewer than 6 fragments on average per sample and SD < 5.14. The cut-off was chosen as it delineates clear groups of background intensity and signal intensity with distinct strengths (Figure 2.5). These binding regions were excluded from further analysis.

Normalization Previous studies (Montgomery et al., 2010; Degner et al., 2012) have shown that appropriate normalization can substantially enhance genetic association signals by removing confounding non-genetic sources of variation. Potential sources of confounding variation include experimental batch effects, GC bias in sequencing library construction and latent unknown technical or biological factors that have systematic effects across large numbers of binding regions. To address these issues, we normalized the raw binding intensity using the following five step approach to generate a normalised adjusted binding intensity (NABI).

1. Rescale by sequence depth.

$$X_{i,j} = \frac{R_{i,j} Mean(S_j)}{S_j}, \ i = 1...M, j = 1...P$$

where $R_{i,j}$ is the raw intensity of the ith binding region of the jth lane, and S_j is the sum of intensity across all binding regions for the jth lane. $R_{i,j}$ is scaled by a factor of the proportion of mean of S across all P lanes over S_j.

2. Remove variance introduced by GC composition. We adjusted for GC bias in sequencing library construction by forming percentile bins for GC composition of all binding regions and normalising the binding intensities within each bin. Since the fragment length is much larger than the motif length, this bias is not strongly influenced by the motif sequence.

$$X_{i,j} = \frac{X_{i,j}}{Median(X_{k,j}; k \text{ same GC bin as } i)}$$

where i, j, k are the indices for binding region, lane, and GC bin respectively.

3. Merge lanes of a same individual by taking the mean. A subset of our samples

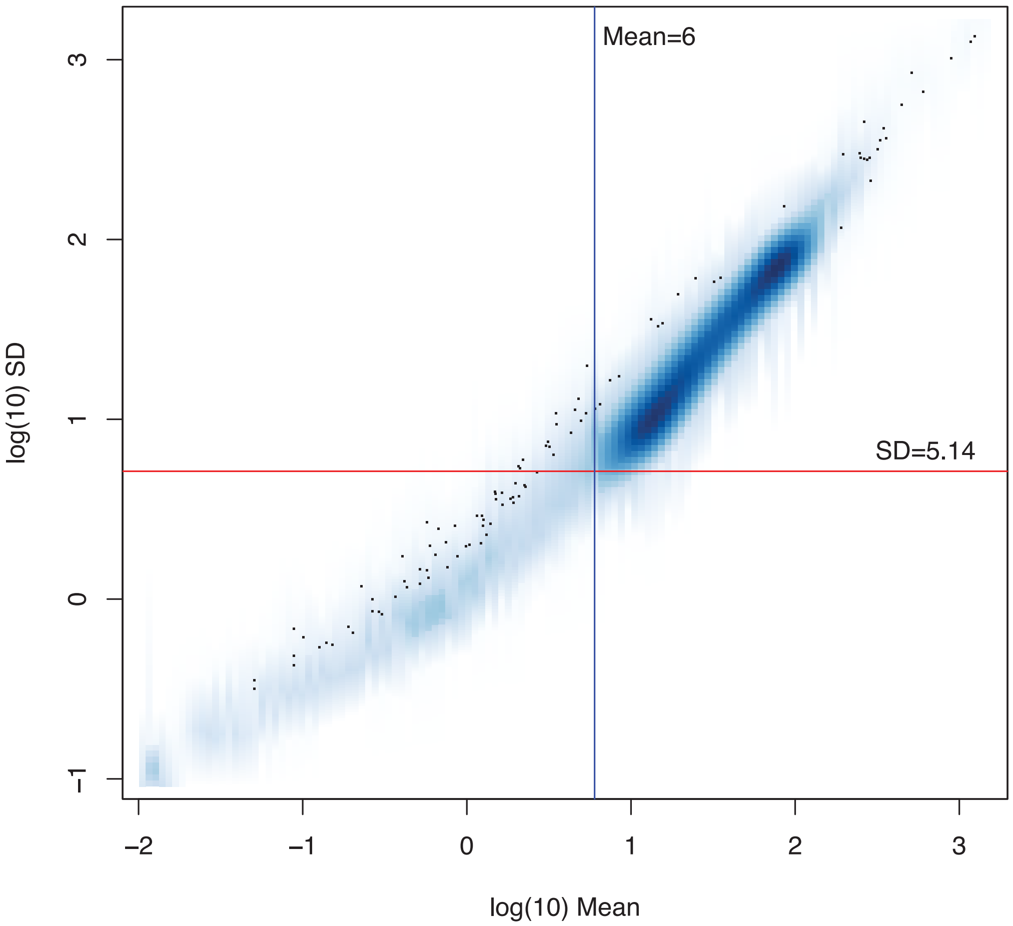

Figure 2.5: Quality control by raw signal intensity and inter cell line variability. For each binding region we counted the overlapping sequencing fragments (identified by a properly paired read pair) and used it as a measure for the raw binding intensity. We plot the log of the variance of the binding intensities across 51 individuals versus the log of the mean of the binding intensities using the R function *smoothScatter*. The degree of blue is proportional to the density of data points. As a Poisson process the mean and variance correlate to each other. There exists a natural cutoff between the lower left tail and the majority at mean 6 and standard deviation 5.14. These lower left tail binding regions are the sites with very low intensity and also low variability. We removed these sites, 4,516 binding regions in total, before further analysis.

were sequenced on multiple lanes and in these cases we took the mean value across lanes as the measurement of the individual.

$$D_{i,l} = Mean(X_{i,j}; j \text{ lanes of } l),\ l = 1...N$$

where $X_{i,j}$ is the measure from the previous step, i, j, l are indices for the binding region, lane and samples, respectively. N is the total number of samples.

4. Centre-scale binding intensity for each binding region. We then scaled the binding intensity for each binding region by subtracting the mean and then dividing by the standard deviation. This transforms the measures of each binding region into zero mean and unit variance, which is needed for the quantile normalization to be less affected by the different variances of different binding regions

$$Z_{i,l} = \frac{D_{i,l} - Mean(D_i)}{StDev(D_i)}$$

where i, l are indices for binding region and sample.

5. Quantile normalize each sample data to a normal distribution. The distribution of binding intensities for each individual is complex. Previous studies have shown that quantile normalization, initially developed for normalising the microarray signals of gene expression, can assist statistical analysis by converting the distributions of each sample to a reference distribution. The linear regression model used to identify QTL in our study assumes a Gaussian distribution of binding measures within each genotype class. We therefore mapped the measures across all binding regions of each sample to the corresponding normal quantiles. This produces a matrix that is essentially a perturbation permutation of the normal quantiles

$$\tilde{Z}_{i,l} = \Phi^{-1}(\frac{\sum_{m=1}^{M} I\{Z_{m,l} < Z_{i,l}\}}{M+1})$$

where Φ is the cumulative normal density function and M is the total number of binding regions. I is an indicator function that returns 1 if the condition is met and 0 otherwise.

6. Remove confounding variation by principal component analysis (PCA). The

Variables	PC1	PC2
Sequencing mapping rate	0.049	0.50
Duplication rate	0.065	0.031
Sequencing depth	0.043	0.012
ChIP batch	0.31	0.097
ChIP batch with sequencing batch regressed out	0.47	0.075
Epstein–Barr virus load	0.12	0.022

Table 2.1: Correlations between PC1, PC2 and the experimental variables. In association tests PC1 was removed.

measures of binding for each individual can be confounded by a number of hidden factors due to either biological or technical factors, or both. We performed PCA and saw that the first factor explained 24.1% of the variance in the data, substantially more than later components (Figure 2.6). Further investigation of this component showed that it was correlated with ChIP batch date, and it was therefore removed (Table 2.1).

2.4 Imputing missing genotypes

Our 51 samples consist of 35 individuals present in the 1000 Genomes Phase 1 release (v3 20101123) (The 1000 Genomes Consortium, 2012), 11 individuals in the 1000 Genomes Pilot, 2 individuals in 1000 Genomes high coverage Trio (NA12891 and NA12892) and 3 individuals in the HapMap III (Stranger et al., 2012). The eleven 1000 Genomes Pilot samples have low coverage. We calculated the genotype likelihood for each of the Phase 1 sites using samtools (Li et al., 2009) and then performed imputation using BEAGLE (Browning and Yu, 2009) and IMPUTE2 (Howie et al., 2009) with the 1000 Genomes Phase 1 data as a reference panel. Using Illumina Omni 2.5M SNP array genotypes (available ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/) as a validation set, we obtained good accuracy from this procedure with a mean non-reference discordance rate of 2.33% and an average genotype dosage R^2 of 0.956. We also imputed the three HapMap III samples, using their genotype data on the Omni 2.5M array as

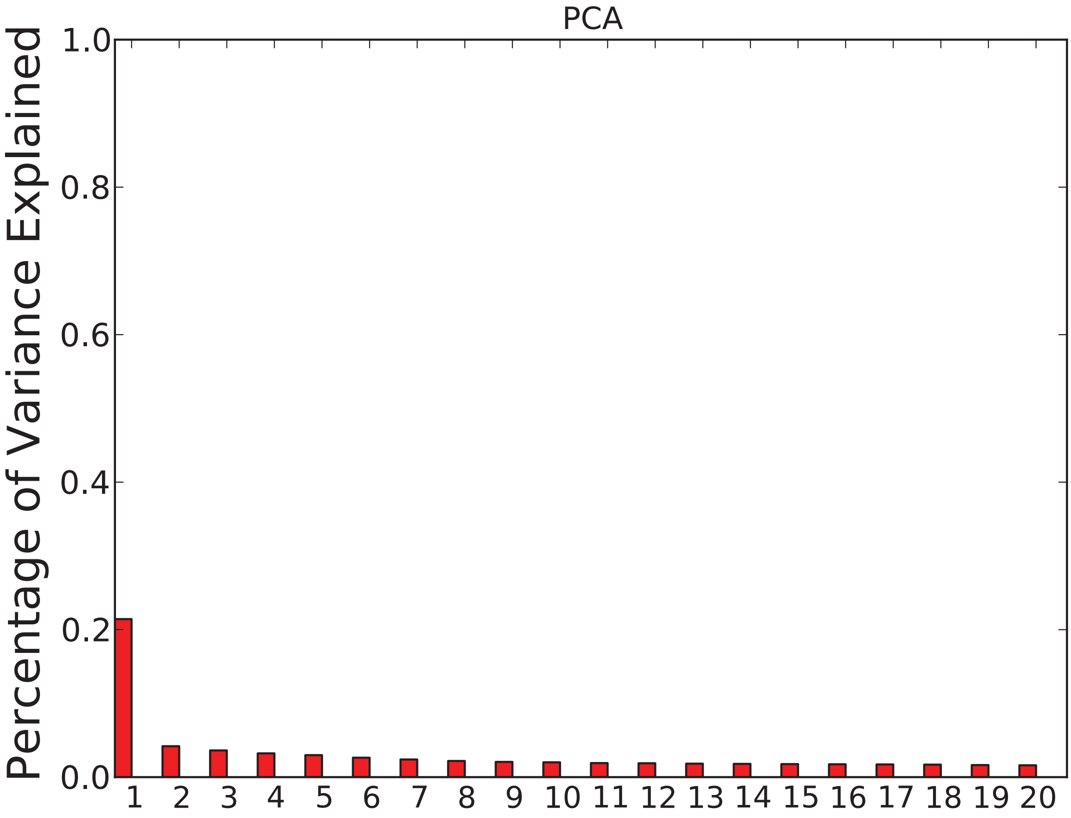

Figure 2.6: Proportion of phenotypic variance explained by each principal component (PC). We performed principal component analysis (PCA) on the normalized data to discover latent factors that explain large proportion of phenotypic variation. We saw that the first principal component explains substantially more variance than the others. When we looked at the correlation between the first principal component and technical and experimental variables, we found that it correlates with ChIP batch at ρ=0.47. The first principal component is removed from the data before further analysis.

the imputation panel and the 1000 Genome Phase 1 as the reference panel. We then integrated data from each source and obtained a consolidated genotype set for all 51 individuals. For association mapping, we filtered variants by requiring >5% minor allele frequency, P value for Hardy-Weinberg Equilibrium (HWE) >1E-4 and position within 50kb to either side of a binding region being mapped. The window size was chosen to be 100kb as we are primarily interested in *cis* regulation but also allowing possibility that there may exist multiple binding sites with variable affinity strengths in the window. Finally, 4,687,317 variants entered analysis, with 4,250,881 SNPs and 436,436 INDELs.

The 1000 Genome Phase 1 release gives a comprehensive ascertainment of the genetic variants. However, it is still possible that some variants private to this study cohort are yet to be found. To address this concern, we performed variant calling for the CTCF binding regions using ChIP-seq data. The calling was done by using samtools mpileup with parameters "-DV -C50 -q 30 -Q 30 -d 10000 –u -l $qtl_regions -b $bam_list -f $reference", followed by BCF tools with parameter "-t $qtl_regions -mv". This is independent from the previous variant calling and gives information private to the ChIP-seq data. We filtered on the quality of the calling by keeping only variants with QUAL score greater than 20. We also kept only the variants that are private to the new call set and are absent in the 1000 Genomes Phase 1 data. In the end, we obtained 4,756 variants are within binding regions with 2,282 SNPs and 2,474 INDELs. It is a small additional quantity compared to the variant set of the 1000 Genome Phase1 release, but is enriched for INDELs (52%). When we conducted the same association scan using only these additional variants, we discovered 55 QTL binding regions associated with 60 variants, out of which only 8 QTL binding regions are new and no variants were found within motif. Thus the effect of this additional variant set is minimum in our QTL scan.

2.5 Association testing

We applied linear regression for association testing. For each binding region, we tested the association between the binding intensities and the genotypes of the variants that are within 50kb of the binding region by linear regression: $y_{il} = \beta_k x_{lk} + \epsilon_{ilk}$, $i \in \{1, ..., 57428\}$ where y_{il} is the normalized binding intensity for the lth individual, x_{lk} is the genetic dosage, represented as the minor allele count, for variant k and individual l, and ϵ is the non-genetic noise term assumed to follow distribution $N(0, \sigma^2)$. The parameters ($\hat{\beta}$,σ^2) can be fitted using maximum likelihood methods. For each loci k, we tested the null hypothesis $\beta = 0$ using test statistics $t = \frac{\hat{\beta}}{\sqrt{var(\hat{\beta})}}$.

We estimated the FDR by a q value method (Storey and Tibshirani, 2003), which establishes P<7.1E-5 as an FDR of 1%. We further filtered the associated SNPs by requiring the P value to be within one order of magnitude to that of the P value of the lead SNP. We report these cluster variants as associated to the target binding region. We also reported results when a more stringent Bonferroni threshold was applied. The threshold was calculated at a significant level of α=0.05 corrected for 13,293,727 tests, which gives 3.8E-9 for the actual threshold.

2.6 Allele specific analysis

Read counts at each allele were counted for the 5.6M SNPs within 50kb of a binding region. Heterozygous SNPs with significant allele-specific CTCF binding were identified. In detail, for each individual at each site, we calculated a binomial P value at all heterozygous SNPs with the null hypothesis that the two allele counts are equal. We then performed multiple testing adjustment for all heterozygous SNPs that have at least 2 reads at each allele and at least 2 reads difference between the two alleles using the Benjamini&Hochberg (Benjamini and Hochberg, 1995) method. Significant allele-specific binding was determined with an FDR 5%.

2.7 Results

Analysis of CTCF binding in 51 genotyped individuals reveals thousands of binding QTLs We performed ChIP-seq on extracted chromatin from genotyped LCLs as previously described (Lee et al., 2012) except that we sequenced the DNA fragments from both ends (Figure 2.7). We quantified binding to binding regions similarly to previous work (Lee et al., 2012) but pooled all the samples and identified a composite set of binding regions with detectable CTCF binding at low threshold. We then counted the sequence fragments that overlap each binding region in each individual, and normalized the signal to correct for systematic biases as in Degner et al., 2012. We discarded binding regions that showed very little inter-individual variance or had only one or two individuals with significant binding scores. Overall, our normalized data showed effectively enhanced signal noise ratio and motivated QTL analysis (Figure 2.7B, 2.9, and 2.11).

To measure the variance due to growth differences between the cells, we grew two individual cell lines as four independent cultures started on four consecutive days. There was higher correlation between these biological replicates from the same individual than between samples from different individuals, although all data sets were modestly correlated as expected for CTCF ChIP-seq (Figure 2.8). We next examined the data to see whether there were any systematic biases between samples. A principal component analysis identified some systematic variance, with a particularly strong first component (explained 24.1% of the variance, Figure 2.6) that on investigation was correlated to known experimental batches. We therefore removed the first principal component, significantly improving the recovery of QTLs (Figure 2.9). This is in general a good practice from previous studies, e.g. Degner et al., 2012, as methods such as PCA could not enhance random noise. We used the resulting normalized adjusted binding intensity (NABI) for subsequent analyses.

To discover QTLs, we correlated SNPs and small biallelic insertion or deletion (INDEL) variants within 50 kb of the binding region with the NABI metric, using a linear model (Table 2.2, example in Figure 2.10). As expected, the majority of variants do not have a significant association with variation in CTCF binding, with the linear model P-value distribution following the expected distribution (>95% of tests, fraction of the overlap between the black line and red line, Figure 2.11). When samples are permuted, the distribution of the test statistic falls on the expected line. Using a nonparametric statistic we saw similar P values (Figure 2.11). Using a Bonferroni adjusted threshold of $P < 3.8E\text{-}9$ we find 509 binding regions with significant QTLs. Using a

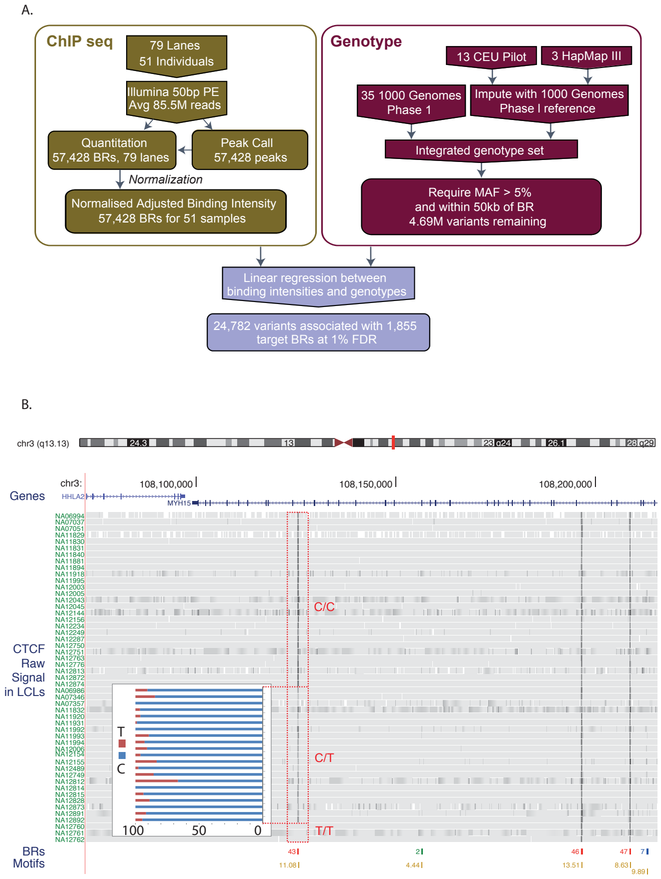

Figure 2.7: A. Flow chart indicating the overall design of the experiment. B. Overview of the binding intensities of a binding site across samples in three genotype groups of the associated SNP. ChIP-seq signal from the samples is aligned as tracks for this region of chromosome 3. The greyness is proportional to fragments mapped at the position, indicating binding intensity, with dark grey indicating high fragment count. Samples are grouped by their genotype at SNP rs936266, C/C, C/T or T/T, respectively. Binding sites were identified, as shown in the binding region track along with the number of samples passing the peak calling threshold. The colours of the binding regions represent the consistency of identifying the binding region across samples. Specifically, red binding regions were identified in 10 or more cell lines, blue binding regions in 5-9 cell lines and green binding regions in 2-4 cell lines. Finally the bottom track shows the corresponding CTCF motifs, with quality score attached to each site. The binding intensity decreases for T heterozygotes and further for T homozygotes. The inset panel shows allele-specific binding for the C and T allele (blue and red, respectively) in the heterozygous individuals (C/T) as percentage of the total count. Binding intensities consistently favour the C allele over the T allele.

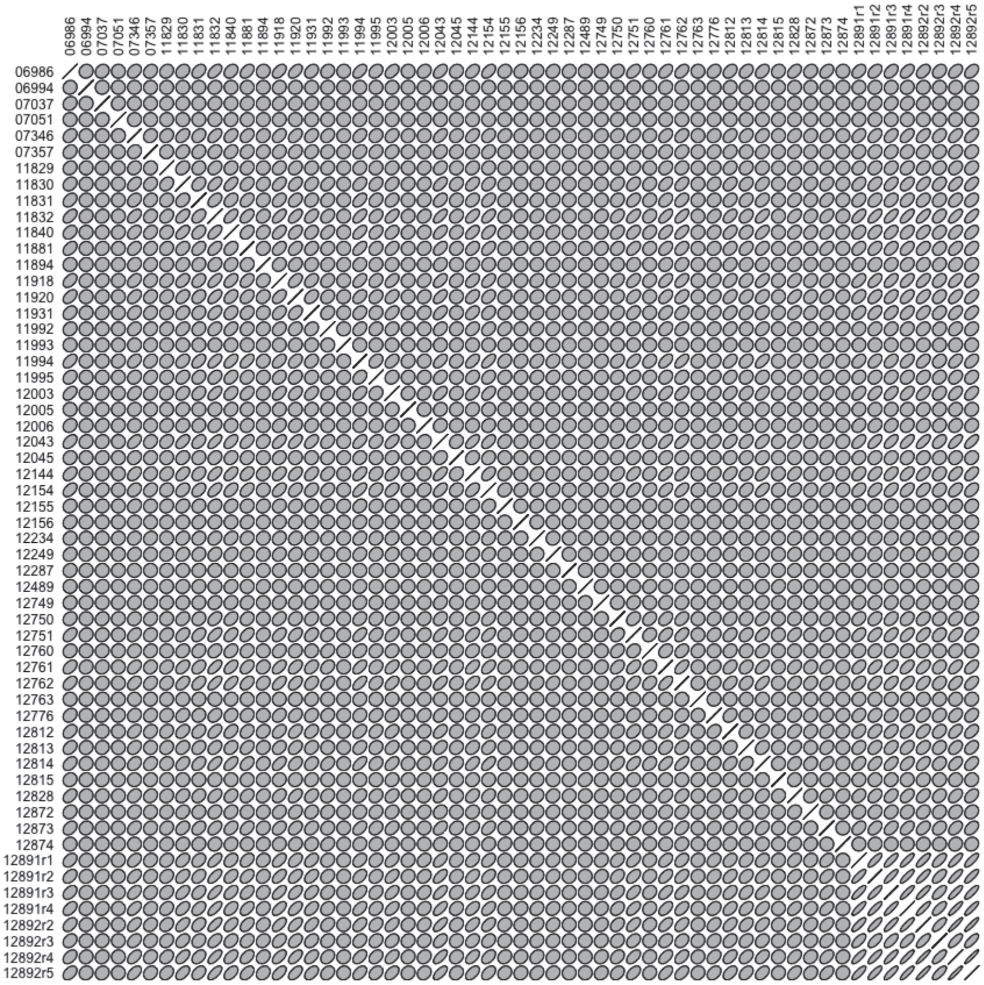

Figure 2.8: Higher correlation within day replicates compared to between different samples. We calculate the pair-wise Spearman correlation among all samples, including the two day-replicates, 12891 and 12892, shown as the last two sets of four samples. A diagonal line in each cell represents perfect correlation whereas a full circle represents no correlation. Increasingly flattened ellipses indicate a greater degree of correlation. When comparing among the day replicates, we obtained a correlation coefficient of 0.83 and 0.82 for GM12891 and GM12892, respectively. We also looked at the mean correlation of all the other samples and found a correlation of 0.17. Therefore we see much higher correlation within day replicates than that of all other samples.

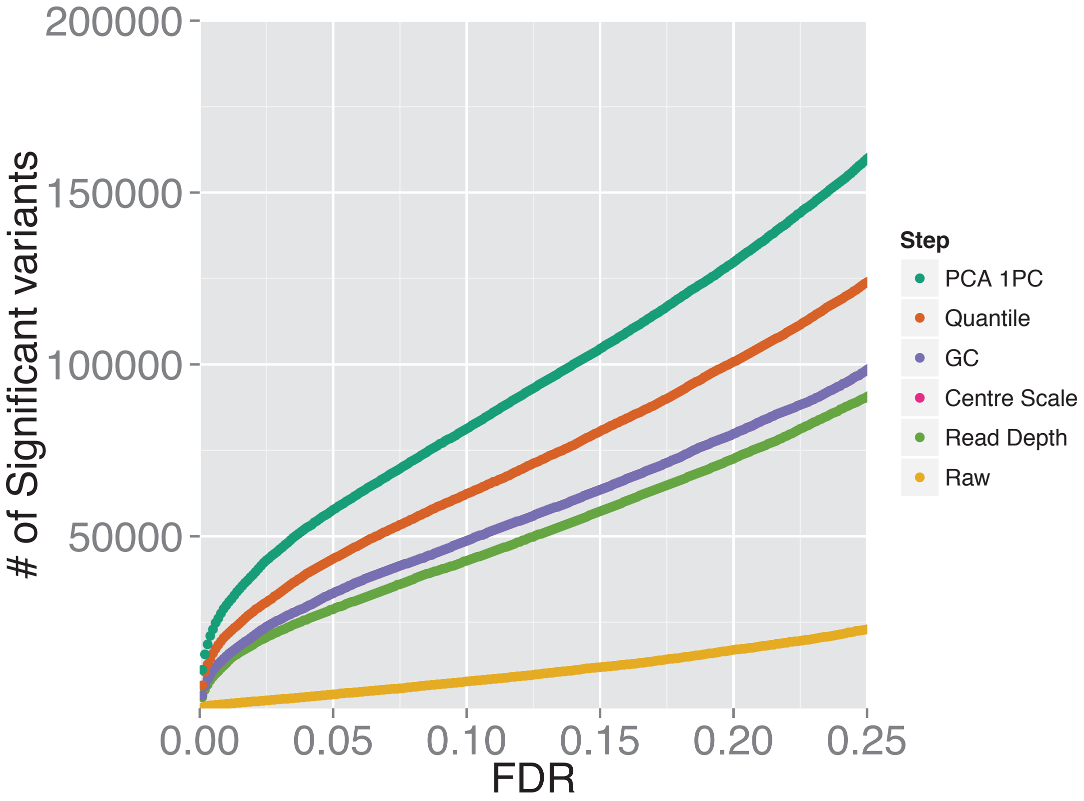

Figure 2.9: The number of significant QTLs found as a function of false discovery rate (FDR), plotted for the raw data and after each stage of the data normalization procedure that we used. We first normalised the binding intensities for each sample by the total read depth for that sample. We then corrected for GC composition by removing the median count of binding regions in the same GC bin (100 bins in total) from each binding region. The measures for each binding region were then centre-scaled by removing the mean and then dividing by the standard deviation (track hidden behind GC as center scale does not affect regression). This was followed by a quantile normalization, which maps the measures of each sample to normal quantiles across all binding regions. Lastly, we removed the first principal component that explains the most global phenotypic variation.

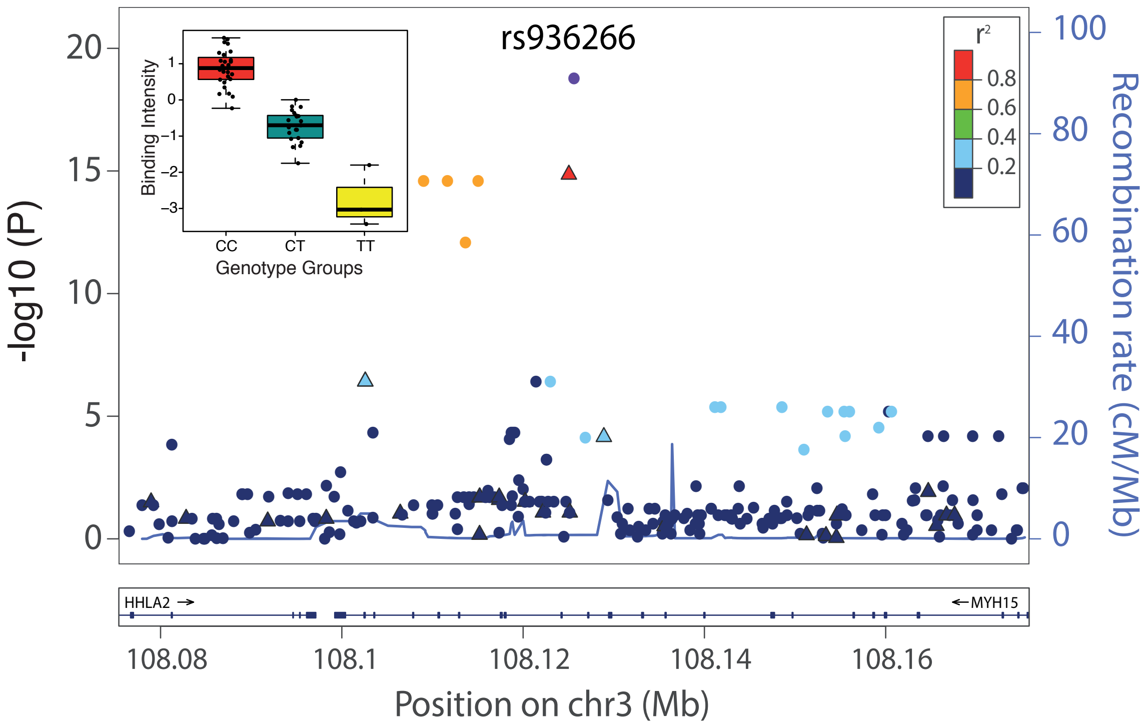

Figure 2.10: An example CTCF QTL. Here shows all associations for all variants in the region of the binding region at chr3:108125397-108125829. SNPs are shown as solid circles and INDELs are shown as triangles, colored by R^2. Inset is boxplot showing the normalized adjusted binding intensity (NABI) for the different possible genotypes of SNP rs936266. Genotype is strongly associated with the binding intensity of the binding region (P=1.69E-19), with the C allele favoring binding.

more liberal False Discovery Rate (FDR) (Storey and Tibshirani, 2003) approach to take advantage of the smaller number of effectively independent tests occurring in these limited *cis*-regions, we discovered 1,837 binding regions (3% of total binding regions) with at least one significant variant at the 1% FDR level; relaxing the threshold to 10% FDR we discover 6,747 binding regions (12% of the total) (Table 2.2).

We chose to focus further analysis on the 1% FDR threshold as this provided ample QTLs from which to derive insights. We only considered one association per binding region, because the small number of samples meant that there was insufficient power

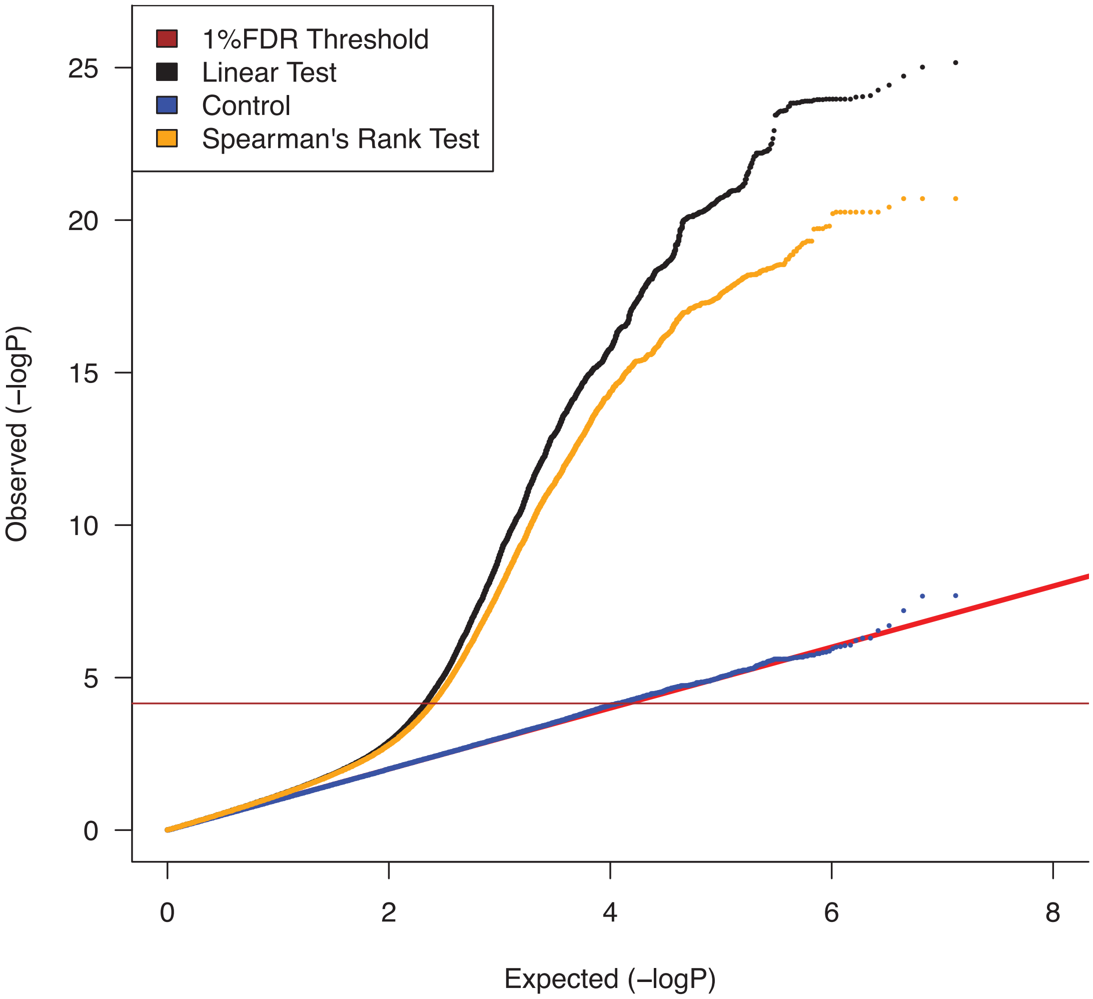

Figure 2.11: A Quantile-Quantile plot showing the distribution of the observed (y-axis) compared to the expected P values(x-axis). The red line is the distribution of the P values from the null model. The brown line on the y-axis shows the 1% FDR level determined by the q value method (Storey and Tibshirani, 2003). Black and blue dots indicate P values from the linear tests and permutation controls, where sample labels are randomly permuted. Association test by linear methods can be inappropriate and give spurious signal if the normality assumption is not met. Although in our normalization procedure the binding measures are mapped to normal quantiles sample-wise, it is still possible that the normality assumption does not hold binding region-wise. To test if this would bias the QTL mapping we performed the same tests using the Spearman's rank method (orange line). We see a slight elevation of the black line, suggesting the rank test is more conservative but would give similar results (1476 out of 1837 QTL binding regions overlap between two tests), and our linear test is mostly appropriate. The subsequent analysis is based on the discovery set from the linear test at a 1% FDR (brown line) threshold.

