# A genetic analysis of molecular traits in skeletal muscle

**D Leland Taylor**

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Jesus College

March 2018

Dedicated to my family

Give me a sharp sense of understanding, a retentive memory, and the ability to grasp things correctly and fundamentally. Grant me the talent of being exact in my explanations and the ability to express myself with thoroughness and charm.

Thomas Aquinas

# Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

This dissertation does not exceed the specified length limit of 60,000 words as defined by the Biology Degree Committee.

D Leland Taylor
March 2018

# Acknowledgements

First and foremost, thank you to my supervisors—Francis Collins and Ewan Birney. It has been my privilege to learn from you. Your inspirational leadership, wisdom, gentle guidance, generosity with time, and enthusiastic encouragement has formed both the way I approach scientific research as well as life in general. I am forever grateful to you both.

Thank you to my thesis advisory committee—John Danesh, Nicole Soranzo, and Ian Dunham—for your active interest in my research, meaningful feedback, and guidance in my maturation as a scientist.

I also wish to express my appreciation to the members of the Collins and Birney laboratories for your friendship, support, and encouragement. Most notably, without the close collaboration of Narisu Narisu, Peter Chines, and Michael Erdos, this research would not have been possible. Thank you to those who processed the tissue biopsies and generated the data analysed in this thesis: Michael Erdos, Amy Swift, and Jackie Idol. Thank to the broader FUSION team—especially Michael Boehnke, Laura Scott, Stephen Parker, Anne Jackson, and Ryan Welch—for your continual support and guidance. In addition, thank you to other collaborators—Francesco Paolo Casale, Oliver Stegle, and David Knowles. It has been a privilege to work with and learn from such brilliant and thoughtful people.

It be remiss of me not to highlight and acknowledge my gratitude to the dedicated staff of the NIH-OxCam program, NHGRI, EMBL-EBI, University of Cambridge, and Jesus College. Without your support this research would not have been possible. Thank you to the Foundation for Advanced Education in the Sciences and the NIH Office of Intramural Training & Education for the support you provide to the NIH trainee community. Special thanks to the International Biomedical Research Alliance for your support of the NIH-OxCam program and to the Albert and Mary Lasker Foundation for funding the Lasker Lessons in Leadership seminar series in partnership with NIH-OxCam. On the European side of the Atlantic, thanks especially to EMBL-EBI for the exceptional support and training environment you provide to all PhD students.

I am particularly grateful for the broader communities of the University of Cambridge, Jesus College, and the town of Cambridge—where the pursuit of knowledge abounds in graceful hospitality and rich history. Special thanks to the Scriptorium study and writing centre at the historic Round Church, where I wrote a large portion of this thesis.

Finally, thank you to my community of family and friends. It is impossible for me to fully articulate how thankful I am for your unceasing support and encouragement.

# Abstract

Genome Wide Association Studies (GWASs) have identified variants associated with disease that promise to deliver insights into disease aetiology. However, because many GWAS variants lie in non-coding genomic regions, it is difficult to define the genes and pathways underlying a GWAS signal. The possibility of linking GWAS variants to molecular traits, combined with the development of high throughput assays, has motivated the mapping of molecular quantitative trait loci (QTLs), genetic associations with molecular traits such as gene expression (eQTLs) and DNA methylation (mQTLs).

The Finland-United States Investigation of NIDDM (FUSION) tissue biopsy study is motivated by the desire to understand the molecular pathogenesis of Type 2 diabetes (T2D), a complex disease where the vast majority of the ~100 independent GWAS loci occur in non-coding regions. To elucidate the molecular mechanisms underlying these signals, we collected skeletal muscle biopsies, a T2D-relevant tissue, from 318 extensively phenotyped individuals who exhibit a range of glucose tolerance levels. From these biopsies, we generated genotype, gene expression, and DNA methylation information, enabling us to directly measure the effects of T2D on molecular traits, and to link non-coding T2D GWAS loci to candidate molecular targets. In this thesis, I present a catalogue of genetic effects on gene expression and DNA methylation. I use this catalogue firstly, to reveal basic biology of the genetic regulators of skeletal muscle molecular traits, and secondly, to identify molecular traits that are relevant to T2D, glycemic, and other complex traits.