Study Parameters	
Traits (Binding Regions)	57,428
Variants	4,687,317
SNPs	4,250,881
INDELs	436,436
Study Results	
Binding Regions	1,837
Variants	24,534
SNPs	22,954
INDELs	1,580
GWAS overlaps	61
eQTL overlaps	366

Table 2.2: Summary statistics of the CTCF QTL scan.

for a conditional analysis for secondary associations in almost all cases. Within this set of associations, the genetic variant accounted for a substantial fraction of the variation in CTCF binding (median R^2 0.38, Figure 2.12). When comparing the effect sizes and the proportion of variance explained between QTL at 1% FDR and 10%FDR, the 1% FDR set has higher values (the average absolute value of beta = 1.1 (0.37-3.39)) than the 10% FDR set (average absolute value of beta = 0.80 (0.26-2.83), Figure 2.13).

We summarized the collective set of variants which might be involved in each binding region association as being the cluster of SNPs within one order of magnitude of the P-value of the lead variant. 24,534 variants were identified in at least one cluster at the 1% FDR level, 13.4 variants on average per binding region (Table 2.2). As expected, these variants were mainly clustered around the target binding region, and when a CTCF binding motif could be identified (1341 of the 1837 cases) and a cluster QTL variant was present in the motif, the frequency was correlated with the information content and the GERP score (Cooper et al., 2005) of the motif (Figure 2.14B), as seen previously (Maurano et al., 2012). This is not driven by any biases

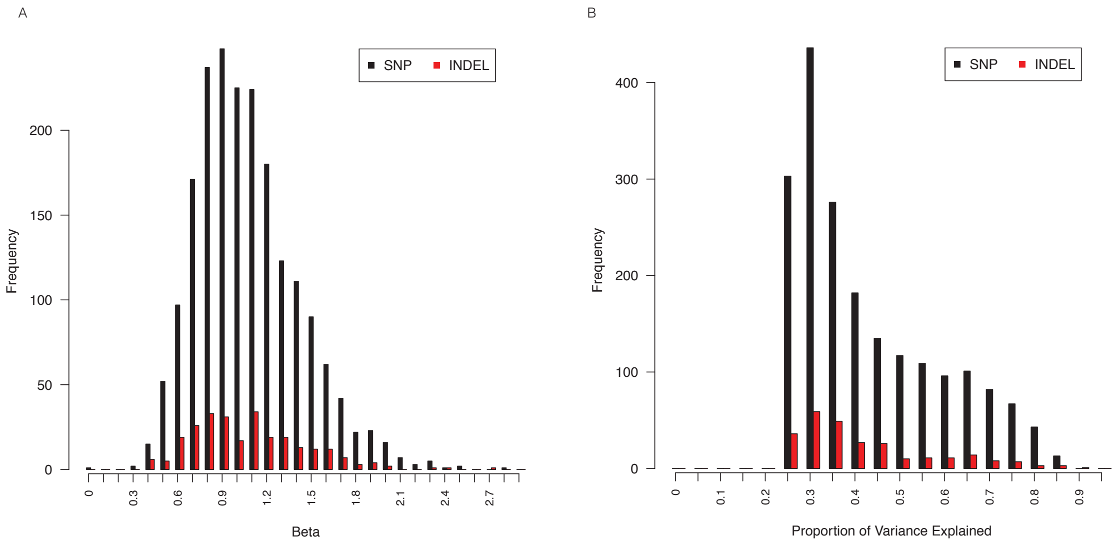

Figure 2.12: The distributions of the effect size (β) and proportion of phenotypic variance explained of the QTL variants.

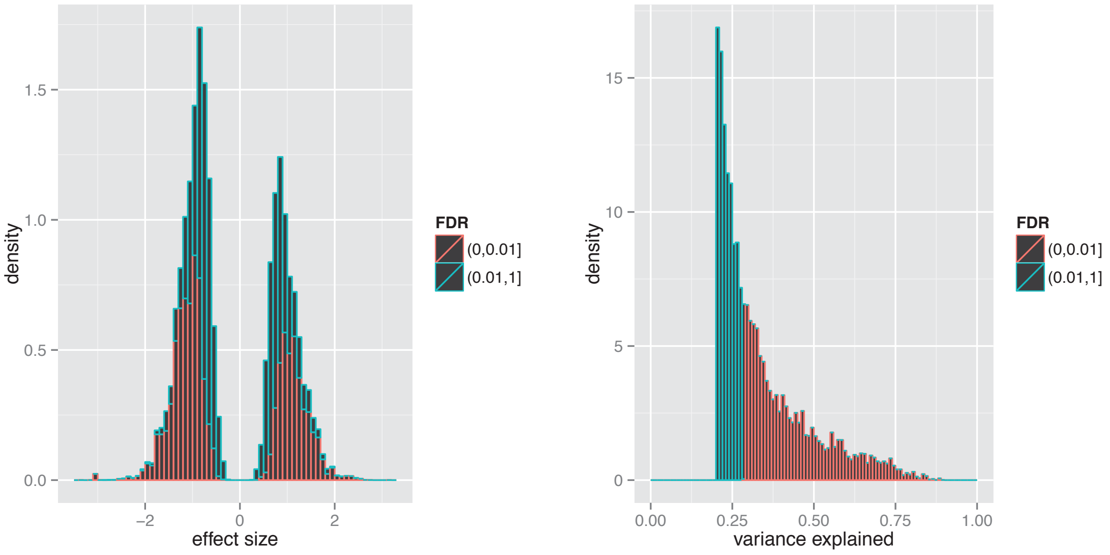

Figure 2.13: Effect sizes and proportion of variance explained of QTLs discovered at 1%FDR and 10%FDR.

Significance	Binding region count	Binding regions with motif	QTL in motif	≤1kb	≤10kb	≤30kb
10% FDR	6,747	5,260	453	1,386	2,583	4,057
1% FDR	1,837	1,341	344	747	1,023	1,199
Bonferroni	509	360	164	258	322	341

Table 2.3: CTCF QTLs with associated variants in different distance ranges.

in the distribution of the variants around the CTCF binding interface (Figure 2.14). However, only a minority of significant binding regions had a QTL candidate within the motif (344/1341), and in only a small majority of cases was there a QTL within 1kb (747/1341), of the binding region (Table 2.3). Considering that out of 45,668 binding regions that contain at least one motif, 2,090 (4.6%) binding regions have at least one variants on its binding motif. The QTL set shows strong enrichment of functional variants that are on motif: out of 1,341 that are motif containing, 344 (25.6%) have QTL variants within the motif.

We explored further the cases where there was no proximal variant in the cluster. There was not a substantial difference in genotype quality around the associated binding regions in these cases compared to binding regions with proximal effects, suggesting that there is not a large missing data problem. When considering all 1000 Genomes Project variants including those with allele frequency below 5%, in 95.5% of these cases, there was a proximal variant within 1kb of the binding region in linkage disequilibrium (LD) with the distal lead variant, where LD was defined as the absolute value of D' $>$ 0.5. In approximately half of these cases the P-value of the proximal association either fell just outside the one order of magnitude threshold to fall in the cluster, or was just under the FDR threshold (Figure 2.16). In the 99 such cases where such a proximal variant was within the CTCF binding motif, the position of the variant was correlated with the information content of the position in the motif (Figure 2.17). Therefore a substantial fraction of the apparently distal cases appear to be explained by proximal cases. However still only a minority can be explained by

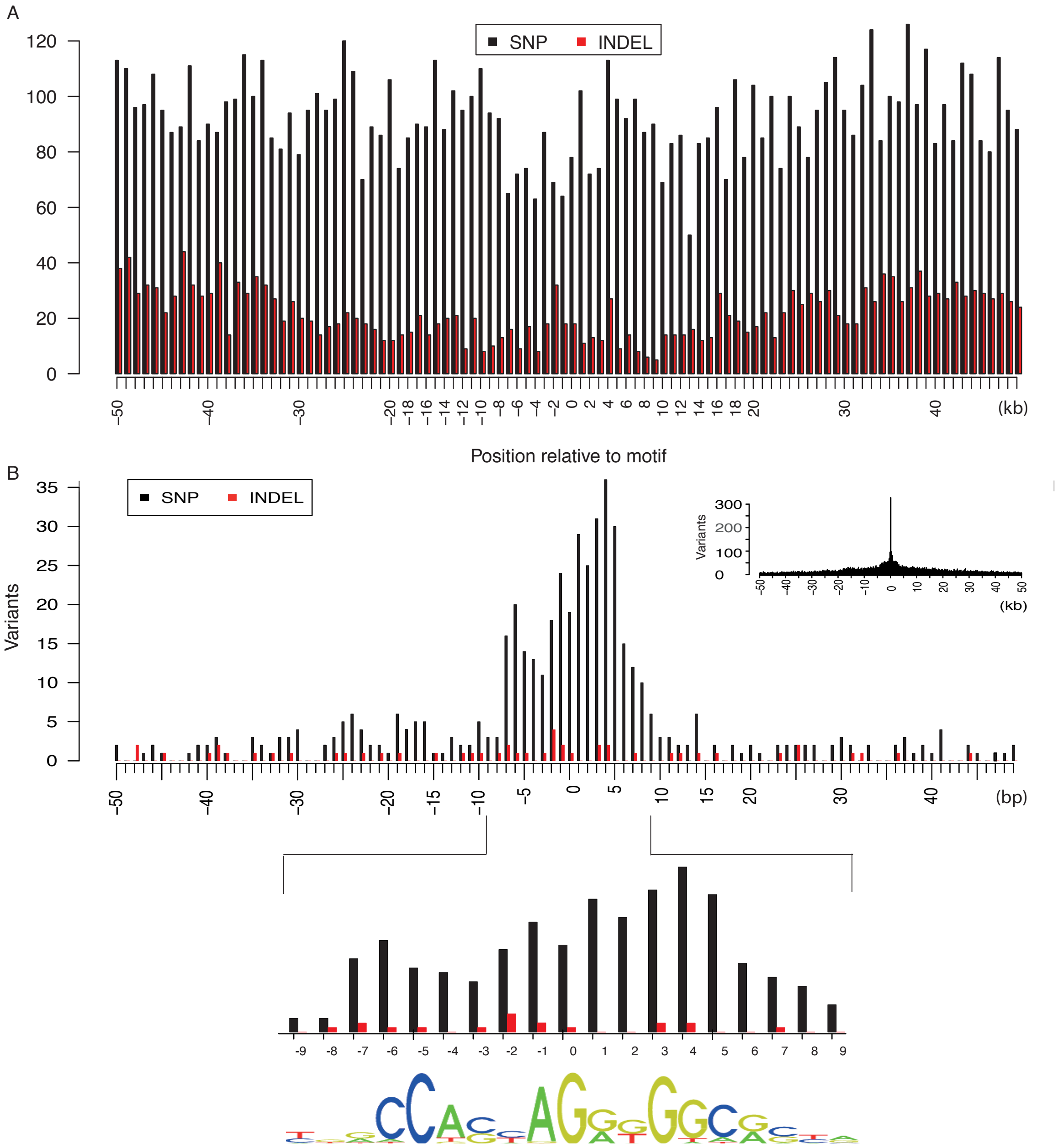

Figure 2.14: A. The distribution of variants (SNPs and INDELs) within a 50kb window over all binding regions that contain a motif. Y axis indicates the number of variants at a given position indicated by X axis with respect to the binding motif. Variants are uniformly distributed throughout the window, except a small reduction at the center corresponding to the high information content of the motif. B. The density of QTL variants with respect to distance from the motif of the associated binding regions. Density plots are shown at kb (inset) and base pair resolution (main plot). SNP and INDEL are shown as black and red bars respectively. For these cases the QTL density correlates with the information content of the motif (Spearman's rank ρ=0.63) shown at the bottom.

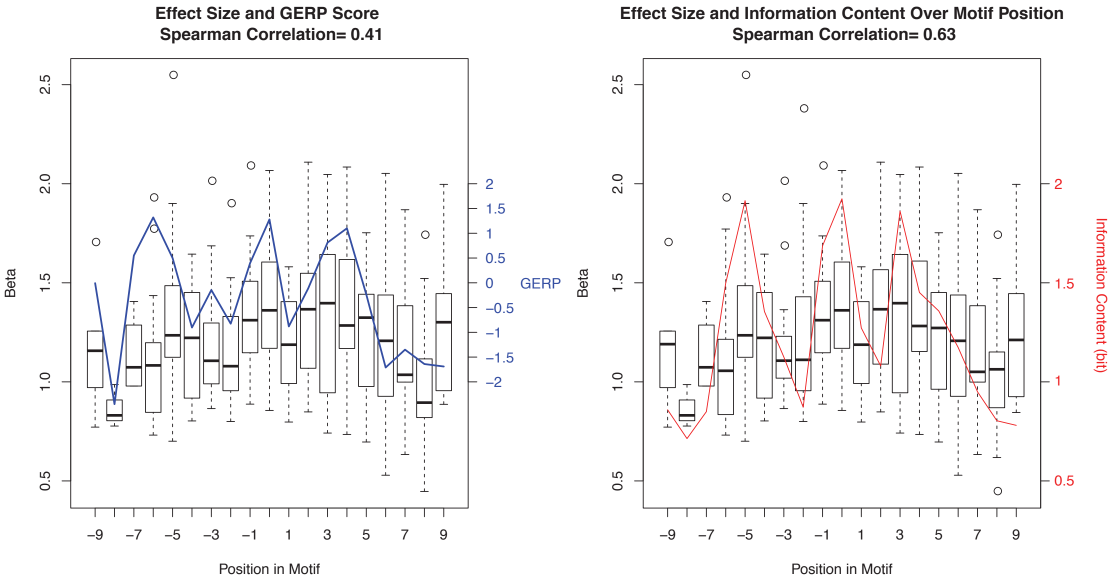

Figure 2.15: The QTL effect size correlates with the information content and the GERP score for the variants present in the motif .

variants in the binding motif itself. We looked at overlapping between the associated variants and other markers including DNase I hypersensitivity sites and transcription factor binding sites. We did not see a clear enrichment towards a particular motif. We also conducted the analysis excluding short INDELs to replicate the more commonplace association analysis using only SNPs. In an INDEL-free analysis we would have missed QTLs in 67 binding regions entirely (~5% of significant binding regions), and for 56 additional binding regions the closest observed explanatory SNP would have been over 1 kb away from the motif inside the peak. For these SNPs, there is usually a short indel with similar direct P-value inside the binding region. We further explored whether another cause for distal QTL effects could be due to the distal variant affecting a second neighbouring binding region of CTCF, which in turn influenced the primary binding region, but there was only one case where we could find any evidence for this model (Figure 2.18).

We additionally investigated the cases where there exist binding interactions between the QTL binding region and the neighboring region. For each of the four categories with sufficient abundance (model 1, 2, 3 and 4 in Figure 2.18), we compared the average signals between the QTL binding region (B1) and the neighboring binding region (B2) for a number of molecular markers using data obtained from the ENCODE project (**?**). We observed distinct patterns of regulatory signals between model 1,2 and model 3,4 (Figure 2.19). We saw that when there exists interactions between two binding regions (model 3,4), active transcription factors, enhancers and active histone markers tend to be more enriched in the QTL binding regions, as shown in red in Figure 2.19. This change is not driven by their distances being closer to the transcription start site (TSS) by chance, measured as the distance to the closest TSS, because the neighboring binding regions have similar distance to the TSS as the QTL binding regions (red and green lines in the density plots, Figure 2.19). We observed corresponding changes in histone modifications (H2AZ, H3k27ac, H3k4me1, H3k4me2 and H3k4me3) depending on the direction of the interactions between two binding regions (Figures 2.19 and Figure 2.20).

The effect size distribution with respect to allele frequency shows increased effect sizes for lower frequency SNPs, with a clear absence of large effects of common alleles (Figure 2.21). There is no statistical difference in effect size distribution between SNP and indel variants (Figure 2.21).

The dual-end sequencing of the ChIP-seq fragments provides the resolution to discover specific binding modes that influence the spatial distribution of the recovered fragments. To analyse this, we characterised ChIP-seq binding regions by metrics that summarised the extent of the peak and the position of the summit on a per individual basis, and used these additional metrics as phenotypes in a quantitative trait analysis using the methods described above. In detail, for each binding region in each sample, we measure the average left end, middle point and right end of sequencing fragments. The variation of the average positions across samples thus reflect variations in binding shapes across samples. Out of all 57,428 binding regions that were tested, we found 25 shifts in peak shape driven by a genetic locus at the 1% FDR. Ten cases were also associated with a change in peak height. An example is shown in Figure 2.22, with the two homozygous genotypes showing the creation of a new associated peak, and merging of a double peak, and from visual inspection the other cases also look as if

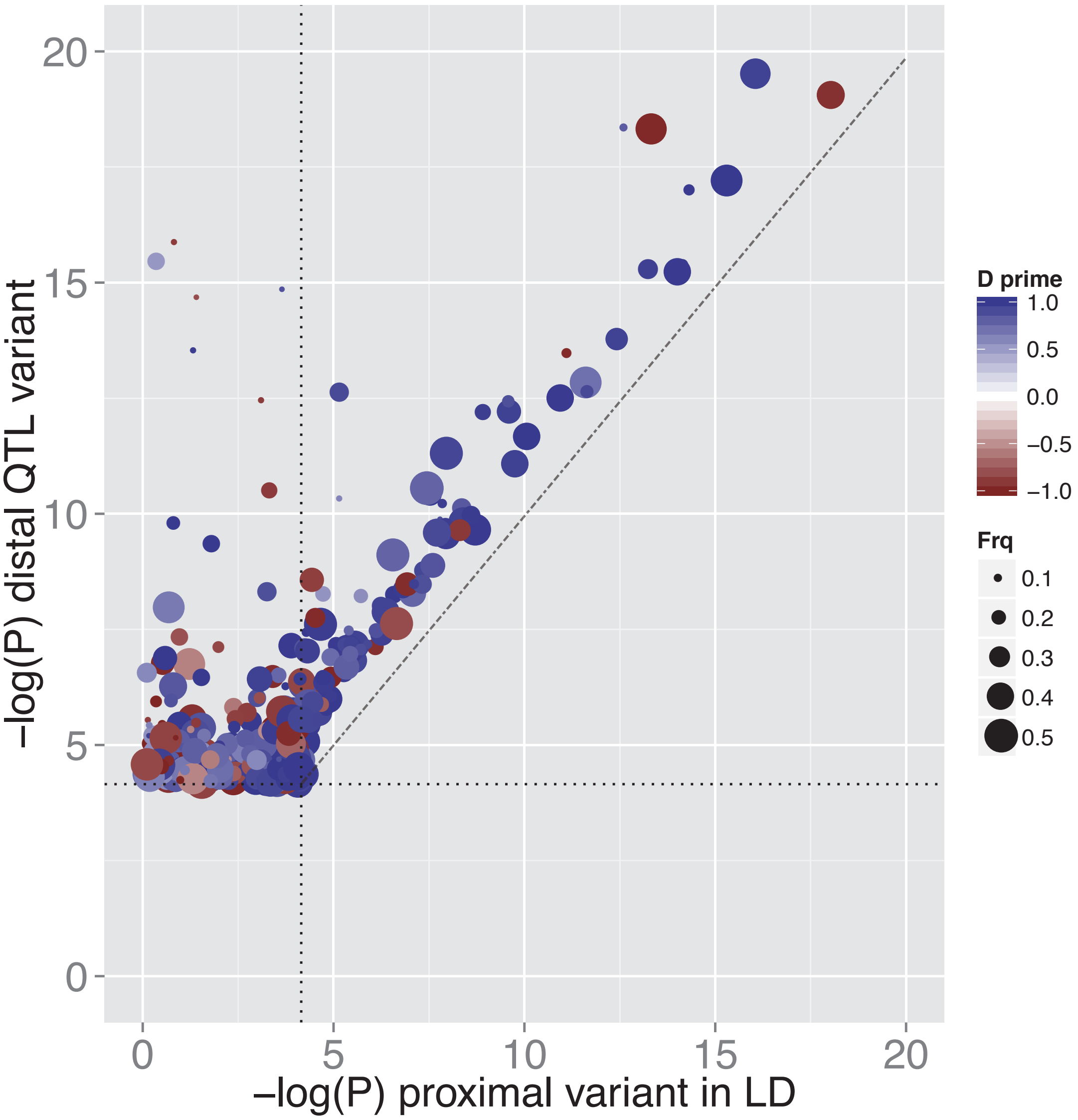

Figure 2.16: P value distribution of the proximal variants. Here the P values from the association between the CTCF binding and the lead distal QTL variants are plotted against that of the proximal variants, which are in LD with the distal QTL variants. The horizontal and vertical dashed lines are the 1% genome wide FDR threshold established in the main analysis. The diagonal line assists to indicate same P values. Each dot is colored by its D' value of LD with its size scaled by the allele frequency of the proximal variant.

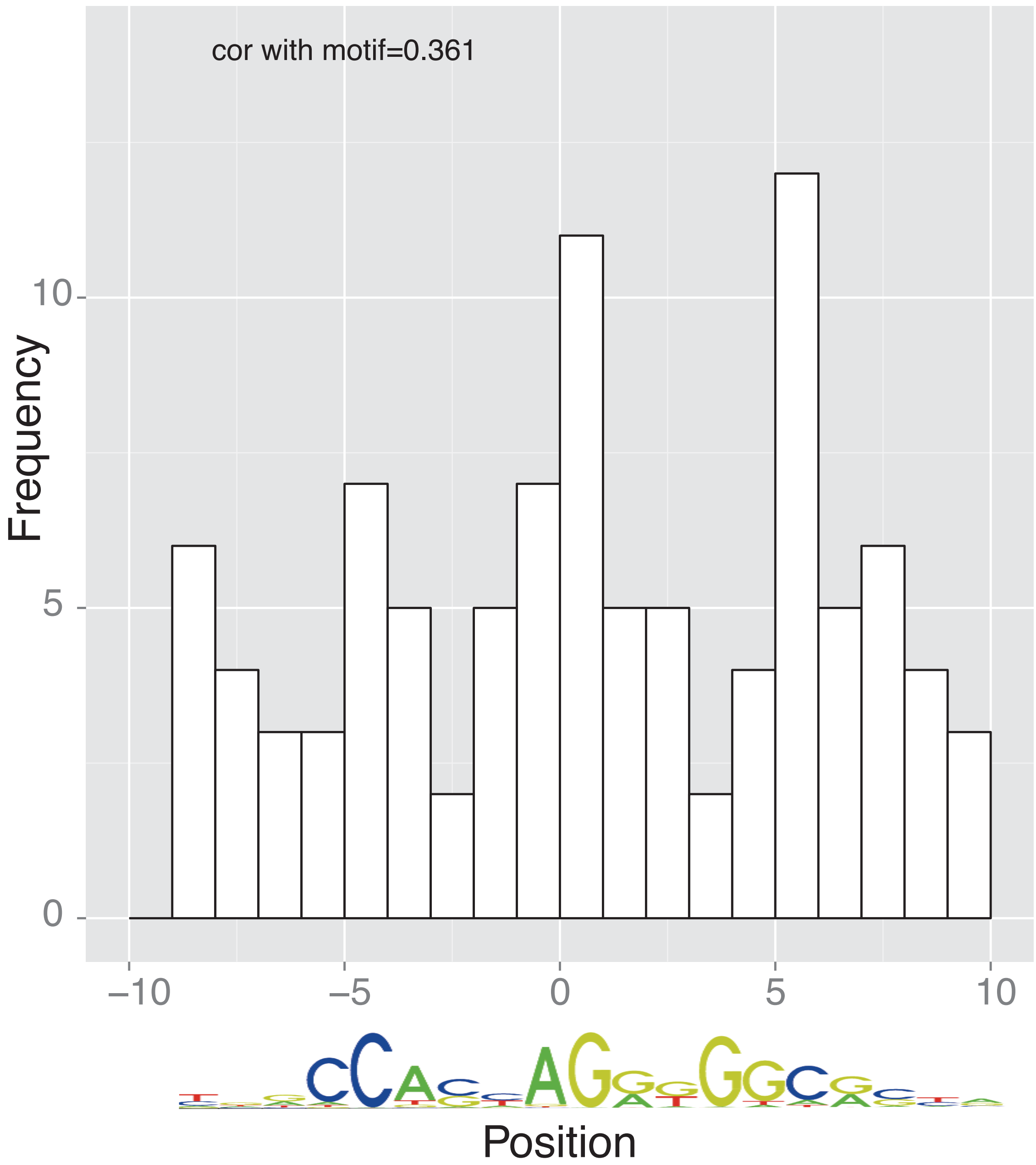

Figure 2.17: Distribution of the proximal variants that are on motif and in LD with the distal lead QTL variants. Here the proximal variants were aligned to the motif positions. We saw a correlation between their distribution and the information content of the motif at $\rho = 0.36$.

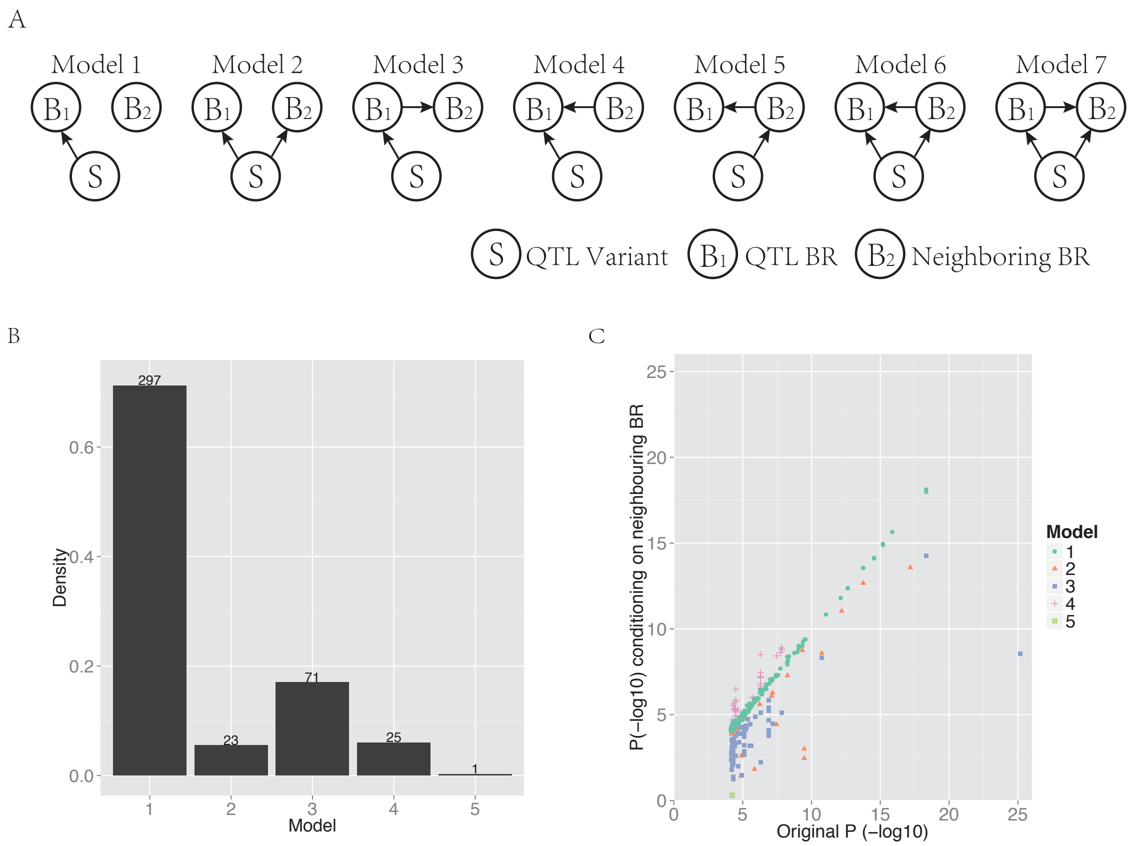

Figure 2.18: Evidence for indirect effects when a second binding region is present in the distal QTL window. Many (75.5%) of our distal QTLs contain a second CTCF binding region in their 50kb *cis*-window. To explore possible causal relationships between the lead variant, the associated binding region(BR1) and the second binding region(BR2) we constructed seven graphical models (A) and compared them using the Bayesian Information Criterion (BIC). In each case we assign the most likely model, chosen as having the lowest BIC (AIC showed same results). The frequency of the chosen models (B) suggests that there is almost never evidence for the association effect of the distal variant being mediated via a secondary binding region. The most frequently preferred model (1) did not involve BR2 at all; for the next most preferred models (3 and 4) there was some evidence of interactions between neighbouring CTCF binding sites, but we could not explain the variant association to BR1 binding via BR2. The only models which support mediation of binding at BR1 via BR2 are 5 and 6, and in only one case do we see one of these being selected. The P value of BR1 when conditioned on BR2 is plotted in (C). We further investigated the enrichment of a range of ENCODE (**?**) signals over the QTL binding region and the neighboring region. We found the association between two binding regions (model 3,4) tend to correlate with the active regulatory signals (Figure 2.19).

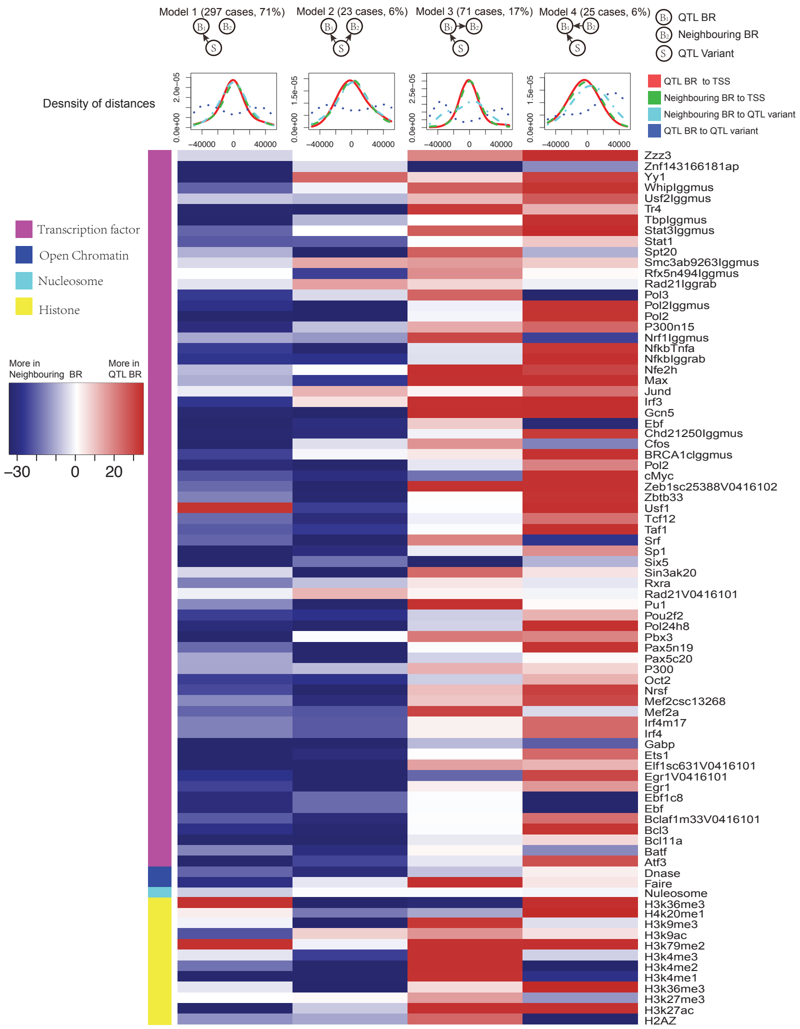

Figure 2.19: The interaction between QTL binding region and neighboring binding region correlates with regulatory events. The distal QTL set is as previously described (Figure 2.18). For each of the four categories with sufficient abundance (model 1, 2, 3 and 4), we compare the average signals between the QTL binding region (B1) and the neighboring binding region (B2) for a number of molecular markers using data obtained from the ENCODE project (Birney et al., 2007). We observed distinct patterns of regulatory signals between model 1,2 and model 3,4. We saw that when there exists interactions between two binding regions (model 3,4), active transcription factors, enhancers and active histone markers tend to be more enriched in the QTL binding regions, as shown in red. This change is not driven by their distances being closer to the transcription start site (TSS) by chance, measured as the distance to the closest TSS, because the neighboring binding regions have similar distance to the TSS as the QTL binding regions (red and green lines in the density plots). Some of the histone modifications (H2AZ, H3k27ac, H3k4me1, H3k4me2 and H3k4me3) swap enrichment direction between model 3 and model 4 depending on the direction of interaction between B1 and B2 (also see Figure 2.20 for more detailed enrichment signals).

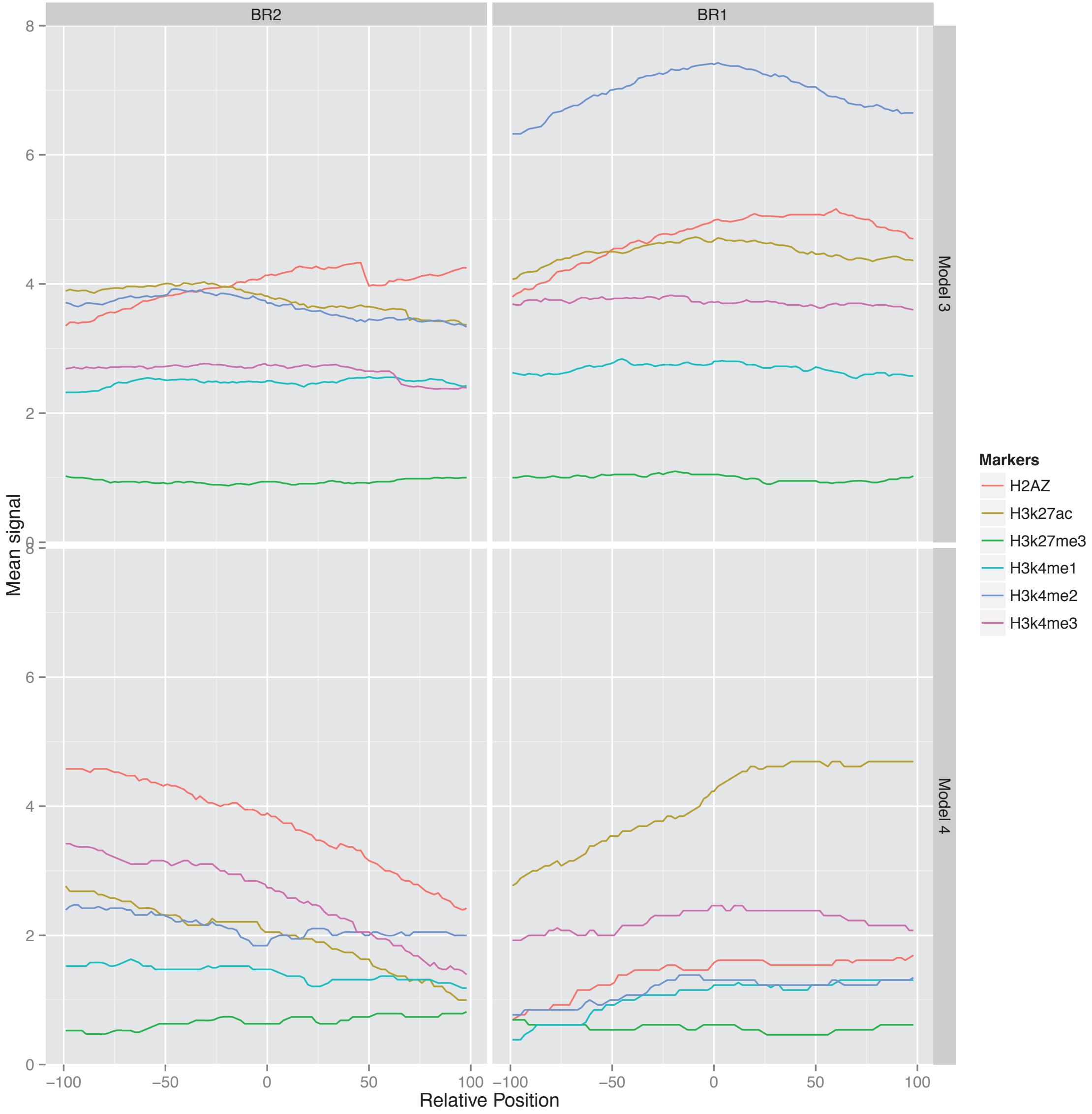

Figure 2.20: Change of histone modifications depending on the interaction models between the QTL binding region and the neighboring binding region (see Figure 2.18 and Figure 2.19 for explanations about the models).

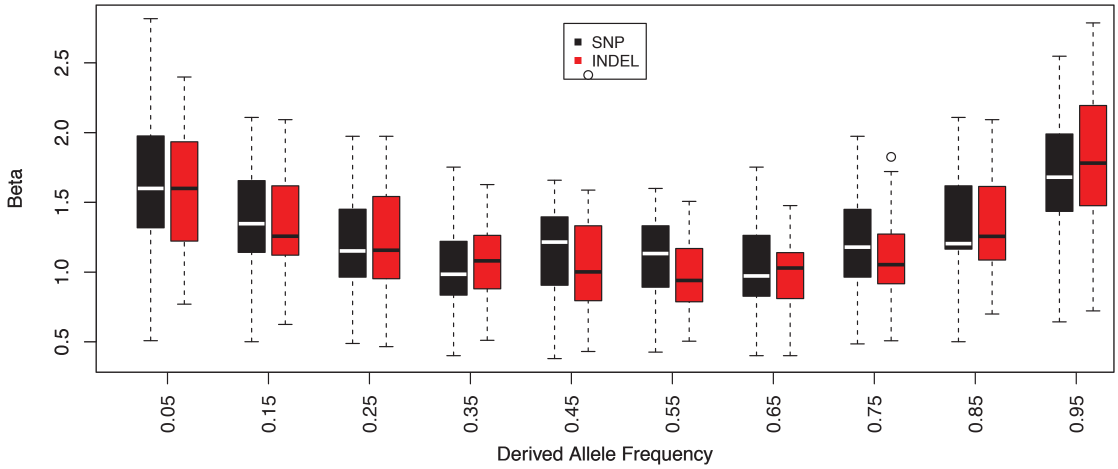

Figure 2.21: Effect size versus derived allele frequency for all CTCF QTLs identified at 1 % FDR.

they can be explained as two CTCF peaks in close proximity, one or both of which is under *cis*-genetic control.

There are 61 CTCF QTL variants that overlap with disease and trait associated variants from other studies (Table 2.4, GWAS Catalog Hindorff et al., 2009). In particular there is a disproportionate overlap with immune system related diseases (20 variants; χ^2 P-value 1.7E-9). This is consistent with the lymphocyte origin of LCLs, and may suggest roles of CTCF binding in the disease phenotypes.

In summary, these results are consistent with previous studies(Kasowski et al., 2010; Maurano et al., 2012; Reddy et al., 2012; Stefflova et al., 2013) that observed substantial variation in transcription factor binding within and between species, only a minority of which could be accounted for by genetic differences in the binding motif. We also found that only 25.7% of our QTLs could be explained by a genetic variant in the motif. The majority of the remainder can be explained by changes within 1kb of the motif, consistent with observations that transcription factor binding differences between mouse strains are more likely if there are genetic differences within 200bp of

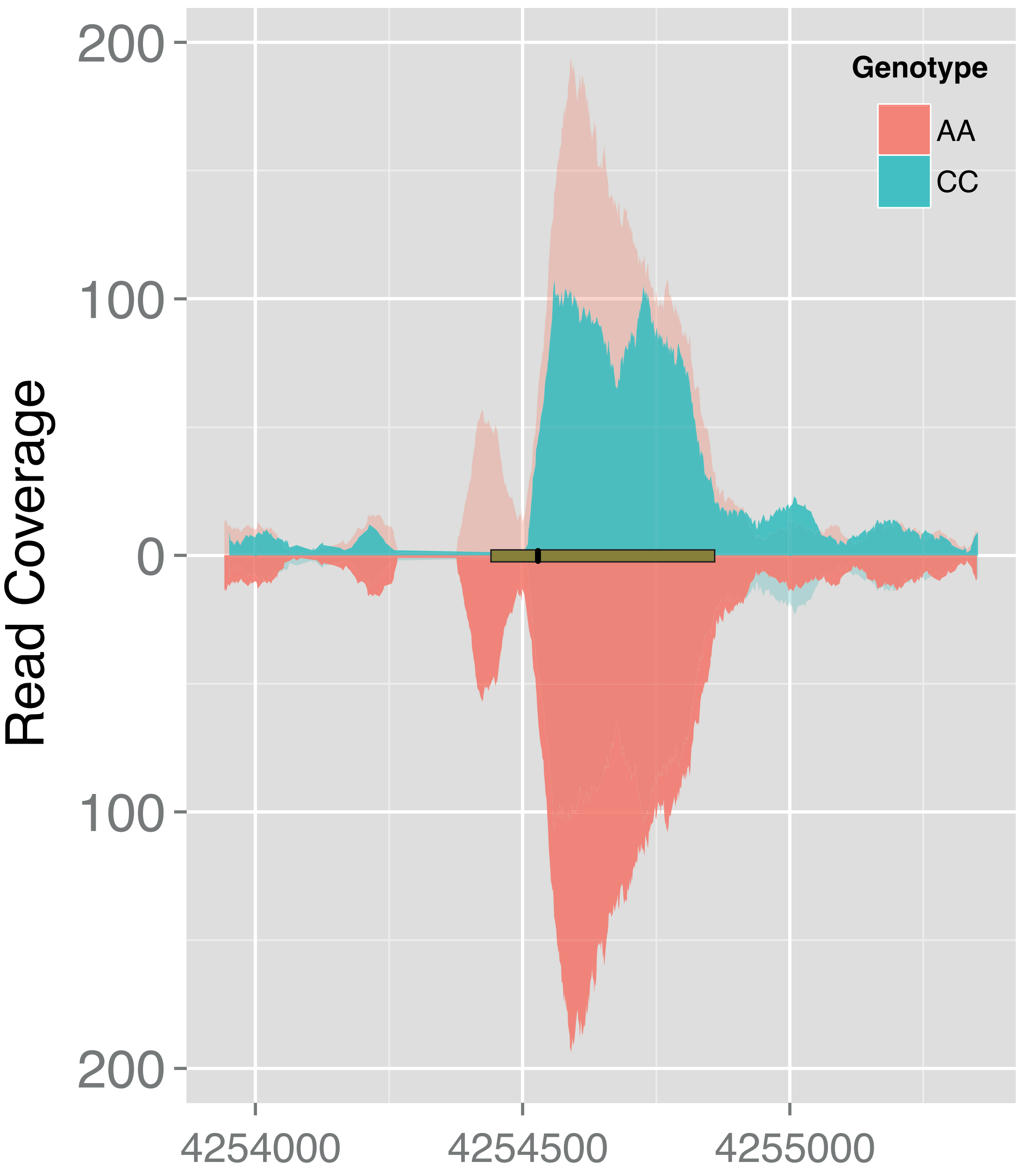

Figure 2.22: Example of CTCF peak shape QTL. Reads for samples in each homozygous genotype group at QTL rs11935835 were merged (AA and CC, respectively), and the average CC genotype profile is plotted above the main axis (in green), and the average AA genotype profile below (in red); each plot is reflected on the other axis in a lighter colour to allow visual comparison. The AA genotype has stronger overall binding, with a second peak to the left, whereas the CC genotype has a double peak. The heterozygote has intermediate profile between these two (not visualized in this figure). The binding region is marked as a brown box with the SNP position marked by a black vertical dash.

the binding site (Heinz et al., 2013). However there remain some genetic associations for which we are not able to identify any proximal candidate, suggesting that longer range influences can make some contribution to CTCF binding. Using published gene expression data for a subset of these samples, we looked at correlations between CTCF bindings and expression levels of nearby genes. We did not see strong correlations between the two, suggesting more complex role of CTCF in influencing genes.

Table 2.4: The overlap between CTCF QTL variants and GWAS variants.

PMID	Disease/Trait	CHR	SNP	-log(P_{GWAS})	Binding region start	-log(P_{QTL})
18204098	Systemic lupus erythematosus	8	rs13277113	10	11339579	4.891
17611496	Asthma	17	rs7216389	10.046	38028921	14.668
21627779	Alzheimer's disease	11	rs1562990	10.398	60018960	15.264
23263486	Urate levels	5	rs17632159	10.398	72431291	5.383
19079260	Body mass index	1	rs2568958	11	72796185	23.913
19079260	Body mass index	1	rs2568958	11	72808237	20.605
19079260	Body mass index	1	rs2568958	11	72794697	20.529
17554300	Type 1 diabetes	12	rs11171739	11	56435260	4.979
22396660	Nephrolithiasis	5	rs11746443	11.046	176797734	4.184
21460841	Alzheimer's disease (late onset)	11	rs4938933	11.097	60018960	14.791
21102463	Crohn's disease	6	rs415890	11.523	167411229	5.477
19430480	Type 1 diabetes	17	rs2290400	12.222	38028921	14.668
23222517	Red blood cell traits	22	rs5749446	12.523	32870620	4.932
21804548	Asthma	12	rs1701704	12.699	56435260	4.931
22561518	Vitiligo	12	rs2456973	13.523	56435260	4.931
22423221	Mean platelet volume	6	rs210134	14.699	33546527	5.42
22700719	Chronic lymphocytic leukemia	6	rs210142	15.046	33546527	5.42
21833088	Multiple sclerosis	16	rs7200786	16.046	11196016	4.233
21829393	Type 1 diabetes autoantibodies	12	rs1701704	17.301	56435260	4.931
17554260	Type 1 diabetes	12	rs2292239	19.699	56435260	5.796
23128233	Inflammatory bowel disease	6	rs1819333	20.155	167411229	5.477
23128233	Inflammatory bowel disease	21	rs7282490	25.699	45659281	4.193
21829393	Type 1 diabetes autoantibodies	12	rs2292239	26.523	56435260	5.796
21149283	Iron status biomarkers	11	rs236918	27	117051957	5.29
22139419	Platelet counts	6	rs210134	35.155	33546527	5.42

PMID	Disease/Trait	CHR	SNP	-log(P_{GWAS})	Binding region start	-log(P_{QTL})
22504420	Bone mineral density	7	rs6959212	37.398	38110179	8.869
21079607	Anorexia nervosa	3	rs6782029	5.046	11661145	4.572
23251661	Obesity-related traits	3	rs1044826	5.097	139072763	4.783
18839057	Attention deficit hyperactivity disorder	16	rs11646411	5.155	82772300	4.259
23192594	Body mass index (interaction)	18	rs11876941	5.301	50906413	9.637
22589738	Subcutaneous adipose tissue	1	rs990871	5.398	72794697	20.061
22589738	Subcutaneous adipose tissue	1	rs990871	5.398	72808237	19.855
22365631	Temperament (bipolar disorder)	21	rs2150410	5.398	40547111	14.838
23319000	Metabolite levels (HVA/MHPG ratio)	2	rs6750634	5.398	50763433	14.085
23251661	Obesity-related traits	17	rs1051424	5.523	57976434	11.83
21998595	Height	6	rs2224391	5.523	5261260	11.231
23319000	Metabolite levels (HVA-5-HIAA Factor score)	8	rs13251954	5.699	29034453	4.16
20195514	Primary tooth development (time to first tooth eruption)	17	rs9674544	6.097	47091576	4.54
20862305	Type 2 diabetes	15	rs1436955	6.155	62417944	4.854
22797727	Renal function-related traits (sCR)	5	rs12654812	6.301	176797734	4.943
21833088	Multiple sclerosis	5	rs4075958	6.301	176797734	4.344
21408207	Systemic lupus erythematosus	8	rs2736340	6.523	11339579	4.891
22451204	Parkinson's disease	2	rs6430538	6.699	135540345	15.085
22797727	Renal function-related traits (eGRFcrea)	5	rs12654812	6.699	176797734	4.943
20228799	Ulcerative colitis	17	rs8067378	7	38028921	14.668
19023125	Brain imaging in schizophrenia (interaction)	5	rs245201	7.046	127169342	5.004
21118971	Small-cell lung cancer	11	rs716274	7.046	103408608	4.285
19079261	Body mass index	1	rs2815752	7.222	72796185	23.913
19079261	Body mass index	1	rs2815752	7.222	72808237	20.605
19079261	Body mass index	1	rs2815752	7.222	72794697	20.529
20596022	Alopecia areata	12	rs1701704	7.523	56435260	4.931
19079260	Weight	1	rs2568958	7.699	72796185	23.913
19079260	Weight	1	rs2568958	7.699	72808237	20.605

PMID	Disease/Trait	CHR	SNP	-log(P_{GWAS})	Binding region start	-log(P_{QTL})
19079260	Weight	1	rs2568958	7.699	72794697	20.529
23128233	Inflammatory bowel disease	5	rs12654812	7.699	176797734	4.943
19165918	Systemic lupus erythematosus	8	rs2618476	7.699	11339579	4.891
20195514	Primary tooth development (number of teeth)	17	rs9674544	7.699	47091576	4.54
18464913	Protein quantitative trait loci	11	rs7112513	8.222	117051957	5.29
19503088	Rheumatoid arthritis	8	rs2736340	8.222	11339579	4.891
20881960	Height	7	rs6959212	8.699	38110179	8.869
19801982	Bone mineral density (spine)	7	rs1524058	9	38110179	9.083
23291587	Behcet's disease	12	rs2617170	9	10563751	5.679
21459883	Dilated cardiomyopathy	1	rs10927875	9	16321009	4.468
18198356	Type 1 diabetes	12	rs1701704	9.046	56435260	4.931
22446961	Kawasaki disease	8	rs2736340	9.046	11339579	4.891
22139419	Mean platelet volume	3	rs10512627	9.301	124339333	9.465
19820697	Hematological parameters	22	rs9609565	9.398	32870620	4.943
23128233	Inflammatory bowel disease	14	rs194749	9.523	69255227	4.287
21943158	Cardiovascular disease risk factors	11	rs508487	9.699	117051957	4.597
22001757	Liver enzyme levels (alkaline phosphatase)	8	rs6984305	9.699	9178038	4.199
23251661	Obesity-related traits	7	rs11976180	5.15490196	143760743	6.02128023
23064961	Dental caries	13	rs735539	5.397940009	21285708	7.303681407
22566498	Response to angiotensin II receptor blocker therapy	11	rs11020821	6.045757491	94234754	5.607190203
22566498	Response to angiotensin II receptor blocker therapy (opposite direction w/ diuretic therapy)	11	rs11020821	5.397940009	94234754	5.607190203
22561518	Vitiligo	11	rs4409785	12.69897	95311198	7.479585622
22001757	Liver enzyme levels (gamma-glutamyl transferase)	1	rs10908458	14.69897	155085124	5.091098739
21833088	Multiple sclerosis	11	rs4409785	6.22184875	95311198	7.479585622
21037568	Hodgkin's lymphoma	2	rs1432295	7.698970004	61066413	7.269570862

PMID	Disease/Trait	CHR	SNP	-log(P_{GWAS})	Binding region start	-log(P_{QTL})
20972438	Bladder cancer	4	rs798766	12.39794001	1731408	5.210110236
20708005	Non-alcoholic fatty liver disease histology (other)	13	rs1305088	5.045757491	29252925	4.210516688
20395239	Optic disc size (cup)	12	rs10858945	5.22184875	90456729	4.268051964
20348956	Urinary bladder cancer	4	rs798766	11	1731408	5.210110236
19343178	Height	7	rs849141	10.52287875	28182178	4.157298252
19197348	Quantitative traits	7	rs2527866	5.522878745	157091071	7.170945424
18759275	Uric acid levels	3	rs6442522	5.301029996	15440342	17.31606037

Allele-specific bias analysis of CTCF binding provides independent confirmation of QTLs This data set represents an excellent resource to directly examine allele-specific biases in TF binding at heterozygous sites in a larger set of individuals than previous studies (McDaniell et al., 2010). Allele-specific binding refers to statistically significant biases in binding to the two alleles in a diploid cell, at sites where a heterozygous polymorphism allows the two alleles to be distinguished. Allele-specific binding thus is an independent way of assessing how genetic variants at binding sites might affect binding variation. Although the two alleles at heterozygous SNPs are normally referred to as the reference or alternate allele (referring to which base is found in the reference genome sequence and which is the alternate base), here we chose to categorize the two alleles as ancestral (shared with chimp) or derived (human specific). This has two advantages. First, any residual effect of biases in aligning sequence reads to the reference allele will be minimized. Second, measuring allele-specific binding in terms of the ancestral and derived allele provides information about how evolutionary changes might affect CTCF binding.

After processing the reads, we identified allele-specific sites using a binomial null model of equal occupancy of both alleles at heterozygous sites, using a 5% FDR corrected threshold, similar as described previously(McDaniell et al., 2010). Allele specific variants were identified using reads pooled across all individuals for each allele. This process identified 589 SNPs that have replicated in at least two individuals showing significant allele-specific bias. We examined the allele counts of all heterozygous individuals at these 589 SNPs. For most sites (91.5%) the allele-specific biases were consistent between individuals, confirming the predominantly genetic basis of allele-specific binding (Figure 2.23). At such sites, the same ancestral or derived allele was preferred for binding across 2 or more individuals.

However, there were 50 (8.5%) sites which showed significant but opposite allele-specific biases between two or more individuals. Six of these 50 sites could potentially be explained by virtue of being close to loci known to be subject to allelic exclusion (the Immunoglobulin heavy chain), a process that affects one allele randomly (see Discussion). One site lies in the KCNQ1 imprinted locus, where the regulatory status depends on parent of origin rather than genotype. The 46 other sites at which the allele- specific binding bias switches between individuals (Appendix Table A.1) could represent new random allelic exclusion loci or imprinted sites, or could arise because the site at which we see allele specificity is incompletely linked with the causal variant (Lappalainen et al., 2013). We tested whether there was a SNP which specifically explained the allele specific switching site; for 28 cases this was the case. We are not able to directly test whether any of these sites could be due to imprinting because parent-of-origin information is not available for the heterozygous alleles of these individuals.

Interestingly, a significant majority (68%, P <1E-16) of the SNPs showed increased binding to the ancestral allele (Figure 2.23). Alignment bias towards the reference allele has been reported before (McDaniell et al., 2010) and because the ancestral allele is more likely to be the reference allele, the increased binding to the ancestral allele could be the result of the alignment bias. To rule out this possibility, we analyzed the cases where the ancestral allele is the alternate allele and found that the binding bias remained towards the ancestral allele (Figure 2.25). Additionally, we repeated the allele-specific analysis after using a variant-aware aligner. The results were largely identical to what we observed as described above, indicating that the preference for the ancestral allele is not a trivial outcome of any alignment bias (Figure 2.26).

The allele-specific signal at binding regions (intra-individual measurements) mostly correlated linearly with the QTL effect size (inter-individual measurements) (Figure 2.24). There were however exceptions to this, and these were mainly cases in which there was an allele-specific signal but not inter-individual QTL. We observed QTLs

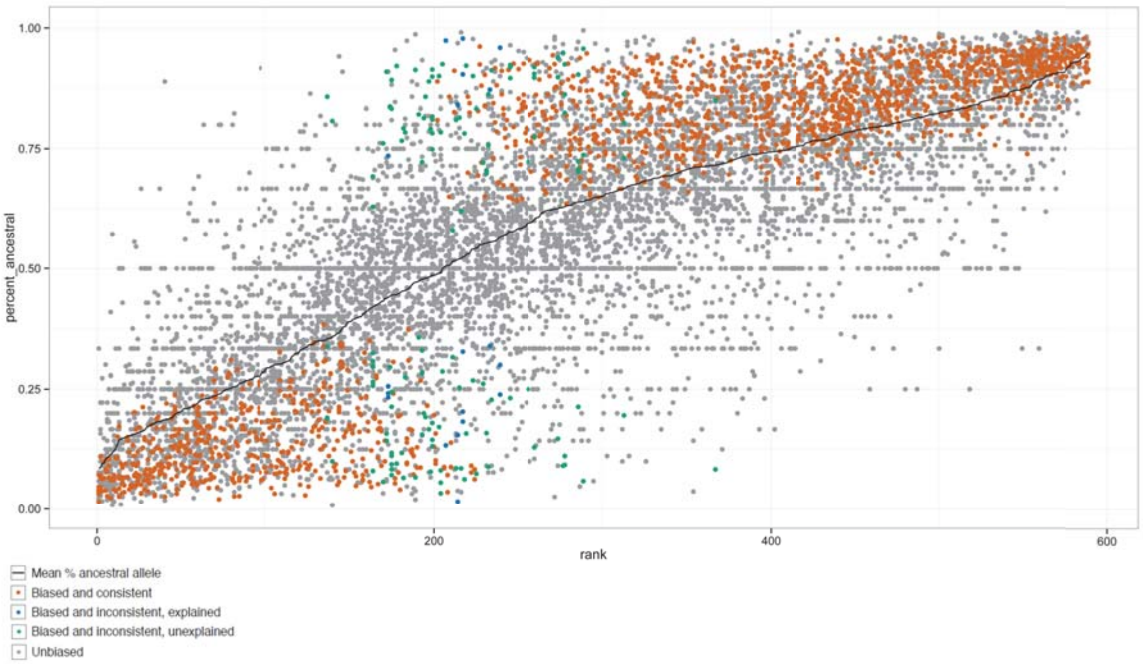

Figure 2.23: Summary of allele-specific analysis. SNP loci that show significant allele-specific CTCF binding in at least 2 samples are included. The y-axis represents the proportion of the total read counts from the ancestral allele. The 589 SNP loci are ordered by mean proportion ancestral allele for all heterozygous samples (black line). Heterozygous samples that do not pass the allele-specificity threshold are shown as gray points. Significant and consistent allele-specific samples (ie. the binding bias is toward the same allele) are represented by orange points. Significant but inconsistent samples are either blue (inconsistency explained by the nature of the site) or green (inconsistency unexplained).

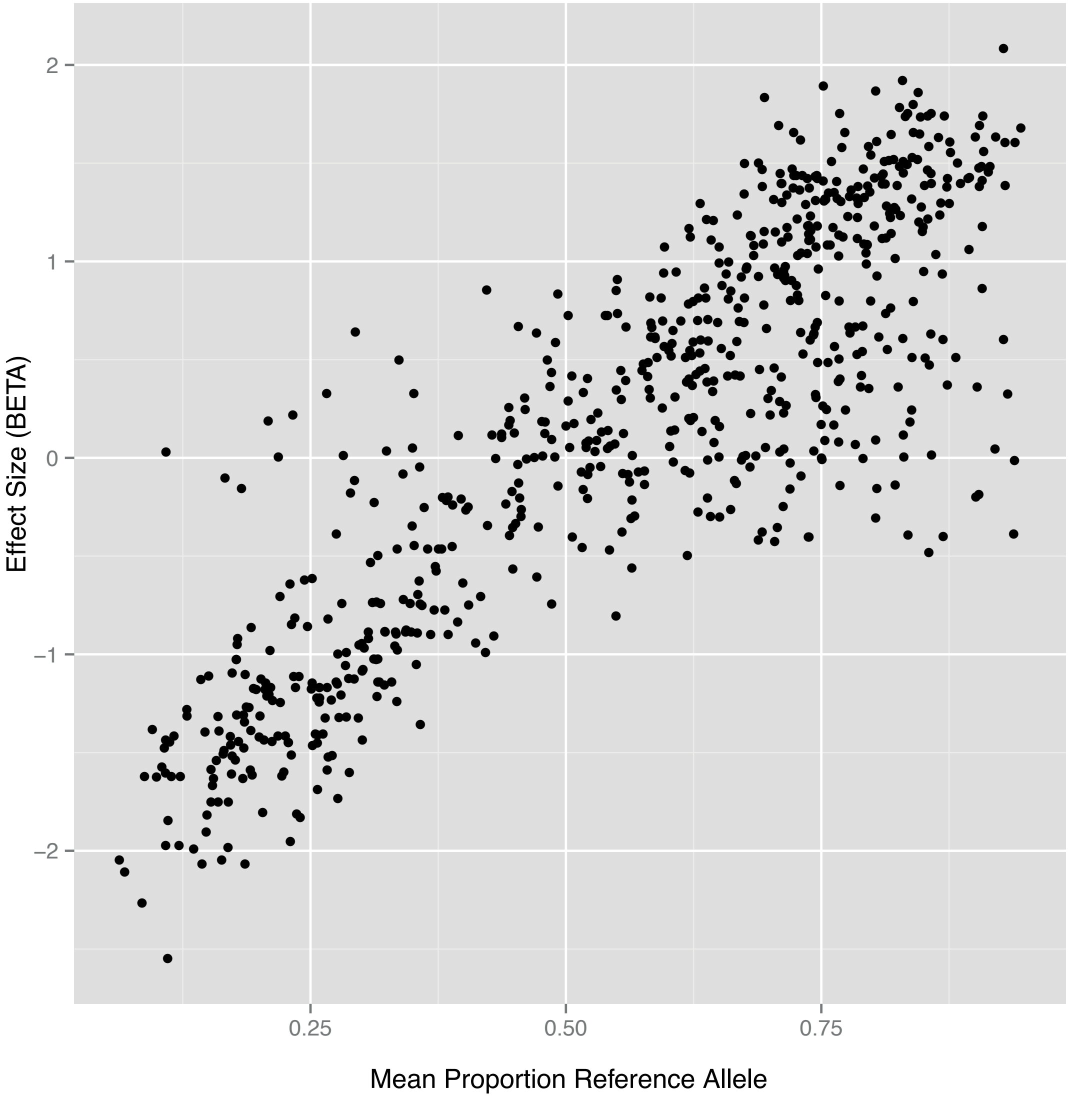

Figure 2.24: Allele-specificity correlates with QTL effect size (β). The mean proportion reference allele count for all heterozygous samples at SNP loci that show significant allele-specificity in at least 2 samples are plotted against the QTL effect size (β) at that locus. Only the β values from associations where the SNP is located within the associated binding region are shown.

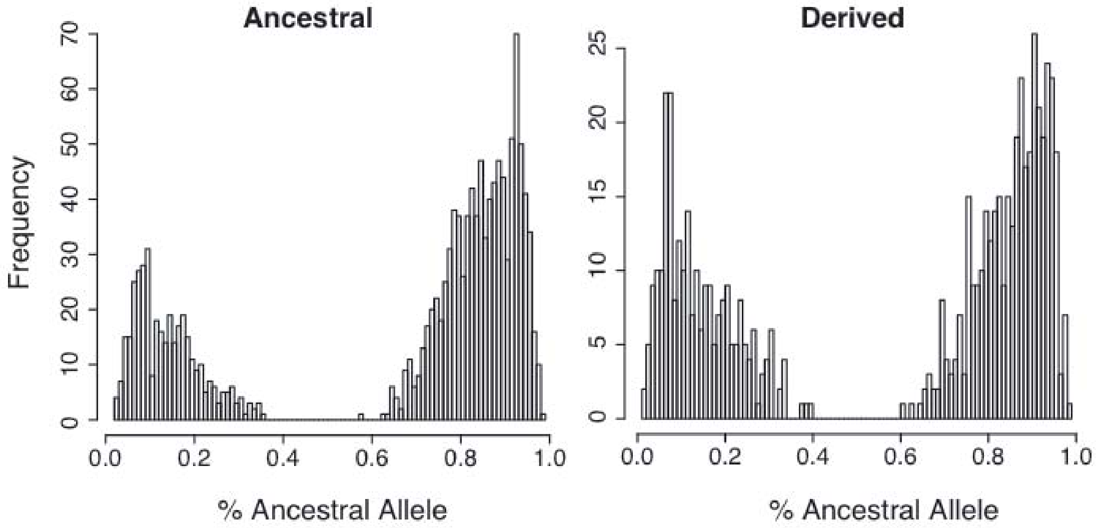

Figure 2.25: Effect of the reference allele. Even when the reference allele is the derived allele (Derived), the binding bias remained towards the ancestral allele.

with strong effect size in binding regions tend to show strong allele-specificity (Figure 2.27).

2.8 Discussion

This study is the first association QTL study performed on transcription factor binding in humans to our knowledge. The single site QTL properties are consistent with and extend other studies such as the family based studies (McDaniell et al., 2010; Maurano et al., 2012), the DNase I QTL (Degner et al., 2012). We find a large number of QTLs, with the majority being within or close to the binding region, and approximately a quarter inside the bound CTCF motif. By using the 1000 Genomes Project cell lines, we can be reasonably confident that we have a full catalog of common variation of which some subset are the causal variants. Using this information we could show that for a large fraction of the associations where the initial analysis suggested a distal variant more than 1kb away, there was a plausible causal candidate also within 1kb of the binding motif. Overall this suggests that, at least for CTCF, the substantial majority (~75%) of common genetic variants in the region with a reasonably strong effect on transcription factor binding lie within 1kb of the binding motif, although only a minority are actually within the motif. This clarifies previous observations that genetic variants contributing to transcription factor binding (CTCF and many others) were typically not in the motif itself (Kasowski et al., 2010; Stefflova et al., 2013) but there was enrichment nearby (Heinz et al., 2013).