In regards to basic biology, I characterise the broader genomic context of QTLs by calculating the enrichment of QTLs in chromatin states across a diverse panel of cell/tissue types. I also identify key skeletal muscle transcription factors (TFs) and classify them as activators or repressors by aggregating the effects of QTLs predicted to perturb TF binding sites. In addition, I characterise the properties of methylation sites associated with gene expression and use inference models to dissect these methylation-expression relationships, classifying cases where the genetic effect is mediated by methylation, expression, or is independent.

I also integrate molecular trait genetics with complex traits. First, I perform a conditional analysis, mapping GWAS variants for T2D and glycemic traits to molecular traits, prioritising disease relevant skeletal muscle molecular traits. Second, recognising QTLs may also be specific to a disease state or environmental context, I leverage the rich phenotyping of participants to map genotype by environment (GxE) effects on gene expression—eQTLs that exhibit effects specific to an environmental context. Altogether, these analyses form a thorough survey of the genetic regulators of skeletal muscle expression and DNA methylation, and provide an important resource for interpreting complex diseases.

# Table of contents

# List of figures

# List of tables

# Nomenclature

### Acronyms / Abbreviations

| | |
|---|---|
| 5caC | 5-carboxylcytosine |
| 5fC | 5-formylcytosine |
| 5hmC | 5-hydroxymethylcytosine |
| 5mC | 5-methylcytosine |
| ABC | ATP binding cassette |
| ACh | acetylcholine |
| AChE | acetylcholinesterase |
| ASE | allele specific expression |
| ATAC-seq | assay for transposase-accessible chromatin with sequencing |
| ATP | adenosine triphosphate |
| BMI | body mass index |
| bp | base pair |
| caQTL | chromatin accessibility quantitative trait locus |
| cDNA | complementary DNA |
| CGI | CpG island |
| ChIP-seq | chromatin immunoprecipitation followed by sequencing |

| | |
|---|---|
| chr | chromosome |
| CIDR | Center for Inherited Disease Research |
| cM | centimorgan |
| CpG | cytosine phosphate guanine |
| DHS | DNase1 hypersensitivity site |
| DIAGRAM | DIAbetes Genetics Replication And Meta-analysis |
| DNA | deoxyribonucleic acid |
| DNAme | DNA methylation |
| DNMT | DNA methyltransferase |
| dsQTL | DNaseI sensitivity quantitative trait locus |
| EBI | EMBL European Bioinformatics Institute |
| EBV | epstein-barr virus |
| EMBL | European Molecular Biology Laboratory |
| EMSA | electrophoretic mobility shift assay |
| ENCODE | Encyclopedia of DNA Elements |
| eQTL | expression quantitative trait locus (gene level expression) |
| eQTM | expression quantitative trait methylation |
| ER | endoplasmic reticulum |
| ERCC | External RNA Controls Consortium |
| ES | embryonic stem (cell) |
| exQTL | exon expression quantitative trait locus |
| F1 | first filial (generation) |
| FDR | false discovery rate |
| FPKM | fragments per kilobase per million reads |

| | |
|---|---|
| FUSION | Finland-United States Investigation of NIDDM (genetics) |
| GoT2D | Genetics of T2D (consortium) |
| GTEx | Genotype-Tissue Expression project |
| GWAS | genome wide association study |
| GxE | genotype by environment effect |
| HDL | high-density lipoprotein |
| HipSci | Human Induced Pluripotent Stem Cell Initiative |
| hQTL | histone quantitative trait locus |
| HRC | Haplotype Reference Consortium |
| IFG | impaired fasting glucose |
| IGT | impaired glucose tolerance |
| iPS | induced pluripotent stem (cell) |
| kb | kilobase pair (1,000 bp) |
| LC-CoA | long chain acyl coenzyme A |
| LCL | lymphoblastoid cell line |
| LD | linkage disequilibrium |
| LDLc | low-density lipoprotein cholesterol |
| LMR | low methylated region |
| m.u. | (genetic) map units |
| MAC | minor allele count |
| MAF | minor allele frequency |
| MAGIC | Meta-Analysis of Glucose and Insulin-related traits Consortium |
| Mb | megabase pair (1,000,000 bp) |
| mESI | muscle expression specificity index |