These results suggest that the regulatory mechanism is not readily explainable by a simple regulation model. When we overlap CTCF QTLs with binding of other transcription factors that are measured by the ENCODE project, CTCF QTL variants that are not within the canonical motif are characterized by a modest enrichment (approximately 2 fold compared to random) of H3K4me3 and other transcription factors, such as PU1, Rad21, Pol II, ZNF1, YY1, and USF1. In these cases it is possible that the effect may be mediated via collaborations between these factors and CTCF. There is a small fraction (1.5%) of CTCF QTLs overlap *cis* eQTLs discovered in previous studies, indicating limited *cis* genetic effect targeting both CTCF binding

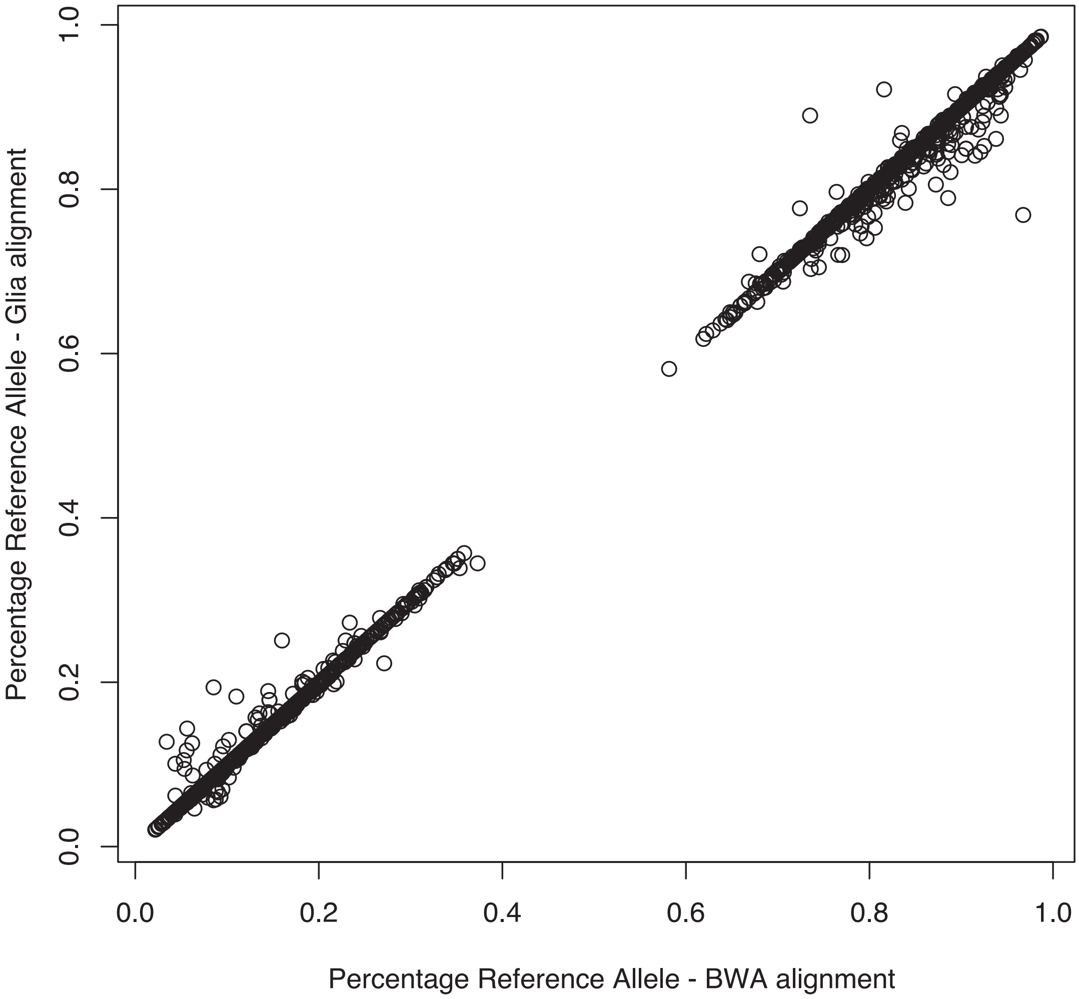

Figure 2.26: Effect of alignment to allele specific analysis. We performed local realignment using a variant aware aligner glia (https://github.com/ekg/glia), which align reads to a variant graph (Lee et al., 2002) built using supplied variants, and compared the allelic bias in our significant allele specific sites between the two alignments. We saw that the effect of local realignment is minimum.

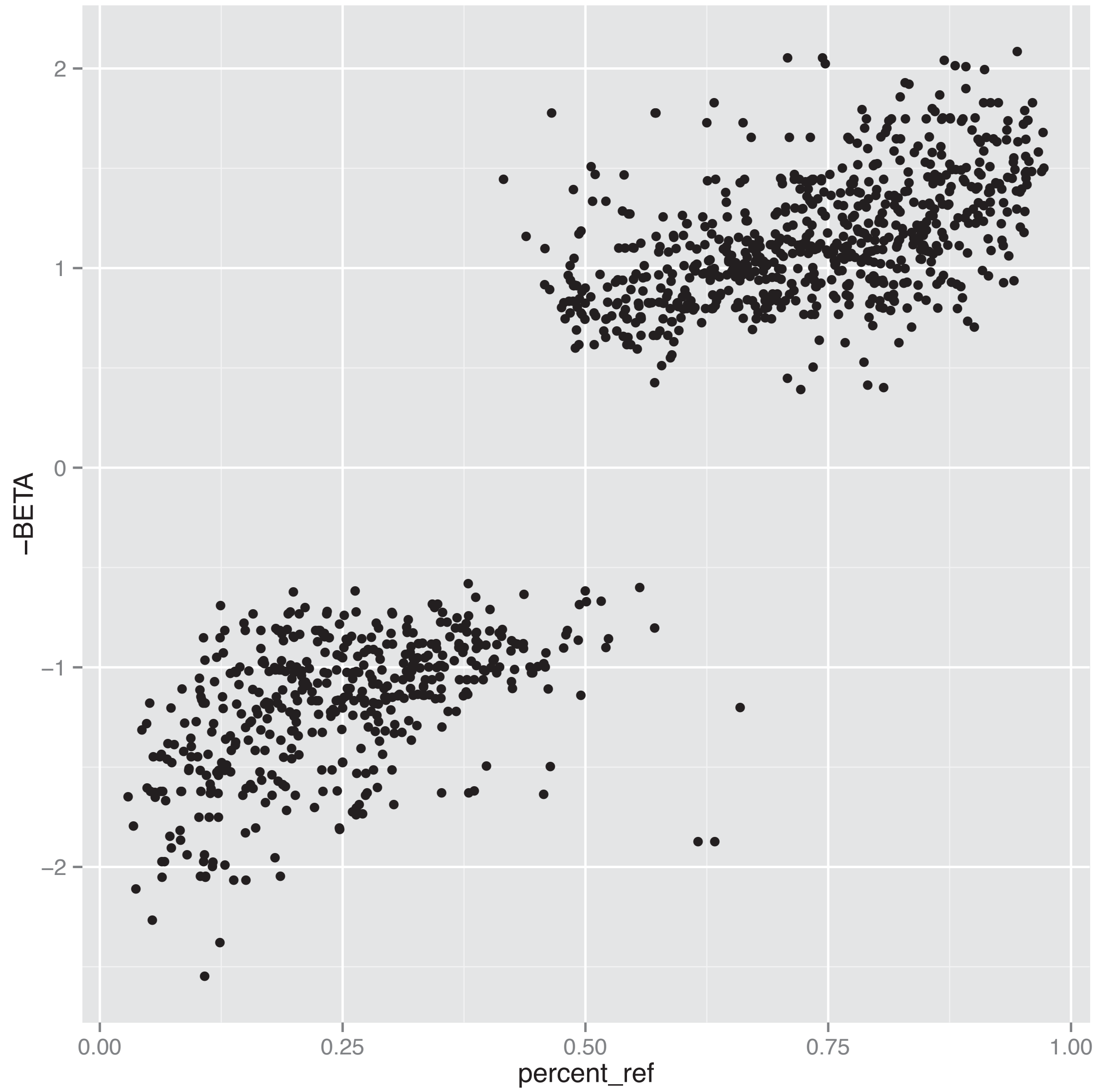

Figure 2.27: No QTLs with strong effect size in binding regions that do not show strong allele specificity. The x-axis shows allele specificity (measured as % reference), and the y-axis shows between-individual effect (β) orientated such that positive is towards reference.

and gene expression simultaneously. However, it is possible that CTCF affect gene expression distantly at a weaker level. Such effect could take place via higher order interactions via 3D chromatin structures such as looping (Bulger and Groudine, 2011). Much more samples may be needed to achieve sufficient power.

The results from this and many other studies suggest that motif adjacent sequences may influence transcription factor activities. This may exist when transcription factor binds to weaker secondary motifs that are close to the canonical motif or distribute around the canonical motif. Even those highly sequence specific transcription factors such as CTCF do not bind to their canonical motif in 100% cases, but instead show a distribution of binding occurring at different positions and alleles. Alternatively, it is also possible that variants that are on the canonical motif are removed from population due to their large effect. On the contrary, the ones in the close vicinity nearby may have weaker but significant phenotypic effects that is below selection (Farh et al., 2014).

We see hundreds of sites showing allele-specific binding. The idea that allele-specific events have similar effects inside one cell as genotypic effects do between individuals is commonplace (McVicker et al., 2013). Here we show that these two effects are well modeled by a linear relationship (at least for this assay), though there is also an interesting subset of allele-specific sites that show no QTL. In contrast there are few QTL loci that overlap binding regions without an allele-specific signal.

As expected, some of the allele specific sites switch specificity between the alleles in different samples, consistent with a nearby, incompletely linked causal allele, random allelic inactivation or parent-of-origin imprinting. Many of these sites can be explained by an incompletely linked nearby locus, highlighting that the causal variant is often not co-incident with the binding region.

Finally with more confident mapping of reads from paired read ChIP-seq data we are able to show that a consistent signal towards reference alleles is in fact predominantly due to a biological effect favoring ancestral alleles (at least for the CTCF transcription factor). This suggests that base pair changes segregating in the population tend to reduce binding of existing sites (rather than create new sites), at least for CTCF, and this is consistent with CTCF motif creation occurring by non-base pair changes, e.g. repeat deposition, as suggested in Schmidt et al. (2012). Similar observations have been made on the allele effect of gene expression, where the new mutations tend to reduce gene expression levels (Chaix et al., 2008).

The understanding of the non-coding variants, which comprise the vast majority of the disease susceptibility variants discovered so far is remains challenging mostly due to our limited knowledge of the regulatory mechanisms in the non-coding regions. This catalog of CTCF QTL sites is part of a growing set of molecular assays that are being examined in outbred individuals (for example, see Kasowski et al., 2010; Degner et al., 2012; Maurano et al., 2012; Kilpinen et al., 2013; McVicker et al., 2013). It provides a specific hypothesis for the 63 disease related loci which overlap these QTLs, and for future overlaps with other molecular, cellular and disease related phenotypes. The gradual unraveling of the different variant effects on different molecular behavior will provide a growing understanding of molecular and physiological processes in health and disease.

Systematic survey using data of multiple cellular events for the same set of samples could offer hints for understanding the underlying biological mechanisms. Much work remains to be done to collect such information. Ideally it should be done *in vivo* such that the data genuinely reflect the biological effect independent of technical interventions. Additionally data of each event should be collected in a time series with events at each time point collected simultaneously . Although this may not be possible with the current technologies, it is most likely to offer the correct information. Such experiments may also pose computational challenges. Methods are needed to deal with very large volumes of data. Meanwhile, the number of cellular events are usually far more numerous than the number of samples, raising a “small n, large p” problem. It is challenging to resolve the causal relationships among these variables. Nevertheless, the regulatory mechanisms largely remain to be understood, only after which it becomes possible to choose the right candidates for therapeutic interventions.

Chapter 3

Using latent factors to enhance power in mapping expression QTLs for ageing

Collaboration note. *This chapter contains work in collaboration with Andrew Brown. The manuscript of this work is accepted by G3 subject to minor revisions. I am the joint lead author with Andrew Brown. My contribution includes processing microarray probe intensities, normalizations, learning both global and pathway factors, association analysis, and manuscript writing.*

3.1 Overview

Ageing is a multifactorial process, reflecting how the physical state of an organism accumulates changes and gene expression has been a field of increasing interest in studying this process. Microarrays and more recent RNA-seq technologies allow the simultaneous quantification of cell population average mRNA abundance for thousands of genes. These technologies have proved useful in providing diagnostic profiles for certain diseases (Reis-Filho and Pusztai, 2011). In the case of ageing, consistent patterns of age-related changes in gene expression have been observed across several tissues and species (Lu et al., 2004), such as over-expression of inflammation and immune-response genes and under-expression of genes involved in energy metabolism in older samples (de Magalhães et al., 2009). Given this commonality of function amongst genes which show age related changes in expression, we decided to investigate ageing in the context of biological knowledge on the function of genes, as provided by pathway annotations.

Array expression experiments generate high dimensional structured data sets, in which there are correlated patterns across large numbers of genes. Some of these are due to known technical or biological effects such as batch effects and cell growth stage, which when not the focus of the analysis can be removed by including them as covariates. However, even after this, there is typically substantial structural correlation. In previous studies, these can be represented by linear components of expression measurements, or factors, that can be inferred using methods such as principal components analysis (PCA) or factor analysis to create global phenotypes (Leek and Storey, 2007; Parts et al., 2011). When the aim is to discover local effects, such as *cis* genetic regulation, these global phenotypes can be treated as nuisance variables and removed from further analysis. This has been seen to increase power in analysis (Montgomery et al., 2010; Pickrell et al., 2010; Stegle et al., 2010). Conversely, if the aim is to differentiate between a case and control condition using expression, then global phenotypes could be more effective classifiers than local phenotypes (Hastie et al., 2000).

A recent study applied factor analysis methods in a two stage procedure to generate phenotypes representing expressions of groups of genes (Stegle et al., 2010). After regressing out global factors, as in Parts et al. (2011), expression levels for groups of functionally related genes, as defined by annotations from pathway databases, were treated as new expression datasets and the same factor analysis methods were used to construct pathway factors. The factors constructed on pathway sets of genes were taken as concise summaries of common expression variation across each pathway. We test these factors values below as phenotypes, and so refer to them as phenotype factors as in some cases just phenotypes.

Here, we apply this method to gene expression data from abdominal skin tissues from 647 samples. Unlike previous studies which have concentrated on genetic variants which regulate multiple genes within a pathway (Parts et al., 2011), we focused on discovering associations between gene expression and an environmental variable age. We obtain our pathway gene sets from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa et al., 2004). Subsequently, by looking for associations between these new pathway phenotypes and age, we discover groups of functionally related genes with a common response to ageing which could be used as biomarkers describing molecular changes with age.

With data from a twin cohort containing both monozygotic and dizygotic twins, we can estimate proportions of variance explained by age, genetic variation, common environmental variation, and unique environmental variation (noise). Stochasticity in gene expression, which will form part of the unique environment component, is believed to play a role in the ageing process (Bahar et al., 2006). By investigating sources of variation within the pathway phenotypes, we find that they are more robust than the expression of individual genes with less unique environment variation. This explains some of our success at discovering associations with age.

3.2 Expression profiling

The data analyzed here are part of the MuTHER project (Multiple Tissue Human Expression Resource - http://www.muther.ac.uk/, Nica et al., 2011) and were downloaded from the ArrayExpress archive, accession no. E-TABM-1140. In summary, the study included 856 Caucasian female individuals (336 monozygotic (MZ) and 520 dizygotic (DZ) twins) recruited from the TwinsUK Adult twin registry (Moayyeri et al., 2013a). The age at sampling ranged from 39 to 85 years with a mean age of 59 years. Punch biopsies (8mm) were taken from relatively photo-protected infra-umbilical skin. Subcutaneous adipose tissue was dissected from each biopsy and the remaining skin tissue was weighed and stored in liquid nitrogen. Expression profiling of this skin tissue was performed using Illumina Human HT-12 V3 BeadChips where 200ng of total RNA was processed according to the protocol supplied by Illumina. All samples were randomized prior to array hybridization and the technical replicates were always hybridised on different beadchips. Raw data were imported to the Illumina

Beadstudio software and probes with fewer than three beads present were excluded. Log2-transformed expression signals were then normalized separately per tissue with quantile normalization of the replicates of each individual followed by quantile normalization across all individuals as previously described (Grundberg et al., 2012). Post-QC expression profiles were subsequently obtained for 647 individuals. The Illumina probe annotations were cross-checked by mapping the probe sequence to the NCBI Build 36 genome with MAQ (Li et al., 2008). Only uniquely mapping probes with no mismatches and either an Ensembl or RefSeq ID were kept for analysis. Probes mapping to genes of uncertain function (LOC symbols) and those encompassing a common SNP (1000G release June 2010) were further excluded, leaving 23,555 probes used in the analysis.

Box 1: Modeling

We model phenotype $y_i$of individual i (age A_i) as follows:

(Full)
$$y_i = \mu + \alpha A_i + \beta_i + \gamma_i + \epsilon_i$$
$$\beta_i \sim N(0, \sigma^2_{FAM})$$
$$\gamma_i \sim N(0, \sigma^2_{MZ})$$
$$\epsilon_i \sim N(0, \sigma^2)$$

(Null)
$$y_i = \mu + \beta_i + \gamma_i + \epsilon_i$$
$$\beta_i \sim N(0, \sigma^2_{FAM})$$
$$\gamma_i \sim N(0, \sigma^2_{MZ})$$
$$\epsilon_i \sim N(0, \sigma^2)$$

To correctly model the twin structure we enforce that $\beta_i = \beta_j$ when i and j are twins, and $\gamma_i = \gamma_j$ when i and j are monozygotic twins (capturing the increased genetic correlation of monozygotic twins).

From the null model we can define heritability (h^2), proportion of environmental variance explained by age (p_a) and the proportion of variance explained by the unique environment (p_e) as:

$$h^2 = \frac{2\sigma^2_{MZ}}{\sigma^2_{FAM}+\sigma^2_{MZ}+\sigma^2+\alpha_i^2 var(A_i)}$$

$$p_a = \frac{\alpha_i^2 var(A_i)}{\sigma^2_{FAM}-\sigma^2_{MZ}+\sigma^2+\alpha_i^2 var(A_i)}$$

$$p_e = \frac{\sigma^2}{\sigma^2_{FAM}+\sigma^2_{MZ}+\sigma^2+\alpha_i^2 var(A_i)}$$

Note that for p_a the genetic variance is removed from the denominator.

3.3 Gene expression pathway factors

In a two step approach, factor analysis methods were first used to discover patterns of common variation across the entire dataset. The software package PEER was applied using the default settings and using technical measurements (experimental batch, RNA quality and concentration) as covariates to create 5 components which in total explained 35.7% of the variation in the dataset. Secondly, the effects of the five global factors together with the technical covariates including batch, RNA concentration and RNA quality were removed from the whole gene expression data sets. The residuals after this process were then grouped to pathway subsets according to the KEGG annotation. For each pathway, we created five pathway phenotypes using PEER with the default settings.

In a two step approach, factor analysis methods were first used to discover patterns of common variation across the entire dataset. The software package PEER (Parts et al., 2011) was applied using the default settings and using technical measurements (experimental batch, RNA quality and concentration) as covariates to create 5 factors, which in total explained 35.7% of the variation in the dataset. For each individual, a factor is a weighted sum of all the gene expression measurements of that individual. The weights are chosen so that the factors iteratively explain the maximum amount of variation in the dataset subject to certain prior assumptions; these factors produce concise summaries of consistent patterns of expression for large numbers of genes.

We then used KEGG pathway annotation (186 pathways) as prior information to group genes into pathways. This allows inference of PEER factors for each pathway that we refer to as phenotype factors, in contrast to the global factors previously described. As before, these factors are weighted sums of gene expression measurements, but in this case only of genes within the pathway. Since global factors have been removed from the dataset prior to calculation of phenotype factors, these factors are unlikely to capture global effects on gene expression, but instead pathway specific patterns of expression. If a large enough module of genes within the pathway is co-expressed then likely one factor will also show the same pattern of co-expression across individuals. Equally, groups of genes could show opposite patterns of expression; this antagonistic gene expression could be reflected as factor values which correlate with one set of genes and are anti-correlated with the other across individuals. Individual genes can contribute positively or negatively to the weighted sum (indicated by the sign of the corresponding weight), meaning that a positive correlation between age and phenotype factor can be induced by negative correlations with individual genes.

We grouped the expression data set into 186 pathway subsets. For each pathway we created five pathway phenotypes using PEER with the default settings. We consider the learnt pathway factor values across individuals as five new phenotypes which can be investigated for associations with age (analysis performed as described in Box 1. An alternative strategy would be to choose different numbers of factors based on the cumulative amount of variance explained. For the sake of simplicity and as a proof of principle, in this analysis we chose to use five factors as they explained a substantial amount of the variance in expression without too large a multiple testing burden.

3.4 Pathway factor and phenotype association

Association tests were performed: i) between each pathway factor and chronological age, and ii) between single genes and chronological age using the linear mixed models defined in Box 1. These models have been implemented by the *lme4* package (Bates et al., 2014) in R (Computing, R Foundation for Statistical Vienna, 2008). For each phenotype a likelihood ratio test of the full model, which includes the age term, and the null model (without modeling age) produced evidence of an age effect. P values produced by this analysis were assessed for significance allowing for multiple testing using a Bonferroni adjusted threshold. A permuted dataset was created, which maintained the twin structure by permuting singletons, dizygotic and monozygotic twins separately and ensuring that twin pairs were kept together.

Significant associations between pathway phenotypes and age were further investigated to trace the particular genes within the pathway driving the signal. We report genes with a Bonferroni significant P value which accounts for the number of genes within the pathway that was tested.

3.5 Heritability analysis

To compute heritability, proportion of environmental variance explained by age, and the proportion of variance explained by unique environment, we fitted the full model from Box 1. Then the genetic component to variation was estimated as twice the additional correlation of MZ twins relative to DZ twins. The environmental component to the phenotype was the sum of the contribution from the fixed age effect, the random noise term, and the shared environmental component, again estimated from the difference between MZ and DZ. Estimates of these proportions are constrained to lie between 0 and 1 inclusive.

3.6 Single-gene based pathway enrichment analysis

We compared the significant pathways found by our factor analysis methods to those found by looking for enrichment of single gene associations with age. Firstly we tested each gene for association with age using the methods described in Box 1 and produced a list of Bonferroni significant genes $P< 0.05$ (this list contained 682 differentially expressed genes). For each pathway, we applied a Fisher's exact test to infer whether the proportion of significantly associated genes within the pathway was greater than would be expected by chance, which is estimated as being proportional to the pathway size. We also investigated whether using an FDR cut-off for significant age associations would produce more significant pathways or power would be diluted by including too many false positives. When re-running the analysis using a less stringent threshold (3,487 genes were associated with age with $\mathrm{FDR}< 0.05$) we found fewer significant pathways, and results correlated less well with the results of the factor based analysis (Spearman correlation of 0.36 ($P=5.1\times10-7$) compared to 0.49 for Bonferroni, $P=2.1 \times 10-12$). A complete list of all significant single gene age associations ($\mathrm{FDR}< 0.05$, 3,487 genes), with estimate of effect size and direction, can be found in Appendix Table A.3.

3.7 Results

The first stage of the analysis was to remove the effect of both known and unknown nuisance variables from the gene expression data. Using PEER software, we estimated five global factors which explained 35.7% of the variation in the complete gene expression data. As the aim of this analysis was to find pathway specific responses to ageing, we treated these global factors as nuisance covariates and regressed these out of the data, together with batch and RNA quality which are known experimental confounders. Data were then divided into subsets of genes within 186 KEGG pathways. For each pathway, five factors were estimated using PEER as described above, which explained on average 17.5% of the residual variation of all genes within this pathway after removing the global factors. For the 186 KEGG pathways, this produced 930 phenotypes which were tested for association with age. In total, 69 significant associations (P<5.38E-5, the Bonferroni adjusted threshold) from 57 distinct pathways were identified. The most significant 20 pathways are listed in Table 3.1, and a list of all 57 significant pathways can be found in Appendix Table A.2.

We also explored an alternative method for finding pathway related to ageing, looking for enrichment in the number of significantly associated genes falling into a particular pathway, analogous to the method used in the DAVID methodology (Huang et al., 2009). This discovered a total of 7 significant pathways (Appendix Table A.3). Thus, applying factor analysis methods to discover significantly associated pathways uncovered eight times as many hits. All pathways discovered by single gene enrichment methods were also discovered using factor analysis. There is strong concordance between P values discovered by the two methods (Spearman correlation = 0.49, P= 2.1 × 10E12). Figure 3.1 shows a Q-Q plot of P values for both methods against the theoretical P values under the complete null hypothesis. We see enrichment of significant P values for both methods, but this is not present when analysing the permuted data with factor analysis methods (green dots). This suggests that age plays a widespread role in the expression of these pathways, as enrichment of P values is not due to invalid model assumptions and can be observed using two different methods.

To investigate which genes drove the significant pathway associations, we exam-

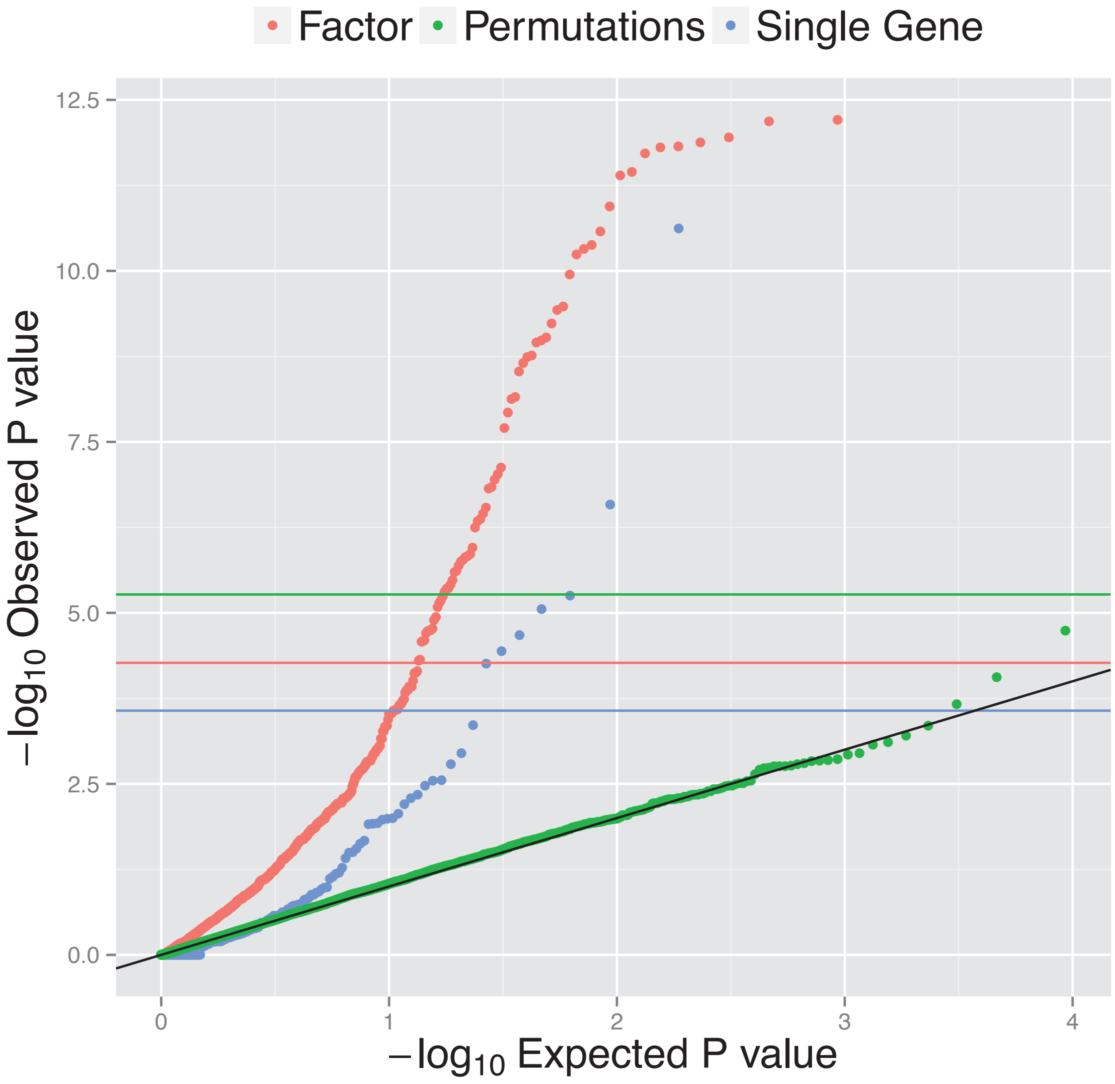

Figure 3.1: Q-Q plot of observed P values against theoretical P values for factor analysis (red dots) and single-gene based methods (in blue). Permutations (in green) shows the results of a combined analysis of 10 permuted datasets. Horizontal lines shown Bonferroni significance thresholds accounting for different numbers of tests (186 tests for single gene measures in blue, 930 for factor analysis in red, and 9300 for the combined 10 permutation analysis in green).

ined how many genes within a significant pathway showed significant age associations (Table 3.1 and Appendix Table A.2). On average 16% of genes within the pathways have P<0.05 after adjusting for the number of genes in the pathway, with a minimum of 1 gene and maximum of 24. The proportion is similar between pathways of different sizes, in contrary to the traditional pathway enrichment analysis, where there is bias towards large pathways.

Table 3.1: List of 20 pathways most significantly associated with age, together with the the number of significantly associated genes ($P < 0.05$, corrected using Bonferroni for the total number of genes in the pathway), the total number of genes, and the heritability of the pathway factor.

KEGG_ID	Pathway	P value of pathway factor	Number of significant genes	Number of genes in pathway	Heritability
00900	Terpenoid Backbone Biosynthesis	6.23E-13	6	13	0.00
00980	Metabolism of Xenobiotics by Cytochrome P450	6.47E-13	6	54	0.09
01040	Biosynthesis of Unsaturated Fatty Acids	1.11E-12	6	17	0.25
00100	Steroid Biosynthesis	1.33E-12	12	14	0.41
00650	Butanoate Metabolism	1.51E-12	8	27	0.39
04146	Peroxisome	1.56E-12	17	64	0.45
00830	Retinol Metabolism	1.93E-12	6	48	0.45
00010	Glycolysis Gluconeogenesis	3.59E-12	12	49	0.42
00051	Fructose and Mannose Metabolism	3.99E-12	8	32	0.32
00290	Valine Leucine and Isoleucine Biosynthesis	1.15E-11	3	11	0.00
00561	Glycerolipid Metabolism	2.63E-11	6	38	0.34
00620	Pyruvate Metabolism	4.20E-11	11	35	0.37
00770	Pantothenate and COA Biosynthesis	4.76E-11	4	16	0.48

00280	Valine Leucine and Isoleucine Degradation	5.79E-11	10	35	0.51
00020	Citrate Cycle TCA Cycle	1.12E-10	8	23	0.43
04916	Melanogenesis	3.34E-10	10	93	0.00
04910	Insulin Signalling Pathway	3.70E-10	13	122	0.45
00565	Ether Lipid Metabolism	5.89E-10	3	27	0.00
00350	Tyrosine Metabolism	9.44E-10	4	32	0.34
00640	Propanoate Metabolism	1.03E-09	6	26	0.59

Different KEGG pathways can contain overlapping sets of genes, as they can describe related biological function. Because of this, our significant associations with age for different pathways could be related due to a common underlying effect on a given set of genes. To explore whether the observed age-associations are unique to their pathway, or common to multiple pathways, we calculated the Spearman correlation between those phenotypes. There are 24 pathway phenotypes with a correlation greater than 0.8 with at least one other phenotype (Appendix Table A.4). These phenotypes frequently relate to metabolism, and form a highly connected set (Figure 3.2). We infer from this that there could be a common effect of age acting on all these phenotype factors.

We next explored how different sources of variation in the different phenotypes analysed here affect our ability to discover age associations. We calculated the heritabilities, the proportion of environmental variance explained by age, and the proportion of variance explained by the unique environment (Box 1) for i) KEGG pathways, ii) global factors (which we have treated as nuisance covariates) and iii) for individual genes (Figure 3.3, global factor histograms are not shown as there are too few phenotypes). The relative differences in sources of variation between global and pathway factors, and individual genes are shown in Figure 3.4. We see that as we move away from local phenotypes (individual genes) to pathway phenotypes and then to global

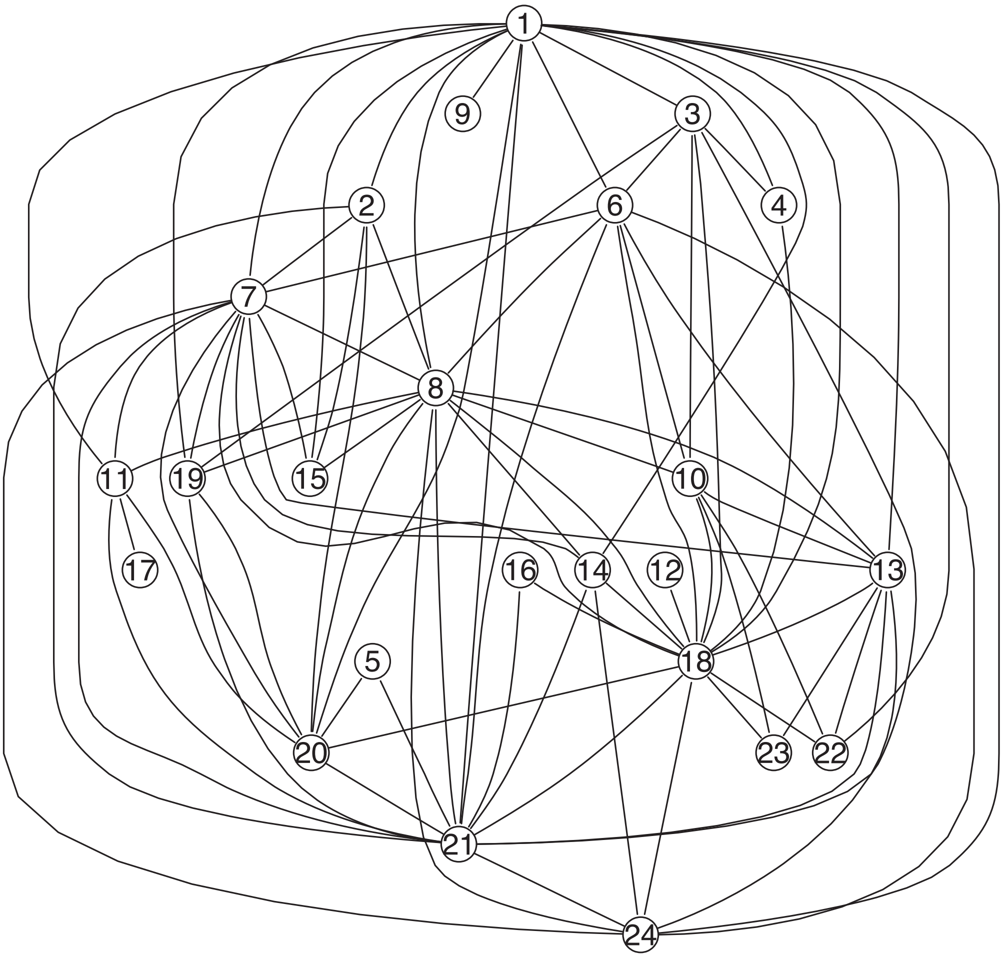

Figure 3.2: Network of connected factor phenotypes. Twenty four of the 69 age-associated factor phenotypes have a Spearman correlation of at least 0.8 with at least one other phenotype. These phenotypes show a highly connected structure, likely meaning there are common age effects driving these associations. A key for identifying which pathways correspond to the nodes can be found in Appendix Table A.4.

phenotypes, the proportion of variation explained by unique environment decreases. This is because that there is a stochastic component to each single gene's expression: by taking a weighted average of a number of genes, we average away this component. If all else were to remain constant, this reduction in stochastic noise would simultaneously increase heritability (as the total variance decreases), and boost the ability to discover associations with biological meaning, such as age. We see in the first panel of Figure 3.4 that the relative contribution of unique environment to pathway phenotypes is smaller than the contribution to genes. This also partly explains the results shown in the second and third panels: a greater proportion of variance is explained by age and genetic factors (heritability) for pathway factors than individual gene measurements.

When considering global factors, as expected the unique environment is greatly reduced. However, there is not a strong influence of ageing and heritability in this case is still moderate. This is likely because age and genetics do not act in a consistent way across large sets of genes. Leek and Storey, 2007 argued that global factors can capture experimental noise and batch effects. This is consistent with our findings. Heritabilities and proportion of variance explained by age for all pathways are reported in Appendix Table A.5.

We further looked for novel genetic associations with these pathway phenotypes, not seen as single gene expression associations. However, this was unsuccessful despite the increased heritability in pathway factors. This is likely due to the genetic architecture of gene regulation. Genes are regulated both in *cis*, where a nearby variant effects the expression of a single gene, and in *trans*, where a long range regulatory effect can hit multiple genes (Grundberg et al., 2012). The genetics of pathway phenotypes is a combination of *cis* effects on individual genes and *trans* effects, potentially affecting multiple genes in the pathway. However, *trans* variants typically have much smaller effect size: the increase in the reliability of pathway phenotypes is insufficient to compensate for the lower power to discover *trans* effects. Thus, the only associations discovered were when single genes loaded heavily enough on a pathway to indirectly reflect the *cis* association.

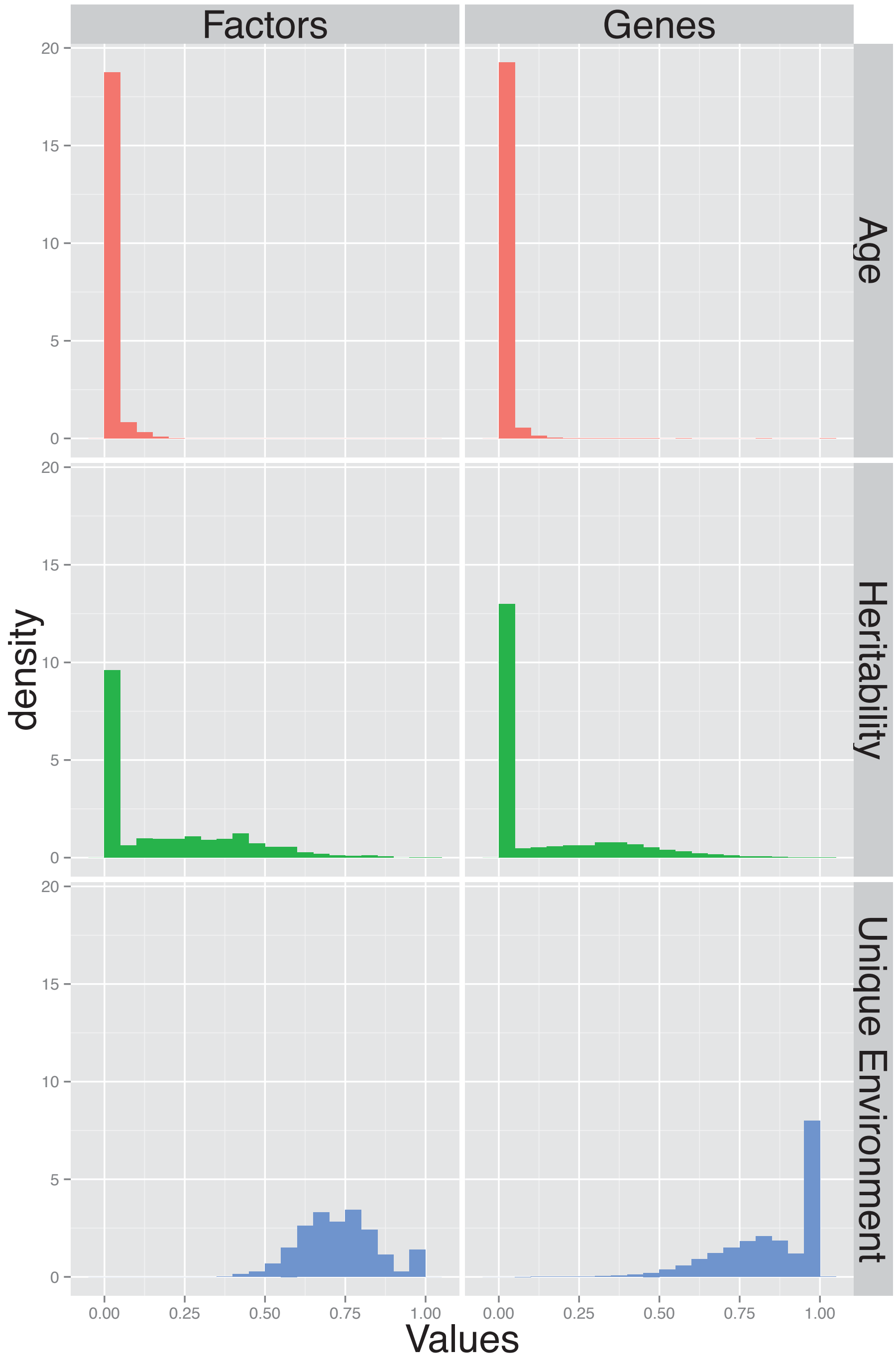

Figure 3.3: Histograms showing the proportion of environmental variation explained by age, heritability, and the proportion of variance explained by the unique environment for pathway factors and the individual gene measurements. The calculations correspond to equations in Box 1. Note that the proportions are not sum to one as they are not normalized by a same denominator: for age the variance explained by the genetic factors is removed.

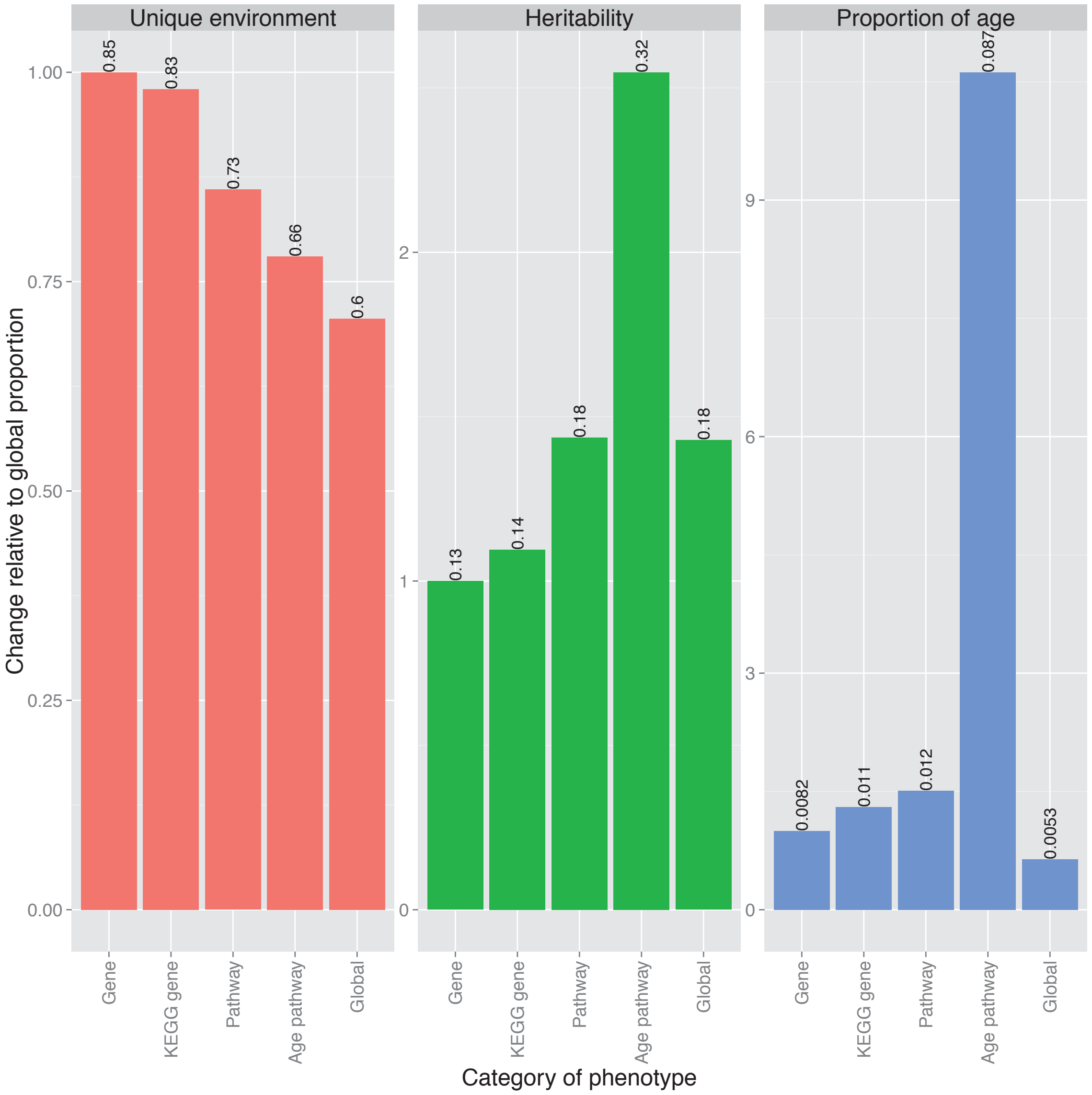

Figure 3.4: The relative importance of sources of variation to global, pathway and gene phenotypes. Measures of variation shown are the proportion of variance explained by unique environment, proportion of variance explained by genetics (heritability) and the proportion of environmental variation explained by age. The five categories are individual genes; genes that are in pathways annotated by KEGG; pathway factors; age associated factors and global factors. To show more clearly the differences in relative importance of these measures to different classes of phenotypes, all proportions are scaled such that contribution to gene phenotypes equals one. Numbers above the bars give the absolute, unscaled proportions.

3.8 Discussion

We have seen that both the heritability and the proportion of environmental variance explained by age is greater for pathway phenotypes than for individual genes. Consistent with this, a previous study found a greater proportion of associations for the pathway phenotypes than using single gene tests using this same dataset (Glass et al., 2013, 23% compared to 7% of phenotypes are significantly associated with age when using the same 0.05 FDR threshold adopted in that paper). This can be explained by our findings on the influence of unique environment on pathway phenotypes relative to single genes.

Stochasticity in gene expression, which contributes to the unique environment component that we measure, has been seen to increase with age. For example, animal model studies (Herndon et al., 2002; Bahar et al., 2006) have reported increased cell-to-cell variation in gene expression with age and tissue specific decline of functions associated to stochastic events. Others have found genes associated with longevity to be strongly regulated in older animals with low levels of stochasticity and higher levels of heritability (McCarroll et al., 2004; Vinuela et al., 2012). The aim of our analysis was to find mean effects, rather than variance effects (though both effects are often seen together). By reducing the unique environment variable component using pathway factor analysis methods, we arguably focus much more on a systematic longevity changes with age rather than the environmental stochasticity. However, it is difficult to make inference about causality with gene expression: we cannot know whether we are observing changes in expression which are driving the ageing process, or markers for it.

Of the 57 significant pathways, we frequently see four types of pathway, all of which have been previously linked with ageing: i) insulin signaling ; ii) sugar and fatty acid metabolism; iii) xenobiotic metabolism; and iv) cancer related pathways.

We find the insulin signaling pathway (hsa04910) to be highly associated with age in our data (P=3.7E-10). Much evidence has accumulated for the influence of the insulin signaling pathway on longevity, originating in *C. elegans*, where lowered insulin/IGF-1 signalling (IIS) can lead to a significant increase in life span (Friedman and Johnson, 1988). This effect has also been seen in the fruit fly *D. melanogaster* (Clancy et al., 2001) and in mice (Holzenberger et al., 2003). Outside of model organisms, it has been observed that variants in FOXO transcription factors related to this pathway can affect longevity in humans (Willcox et al., 2008), although its gene expression does not show significant association with age.

In addition to those related to insulin, our list of age-associated pathways includes many that are involved in metabolism or glycolosis. Examples of these include biosynthesis of unsaturated fatty acids (hsa00980), butanoate metabolism (hsa00650), glycolysis gluconeogenesis (hsa00010), fructose and mannose metabolism (hsa00051) and valine leucine and isoleucine biosynthesis (hsa00290) ($P \leq 1.15E\text{-}11$). It has previously been suggested that metabolism related pathways play roles in ageing and ageing related diseases(Barzilai et al., 2012). In particular, Houtkooper et al. (2011) showed that glucose and compounds involved in the metabolism of glucose were biomarkers of ageing in liver and muscle tissue in mice.

Other ageing related pathways include those involved in the metabolism of xenobiotics allow cells to deactivate and excrete unexpected compounds. One example is glutathione metabolism (hsa00480, P=1.45E-7), a well known anti-oxidant which protects against cell damage by reactive oxygen species (Pompella et al., 2003).

Finally, similarities between cancer and ageing have been noticed (Finkel et al., 2007). For example, cellular senescence, when a cell loses the ability to divide, can form a break on cancer development; clearing such cells can delay the development of age-associated disorders (Baker et al., 2011). There are a number of pathways in our list that have been linked to cancer, in particular skin cancer, possibility because this was done using skin tissue. These include melanogenesis (hsa04916, P=3.34E-10), the PPAR signaling pathway (hsa03320, P=1.83E-9), the hedgehog signaling pathway (hsa04340, P=1.12E-7) and glioma (hsa05214, P=4.26E-7)

In addition to age, other phenotypes have been linked to expression patterns of multiple genes. For example, BMI has been linked to expression patterns in adipose tissue of multiple genes within a group which share a common *trans* master regulator, and such phenotypes could mediate between expression and diseases such as type 2 diabetes (Small et al., 2011). Principal components and factor analysis has also been suggested as a way to build classifiers for binary traits (Hastie et al., 2000), perhaps to predict prognosis of disease from gene expression data. The ability of pathway phenotypes to provide reliable measures of expression with direct biological interpretation means they could also be applied in both situations, to understand the relationship between expression and such phenotypes.

Our analysis shows that factor analysis applied to gene expression data effectively reduces stochastic noise in summaries of gene expression patterns, giving more power to discover associations. These phenotypes are substantially more heritable than individual genes. Using them we can improve our ability to identify biological processes underpinning ageing. This is consistent with the idea that removing latent factors that exert broad effects on gene expressions increases power in associations. We show that the same idea can be used to create pathway factors that are robust and interpretable. Finally, our analysis reveals pathways that have been seen to be important in longevity from a number of previous studies, as well as novel pathways that can be further investigated.

Chapter 4

Measuring telomere length from sequence data

Collaboration Note. *The method developed in this work was designed by Richard Durbin and implemented and evaluated by myself. The study uses data collected from the TwinsUK cohort.*

4.1 Overview

Telomeres cap the ends of chromosomes and are critical for the maintenance of genome integrity. In humans, telomeres comprise sequences of 5-15kb TTAGGG tandem repeats and their telomere binding proteins (Samassekou et al., 2010). In the absence of telomerase or the alternative lengthening pathways (Henson et al., 2002), telomeres undergo progressive attrition, which ultimately leads to replicative senescence or apoptosis. Thus, telomere length is an indicator of replicative history and replicative potential — two features of great importance to human health and disease (Blasco, 2005).

Standard methods for telomere length measurement are generally classified into three categories: (i) Southern blot analysis of the terminal restriction fragments that measures the average length (mTRF) and length distribution of telomeres in a sample of cells (Kimura et al., 2010); (ii) methods that examine variation in telomere length between chromosomes and cells, i.e., fluorescence in situ hybridization (FISH) techniques, including Q-FISH (Martens et al., 1998) and Flow-FISH (Baerlocher et al., 2006); and (iii) quantitative PCR (qPCR)-based techniques that measure telomere DNA content in relative units (compared to single gene DNA) (Cawthon, 2009).

Next-generation sequencing has now provided an opportunity to obtain genomic information cost effectively in large scale. Shotgun sequence data contains sequencing reads from the telomeres just as any other region of the genome. However, little information about the telomeres can be gained from standard alignments of these reads to the reference sequence. This is because the repetitive nature of the telomeric regions means that it is not possible to assign with confidence the exact origins of the reads, and also because in the human reference sequence (build GRCh37) the ends of most chromosomes are simply stretches of Ns, representing unknown nucleotides.

Instead, previous studies (Castle et al., 2010) have shown that information on telomere length is contained in the number of telomere motif copies (TTAGGG or CCCTAA) found in reads. Parker et al. (2012) applied this idea to cancer samples. However, cancer samples typically suffer from aneuploidy, complicating the validation of their results by method such as qPCR (it relies on normalising against a unit copy region). This may be the reason why the measures in Parker et al. (2012) only converge to a low resolution telomere status, defined as either gain, no change or loss relative to normal control. Additionally, the vast majority of the samples were pediatric with mean age 7.5 years, and they did not demonstrate a relationship between age and their sequence-based telomere length measurement.

Here, we further examine the relationship between reads containing telomere repeat sequence and telomere length, and describe software for estimating telomere length based on genome-wide sequence data. We demonstrate our method on 260 leukocyte samples (aged 27 -74 years, mean age 51 years) from the TwinsUK cohort (Moayyeri et al., 2013b) that have both Illumina 100bp paired-end whole genome sequence and telomere length measurements using Southern blot mTRFs. We also investigate 96 samples from the 1000 Genomes Project (The 1000 Genomes Consortium, 2010) that have both whole genome and exome data.

4.2 Study samples and data

The 260 UK10K individuals investigated in this study were all female aged 27 - 74 years (mean age 51 years) from the TwinsUK cohort (Moayyeri et al., 2013b, http://www.twinsuk.ac.uk/). Except for 5 pairs of dizygous twins, the rest were all unrelated. Leukocyte telomere lengths of these individuals as mTRFs were measured using Southern blot. Whole genome sequencing was conducted using the Illumina HiSeq technology, yielding sequencing reads with coverage ranging from 4X to 16.6X (average 6.5X, pooled across lanes). Twelve individuals with a much higher read duplication rate (more than 3 fold that of other samples) were excluded from the rest of the analysis since they gave outlier results (Figure 4.1).

Sequence data are available from the European Genome-phenome Archive (EGA) study number EGAS00001000108, submitted by UK10K (http://www.uk10k.org). The 1000 Genomes Project sequence data were downloaded from http://www.1000genomes.org.

4.3 Estimating telomere length from whole genome sequence data.

4.3.1 Estimator

We first examined the frequency of reads from the TwinsUK dataset with different numbers of copies of TTAGGG and also each non-cyclical permutation of TTAGGG as a control. The frequencies of all non-TTAGGG hexamers showed a monotonic decay as the number of repeat units increased, with none occurring in a read more than eleven times (Figure 4.2). In contrast, beyond seven repeats there was an increase in the number of reads containing TTAGGG. We defined reads as telomeric if they contained$k$ or more TTAGGG repeats, with a default threshold value of $k = 7$, values higher than which do not increase performance substantially. These can then be translated into an estimate of the physical length via a size factor s and a constant length c in $l = t_k g/(46s)$, where l is the length estimate, t_k is the number of telomeric reads at threshold k, g is the genome length and s is the total number of reads. The

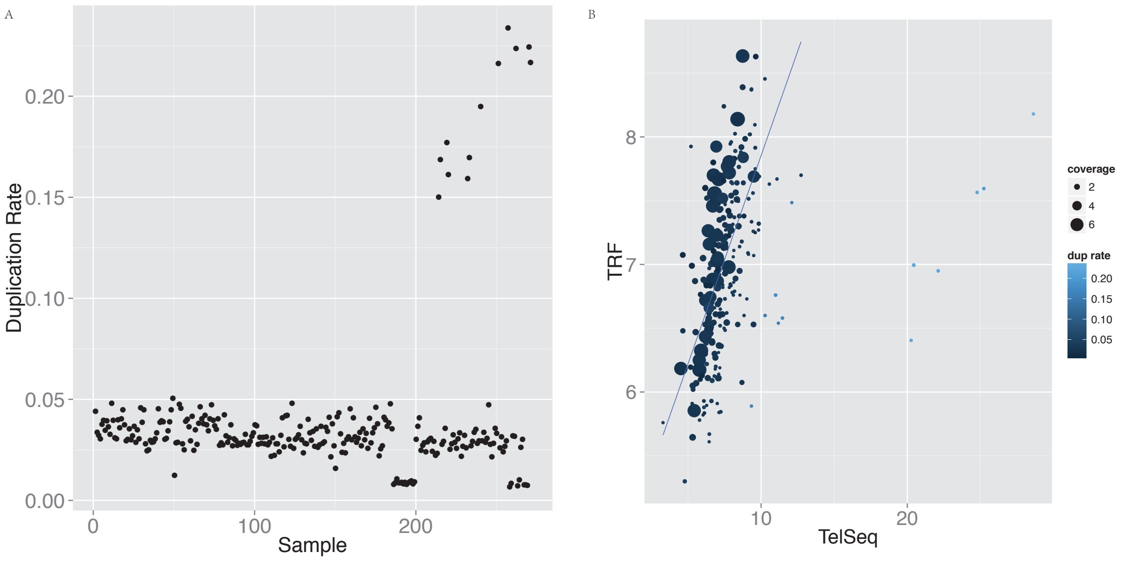

Figure 4.1: The effect of duplication rate and coverage to TelSeq performance. In essence, TelSeq relies on sampling of genomic regions from a sequencing library. Coverage and duplication thus affect the translation of a relative measure into an absolute one. Low coverage indicates insufficient sampling and thus results in high variation in estimation (Figure 4.4) while high duplication suggests over enrichment of certain genomic regions and thus changes the translation factor c. In whole genome sequencing high duplication rate indicates low library complexity and loss of information. Twelve of our samples were found to have an exceptionally high duplication rate (>3 fold greater than the rest, panel A), and were outliers when regressing against mTRF (panel B). We based our evaluation on samples with duplication rate below 10%, which is typically what is expected for whole genome sequencing.

factor of 46 corresponds to number of telomere ends 46(23×2).

Studies have shown that DNA molecules in a sequencing library are not sampled and sequenced with equal probability, but instead are subject to biases due to different molecular properties such as GC composition - a high value of which favors more amplification in the PCR step (Dohm et al., 2008). This results in different representations of genome regions and makes defining s as the total read number not a good estimate. Instead, we define s as a fraction of all reads within a specific GC composition range, and similarly g as the length of genome for which 100bp segment lie within the same GC range. The range was chosen to be close to the telomeric GC composition, which is 50% at the TTAGGG dense regions (see Figure 4.3 for results for other GC composition ranges).

Considering the GC composition removed an important source of experimental error; and effectively increased the signal by nearly two-fold, as measured by the correlation between experimental estimates (Figure 4.3). This method is implemented in a program TelSeq which reads one or more BAM files and returns a report with one row per read group present in the input.

To calculate g we divide the reference sequence into 100bp consecutive bins and add 100bp to g if the GC composition of the bin is within the range.

Association to age and mTRF The Pearson's Correlation Coefficient was calculated using the *cor* function of the R language (Computing, R Foundation for Statistical Vienna, 2008, http://www.r-project.org/). The regression between age and TelSeq and between age and mTRF was calculated using the *lm* function of R in models *lm(age ~ telseq)* and *lm(age ~ mTRF).* Two measures were also included in one model *lm(age~telseq + mTRF)* as two independent fixed effects. A t-test was done for each of the two regression coefficient (β) against null hypothesis $\beta = 0$, the results of which can be seen in the output of the summary function.

Calculating the variance explained To compute the proportion of variance of age explained, we used the *cor* function in R *cor(age, mTRF, method="pearson")^2.* To compute the additional variance that can be explained by mTRF while controlling

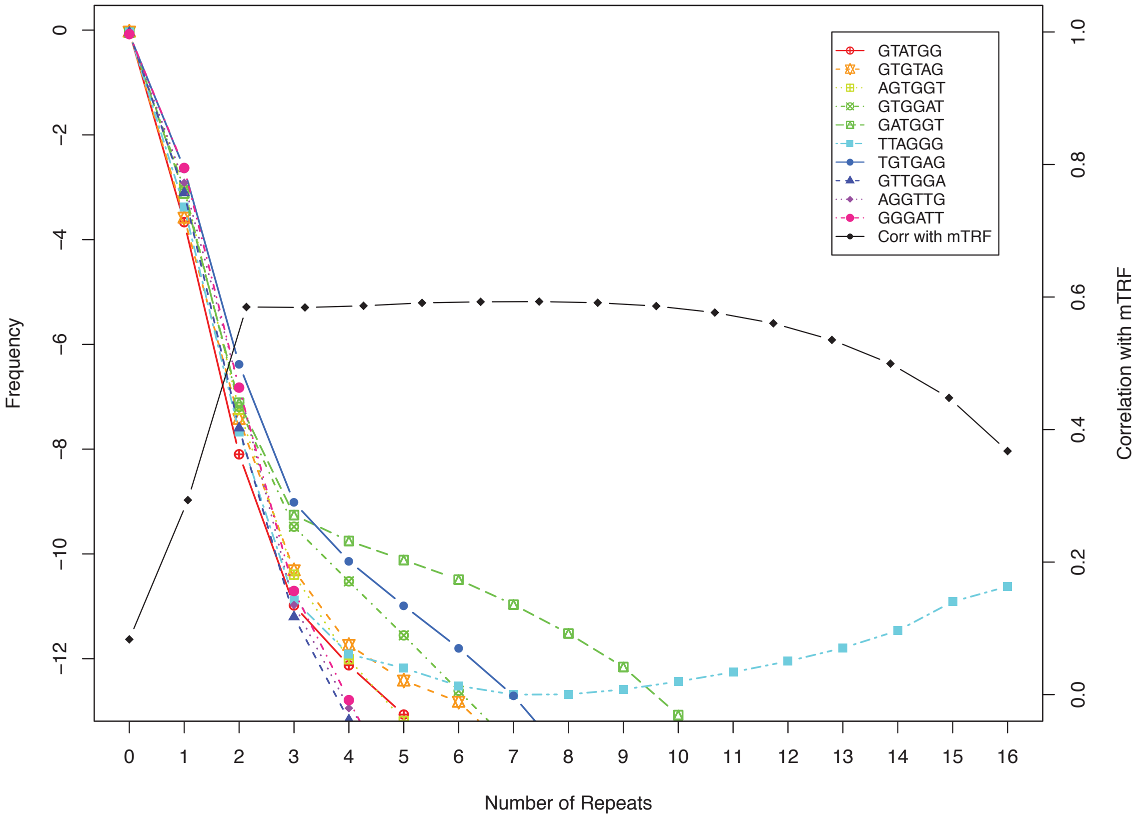

Figure 4.2: Identification of telomeric reads. In cyan the log scale frequencies of reads with different numbers of TTAGGG repeats averaged across the 260 TwinsUK samples, with corresponding plots for permutations of TTAGGG in other colours. In black the correlation of TelSeq to mTRF as a function of the threshold k for the number of repeats per read used in the TelSeq measurement.

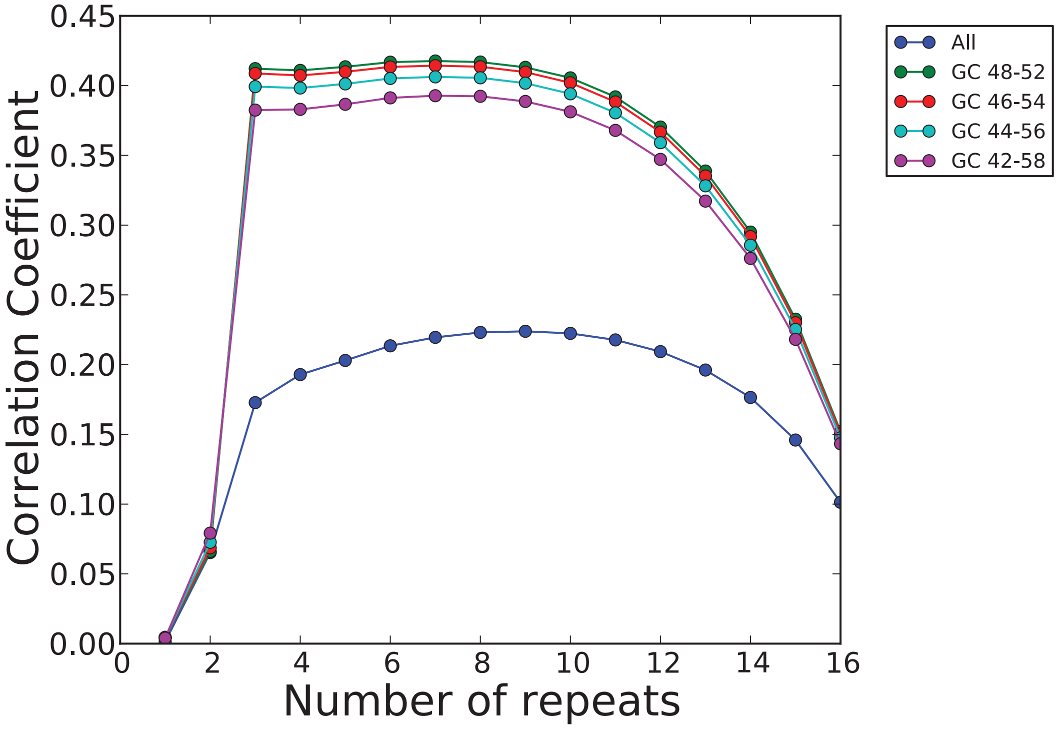

Figure 4.3: Normalising by reads with similar GC improves the performance of TelSeq. It is known that read abundance in a sequencing library is affected by the GC composition of a read, a bias primarily introduced in the PCR step where high GC reads get amplified more often due to their high molecular affinity. Thus, using reads with similar GC content as background accounts for this molecular property and reflects the signal to noise ratio more accurately. To demonstrate this we evaluated the performance of TelSeq, as measured by the correlation with mTRF, when normalised by reads from different GC groups, 42%-58% (purple), 44%-56% (light green), 46%-54% (red), 48%- 52% (dark green) as well as by all reads (blue). The result showed that there was a gradual increase to the correlation when GC range approaches 50%. And in all these cases, the correlation was much higher than that when all reads were used from a library. Here the analysis was done for the whole range of threshold k, the number of TTAGGG repeats in a read.

for TelSeq, we firstly obtained the residuals from a regression between age and TelSeq (*x<-lm(age~TelSeq)$residuals*); and then used the residuals to compute the additional variation explained (*cor(x,mTRF)^2*). The same procedure was done for TelSeq.