| | |
|---|---|
| MeSS | methylation specificity score |
| Meth | methylated |
| MHC | major histocompatibility complex |
| miRNA | microRNA |
| MODY | maturity onset diabetes of the young |
| mQTL | methylation quantitative trait locus |
| MR | Mendelian randomisation |
| mRNA | messenger RNA |
| ncRNA | non-coding RNA |
| NGS | next generation sequencing |
| NGT | normal glucose tolerance |
| NHGRI | National Human Genome Research Institute |
| NIDDM | non-insulin-dependent diabetes mellitus |
| NIH | National Institutes of Health |
| NISC | NIH Intramural Sequencing Center |
| OGTT | oral glucose tolerance test |
| OMIM | Online Mendelian Inheritance in Man |
| PC | principal component |
| PCA | principal component analysis |
| PCR | polymerase chain reaction |
| Pol II | RNA polymerase II |
| PWM | position weight matrix |
| QC | quality control |
| QTL | quantitative trait locus |

| | |
|---|---|
| RCT | randomised controlled trial |
| reQTL | (environmental) response expression quantitative trait locus |
| RFLP | restriction fragment length polymorphism |
| RIN | RNA integrity number |
| RMSE | root mean square error |
| RNA | ribonucleic acid |
| RPKM | reads per kilobase per million reads |
| rRNA | ribosomal RNA |
| SkMC | skeletal muscle cell (cultured) |
| SNP | single nucleotide polymorphism |
| sQTL | splicing quantitative trait locus |
| SR | sarcoplasmic reticulum |
| T2D | type 2 diabetes |
| T2D-GENES | T2D Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples |
| TC | total cholesterol |
| TET | ten-eleven translocation (protein) |
| TF | transcription factor |
| TFBS | transcription factor binding site |
| TG | triglyceride |
| TPM | transcripts per million |
| tRNA | transfer RNA |
| tSNP | transcribed single nucleotide polymorphism |
| TSS | transcription start site |

| | |
|---|---|
| Umeth | unmethylated |
| WGBS | whole genome bisulfite sequencing |
| WGS | whole genome sequencing |
| WTCCC | Wellcome Trust Case Control Consortium |

# Chapter 1

# Introduction

## 1.1 Preface

This thesis describes a genetic analysis of gene expression and DNA methylation (DNAme) in skeletal muscle. This is one of the largest studies of DNA methylation in skeletal muscle, and the only study with extensive phenotypic characterisation of the tissue donors. Unless specified otherwise, as in segments of Chapter 2, Chapter 4, and Chapter 6, I conducted the research presented in this thesis. In Chapter 2, I describe the quality control measures taken in this study. In Chapter 3, I characterise the relationship between expression and DNAme. In Chapter 4, I analyse maps of genetic regulators of gene expression and DNAme. In Chapter 5, I analyse the effects of disease-associated genetic variation on gene expression and DNAme. Finally in Chapter 6, I map context-specific genetic effects on gene expression.

This study was motivated by the desire to understand the molecular effects of genetic variants associated with risk of type 2 diabetes (T2D) and T2D-related traits. In this introductory chapter, I describe the underlying motivation for this study in the context of T2D genetics. Since T2D genetics has advanced rapidly in the last decade with the advent of studies on common genetic variation, I begin this chapter by giving a brief overview of the history of genetics. Following this overview, I describe T2D pathophysiology and T2D genetics, highlighting the motivations for this study. Finally, I position this study in relation to the larger field of molecular quantitative trait genetics.

## 1.2 A brief history of genetics

### 1.2.1 Etymology

The use of the term genetics to describe a scientific discipline was first suggested by William Bateson in a letter dating 18 April 1905 to Adam Sedgwick, discussing the founding of a "Professorship relating to Heredity and Variation" [25, p. 93] at the University of Cambridge [72]. In an attempt to compress these ideas and describe this emerging discipline, Bateson suggested the word *genetics*, presumably drawn from the Greek word 'genos' (plural 'genē') referring to a group claiming common descent or kinship [75].[1] One year later, at the Royal Horticultural Society's 1906 International Conference on Plant Hybridisation, the term was popularised when Bateson proposed using genetics to describe the new science of heredity based on the laws of an Austrian monk, Gregor Mendel. The proposal was met with enthusiasm and the proceedings were published as a the "Report of the Third International Conference 1906 on Genetics" [324].