4.3.2 Simulation

We employed simulated datasets to investigate the effect of sequencing coverage. This was also to discover the minimum amount of sequence required for reasonable length estimation. We chose the reference sequence (GRCh37) of human chromosome 1 as the sequence source, but with 30kb nucleotides (including unknown nucleotide Ns) removed from each end and replaced with telomere repeat sequences (TTAGGG) of the same length. We then simulated Illumina pair-end reads using the software SimSeq (https://github.com/jstjohn/SimSeq, parameters -1 100 -2 100 –insert_size 500 –insert_stdev 200) with sequencing coverage in individual BAMs varying from 0.2X (498,501 reads) to 10X(24,925,063 reads) in 0.2X increments (Figure 4.4). For each setting we repeated the simulation 5 times and generated 255 BAMs in total. We then applied TelSeq to estimate telomere lengths of these BAMs. TelSeq predicted a length of 29.4kb on average with 1.47kb standard deviation (5% of mean). Significant higher variation was seen when coverage was below 2.5X (F=10.5, P=2.2E-16 in the F test) when compared to results from the higher coverage BAMs (Figure 4.4). For BAMs with >2.5X coverage, TelSeq predicted telomere length to be 29.5kb with 0.71kb standard deviation (2.4% of mean).

4.3.3 Results

When TelSeq was applied to the TwinsUK data, the estimates of leukocyte telomere length (LTL) correlated well with the mTRFs measurements across a range of choices of k, with correlation $\rho = 0.60$ at the default threshold $k = 7$ (P<10E−16; Figure 4.5A). We next examined the relationship between the TelSeq-based LTLs and age of the donors. Given the wide inter-individual variation in LTLs for persons of the same age and the impact of environmental factors on this parameter, the correlation between LTL measurements and age in cross-sectional studies, including TwinsUK,

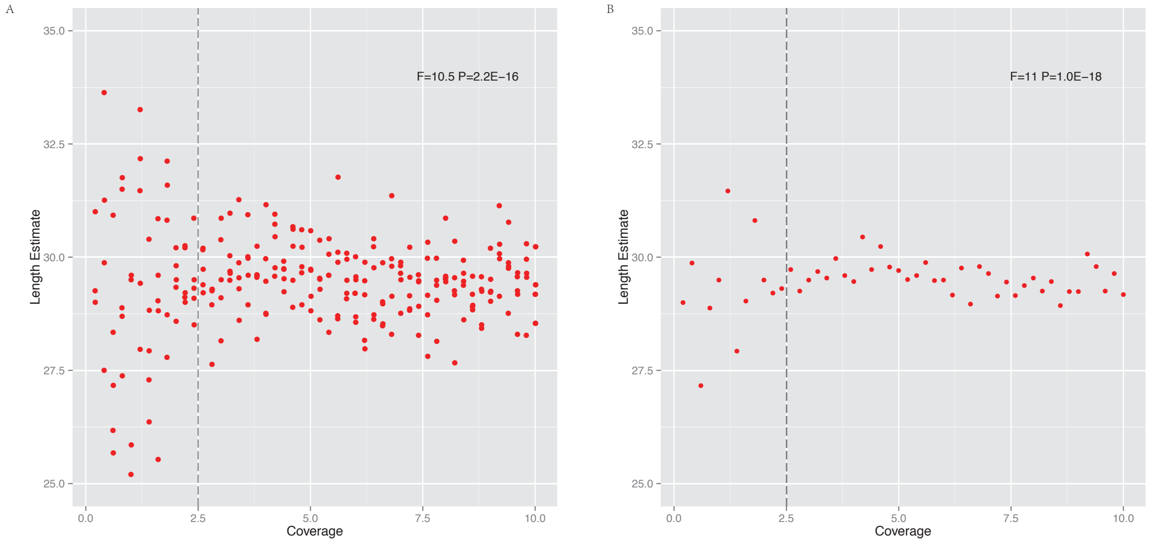

Figure 4.4: The effect of sequencing coverage on TelSeq measurement, assessed by simulation. A group of BAMs were simulated using software SimSeq (https://github.com/jstjohn/SimSeq). Sampling noise is substantially higher when the coverage is below 2.5X (mean=29.4kb, variation=5% of mean), compared to when coverage is above 2.5X (mean=29.5kb, variation=2.4% of mean) (**A**). The mean estimates are close to the true value 30kb independent of coverage. When using the weighted average of 5 BAMs for each coverage group (**B**), the variation is much smaller (1% of mean). This is justified theoretically by the relationship $X \sim N(\mu, \sigma^2)$, $\bar{X} \sim N(\mu, \sigma^2/n)$, where n is the sample size. In real experiments, ideally estimates should be obtained from multiple libraries across multiple lanes for a sample. The coefficient of variation across lanes per sample is on average 3.2% (Figure 4.7).

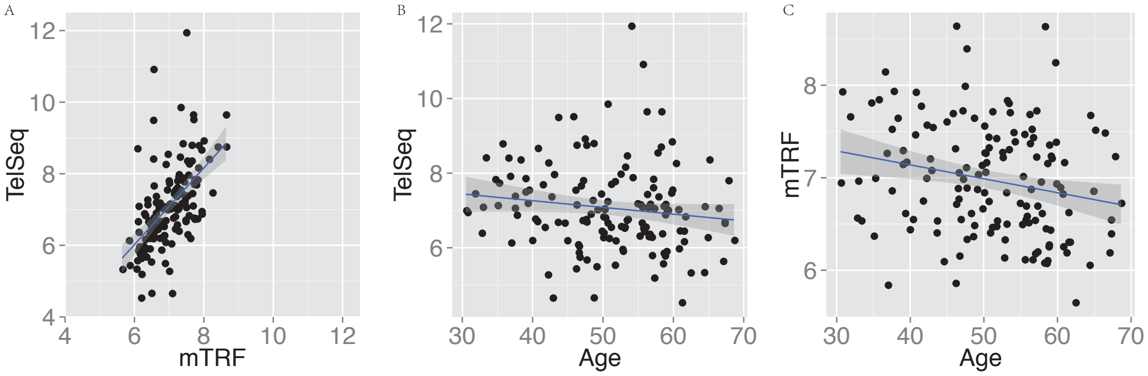

Figure 4.5: Comparison of TelSeq with experimental measure and age in TwinsUK samples. (A) TelSeq estimate of average telomere length plotted against mTRF estimate; TelSeq (B) and mTRF (C) estimates plotted against age. All average length estimates in kilobases and ages in years.

is usually modest (Valdes et al., 2005; Broer et al., 2013). Nevertheless, since the relationship between measurement and donor age depends on the true LTL value, the correlation provides a means for independent assessment of the informativeness of different experimental techniques for estimating LTL. The TelSeq measurement displayed correlation of ρ=-0.24 (explaining 6.5% variance of age, Figure 4.5B) with age, comparable to that of mTRF (Figure 4.5C; ρ=-0.26, explaining 7.5% variance of age). The difference between -0.24 and -0.26 is not significant in a *t*-test using a standard deviation derived by bootstrapping (P=0.79, Figure 4.6). The coefficient of multiple correlation between age and both LTL and mTRF was higher than either individual correlation (ρ=- 0.34, explaining 9% variance of age); both measurements contributed significantly to the underlying linear regression model, (P=0.016, *t*-test for the TelSeq term; P=0.009, *t*-test for the mTRF term). This implies that neither TelSeq nor mTRF captured all the information available, and that TelSeq contains additional information independent from that provided by mTRF.

Comparing the correlation coefficients with age by the two methods To test whether the difference is significant in the strength of associations between age and each of two measures, ρ = -0.24 for TelSeq and ρ = -0.26 for mTRF, we conducted bootstrapping using R (*sample(sample_index,sample_size,replace=TRUE))*) sampling our cohort 1000 times, from which we obtained an estimate for the standard deviation of ρ for mTRF (0.052) and TelSeq(0.056). We can then compute the t statistic $t = (\rho_{telseq} - \rho_{mTRF})/sqrt(s^2_{telseq} + s^2_{mTRF})$ for hypothesis testing (Figure 4.6).

Coefficient of variation A subset of our samples were sequenced on multiple lanes in separate runs. They can be considered as technical replicates and used to assess the variability of TelSeq measures. The coefficient of variation (CV) was computed as the ratio of the standard deviation (SD) to the mean across the technical replicates for each sample. We selected 110 samples that were sequenced on more than ten lanes to evaluate the CV and observed an average value of 3.17% with 0.98% standard deviation (4.7), comparable to or smaller than that from the experimental measurements (Kimura and Aviv, 2011).

Interestingly, when lanes analyzed separately and the telomere length estimate calculated as the mean across lanes, weighted by lane yield, the sampling error was further reduced and the correlation with mTRF was stronger (ρ=0.62 with mTRF when merged as opposed to ρ=0.60).

Difference in length estimates Notably, the TelSeq estimate of telomere length was consistently shorter than the mTRF estimate(mean 5.63kb compared to 6.97kb), and the mean rate of shortening per year was consistently greater (34.5bp/year against 19.8bp/year) (Figure 4.5B, Figure 4.5C). The mTRF measurements reflect the average distance from a restriction enzyme site (HinfI/RsaI or HphI/MnlI) to the end of a chromosome, and hence overestimate the canonical region of the telomeres of TTAGGG repeats only. Kimura and Aviv (2011) obtained a similar figure of around 1kb for the additional sub-telomeric length included in an mTRF measurement. The difference between the TelSeq and mTRF estimates changes as the TelSeq threshold k changes, reflecting inclusion of different amounts of subtelomeric sequence (Figure 4.8); although the correlation between TelSeq and mTRF remains similar across a range of values of k (Figure 4.2).

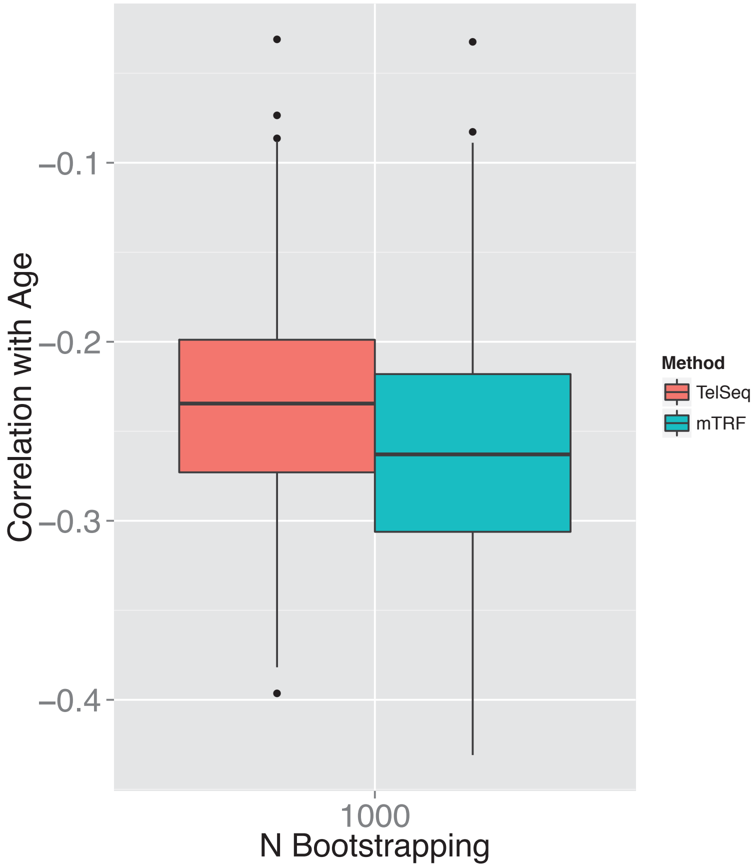

Figure 4.6: Compare correlation coefficient obtained from mTRF and TelSeq. To compare the correlation coefficients between age and telomere length estimates from TelSeq and mTRF, we conducted 1000 bootstraps with replacement from the data set to obtain an estimate of the standard deviations of the correlation estimates ρ. We can then perform a t-test for whether the difference between the observed values -0.24 and -0.26 is significant. The result gave t=0.26, P=0.79, which suggest no statistical difference between the coefficients obtained from the two measurements.

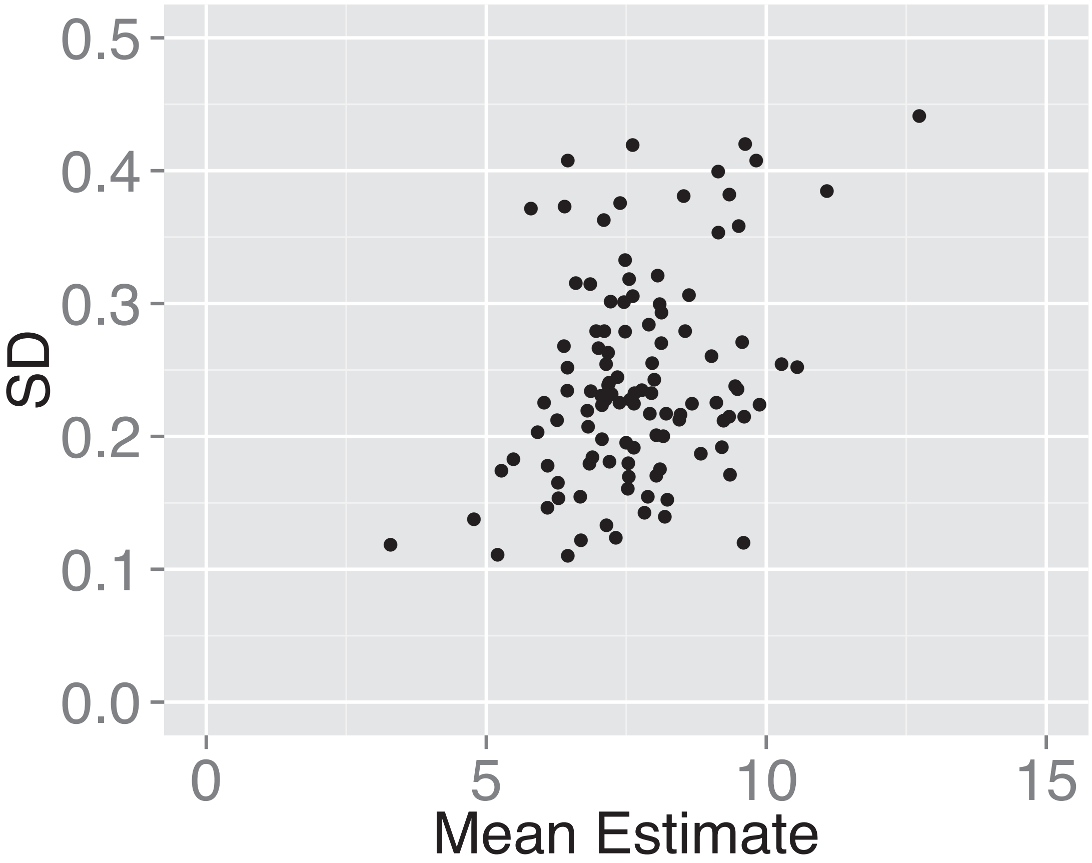

Figure 4.7: Sequencing lane variation in TelSeq measures. For each sample that was sequenced on more than ten lanes, the standard deviation of the length estimates across lanes is plotted against the mean length estimate. The coefficient of variation (CV), defined as the ratio of the standard deviation to the mean, varies between 1.3% and 6.4%, with mean 3.17% and standard deviation 0.98%.

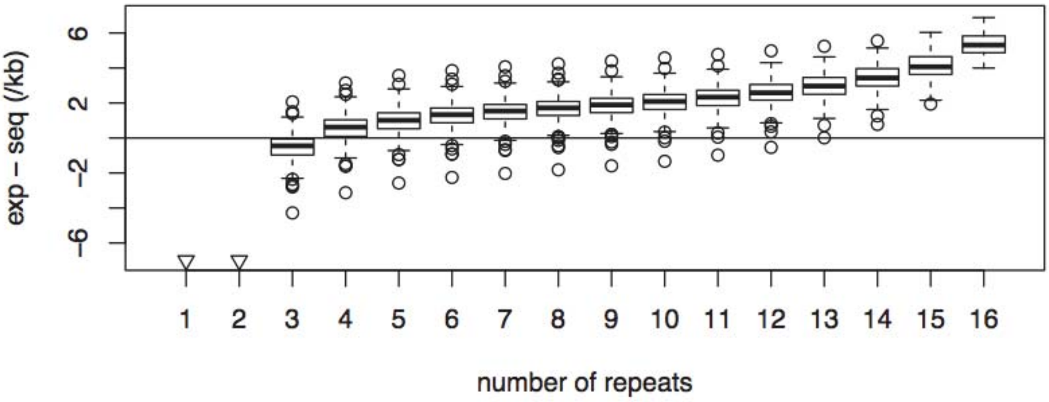

Figure 4.8: The mTRF measurement is longer than TelSeq estimates across a range of values for the choices of TelSeq threshold (k). The difference between mTRF and TelSeq is 1.49kb at k=7, and 5.34kb at k=16. The difference reflects the fact that mTRF measures the average distance from subtelomeic regions, where the excision sites of restriction enzymes exist, to the chromosome ends, while TelSeq approaches include only the ends when choosing a large k. Measurements of two methods correlate with age similarly, suggesting they both capture the information of telomere shortening with age.

4.4 Estimating telomere length from exome sequence data.

In addition to whole genome sequence data, a large number of samples have exome sequence data collected by enrichment of whole genome shotgun sequencing libraries using capture reagents. In theory, if the exome capture works perfectly, it would not be possible to use these data for our method. However, in practice with current technology a typical exome sequencing output contains some fraction (typically 10-50%) of sequence that is off-target, i.e. not exonic. This fraction represents information on the rest of the genome and can be used to estimate relative telomere length by our method. To test this approach, we selected 96 samples from the 1000 Genomes Project pilot that have matched whole genome and exome sequence and applied TelSeq to both data sets. We found that when we classify telomeric reads as those containing more than three TTAGGG hexamers, estimates of telomere length from the two data sets started to be tightly correlated (Figure 4.9). Using our default threshold of k=7, the two measures have a Spearman's Rank correlation coefficient 0.78. This result suggests that TelSeq can effectively work with exome data, which substantially extends its potential applications.

4.5 Applications of the method

Mutations in POT1 gene predispose to melanoma Robles-Espinoza et al. (2014) performed exome sequencing on pedigrees recruited in the UK, Netherlands and Australia with melanoma cases looking for variants that are explanatory to the disease. Four loss of function variants in the protection of telomeres 1 gene (POT1) were identified as cosegregating with melanoma cases in family UF20 (See Figure 4.10A for the pedigree with melanoma cases (arrowed) and missense mutations in POT1 at p.Tyr89Cys). The mutation disrupts the interaction between POT1 and single-stranded DNA and led to elongated telomere length (Robles-Espinoza et al., 2014). Telomere length information is thus an important phenotype to this study.

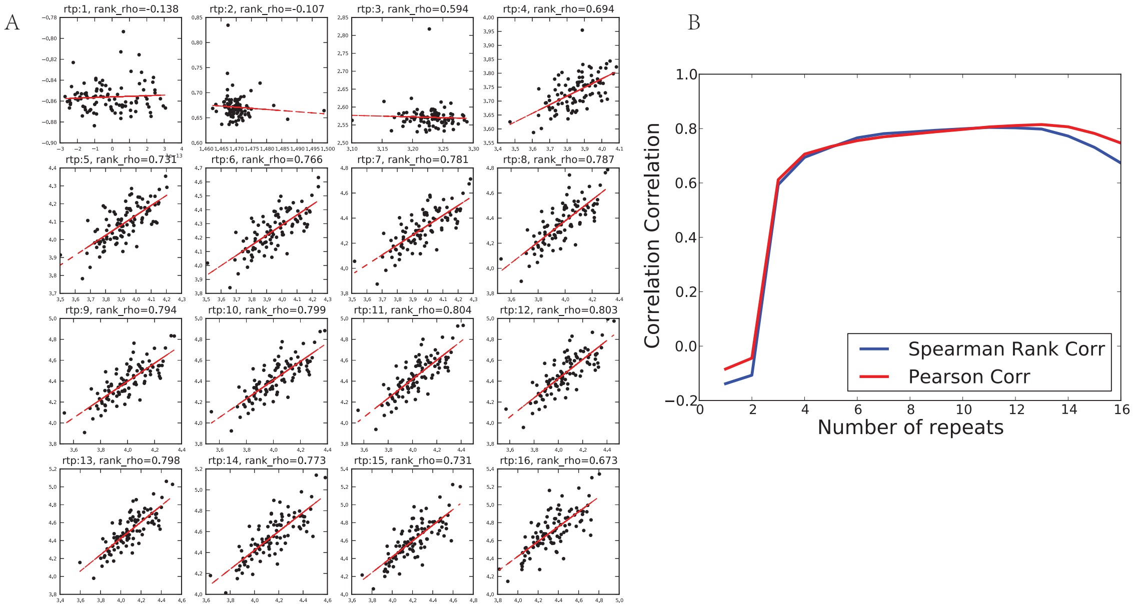

Figure 4.9: TelSeq estimates from exome data are highly correlated with those from whole genome data in 96 samples from the 1000 Genomes Project with matched whole genome sequences and exome sequence data. A. Scatter plots for TelSeq estimates from matched whole genome sequence and exome sequence at different thresholds of k, the amount of TTAGGG repeats in a read. Panels are organised from left to right, top to bottom as k increases from 1 to 16, where in each plot X axis is the estimates from the whole genome sequences and y axis is the estimates for the matched exome sequences. A correlation coefficient is calculated for each panel and plotted in B. The two measurements start becoming tightly correlated with each other when $k>= 3$.

Telomere lengths of the cases along with 38 controls that have wild type POT1 gene were measured using the qPCR method (Figure 4.10B) and Telseq (Figure 4.10C). Two methods show consistent signal that the cases with mutated POT1 gene have much longer telomere than the controls ($P < 0.00019$).

4.6 Conclusion

In conclusion, we have demonstrated an approach for measuring telomere length using whole genome or exome sequencing data. This is the first study to our knowledge to evaluate in detail the relationship between the frequency of telomere repeat sequence in shotgun sequence data and telomere length, and also to validate extensively with experimental measurements in a representative large sample cohort with a wide range of ages. There are some limitations to TelSeq, such as it is not able to obtain individual telomere length for chromosome arms. Nevertheless, Telseq allows any cohort with existing genomewide sequence data, including increasingly many cancer genomics and epidemiological cohort studies, to produce a validated measure of the average telomere length at effectively no cost, with no need for the further sample collection and experimental procedures required by other methods of ascertaining telomere length.

4.7 Software implementation

Telseq is implemented in C++. It uses BamTools (Barnett et al., 2011) to read BAM files. The source code is licensed under GNU General Public License Version 3 and is freely available online (https://github.com/zd1/telseq). To compile, a recent version of GNU Compiler Collection (GCC) is recommended (Version 4.8 or above).

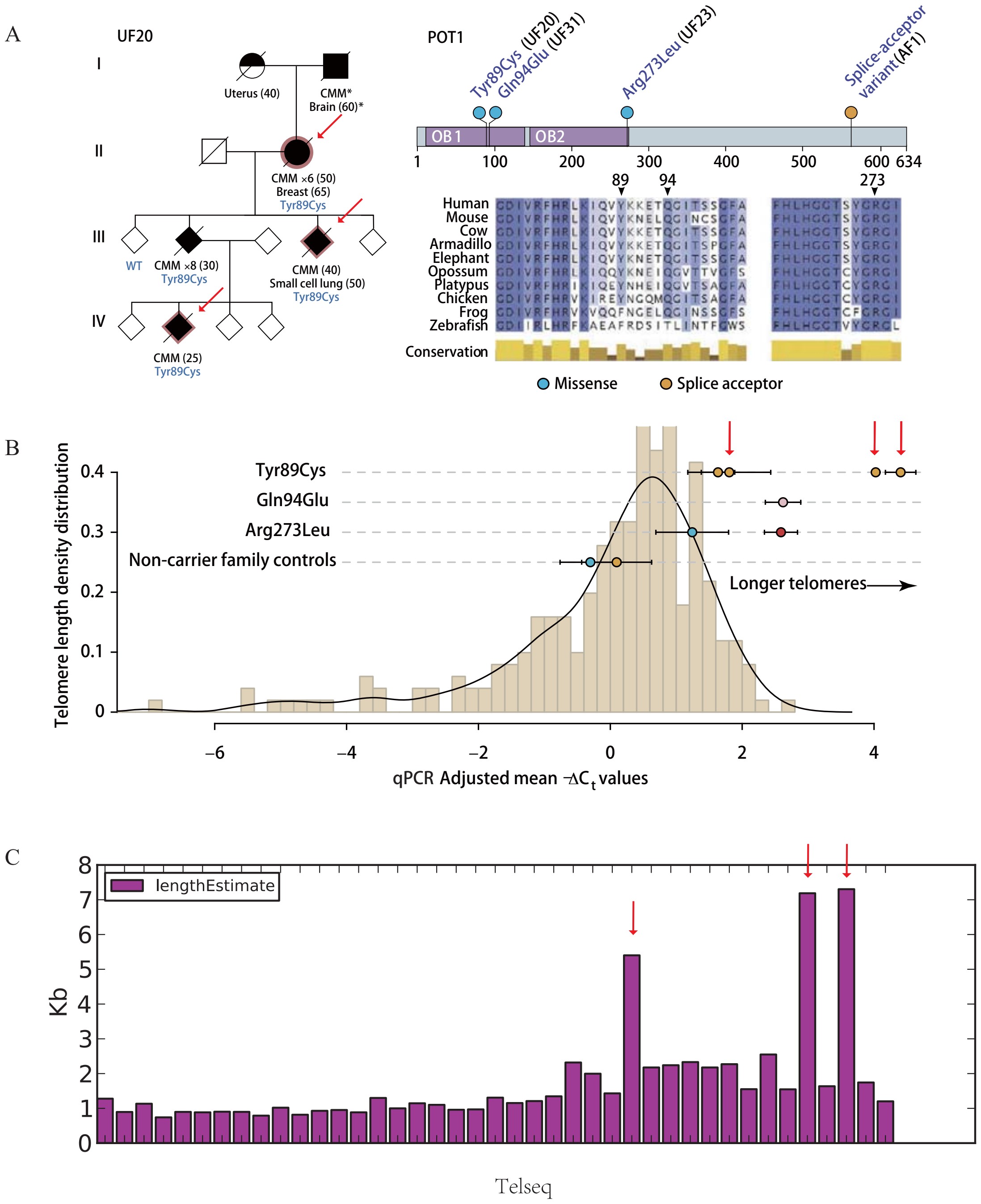

Figure 4.10: Measuring telomere lengths in melanoma cases. Mutations in the Protection of Telomeres 1 gene (POT1) were found transmitted in melanoma cases in pedigree UF20 (A). The telomere length estimates were obtained independently using a qPCR approach and Telseq (B and C). The cases that red-arrowed and compared against controls that have wild type POT1. Both methods indicate longer telomeres in the three cases. Panel A and B are adapted from Figure 2 in Robles-Espinoza et al., 2014.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The application of association genetic methods to cellular traits is an important area of research with great potential to reveal new information about cellular functions, and help interpret the genetics of diseases and other whole organism traits. This thesis contributes to the understanding of how DNA variation regulates transcription factor binding by investigating a key transcription factor CTCF. In chapter 2, I described a study that performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) for lymphoblastoid cell lines collected from 51 HapMap individuals. We have identified tens of thousands of bindings sites with the vast majority showing large inter individual variations. Most of the binding sites are identified with a matched CTCF canonical motif, reflecting sequence specificity of the transcription factor. To reveal the genetic contribution to such variation, we performed QTL analysis using a linear additive model, and focused on *cis* regulation within 50kb genomic window. Our results suggest a strong genetic basis for variation at many binding regions. The distributions of the physical locations as well as the corresponding effect sizes of the QTL variants show clear tracking on the information content with the motif, which suggests that mutations at the DNA-protein binding interface exert a functional impact on binding proportional to the predicted binding affinity of the nucleotide. It also shows strong correlation with sequence conservation, which supports the functional impact of the variants. Interestingly, there are binding regions whose variation of binding intensity appears to be only attributable to variants that are distal to the sequence motif. We recognized that in a substantial fraction of these cases there are variants within the motif but with low frequency (<5% MAF). Particularly, these low frequency variants are often in linkage disequilibrium with the lead QTL variants that are distal, indicating a phenomenon that the lead QTL variant is actually tagging a putative functional on-motif variant but with lower frequency. Taken this together, we show that the majority of CTCF binding QTLs have genetic regulatory variants in close vicinity.

Also in this work, we suggested a novel discovery of CTCF binding patterns on the X chromosome (See Ding et al., 2014 for details). Particularly we showed there exist three different types of binding in regions on the X chromosome: ones that only bind on the inactive X; ones that bind on both X; and ones that only bind to X in females. Also CTCF binding shows much stronger correlations between nearby sites on X than on autosome, suggesting important roles of CTCF in X inactivation. Together these suggest that CTCF maybe involved in a large scale chromatin remodeling associated with inactivation, although the specific functions of the different types of sites are yet to be found.

The resulting data from this study also contribute to the growing cellular information accumulating around the HapMap cell lines. Currently most public data on transcription factor binding are on a handful of samples, quite often only one cell line. This study provides a precious resource for scientists to try to understand the global picture of cellular regulation by looking at both intra-individual and inter-individual variations. Statistical methods that adjust for noise in cellular assays may also be able to use it to better model noises in transcription factor binding assays.

In chapter 3 we tackle the problems of performing phenotype association using noisy gene expression data. The expression levels of functionally related genes, such as genes in a signaling pathway, tend to correlate to each other for a biological reason. Taken this common variation, feature extraction methods can effectively extract signal out of noise present in data from individual genes. We applied factor analysis firstly to remove systematic noise in the entire gene expression data, then to each individual pathway to extract pathway factors as our new phenotypes. We showed that these pathway factors are substantially more heritable than individual genes, using estimates from the twins data in our study. Using ageing phenotypes, our approach revealed a number of pathways known to be related to ageing as well as new pathways that are candidates for further investigation.

In this thesis, I also worked on developing new methods for important cellular traits. In chapter 4, I described a novel method that uses whole genome as well as exome data to estimate telomere length. It is particularly attractive to the large sequencing cohorts generated from cancer, epidemiology and ageing studies. Many important questions about the role of telomere length regulation can be tackled as the phenotype data can be made available by the new method. I have already applied it in a melanoma study which discovered that mutations in POT1 gene predispose to the disease and the mutated POT1 gene is associated with longer telomeres.

5.2 Future Work

Genomic DNA stores functional information that makes diverse biological systems in various cells and organisms. In addition to the protein-coding sequences, there is also a group of sequences that can indirectly influence the expression of genes as regulators. These sequences are key in achieving the complexity of the organisms. For example, having hugely different number of cells and cell types, *Caenorhabditis elegans* (around a thousand cells) and humans (trillions of cells) have quite similar amount of protein-coding genes (around 20 to 25 thousands).

A large volume of research efforts have been made to understand the regulatory mechanisms. The genomic regions that are immediately upstream of the transcription start sites have been found critical for controlling gene expression (See review Wittkopp and Kalay, 2011). In addition to these regions, one important phenomenon of gene regulation in higher order organisms is that the regulatory machinery needs not always be proximal. Increasing amount of studies have shown that there exists functional elements that are spatially distal to genes but are capable of influencing gene expression

(See review Bulger and Groudine, 2011). This is possible as chromatin is in a three dimensional space, where elements that are distal to their target genes can be brought close by higher order structures. CTCF is one of those special proteins that is involved in such function. Systematic studies on its bindings including this thesis show that its binding is regulated by genetic factors. Although nearly half of the QTL variants are close to the canonical motif, many are not. Only a minority of QTL variants are within the binding motif affecting binding directly. It is possible that the genetic effects of the proximal variants are mediated by collaborative factors. These requires investigation of binding patterns of more transcription factors in a group of individuals to search for correlation of bindings between them. Also, importantly, a higher resolution map, ideally at single base pair, is needed for identifying such interactions by knowing exactly where they bind. The current standard ChIP-seq experiments produce binding peaks at a resolution of a few hundred base pairs, much larger than the binding interface. Methods such as the enrichment analysis used in this case may not be reliable for individual events. Some of the recent technologies, such as Capture-C (Hughes et al., 2014), shows some promising directions by producing base pair resolution for interactions between a pair of factors in a *cis* window. Meanwhile, it is also possible that transcription factors not always bind to their canonical motifs, but is subject to a probability as a stochastic process. The bindings may occur at sites with weaker motif pattern that are yet to be discovered. This could be more confidently identified if we know the exact positions of binding, which gives much better signal noise ratio than searching for motifs in long sequences.

Future studies should also extend beyond lymphoblastoid cell lines to more tissues such as neurons, muscles etc and attempt to collect measurement *in vivo*. By doing so the chance of revealing the genuine biological mechanisms are much elevated. The diversity of regulation in these tissues provide a natural platform for comparing binding patterns of transcription factors and expression patterns of genes. Associating these patterns with the developmental features of these tissues may give important information on the regulatory mechanisms.

In this and many other studies, QTL analysis and allele specific analysis are done separately. This is because *cis* regulation patterns can be captured both by comparing between individuals, where individuals with AA, AB and BB genotypes should have different phenotypic levels if the locus is causal, or comparing within individuals, where in individuals with AB genotype the signals from A and B alleles should be different. This can be combined into a joint haplotype test that increases the statistical power (see methods in McVicker et al., 2013). Additionally, quite often the loci showing an allele specific signal are not the causal loci themselves, supported by the observations of conflicting signal directions of allele specific alleles between different individuals. Some initial methods have been developed to search in the local region to find variant that maximize a test score for consistent allelic imbalance (Lappalainen et al., 2013). Methods that extend towards these two directions only just begin to emerge and there is much rooms to develop them further. One approach could be parametrize phenotype and genotype value by haplotype instead of by individual. An individual that is homozygous and has a phenotypic value of 10 would be encoded as two entries in a (genotype, phenotype) format as (0,5), (0,5) or (1,5), (1,5) depending on whether the genotype is 0 or 1. And an individual that is heterozygous would be encoded as (0, allelic value for genotype 0) and (1, allelic value for genotype 1), taking the actual measured allelic value from these individuals. Associations can be tested using a linear model linking the two variables. This way the inter-individual QTL test and the intra-individual allele test can be combined. One caveat of such encoding is that the two entries of an individual are not independent. This is however similar to correcting for population structures, such as for monozygous twins. One could use a linear mixed model to correct for such structure using a relationship matrix.

As more and more functional data are accumulated around study samples, ideally one would want to model them together, looking for not only the effect of genotypes on each individual phenotype, but also on the network of them, learning their interactions and eventually the causality directions. Taking advantages of improved phasing algorithms, studies have started to investigate the phenotypes at a haplotype level, which extends to a larger genomic regions beyond a single SNP, and allows to model many other phenotypes that occur on the same haplotype. For example, the CTCF data generated in chapter 2 can be combined with RNA-seq data published for the same individuals for such analysis. One could model the allelic effects of CTCF binding and gene expression using two binomial variables, and see if they behave independently. Interestingly, haplotype that links two traits can be broken by recombination, forming recombinant samples. This gives an opportunity to distinguish the causal signals if any, because if the direction of association appear both in non-recombinants and recombinants, it suggests that two traits are causally linked.

The algorithm and software developed in chapter 4 have already generated strong interest in cancer and ageing studies. I am involved in two ongoing cancer studies on prostate cancer and melanoma while I am writing this thesis. In the ageing context, some recent work shows strong parental age effect, particularly paternal effect, on the health and disease status of their offspring (Kong et al., 2012, Goriely and Wilkie, 2012). Such paternal effects can be associated with telomere lengths as longer telomeres are transmitted to offspring by older fathers. Telomere lengths can be estimated from trio sequence data that some already become available from epidemiology studies. The age at conception can be easily worked out if the ages of the trio are known. It is also of interest to understand the heritability of telomere length, which can also be worked out using these data. Additionally, genome wide association analysis can reveal the genetic loci that are associated with the variation of telomere lengths, which will be very relevant to the context of ageing, cancer and a number of other diseases.

References

Abzhanov, A., M. Protas, B. R. Grant, P. R. Grant, and C. J. Tabin (2004). Bmp4 and morphological variation of beaks in Darwin's finches. *Science (New York, N.Y.) 305*(5689), 1462–1465. 1.3.4

Altshuler, D., J. N. Hirschhorn, M. Klannemark, C. M. Lindgren, M. C. Vohl, J. Nemesh, C. R. Lane, S. F. Schaffner, S. Bolk, C. Brewer, et al. (2000). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics 26*(1), 76–80. 1.2.1

Baerlocher, G. M., I. Vulto, G. de Jong, and P. M. Lansdorp (2006). Flow cytometry and FISH to measure the average length of telomeres (flow FISH). *Nature Protocols 1*(5), 2365–2376. 4.1

Bahar, R., C. H. Hartmann, K. A. Rodriguez, A. D. Denny, R. A. Busuttil, M. E. T. Dollé, R. B. Calder, G. B. Chisholm, B. H. Pollock, C. A. Klein, et al. (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature 441*(7096), 1011–1014. 3.1, 3.8

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research 37*, 202–208. 2.2

Baker, D. J., T. Wijshake, T. Tchkonia, N. K. LeBrasseur, B. G. Childs, B. van de Sluis, J. L. Kirkland, and J. M. van Deursen (2011). Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders. *Nature 479*, 232–236. 3.8

Balding, D. J., M. Bishop, and C. Cannings (2008). *Handbook of Statistical Genetics*. Handbook of Statistical Genetics: Third Edition. Chichester, UK: John Wiley & Sons, Ltd. 1.2.2.2

Barnett, D. W., E. K. Garrison, A. R. Quinlan, M. P. Strmberg, and G. T. Marth (2011, June). Bamtools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics 27*(12), 1691–1692. 4.7

Barzilai, N., D. M. Huffman, R. H. Muzumdar, and A. Bartke (2012). The Critical Role of Metabolic Pathways in Aging. *Diabetes 61*(6), 1315–1322. 3.8

Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods*. Wiley Series in Probability and Statistics. Wiley-Blackwell. 1.3.3

Bates, D., M. Maechler, B. Bolker, and S. Walker (2014). *lme4: Linear mixed-effects models using Eigen and S4*. 3.4

Bell, A. C. and G. Felsenfeld (2000, May). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature 405*(6785), 482–485. 2.1

Bell, A. C., A. G. West, and G. Felsenfeld (1999, August). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell 98*(3), 387–396. 2.1

Bell, J. T., A. a. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi, J. F. Degner, Y. Gilad, and J. K. Pritchard (2011, January). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology 12*(1), R10. 1.3.6

Bell, J. T. and T. D. Spector (2012, October). DNA methylation studies using twins: what are they telling us? *Genome Biology 13*(10), 172. 1.3.6

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statisitical Society, Series B 57*(1), 289–300. 1.2.3, 2.2, 2.6

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature 456*(7218), 53–59. 1.3.2

Birney, E., J. a. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, et al. (2007, June). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature 447*(7146), 799–816. 2.19

Bishop, C. (2006). *Pattern recognition and machine learning*, Volume 4 of *Information science and statistics*. Springer. 1.3.3

Blasco, M. A. (2005, August). Telomeres and human disease: ageing, cancer and beyond. *Nature Reviews. Genetics 6*(8), 611–622. 4.1

Borneman, A. R., T. a. Gianoulis, Z. D. Zhang, H. Yu, J. Rozowsky, M. R. Seringhaus, L. Y. Wang, M. Gerstein, and M. Snyder (2007, August). Divergence of transcription factor binding sites across related yeast species. *Science (New York, N.Y.) 317*(5839), 815–819. 1.3.5

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, et al. (2005, September). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell 122*(6), 947–956. 1.3.5

Broer, L., V. Codd, D. R. Nyholt, J. Deelen, M. Mangino, G. Willemsen, E. Albrecht, N. Amin, M. Beekman, E. J. C. de Geus, et al. (2013, October). Meta-analysis of telomere length in 19,713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect. *European Journal of Human Genetics : EJHG 21*(10), 1163–8. 4.3.3

Brookes, A. J. (1999). The essence of SNPs. *Gene 234*, 177–186. 1.1.3

Browning, B. L. and Z. Yu (2009). Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies. *American Journal of Human Genetics 85*(6), 847–861. 2.4

Bulger, M. and M. Groudine (2011, February). Functional and mechanistic diversity of distal transcription enhancers. *Cell 144*(3), 327–339. 2.8, 5.2

Campos, E. I. and D. Reinberg (2009, January). Histones: annotating chromatin. *Annual Review of Genetics 43*(1), 559–599. 1.3.6

Castle, J. C., M. Biery, H. Bouzek, T. Xie, R. Chen, K. Misura, S. Jackson, C. D. Armour, J. M. Johnson, C. A. Rohl, et al. (2010). DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. *BMC Genomics 11*(1), 244. 4.1

Cawthon, R. M. (2009, February). Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Research 37*(3), e21–e21. 4.1

Chaix, R., M. Somel, D. P. Kreil, P. Khaitovich, and G. A. Lunter (2008). Evolution of primate gene expression: Drift and corrective sweeps? *Genetics 180*, 1379–1389. 2.8

Cheung, V. G., R. R. Nayak, I. X. Wang, S. Elwyn, S. M. Cousins, M. Morley, and R. S. Spielman (2010, January). Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biology 8*(9), e1000480. 1.3.4

Chu, Y. and D. R. Corey (2012, August). RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics 22*(4), 271–4. 1.3.2

Clancy, D. J., D. Gems, L. G. Harshman, S. Oldham, H. Stocker, E. Hafen, S. J. Leevers, and L. Partridge (2001). Extension of life-span by loss of CHICO, a Drosophila insulin receptor substrate protein. *Science (New York, N.Y.) 292*(5514), 104–106. 3.8

Computing, R Foundation for Statistical Vienna, A. (2008). R: A language and environment for statistical computing. 3.4, 4.3.1

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, et al. (2013, February). Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.) 339*(6121), 819–23. 1.3.5

Cooper, G. M., E. a. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research 15*(7), 901–913. 2.7

Core, L. J., J. J. Waterfall, and J. T. Lis (2008, December). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.) 322*(5909), 1845–1848. 1.3.2

Cuddapah, S., R. Jothi, D. E. Schones, T. Y. Roh, K. Cui, and K. Zhao (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research 19*(1), 24–32. 2.1

Cusanovich, D. A., B. Pavlovic, J. K. Pritchard, and Y. Gilad (2014). The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genetics 10*(3), 1–30. 1.3.5

de Magalhães, J. a. P., J. a. Curado, and G. M. Church (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics 25*(7), 875–881. 3.1

Degner, J. F., A. a. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, G. E. Crawford, et al. (2012, February). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature 482*(7385), 390–394. 1.2.3, 2.1, 2.3, 2.7, 2.7, 2.8

Dimas, A. S., S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, M. Gutierrez Arcelus, M. Sekowska, et al. (2009, September). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science (New York, N.Y.) 325*(5945), 1246–1250. 1.3.4

Ding, Z., M. Mangino, A. Aviv, UK10 Consortium, T. Spector, and R. Durbin (2014, January). Estimating telomere length from whole genome sequence data. *Nucleic Acids Research 42*(9), doi: 10.1093/nar/gku181. 1.4

Ding, Z., Y. Ni, S. Timmer, er W, B.-K. Lee, A. Battenhouse, S. Louzada, ra, F. Yang, I. Dunham, et al. (2014, November). Quantitative Genetics of CTCF Binding Reveal Local Sequence Effects and Different Modes of X-Chromosome Association. *PLoS Genetics 10*(11), e1004798. 1.4, 2, 5.1

Dixon, A. L., L. Liang, M. F. Moffatt, W. Chen, S. Heath, K. C. C. Wong, J. Taylor, E. Burnett, I. Gut, M. Farrall, et al. (2007, October). A genome-wide association study of global gene expression. *Nature Genetics 39*(10), 1202–1207. 1.3.4

Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren (2012, May). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature 485*(7398), 376–380. 1.3.6

Dohm, J. C., C. Lottaz, T. Borodina, and H. Himmelbauer (2008, August). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research 36*(16), e105—-e105. 4.3.1

Dostie, J. and J. Dekker (2007). Mapping networks of physical interactions between genomic elements using 5C technology. *Nature Protocols 2*(4), 988–1002. 1.3.2

Dueck, D., Q. D. Morris, and B. J. Frey (2005). Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics 21*, i144–i151. 1.3.3

Ekelund, J., D. Lichtermann, I. Hovatta, P. Ellonen, J. Suvisaari, J. D. Terwilliger, H. Juvonen, T. Varilo, R. Arajärvi, M. L. Kokko-Sahin, et al. (2000). Genome-wide scan for schizophrenia in the Finnish population: evidence for a locus on chromosome 7q22. *Human Molecular Genetics 9*(7), 1049–1057. 1.2.1

Elnitski, L., V. X. Jin, P. J. Farnham, and S. J. Jones (2006, December). Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Research 16*(12), 1455–1464. 1.3.5

Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir, et al. (2008). Genetics of gene expression and its effect on disease. *Nature 452*(7186), 423–428. 1.3.3

Farh, K. K.-H., A. Marson, er, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. H. Ryan, et al. (2014). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2.8

Finkel, T., M. Serrano, and M. A. Blasco (2007). The common biology of cancer and ageing. *Nature 448*(7155), 767–774. 3.8

Fisher, R. A. (1919). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh 52*(02), 399–433. 1.1.2

Forrest, A. R. R., H. Kawaji, M. Rehli, J. K. Baillie, M. J. L. de Hoon, T. Lassmann, M. Itoh, K. M. Summers, H. Suzuki, C. O. Daub, et al. (2014, March). A promoter-level mammalian expression atlas. *Nature 507*(7493), 462–70. 1.3.4

Franke, A., D. P. B. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, et al. (2010, November). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics 42*(12), 1118–1125. 1.2.5

Friedman, D. B. and T. E. Johnson (1988). Three mutants that extend both mean and maximum life span of the nematode, Caenorhabditis elegans, define the age-1 gene. *Journal of Gerontology 43*(4), B102–B109. 3.8

Furey, T. S. (2012, December). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics 13*(12), 840–852. 1.3.2

Fusi, N., O. Stegle, and N. D. Lawrence (2012, January). Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Computational Biology 8*(1), e1002330. 1.3.3

Gaffney, D. J., J.-B. Veyrieras, J. F. Degner, R. Pique-Regi, A. a. Pai, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard (2012, January). Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology 13*(1), R7. 1.3.7

Gärtner, C. F. (1849). *Versuche und Beobachtungen über die Bastarderzeugung im Pflanzenreich.* Stuttgart. 1.1.1

Gibson, G. (2008). The environmental contribution to gene expression profiles. *Nature Reviews. Genetics 9*(8), 575–581. 1.3.3

Glass, D., A. Viñuela, M. N. Davies, A. Ramasamy, L. Parts, D. Knowles, A. A. Brown, A. K. Hedman, K. S. Small, A. Buil, et al. (2013). Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biology 14*(7), R75. 3.8

Goriely, A. and A. O. Wilkie (2012, February). Paternal Age Effect Mutations and Selfish Spermatogonial Selection: Causes and Consequences for Human Disease. *The American Journal of Human Genetics 90*(2), 175–200. 5.2

Göring, H. H. H., J. E. Curran, M. P. Johnson, T. D. Dyer, J. Charlesworth, S. A. Cole, J. B. M. Jowett, L. J. Abraham, D. L. Rainwater, A. G. Comuzzie, et al. (2007, October). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genetics 39*(10), 1208–1216. 1.3.4

Grant, C. E., T. L. Bailey, and W. S. Noble (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics 27*(7), 1017–1018. 2.2

Grundberg, E., K. S. Small, A. s. K. Hedman, A. Nica, ra C, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, et al. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics 44*(10), 1084–1089. 1.3.4, 3.2, 3.7

Gudmundsson, J., P. Sulem, V. Steinthorsdottir, J. T. Bergthorsson, G. Thorleifsson, A. Manolescu, T. Rafnar, D. Gudbjartsson, B. A. Agnarsson, A. Baker, et al. (2007, August). Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nature Genetics 39*(8), 977–983. 1.2.5

Guelen, L., L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, et al. (2008, June). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature 453*(7197), 948–951. 1.3.6

Gusella, J. F., N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, and A. Y. Sakaguchi (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature 306*(5940), 234–238. 1.2.1

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics 8*(21), 299–309. 1.1.3

Hastie, T., R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein, and P. Brown (2000). Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology 1*(2). 1.3.3, 3.1, 3.8

Heinz, S., C. E. Romanoski, C. Benner, K. a. Allison, M. U. Kaikkonen, L. D. Orozco, and C. K. Glass (2013, November). Effect of natural genetic variation on enhancer selection and function. *Nature 503*(7477), 487–92. 2.1, 2.7, 2.8

Helgadottir, A., G. Thorleifsson, A. Manolescu, S. Gretarsdottir, T. Blondal, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, A. Baker, A. Palsson, et al. (2007, June). A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science (New York, N.Y.) 316*(5830), 1491–1493. 1.2.5

Henson, J. D., A. A. Neumann, T. R. Yeager, and R. R. Reddel (2002). Alternative lengthening of telomeres in mammalian cells. *Oncogene 21*, 598–610. 4.1

Herndon, L. A., P. J. Schmeissner, J. M. Dudaronek, P. A. Brown, K. M. Listner, Y. Sakano, M. C. Paupard, D. H. Hall, and M. Driscoll (2002). Stochastic and genetic factors influence tissue-specific decline in ageing C. elegans. *Nature 419*(6909), 808–814. 3.8

Hindorff, L. a., P. Sethupathy, H. a. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. a. Manolio (2009, June). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America 106*(23), 9362–9367. 1.2.5, 2.7

Holzenberger, M., J. Dupont, B. Ducos, P. Leneuve, A. Géloën, P. C. Even, P. Cervera, and Y. Le Bouc (2003). IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature 421*(6919), 182–187. 3.8

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*(6), 417–441. 1.3.3

Houtkooper, R. H., C. Argmann, S. Houten, er M, C. Cantó, E. H. Jeninga, P. A. Andreux, C. Thomas, R. Doenlen, K. Schoonjans, et al. (2011). The metabolic footprint of aging in mice. 3.8

Howie, B. N., P. Donnelly, and J. Marchini (2009, June). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics 5*(6), e1000529. 2.4

Huang, D. W., B. T. Sherman, and R. A. Lempicki (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols 4*(1), 44–57. 3.7

Hughes, J. R., N. Roberts, S. McGowan, D. Hay, E. Giannoulatou, M. Lynch, M. De Gobbi, S. Taylor, R. Gibbons, and D. R. Higgs (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics 46*, 205–12. 5.2

Hwang, W. Y., Y. Fu, D. Reyon, M. L. Maeder, S. Q. Tsai, S, J. D. er, R. Peterson, all T, J.-R. J. Yeh, et al. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology 31*(3), 227–9. 1.3.5

Hyun, M. K., C. Ye, and E. Eskin (2008). Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics 180*(4), 1909–1925. 1.3.3

Ingolia, N. T. (2014, March). Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews. Genetics 15*(3), 205–13. 1.3.2

International Cancer Genome Consortium (2010, April). International network of cancer genome projects. *Nature 465*(7291), 966–966. 1.1.3

Jeffreys, A. J., R. Neumann, M. Panayi, S. Myers, and P. Donnelly (2005, June). Human recombination hot spots hidden in regions of strong marker association. *Nature Genetics 37*(6), 601–606. 1.1.3

Jostins, L., S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. a. Anderson, et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature 491*(7422), 119–24. 1.2.5

Kamb, A., D. Shattuck-Eidens, R. Eeles, Q. Liu, N. A. Gruis, W. Ding, C. Hussey, T. Tran, Y. Miki, and J. Weaver-Feldhaus (1994). Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nature Genetics 8*(1), 23–26. 1.2.5

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research 32*(90001), D277–D280. 1.3.8, 3.1

Kang, H. M., J. H. Sul, S. K. Service, N. a. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin (2010, March). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics 42*(4), 348–354. 1.2.2.4

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin (2008, February). Efficient control of population structure in model organism association mapping. *Genetics 178*(3), 1709–1723. 1.2.2.4

Kasowski, M., F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S. M. Waszak, L. Habegger, J. Rozowsky, M. Shi, A. Urban, et al. (2010, April). Variation in transcription factor binding among humans. *Science (New York, N.Y.) 328*(5975), 232–235. 1.3.5, 1.3.5, 2.1, 2.7, 2.8

Kilpinen, H., S. M. Waszak, A. R. Gschwind, S. K. Raghav, R. M. Witwicki, A. Orioli, E. Migliavacca, M. Wiederkehr, M. Gutierrez-Arcelus, N. I. Panousis, et al. (2013, November). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science (New York, N.Y.) 342*(6159), 744–7. 2.8

Kimura, M. and A. Aviv (2011, July). Measurement of telomere DNA content by dot blot analysis. *Nucleic Acids Research 39*(12), e84—-e84. 4.3.3, 4.3.3

Kimura, M., R. C. Stone, S. C. Hunt, J. Skurnick, X. Lu, X. Cao, C. B. Harley, and A. Aviv (2010, September). Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. *Nature Protocols 5*(9), 1596–1607. 4.1

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure (2014, January). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics 46*(3), 310–5. 1.3.5

Kleinjan, D. A. and V. van Heyningen (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *American Journal of Human Genetics 76*, 8–32. 1.3.4

Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, et al. (2012, August). Rate of de novo mutations and the importance of father's age to disease risk. *Nature 488*(7412), 471–475. 5.2

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, et al. (2002, June). A high-resolution recombination map of the human genome. *Nature Genetics 31*(3), 241–247. 1.1.3

Krueger, F., B. Kreck, A. Franke, and S. R. Andrews (2012, February). DNA methylome analysis using short bisulfite sequencing data. *Nature Methods 9*(2), 145–151. 1.3.2

Krueger, G. G., R. G. Langley, C. Leonardi, N. Yeilding, C. Guzzo, Y. Wang, L. T. Dooley, and M. Lebwohl (2007, February). A human interleukin-12/23 monoclonal antibody for the treatment of psoriasis. *The New England Journal of Medicine 356*(6), 580–592. 1.2.5

Kunarso, G., N.-Y. Chia, J. Jeyakani, C. Hwang, X. Lu, Y.-S. Chan, H.-H. Ng, and G. Bourque (2010, July). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics 42*(7), 631–634. 1.3.5, 2.1

Laajala, T. D., S. Raghav, S. Tuomela, R. Lahesmaa, T. Aittokallio, and L. L. Elo (2009, January). A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics 10*(1), 618. 1.3.5

Lander, E. S. and S. Botstein (1989, January). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics 121*(1), 185. 1.1.3

Lappalainen, T., M. Sammeth, M. R. Friedländer, P. A. C. 't Hoen, J. Monlong, M. a. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature 501*, 506–11. 1.3.4, 2.7, 5.2

Lawlor, D. A., R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. D. Smith (2008, April). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine 27*(8), 1133–1163. 1.3.8

Leal, S. M. (1998). *Genetics and Analysis of Quantitative Traits.* Sinauer Associates. 1.2.2.2

Lee, B. K., A. A. Bhinge, A. Battenhouse, R. M. McDaniell, Z. Liu, L. Song, Y. Ni, E. Birney, J. D. Lieb, T. S. Furey, et al. (2012). Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Research 22*, 9–24. 2.1, 2.2, 2.7

Lee, C., C. Grasso, and M. F. Sharlow (2002, March). Multiple sequence alignment using partial order graphs. *Bioinformatics (Oxford, England) 18*(3), 452–464. 2.26

Leek, J. T. and J. D. Storey (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics 3*(9), 1724–1735. 1.3.3, 3.1, 3.7

Lewis, E. B. (1945, March). The Relation of Repeats to Position Effect in Drosophila Melanogaster. *Genetics 30*(2), 137–166. 1.3.4

Li, H. and R. Durbin (2009, July). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics 25*(14), 1754–1760. 2.2

Li, H., H, B. saker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al. (2009). The Sequence Alignment / Map format and SAMtools. *Bioinformatics 25*(16), 2078–2079. 2.4

Li, H., J. Ruan, and R. Durbin (2008, November). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research 18*(11), 1851–1858. 3.2

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.) 326*(5950), 289–293. 1.3.2

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics (Oxford, England) 18*(1), 51–60. 1.3.3

Lindblad-Toh, K., M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, et al. (2011, October). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature 478*(7370), 476–482. 2.1

Listgarten, J., C. Lippert, and D. Heckerman (2012). Fast-LMM-Select for confounding from spatial structure and rare variants. *Nature Genetics*. 1.2.2.4

Listgarten, J., C. Lippert, C. M. Kadie, R. I. Davidson, E. Eskin, and D. Heckerman (2012, June). Improved linear mixed models for genome-wide association studies. *Nature Methods 9*(6), 525–526. 1.2.2.4

Lu, T., Y. Pan, S.-Y. Kao, C. Li, I. Kohane, J. Chan, and B. A. Yankner (2004). Gene regulation and DNA damage in the ageing human brain. *Nature 429*(6994), 883–891. 3.1

Luo, D. F., M. M. Bui, A. Muir, N. K. Maclaren, G. Thomson, and J. X. She (1995). Affected-sib-pair mapping of a novel susceptibility gene to insulin-dependent diabetes mellitus (IDDM8) on chromosome 6q25-q27. *American Journal of Human Genetics 57*, 911–919. 1.2.1

MacDonald, M. E., A. Novelletto, C. Lin, D. Tagle, G. Barnes, G. Bates, S. Taylor, B. Allitto, M. Altherr, and R. Myers (1992, May). The Huntington's disease candidate region exhibits many different haplotypes. *Nature Genetics 1*(2), 99–103. 1.2.1

Manolio, T. a., L. D. Brooks, and F. S. Collins (2008). A HapMap harvest of insights into the genetics of common disease. *Journal of Clinical Investigation 118*(5), 1590–1605. 1.2.5

Marguerat, S., A. Schmidt, S. Codlin, W. Chen, R. Aebersold, and J. Bähler (2012). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell 151*(3), 671–683. 1.3.7

Martens, U. M., J. M. Zijlmans, S. S. Poon, W. Dragowska, J. Yui, E. A. Chavez, R. K. Ward, and P. M. Lansdorp (1998). Short telomeres on human chromosome 17p. *Nature Genetics 18*(1), 76–80. 4.1

Mathew, C. G. and C. M. Lewis (2004). Genetics of inflammatory bowel disease: progress and prospects. *Human Molecular Genetics 13 Spec No*(90001), R161–R168. 1.2.1

Maurano, M. T., H. Wang, T. Kutyavin, and J. a. Stamatoyannopoulos (2012, March). Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genetics 8*(3), e1002599. 1.2.3, 2.1, 2.7, 2.7, 2.8

McCarroll, S. a., C. T. Murphy, S. Zou, S. D. Pletcher, C.-S. Chin, Y. N. Jan, C. Kenyon, C. I. Bargmann, and H. Li (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genetics 36*(2), 197–204. 3.8

McDaniell, R., B.-K. Lee, L. Song, Z. Liu, A. P. Boyle, M. R. Erdos, L. J. Scott, M. A. Morken, K. S. Kucera, A. Battenhouse, et al. (2010, April). Heritable individual-specific and allele-specific chromatin signatures in humans. *Science (New York, N.Y.) 328*(5975), 235–239. 1.3.5, 1.3.6, 2.1, 2.7, 2.7, 2.8

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham (2010, August). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics 26*(16), 2069–2070. 1.3.5

McVean, G. (2009, October). A genealogical interpretation of principal components analysis. *PLoS Genetics 5*(10), e1000686. 1.3.3

McVicker, G., B. van de Geijn, J. F. Degner, C. E. Cain, N. E. Banovich, A. Raj, N. Lewellen, M. Myrthil, Y. Gilad, and J. K. Pritchard (2013, November). Identification of genetic variants that affect histone modifications in human cells. *Science (New York, NY) 342*(6159), 747–749. 1.2.3, 1.3.6, 2.1, 2.8, 5.2

Mein, C. A., L. Esposito, M. G. Dunn, G. C. Johnson, A. E. Timms, J. V. Goy, A. N. Smith, L. Sebag-Montefiore, M. E. Merriman, A. J. Wilson, et al. (1998). A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nature Genetics 19*(3), 297–300. 1.2.1

Mendel, G. (1865). Versuche über Pflanzenhybriden. *Verhandlungen Des Naturforschenden Vereines in Brünn, Bd.* 1.1.1

Merkenschlager, M. and D. T. Odom (2013, March). CTCF and cohesin: Linking gene regulatory elements with their targets. *Cell 152*(6), 1285–1297. 2.1

Moayyeri, A., C. J. Hammond, A. M. Valdes, and T. D. Spector (2013a). Cohort profile: Twinsuk and healthy ageing twin study. *International Journal of Epidemiology 42*(1), 76–85. 3.2

Moayyeri, A., C. J. Hammond, A. M. Valdes, and T. D. Spector (2013b). Cohort profile: Twinsuk and healthy ageing twin study. *International Journal of Epidemiology 42*(1), 76–85. 4.1, 4.2

Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis (2010, January). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature 464*(7289), 773–777. 2.1, 2.3, 3.1

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine 13*, 47–60. 1.2.2.3

Morley, M., M. Morley, C. M. Molony, C. M. Molony, T. M. Weber, T. M. Weber, J. L. Devlin, J. L. Devlin, K. G. Ewens, K. G. Ewens, et al. (2004, August). Genetic analysis of genome-wide variation in human gene expression. *Nature 430*(7001), 743–7. 1.3.4

Morris, A. P., B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segrè, V. Steinthorsdottir, R. J. Strawbridge, H. Khan, H. Grallert, A. Mahajan, et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics 44*, 981–990. 1.2.5

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly (2005, October). A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, N.Y.) 310*(5746), 321–324. 1.1.3

Needham, J. (1942). *New paths in genetics*, Volume 34. Harper & Brothers. 1.3.4

Nejentsev, S., N. Walker, D. Riches, M. Egholm, and J. A. Todd (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science (New York, N.Y.) 324*, 387–389. 1.2.5

Nica, A., ra C, L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, et al. (2011, February). The architecture of gene regulatory variation across multiple human tissues: The MuTHER study. *PLoS Genetics 7*(2), e1002003. 1.3.3, 3.2

Noonan, J. P. and A. S. McCallion (2010, September). Genomics of long-range regulatory elements. *Annual Review of Genomics and Human Genetics 11*(1), 1–23. 2.1

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, et al. (2008). Genes mirror geography within Europe. *Nature 456*(7219), 98–101. 1.3.3

Novembre, J. and M. Stephens (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics 40*(5), 646–649. 1.3.3

Odom, D. T., R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, er, C. M. Conboy, D. K. Gifford, et al. (2007, June). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics 39*(6), 730–732. 1.3.5

Park, P. J. (2009, September). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews. Genetics 10*(10), 669–680. 1.3.5

Parker, M., X. Chen, A. Bahrami, J. Dalton, M. Rusch, G. Wu, J. Easton, N.-K. Cheung, M. Dyer, E. R. Mardis, et al. (2012). Assessing telomeric DNA content in pediatric cancers using whole-genome sequencing data. *Genome Biology 13*, R113. 4.1

Parts, L., O. Stegle, J. Winn, and R. Durbin (2011). Joint genetic analysis of gene expression data with inferred cellular phenotypes. *PLoS Genetics 7*. 3.1, 3.3

Pearson, K. (1901, November). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*(11), 559–572. 1.3.3

Petretto, E., J. Mangion, N. J. Dickens, S. A. Cook, M. Kumaran, e K, H. Lu, J. Fischer, H. Maatz, V. Kren, et al. (2006, October). Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genetics 2*(10), 1625–1633. 1.3.4

Phillips, J. E. and V. G. Corces (2009). CTCF: Master Weaver of the Genome. 2.1

Pickrell, J. K., J. C. Marioni, A. a. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard (2010, April). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature 464*(7289), 768–772. 2.1, 3.1

Pompella, A., A. Visvikis, A. Paolicchi, V. De Tata, and A. F. Casini (2003). The changing faces of glutathione, a cellular protagonist. In *Biochemical Pharmacology*, Volume 66, pp. 1499–1503. 3.8

Pournara, I. and L. Wernisch (2007). Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics 8*(1), 61. 1.3.3

Price, A. L., A. Helgason, G. Thorleifsson, S. a. McCarroll, A. Kong, and K. Stefansson (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics 7*(2), e1001317. 1.3.4

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006, August). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet 38*(8), 904–909. 1.3.3

Quinlan, A. R. and I. M. Hall (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics 26*(6), 841–842. 2.3

Rakyan, V. K., H. Beyan, T. A. Down, M. I. Hawa, S. Maslau, D. Aden, A. Daunay, F. Busato, C. A. Mein, B. Manfras, et al. (2011, September). Identification of type 1 Diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genetics 7*(9), e1002300. 1.3.6

Reddy, T. E., J. Gertz, F. Pauli, K. S. Kucera, K. E. Varley, K. M. Newberry, G. K. Marinov, A. Mortazavi, B. A. Williams, L. Song, et al. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research 22*, 860–869. 2.1, 2.7

Reis-Filho, J. S. and L. Pusztai (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet 378*(9805), 1812–23. 3.1

Rhee, H. S. and B. F. Pugh (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell 147*(6), 1408–1419. 1.3.2

Riordan, J. R., J. M. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, and J. L. Chou (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science (New York, N.Y.) 245*(4922), 1066–1073. 1.2.1

Robles-Espinoza, C. D., Harl, M. , A. J. Ramsay, L. G. Aoude, V. Quesada, Z. Ding, K. a. Pooley, A. L. Pritchard, J. C. Tiffen, et al. (2014). POT1 loss-of-function variants predispose to familial melanoma. *Nature Genetics 46*, 478–81. 1.4, 4.5, 4.10

Samassekou, O., M. Gadji, R. Drouin, and J. Yan (2010, September). Sizing the ends: Normal length of human telomeres. *Annals of Anatomy 192*(5), 284–291. 4.1

Sandelin, A., W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research 32*(90001), D91–D94. 2.2

Sanguinetti, G., G. Sanguinetti, N. D. Lawrence, N. D. Lawrence, M. Rattray, and M. Rattray (2006). Probabilistic inference of transcription factor cocentrations and gene-specific regulatory activities. *Bioinformatics 22*, 2775–2781. 1.3.3

Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics 37*, 710–717. 1.3.8

Schena, M., D. Shalon, R. W. Davis, and P. O. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.) 270*(5235), 467–470. 1.3.2