### 1.2.2 Mendel's laws of inheritance

This new field of genetics was founded on the seminal work by Gregor Mendel who, analysing how traits were passed on from parent to offspring in peas (*Pisum sativum*), deduced a series of laws describing inheritance. In his now famous experiments, Mendel treated wrinkled and round pea seeds as traits and crossed wrinkled peas with round peas. Contrary to the intermediate wrinkled/round pea hybrid that may have been expected, Mendel observed only round peas in the resulting offspring or first filial (F1) generation. When Mendel crossed F1 peas with each other, he found the wrinkled trait reappeared at a roughly 1:3 ratio in the F2 generation—one wrinkled to every three round. Remarkably, the *recessive* wrinkled trait was not lost in the F1 generation, even though the round trait was completely *dominant* in the F1 generation. Mendel experimented with seven other pea traits such as pea colour and stem length, and even combined pairs of traits. Based on his observations, he published a famous set of laws in 1866, laying the foundational principles of inheritance [249]:

[1] Some have also drawn a link to the Greek word 'genesis' meaning origin [284, 283]

1. Two alleles are present for any given trait and are passed on to the offspring at random (law of segregation).
2. Separate traits are inherited independently (law of independent assortment).
3. Recessive alleles are masked by dominant alleles (law of dominance).

Unfortunately, Mendel's ideas went unappreciated by the scientific community at the time.

### 1.2.3 Rediscovery of Mendel and the biometrician dispute

Nearly 30 years later, in 1900, Mendel's research was rediscovered and separately replicated by three botanists: Hugo de Vries, Erich von Tschermak, and Carl Correns [374, pp. 25-29]. However, it was the aforementioned William Bateson who played a central role popularising Mendel's ideas (reviewed in [127]). In May 1900, Bateson presented Mendel's ideas and de Vries' work to the Royal Horticultural Society in London [374, p. 30]. Later in 1902, Bateson published *Mendel's Principles of Heredity: A Defense* [26]. Bateson wrote this book in response to a paper published earlier that year by W.F.R. Weldon [418] critiquing Mendel's laws [127].[2]

Weldon was part of an intellectual community known as the biometricians, who can be traced back to Francis Galton, the half-cousin of Charles Darwin. Inspired by Darwin's *The Origin of Species* [73], one of Galton's life passions was to understand the heredity of variation in humans [292, p. 86]. This passion was likely motivated by his unfortunate desire to apply breeding principles to the human race [116], which lead to his coining of the term 'eugenics' in 1883, stemming from the Greek 'eugenes' which means good in stock or "hereditarily endowed with noble qualities" [119].[3] To that end Galton applied himself to the study of biometrics, quantifying a myriad of human traits including height, reaction time, strength of pull, and even intelligence [127].

By all accounts, Galton was an innovator. In his career, he established many concepts commonly used today. He helped pioneer the use of fingerprints in forensics [368]. In

[2] In this same book and an earlier lecture, Bateson introduced terms like allelomorph, meaning alternative forms of a Mendelian factor, as well as homozygote and heterozygote, describing individuals carrying the same or different alleles at a given locus.

[3] As described earlier, 'genos' refers to a kinship group, and 'eu-' is simply a prefix meaning good.

statistics, he developed concepts like regression towards the mean [116] and independently rediscovered the idea of a correlation coefficient [118]. He even laid the foundations for the concept of using of twins to study the "powers nature and nurture" [117],[4] a term he popularised presumably from Shakespeare's *The Tempest* Act 4 Scene 1 [346].