Schmidt, D., P. C. Schwalie, C. S. Ross-Innes, A. Hurtado, G. D. Brown, J. S. Carroll, P. Flicek, and D. T. Odom (2010, May). A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Research 20*(5), 578–588. 2.1

Schmidt, D., P. C. Schwalie, M. D. Wilson, B. Ballester, A. Gonalves, C. Kutter, G. D. Brown, A. Marshall, P. Flicek, and D. T. Odom (2012, January). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell 148*(1-2), 335–348. 2.8

Schmidt, D., M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown, A. Marshall, C. Kutter, S. Watt, C. P. Martinez-Jimenez, S. Mackay, et al. (2010, May). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.) 328*(5981), 1036–1040. 1.3.5, 2.1

Scott, L. J., K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines, A. U. Jackson, et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science (New York, N.Y.) 316*(5829), 1341–1345. 1.2.5

Segal, E. and J. Widom (2009). From DNA sequence to transcriptional behaviour: a quantitative approach. *Nature Reviews. Genetics 10*(7), 443–456. 1.3.5, 1.3.6

Shivaswamy, S., A. Bhinge, Y. Zhao, S. Jones, M. Hirst, and V. R. Iyer (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biology 6*, 0618–0630. 2.2

Sie, L., S. Loong, and E. K. Tan (2009, July). Utility of lymphoblastoid cell lines. *Journal of Neuroscience Research 87*(9), 1953–1959. 2.1

Simonis, M., P. Klous, I. Homminga, R.-J. Galjaard, E.-J. Rijkers, F. Grosveld, J. P. P. Meijerink, and W. de Laat (2009). High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. *Nature Methods 6*(11), 837–842. 1.3.2

Skelly, D. A., J. Ronald, and J. M. Akey (2009). Inherited variation in gene expression. *Annual Review of Genomics and Human Genetics 10*(1), 313–332. 1.3.4

Small, K. S., A. K. Hedman, E. Grundberg, A. Nica, ra C, G. Thorleifsson, A. Kong, U. Thorsteindottir, S.-Y. Shin, H. B. Richards, et al. (2011, June). Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature Genetics 43*(6), 561–564. 1.3.4, 3.8

Sopher, B. L., P. D. Ladd, V. V. Pineda, R. Libby, ell T, S. M. Sunkin, J. B. Hurley, C. Thienes, t P, Gaasterl, et al. (2011, June). CTCF Regulates Ataxin-7 Expression through Promotion of a Convergently Transcribed, Antisense Noncoding RNA. *Neuron 70*(6), 1071–1084. 2.1

Southern, E. M. (1992). Detection of specific sequences among DNA fragments separated by gel electrophoresis. 1975. *Biotechnology (Reading, Mass.) 24*, 122–139. 1.3.2

Spielman, R. S., L. A. Bastone, J. T. Burdick, M. Morley, W. J. Ewens, and V. G. Cheung (2007, March). Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics 39*(2), 226–231. 2.1

Splinter, E., H. Heath, J. Kooren, R. J. Palstra, P. Klous, F. Grosveld, N. Galjart, and W. De Laat (2006, September). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes and Development 20*(17), 2349–2354. 2.1

Stedman, W., H. Kang, S. Lin, J. L. Kissil, M. S. Bartolomei, and P. M. Lieberman (2008, March). Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *The EMBO Journal 27*(4), 654–666. 2.1

Stefflova, K., D. Thybert, M. D. Wilson, I. Streeter, J. Aleksic, P. Karagianni, A. Brazma, D. J. Adams, I. Talianidis, J. C. Marioni, et al. (2013, August). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell 154*(3), 530–540. 2.1, 2.7, 2.8

Stegle, O., L. Parts, R. Durbin, and J. Winn (2010, May). A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology 6*(5), 1–11. 1.3.3, 3.1

Storey, J. D. and R. Tibshirani (2003, August). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America 100*(16), 9440–9445. 1.2.3, 2.5, 2.7, 2.11

Stranger, B. E., S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, et al. (2012, April). Patterns of Cis regulatory variation in diverse human populations. *PLoS Genetics 8*(4), e1002639. 1.3.4, 2.4

Stranger, B. E., A. Nica, ra C, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, et al. (2007). Population genomics of human gene expression. *Nature Genetics 39*(10), 1217–1224. 1.3.4, 1.3.4, 2.1

Suzuki, H., A. R. R. Forrest, E. van Nimwegen, C. O. Daub, P. J. Balwierz, K. M. Irvine, T. Lassmann, T. Ravasi, Y. Hasegawa, M. J. L. de Hoon, et al. (2009, May). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics 41*(5), 553–562. 1.3.5

Taiwo, O., G. Wilson, T. Morris, S. Seisenberger, W. Reik, D. Pearce, S. Beck, and L. Butcher (2012, April). Methylome analysis using MeDIP-seq with low DNA concentrations. *Nature Protocols 7*(4), 617–636. 1.3.2

Taub, M. A., H. Corrada Bravo, and R. A. Irizarry (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome Medicine 2*, 87. 1.3.2

The 1000 Genomes Consortium (2010, October). A map of human genome variation from population-scale sequencing. *Nature 467*(7319), 1061–1073. 1.1.3, 2.1, 4.1

The 1000 Genomes Consortium (2012, November). An integrated map of genetic variation from 1,092 human genomes. *Nature 491*(7422), 56–65. 2.4

The Fantom Consortium (2014, March). A promoter-level mammalian expression atlas. *Nature 507*(7493), 462–470. 1.3.5, 1.3.7

The International HapMap 3 Consortium (2010, September). Integrating common and rare genetic variation in diverse human populations. *Nature 467*(9), 52–8. 1.1.3

Tipping, M. E. and C. M. Bishop (1997). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61*, 611–622. 1.3.3

Tompa, M., N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. Favorov, er V, M. C. Frith, Y. Fu, et al. (2005, January). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology 23*(1), 137–144. 1.3.5

Tsui, L. C., M. Buchwald, D. Barker, J. C. Braman, R. Knowlton, J. W. Schumm, H. Eiberg, J. Mohr, D. Kennedy, and N. Plavsic (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science (New York, N.Y.) 230*(4729), 1054–1057. 1.2.1

Tuch, B. B., D. J. Galgoczy, A. D. Hernday, H. Li, and A. D. Johnson (2008, February). The evolution of combinatorial gene regulation in fungi. *PLoS Biology 6*(2), 0352–0364. 1.3.5

Valdes, A. M., T. Andrew, J. P. Gardner, M. Kimura, E. Oelsner, L. F. Cherkas, A. Aviv, and T. D. Spector (2005, August). Obesity, cigarette smoking, and telomere length in women. *Lancet 366*(9486), 662–664. 4.3.3

van de Nobelen, S., M. Rosa-Garrido, J. Leers, H. Heath, W. Soochit, L. Joosen, I. Jonkers, J. Demmers, M. van der Reijden, V. Torrano, et al. (2010, January). CTCF regulates the local epigenetic state of ribosomal DNA repeats. *Epigenetics & Chromatin 3*(1), 19. 2.1

Veyrieras, J. B., S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, and J. K. Pritchard (2008, March). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics 4*(3), 452–64. 1.3.4

Vinuela, A., L. B. Snoek, J. A. G. Riksen, and J. E. Kammenga (2012). Aging Uncouples Heritability and Expression-QTL in Caenorhabditis elegans. *G3; Genes/Genomes/Genetics 2*, 597–605. 3.8

Wellcome Trust Case Control Consortium (2007, June). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature 447*(7145), 661–678. 1.2.2, 1.2.5

Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research 42*. 1.2.2

Willcox, B. J., T. A. Donlon, Q. He, R. Chen, i, J. S. Grove, K. Yano, K. H. Masaki, D. C. Willcox, B. Rodriguez, et al. (2008). FOXO3A genotype is strongly associated with human longevity. *Proceedings of the National Academy of Sciences of the United States of America 105*(37), 13987–13992. 3.8

Williams, N. M., M. I. Rees, P. Holmans, N. Norton, A. G. Cardno, L. A. Jones, K. C. Murphy, S, R. D. ers, G. McCarthy, et al. (1999). A two-stage genome scan for schizophrenia susceptibility genes in 196 affected sibling pairs. *Human Molecular Genetics 8*(9), 1729–1739. 1.2.1

Wilson, M. D., N. L. Barbosa-Morais, D. Schmidt, C. M. Conboy, L. Vanes, V. L. J. Tybulewicz, E. M. C. Fisher, S. Tavaré, and D. T. Odom (2008, October). Species-specific transcription in mice carrying human chromosome 21. *Science (New York, N.Y.) 322*(5900), 434–438. 1.3.5

Wittkopp, P. J., B. K. Haerum, and A. G. Clark (2004, July). Evolutionary changes in cis and trans gene regulation. *Nature 430*(6995), 85–88. 1.3.4

Wittkopp, P. J. and G. Kalay (2011). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. 5.2

Wray, G. a. and G. a. Wray (2007, March). The evolutionary significance of. *Genetics 8*(3), 206–216. 1.3.4

Yamazaki, K., D. McGovern, J. Ragoussis, M. Paolucci, H. Butler, D. Jewell, L. Cardon, M. Takazoe, T. Tanaka, T. Ichimori, et al. (2005). Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Human Molecular Genetics 14*, 3499–3506. 1.2.5

Yusufzai, T. M., H. Tagami, Y. Nakatani, and G. Felsenfeld (2004, January). CTCF Tethers an Insulator to Subnuclear Sites, Suggesting Shared Insulator Mechanisms across Species. *Molecular Cell 13*(2), 291–298. 2.1

Zhang, X., E. L. Moen, C. Liu, W. Mu, E. R. Gamazon, S. M. Delaney, C. Wing, L. A. Godley, M. E. Dolan, and W. Zhang (2014, June). Linking the genetic architecture of cytosine modifications with human complex traits. *Human Molecular Genetics 1200*(22), ddu313–. 1.3.6

Zhang, Z., E. Ersoz, C.-Q. Lai, R. J. Todhunter, H. K. Tiwari, M. a. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, et al. (2010, April). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics 42*(4), 355–360. 1.2.2.4

Zheng, W., H. Zhao, E. Mancera, L. M. Steinmetz, and M. Snyder (2010, April). Genetic analysis of variation in transcription factor binding in yeast. *Nature 464*(7292), 1187–1191. 1.3.5, 1.3.5

Zhou, X. and M. Stephens (2012, July). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics 44*(7), 821–824. 1.2.2.4

Zollner, S. and J. K. Pritchard (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *American Journal of Human Genetics 80*(4), 605–615. 1.2.4

Appendix A

Supplementary Tables

Table A.1: Sites with Random Allelic Bias.

chrm	asSNP__pos	explanation	conconrdSNP__pos	score	pval	nhets
3	195380708	linked to putative rSNP	195343986	0.413746355	0.012	10
13	24473876	linked to putative rSNP	24429097	0.342206506	0.001	19
2	79226788	linked to putative rSNP	79234011	0.303255216	0.001	8
6	11656064	linked to putative rSNP	11641470	0.388698152	0.001	13
7	22893706	linked to putative rSNP	22892011	0.1773247	0.001	17
6	46378261	unexplained	46372838	0.131808112	0.298	23
12	562651	linked to putative rSNP	553733	0.313721634	0.001	13
12	562580	unexplained	536056	0.353706121	0.052	16
9	104211370	unexplained	104211370	0.056373159	0.904	9
10	52420020	unexplained	52463140	0.368795666	0.07	8
15	51166142	linked to putative rSNP	51127650	0.341124891	0.001	14
12	562661	linked to putative rSNP	573870	0.386307382	0.033	10
10	57391147	unexplained	57424504	0.438082956	0.056	11
14	106598644	unexplained, IgH	106619371	0.172644288	0.172	8
14	106627178	unexplained, IgH	106625322	0.241054365	0.16	10
5	171889524	linked to putative rSNP	171903769	0.475113379	0.028	9
8	80979777	linked to putative rSNP	80977521	0.165461923	0.026	20
7	141437957	linked to putative rSNP	141429029	0.360188346	0.001	11
5	110867634	linked to putative rSNP	110839262	0.172184058	0.001	8
10	32125803	unexplained	32164980	0.241839254	0.518	9
12	8086062	linked to putative rSNP	8086083	0.244866078	0.001	13
8	16870536	linked to putative rSNP	16845136	0.382392542	0.024	9
11	2554149	within imprinted gene KCNQ1	2552450	0.17695326	0.608	8
18	43303114	unexplained	43302764	0.278125778	0.098	8
3	34021986	linked to putative rSNP	34014691	0.321042367	0.046	9

Table A.1: Sites with Random Allelic Bias.

14	106626958	unexplained, IgH	106625322	0.251414942	0.144	9
18	61009609	linked to putative rSNP	61019692	0.424321072	0.002	9
21	39084764	unexplained	39113085	0.438694795	0.097	8
11	362099	linked to putative rSNP	344035	0.43792284	0.001	8
10	71168561	linked to putative rSNP	71119208	0.318892421	0.001	11
6	14397660	linked to putative rSNP	14357172	0.294558813	0.001	12
13	112092991	unexplained	112092991	0.288589199	0.054	12
1	40138507	linked to putative rSNP	40151426	0.296879492	0.003	11
16	75498793	unexplained	75524371	0.312558616	0.154	10
9	114360287	linked to putative rSNP	114357659	0.138953631	0.001	9
12	84220073	linked to putative rSNP	84262595	0.392129301	0.012	10
3	46484283	unexplained	46484283	0.353244439	0.163	7
10	102295658	unexplained	102298664	0.089200481	0.066	12
19	38042814	unexplained	38070460	0.375844732	0.129	9
7	142420355	linked to putative rSNP	142379936	0.14320426	0.002	17
10	134717856	unexplained	134671822	0.29972073	0.091	8
10	11800229	linked to putative rSNP	11801353	0.474904024	0.009	8
12	67835821	linked to putative rSNP	67803505	0.29289579	0.027	12
2	231807181	linked to putative rSNP	231769231	0.269179103	0.001	8
6	100270972	unexplained	100249423	0.31902441	0.188	9
10	835055	linked to putative rSNP	859270	0.361235687	0.028	15
19	19453560	unexplained	19445856	0.4082047	0.132	9
1	171220925	unexplained	171262373	0.410678075	0.147	9
1	248100467	linked to putative rSNP	248059456	0.216975022	0.026	17
12	108279247	unexplained	108263228	0.290672468	0.052	17

Table A.2: List of all pathways significantly associated with age, together with the number and proportion of significantly associated genes ($P < 0.05$, corrected using Bonferroni for the total number of genes in the pathway).

KEGG ID	Pathway	P value	Number of genes	Proportion of genes
00900	Terpenoid Backbone Biosynthesis	6.23E-13	chrm	asSNP_pos
00980	Metabolism of Xenobiotics By Cytochrome P450	6.47E-13	3	195380708
01040	Biosynthesis of Unsaturated Fatty Acids	1.11E-12	13	24473876
00100	Steroid Biosynthesis	1.33E-12	2	79226788

KEGG ID	Pathway	P value	Number of genes	Proportion of genes
00650	Butanoate Metabolism	1.51E-12	6	11656064
04146	Peroxisome	1.56E-12	7	22893706
00830	Retinol Metabolism	1.93E-12	6	46378261
00010	Glycolysis Gluconeogenesis	3.59E-12	12	562651
00051	Fructose and Mannose Metabolism	3.99E-12	12	562580
00290	Valine Leucine and Isoleucine Biosynthesis	1.15E-11	9	104211370
00561	Glycerolipid Metabolism	2.63E-11	10	52420020
00620	Pyruvate Metabolism	4.20E-11	15	51166142
00770	Pantothenate and COA Biosynthesis	4.76E-11	12	562661
00280	Valine Leucine and Isoleucine Degradation	5.79E-11	10	57391147
00020	Citrate Cycle TCA Cycle	1.12E-10	14	106598644
04916	Melanogenesis	3.34E-10	14	106627178
04910	Insulin Signalling Pathway	3.70E-10	5	171889524
00565	Ether Lipid Metabolism	5.89E-10	8	80979777
00350	Tyrosine Metabolism	9.44E-10	7	141437957
00640	Propanoate Metabolism	1.03E-09	5	110867634
04530	Tight Junction	1.12E-09	10	32125803
00030	Pentose Phosphate Pathway	1.74E-09	12	8086062
03320	PPAR Signalling Pathway	1.83E-09	8	16870536
00630	Glyoxylate and Dicarboxylate Metabolism	2.22E-09	11	2554149

KEGG ID	Pathway	P value	Number of genes	Proportion of genes
00982	Drug Metabolism Cytochrome P450	2.93E-09	18	43303114
00260	Glycine Serine and Threonine Metabolism	7.02E-09	3	34021986
00140	Steroid Hormone Biosynthesis	7.49E-09	14	106626958
00380	Tryptophan Metabolism	1.17E-08	18	61009609
04930	Type II Diabetes Mellitus	1.98E-08	21	39084764
05412	Arrhythmogenic Right Ventricular Cardiomyopathy Arvc	7.44E-08	11	362099
00052	Galactose Metabolism	9.27E-08	10	71168561
04340	Hedgehog Signaling Pathway	1.12E-07	6	14397660
00480	Glutathione Metabolism	1.45E-07	13	112092991
00532	Glycosaminoglycan Biosynthesis Chondroitin Sulfate	1.53E-07	1	40138507
04920	Adipocytokine Signaling Pathway	2.87E-07	16	75498793
05214	Glioma	4.26E-07	9	114360287
05322	Systemic Lupus Erythematosus	4.56E-07	12	84220073
05414	Dilated Cardiomyopathy	5.64E-07	3	46484283
00410	Beta Alanine Metabolism	1.11E-06	10	102295658
00330	Arginine and Proline Metabolism	1.39E-06	19	38042814
04510	Focal Adhesion	1.47E-06	7	142420355
00340	Histidine Metabolism	1.53E-06	10	134717856

KEGG ID	Pathway	P value	Number of genes	Proportion of genes
04360	Axon Guidance	1.66E-06	10	11800229
04060	ECM Receptor Interaction	1.77E-06	12	67835821
04150	MTOR Signaling Pathway	2.02E-06	2	231807181
04270	Vascular Smooth Muscle Contraction	3.31E-06	6	100270972
00071	Fatty Acid Metabolism	3.84E-06	10	835055
04142	Lysosome	4.43E-06	19	19453560
00983	Drug Metabolism Other Enzymes	5.71E-06	1	171220925
00040	Pentose and Glucuronate Interconversions	6.49E-06	1	248100467
05416	Viral Myocarditis	1.16E-05	12	108279247
1000	Amino Sugar and Nucleotide Sugar Metabolism	1.70E-05	7	0.179
05217	Basal Cell Carcinoma	1.80E-05	10	0.192
00510	N-Glycan Biosynthesis	1.82E-05	7	0.175
04260	Cardiac Muscle Contraction	1.83E-05	5	0.0847
05216	Thyroid Cancer	1.99E-05	8	0.364
05120	Epithelial Cell Signaling In Helicobacter Pylori Infection	4.85E-05	11	0.186

Table A.3: List of the seven pathways which were signicantly associated with age, discovered by looking for enrichment of single gene age associations.

KEGG ID	Pathway	P value

650	Butanoate Metabolism	8.86E+06
4060	ECM Receptor Interaction	3.64E+05
4146	Peroxisome	2.61E+07
620	Pyruvate Metabolism	5.49E+05
100	Steroid Biosynthesis	2.39E+11
900	Terpenoid Backbone Biosynthesis Valine Leucine and Isoleucine	2.13E+05
290	Degradation	5.58E+06

Table A.4: Key showing which pathways correspond to which nodes in Figure 3.2, and the maximum Spearman correlation of that phenotype with any of the others representing pathways.

Node	Pathway	Maximum rho with other phenotype
1	Glycolysis Gluconeogenesis	0.91
2	Citrate Cycle TCA Cycle	0.90
3	Pentose Phosphate Pathway	0.84
4	Fructose and Mannose Metabolism	0.84
5	Beta Alanine Metabolism	0.85
6	Glutathione Metabolism	0.85
7	Pyruvate Metabolism	0.81
8	Butanoate Metabolism	0.94
9	Drug Metabolism Cytochrome P450	0.84
10	Biosynthesis of Unsaturated Fatty Acids	0.92
11	Fatty Acid Metabolism	0.87
12	Glyoxylate and Dicarboxylate Metabolism	0.80
13	Glycerolipid Metabolism	0.90
14	Terpenoid Backbone Biosynthesis	0.90
15	Valine Leucine and Isoleucine Biosynthesis	0.84
16	Pantothenate and COA Biosynthesis	0.85
17	Tryptophan Metabolism	0.82
18	Peroxisome	0.92
19	Insulin Signaling Pathway	0.84
20	Propanoate Metabolism	0.92
21	Valine Leucine and Isoleucine Degradation	0.94
22	Retinol Metabolism	0.90
23	Steroid Hormone Biosynthesis	0.84
24	Steroid Biosynthesis	0.90

Table A.5: Heritability and proportion of variance explained by age for all pathways. Value reported is for the pathway phenotype most significantly associated with ageing.

KEGG ID	Pathway	Heritability	Proportion
00900	Terpenoid Backbone Biosynthesis	1.53E-11	0.0898
00980	Metabolism of Xenobiotics By Cytochrome P450	0.0904	0.0986
01040	Biosynthesis of Unsaturated Fatty Acids	0.253	0.11
00100	Steroid Biosynthesis	0.406	0.143
00650	Butanoate Metabolism	0.39	0.137
04146	Peroxisome	0.453	0.152
00830	Retinol Metabolism	0.449	0.149
00010	Glycolysis Gluconeogenesis	0.417	0.14
00051	Fructose and Mannose Metabolism	0.316	0.109
00290	Valine Leucine and Isoleucine Biosynthesis	2.61E-12	0.0771
00561	Glycerolipid Metabolism	0.337	0.113
00620	Pyruvate Metabolism	0.368	0.117
00770	Pantothenate and COA Biosynthesis	0.477	0.136
00280	Valine Leucine and Isoleucine Degradation	0.51	0.147
00020	Citrate Cycle TCA Cycle	0.436	0.126
04916	Melanogenesis	2.23E-16	0.0708
04910	Insulin Signaling Pathway	0.453	0.121
00565	Ether Lipid Metabolism	1.13E-15	0.064

KEGG ID	Pathway	Heritability	Proportion
00350	Tyrosine Metabolism	0.342	0.0975
00640	Propanoate Metabolism	0.591	0.157
04530	Tight Junction	0.103	0.0751
00030	Pentose Phosphate Pathway	0.291	0.0831
03320	PPAR Signaling Pathway	0.235	0.0777
00630	Glyoxylate and Dicarboxylate Metabolism	0.275	0.0836
00982	Drug Metabolism Cytochrome P450	0.248	0.0811
00260	Glycine Serine and Threonine Metabolism	0.599	0.141
00140	Steroid Hormone Biosynthesis	0.655	0.167
00380	Tryptophan Metabolism	0	0.0491
04930	Type II Diabetes Mellitus	0.594	0.13
05412	Arrhythmogenic Right Ventricular Cardiomyopathy Arvc	0.241	0.0674
00052	Galactose Metabolism	3.40E-11	0.0504
04340	Hedgehog Signaling Pathway	0.375	0.08
00480	Glutathione Metabolism	0.415	0.0804
00532	Glycosaminoglycan Biosynthesis Chondroitin Sulfate	0.273	0.0682
04920	Adipocytokine Signaling Pathway	1.30E-20	0.0475
05214	Glioma	0.102	0.0466

KEGG ID	Pathway	Heritability	Proportion
05322	Systemic Lupus Erythematosus	8.17E-17	0.045
05414	Dilated Cardiomyopathy	0.532	0.0867
00410	Beta Alanine Metabolism	0.709	0.14
00330	Arginine and Proline Metabolism	1.70E-16	0.0402
04510	Focal Adhesion	0.397	0.0669
00340	Histidine Metabolism	0.519	0.0874
04360	Axon Guidance	0.606	0.0995
04060	ECM Receptor Interaction	0.792	0.196
04150	MTOR Signaling Pathway	0.219	0.0511
04270	Vascular Smooth Muscle Contraction	0.27	0.0542
00071	Fatty Acid Metabolism	0.823	0.204
04142	Lysosome	0.566	0.0804
00983	Drug Metabolism Other Enzymes	0	0.0322
00040	Pentose and Glucuronate Interconversions	0.562	0.0792
05416	Viral Myocarditis	0.569	0.0815
00520	Amino Sugar and Nucleotide Sugar Metabolism	0.453	0.0577
05217	Basal Cell Carcinoma	0.593	0.0799
00510	N Glycan Biosynthesis	5.87E-16	0.0313
04260	Cardiac Muscle Contraction	8.30E-13	0.0312
05216	Thyroid Cancer	2.56E-09	0.0332

KEGG ID	Pathway	Heritability	Proportion
05120	Epithelial Cell Signaling In Helicobacter Pylori Infection	0.652	0.0859
04060	Cytokine Cytokine Receptor Interaction	3.51E-17	0.0276
00120	Primary Bile Acid Biosynthesis	1.69E-16	0.0265
00190	Oxidative Phosphorylation	1.41E-11	0.0268
00760	Nicotinate and Nicotinamide Metabolism	0.401	0.0433
00360	Phenylalanine Metabolism	0.711	0.088
00512	O Glycan Biosynthesis	1.78E-18	0.0253
05213	Endometrial Cancer	0.428	0.0408
00250	Alanine Aspartate and Glutamate Metabolism	0.526	0.0507
00564	Glycerophospholipid Metabolism	0	0.0231
04012	ERBB Signaling Pathway	0.121	0.0253
05211	Renal Cell Carcinoma	3.64E-11	0.0237
02010	ABC Transporters	0.506	0.0454
04710	Circadian Rhythm Mammal	0.0407	0.0292
05222	Small Cell Lung Cancer	1.03E-17	0.024
04062	Chemokine Signaling Pathway	0.124	0.0277
00590	Arachidonic Acid Metabolism	0.141	0.027
04610	Complement and Coagulation Cascades	0.504	0.0453

KEGG ID	Pathway	Heritability	Proportion
03022	Basal Transcription Factors	0.537	0.0424
00600	Sphingolipid Metabolism	8.68E-19	0.0219
05410	Hypertrophic Cardiomyopathy Hcm	3.30E-13	0.0147
04912	GNRH Signaling Pathway	3.11E-16	0.0187
04720	Long Term Potentiation	0	0.0183
03050	Proteasome	0.425	0.0314
04620	JAK Stat Signaling Pathway	0.503	0.0382
05330	Allograft Rejection	0	0.016
03450	Non Homologous End Joining	0.132	0.0199
05320	Autoimmune Thyroid Disease	0	0.0156
03060	Protein Export	0.235	0.0197
03420	Nucleotide Excision Repair	3.19E-14	0.0178
00660	Alpha Linolenic Acid Metabolism	0.458	0.0311
04144	Endocytosis	0.0714	0.0181
05010	Alzheimers Disease	0.0757	0.0172
00591	Linoleic Acid Metabolism	3.00E-11	0.0159
00240	Pyrimidine Metabolism	6.42E-13	0.0152
00270	Cysteine and Methionine Metabolism	0.00281	0.0162
03410	Base Excision Repair	0.377	0.0219
04722	Neurotrophin Signaling Pathway	4.88E-18	0.0152

KEGG ID	Pathway	Heritability	Proportion
04070	Phosphatidylinositol Signaling System	0.312	0.0207
04960	Aldosterone Regulated Sodium Reabsorption	3.36E-15	0.0142
05130	Pathogenic Escherichia Coli Infection	0.158	0.0158
04310	WNT Signaling Pathway	0.176	0.0174
00562	Inositol Phosphate Metabolism	3.24E-16	0.0138
05221	Acute Myeloid Leukemia	0.472	0.0268
00071	Selenoamino Acid Metabolism	3.71E-10	0.0137
04742	Taste Transduction	0.149	0.0174
00531	Glycosaminoglycan Degradation	2.23E-19	0.0135
05340	Primary Immunodeficiency	0	0.0133
04640	Hematopoietic Cell Lineage	2.35E-16	0.0132
05310	Asthma	0.331	0.0183
04620	TGF Beta Signaling Pathway	1.72E-18	0.0131
00860	Porphyrin and Chlorophyll Metabolism	9.84E-16	0.0124
04612	Antigen Processing and Presentation	2.03E-11	0.0129
05010	Parkinsons Disease	4.25E-09	0.012
00790	Folate Biosynthesis	1.07E-11	0.0119
00500	Starch and Sucrose Metabolism	0.429	0.0111
05223	Non Small Cell Lung Cancer	0	0.0115

KEGG ID	Pathway	Heritability	Proportion
03030	DNA Replication	0	0.0116
04622	RIG I Like Receptor Signaling Pathway	0	0.0117
04666	FC Gamma R Mediated Phagocytosis	0.747	0.0415
04514	Cell Adhesion Molecules CAMS	0.278	0.016
03430	Mismatch Repair	7.18E-17	0.011
03010	Ribosome	8.63E-19	0.0108
05220	Chronic Myeloid Leukemia	0.333	0.0164
00910	Nitrogen Metabolism	0	0.0106
04330	Notch Signaling Pathway	0.585	0.0251
04520	Adherens Junction	1.15E-09	0.0107
05210	Colorectal Cancer	0.289	0.0141
03018	RNA Degradation	1.03E-13	0.00998
03440	Homologous Recombination	0	0.0093
00920	Sulfur Metabolism	0.121	0.011
00310	Lysine Degradation	0.446	0.0166
04662	B Cell Receptor Signaling Pathway	0.494	0.0183
00430	Taurine and Hypotaurine Metabolism	8.53E-13	0.00891
04964	Proximal Tubule Bicarbonate Reclamation	0.456	0.0163
04614	Renin Angiotensin System	0.556	0.0183
00970	Aminoacyl tRNA Biosynthesis	0.107	0.0102

KEGG ID	Pathway	Heritability	Proportion
04672	Intestinal Immune Network For IGA Production	0	0.00883
04810	Regulation of Actin Cytoskeleton	0.215	0.0104
05215	Prostate Cancer	1.55E-09	0.00719
00563	Glycosylphosphatidylinositol Gpi Anchor Biosynthesis	0	0.00816
04660	NOD Like Receptor Signaling Pathway	0	0.00828
04540	Gap Junction	0.121	0.0096
00903	Limonene and Pinene Degradation	4.80E-12	0.00822
05200	Pathways In Cancer	0.275	0.0119
04660	Toll Like Receptor Signaling Pathway	8.13E-17	0.00782
04730	Long Term Depression	0.128	0.00885
04020	Calcium Signaling Pathway	0.148	0.00936
04320	Dorso Ventral Axis Formation	0.271	0.00857
05110	Vibrio Cholerae Infection	0.353	0.011
04115	P53 Signaling Pathway	1.07	-0.0975
04962	Vasopressin Regulated Water Reabsorption	0.331	0.0107
04670	Leukocyte Transendothelial Migration	0.248	0.00871
03020	RNA Polymerase	2.52E-16	0.00609
04664	FC Epsilon RI Signaling Pathway	0.35	0.00908
04140	Regulation of Autophagy	0	0.00509

KEGG ID	Pathway	Heritability	Proportion
05010	Huntingtons Disease	0.894	0.0529
00670	One Carbon Pool By Folate	9.11E-13	0.00564
04660	T Cell Receptor Signaling Pathway	0.487	0.0103
00740	Riboflavin Metabolism	0.252	0.00627
00533	Glycosaminoglycan Biosynthesis Keratan Sulfate	0	0.00452
00230	Purine Metabolism	3.84E-18	0.00462
04130	Snare Interactions In Vesicular Transport	1.20E-17	0.00475
05020	Prion Diseases	0.272	0.0059
05219	Bladder Cancer	0.229	0.00531
03040	Spliceosome	0.224	0.00573
04010	Mapk Signaling Pathway	0.221	0.00506
00534	Glycosaminoglycan Biosynthesis Heparan Sulfate	1.40E-18	0.00416
00604	Glycosphingolipid Biosynthesis Ganglio Series	0	0.00372
04940	Type I Diabetes Mellitus	0.446	0.00735
04623	Cytosolic DNA Sensing Pathway	0.431	0.00706
05332	Graft Versus Host Disease	0.432	0.00691
04740	Olfactory Transduction	0	0.0035
04110	Cell Cycle	5.02E-18	0.00369
00511	Other Glycan Degradation	1.07E-24	0.00321
05140	Leishmania Infection	0.136	0.00381

KEGG ID	Pathway	Heritability	Proportion
04914	Progesterone Mediated Oocyte Maturation	1.82E-19	0.00322
04120	Ubiquitin Mediated Proteolysis	2.55E-15	0.00315
00604	Glycosphingolipid Biosynthesis Globo Series	0	0.00271
00601	Glycosphingolipid Biosynthesis Lacto and Neolacto Series	0.213	0.00341
04370	VEGF Signaling Pathway	0.192	0.00362
00053	Ascorbate and Aldarate Metabolism	0	0.00197
04650	Natural Killer Cell Mediated Cytotoxicity	4.16E-19	0.00222
05212	Pancreatic Cancer	5.99E-48	0.00212
04114	Oocyte Meiosis	1.82E-11	0.00201
04210	Apoptosis	0.632	0.00523
05218	Melanoma	0.349	0.00284
04080	Neuroactive Ligand Receptor Interaction	1.76E-17	0.00158
05014	Amyotrophic Lateral Sclerosis ALS	0	0.00102
04950	Maturity Onset Diabetes of The Young	8.21E-12	0.000707