Importantly, Galton also developed a theory of inheritance, the Law of Ancestral Heredity [120, 48], which states the heritable component of a trait can be calculated as a continuous series, such that parents contribute one-half of the total heritage (0.5), grandparents one-quarter $(0.5)^2$, great-grandparents one-eighth $(0.5)^3$, and so on. Thus, the Law of Ancestral Heredity separates the contributions of each ancestor to the total heritage. This theory of inheritance was pitted against Mendel's theory of inheritance, which led to a heated and at times personal dispute between the Mendelians, lead by Bateson, and the biometricians, lead by W.F.R. Weldon and the famous Karl Pearson (reviewed in [127]).

Over time, the dispute came to center on the mechanism of heritability for continuous, *quantitative* traits (reviewed in [127, 298]). On one side, the Mendelians prescribed a set of laws that elegantly described inheritance patterns of discrete, *qualitative* traits. Furthermore, these laws were beginning to be linked to biological mechanisms (see Section 1.2.4 below). On the other hand, the biometricians argued that quantitative traits are inherently continuous and therefore Mendel's qualitative laws could not possibly apply. Moreover, the biometricians questioned the assumption that many qualitative traits were really discrete, and did not actually exist on a continuum. This continuum may be readily observable (e.g., [127, p. 72]) or may potentially lie in a latent *liability distribution*, such that only segments of this distribution (e.g., the extremes) present with a phenotype [291].

Although G. Udny Yule suggested Mendelian principles may underlie the Law of Ancestral Heredity in 1902 [437], the feud continued to simmer until finally Ronald Fisher concisely ended the debate in 1918 with a paper [99] that became the cornerstone for the modern field of *quantitative genetics*. In his seminal paper, Fisher demonstrated a normal distribution can emerge from the sum of multiple Mendelian (i.e., genetic) factors of small, roughly equal effects that are individually inherited in a Mendelian fashion (see also [298, Box 1]). Fisher's ingenious insight, made in the 20th century, is still highly relevant today, as modern genetics research has shown that the genetic architecture of many complex traits is indeed highly polygenic, consisting of many variants of small effect size (see Section 1.2.6 and Section 1.3.2.3 below).

[4] Although important, this was not the classical twin study of today, as the mechanism of inheritance had not fully been worked out [309].

### 1.2.4 Chromosomal theory and linkage

Returning to the rediscovery of Mendel, part of the interest in Mendel's ideas stemmed from emerging research in the cytology community. At the time, cytologists were captivated by structures that appeared around cell division and, when stained, looked like coloured (khrōma) bodies (sōma), termed *chromosomes* [281]. Most intriguingly, during meiosis, these chromosomes, which appear to act independently of one another, moved through a beautifully choreographed dance where they pair and equally segregate into gametes. Around 1902, Walter Sutton and Theodor Boveri independently recognised such behaviour of chromosomes during meiosis closely paralleled the properties of Mendel's particles, and posed the chromosomal theory of inheritance, or the Boveri-Sutton chromosome theory, that states chromosomes are the material of inheritance [377, 378, 42].

The chromosomal theory of inheritance ignited vigorous debate, as there were significant, unresolved issues in synthesising Mendel's laws with chromosomal theory. Specifically, Mendel's second law of independent assortment states that traits are inherited independently. This law raised many questions. Is each chromosome an independent entity? If chromosomes truly house or are themselves Mendel's particles, how can traits be independent given that there are more traits than chromosomes, and therefore some chromosomes must have multiple traits? Was Mendel simply lucky in choosing traits that localised on different chromosomes in peas?

Troubled by this incomplete synthesis, Thomas Morgan studied the chromosomal theory of inheritance in relationship to Mendel's laws of inheritance. Analysing fruit fly (*Drosophila melanogaster*) trait inheritance patterns, Morgan observed that recombinant (non-parental) allele combinations in the F2 generation of a test cross did not always follow the 50% recombinant rate predicted by Mendel's second law of independent assortment. Informed by the cytologists' microscopy images, Morgan hypothesised that information was exchanged between chromosomes during meiosis, termed crossover, and that the recombination rates reflected the likelihood of crossover events which changed according to the chromosomal distance between the loci of two alleles [262]. A talented student of Morgan's, Alfred Sturtevant, conclusively proved this theory of "loci along a linear structure" in 1913 by integrating recombinant rates of multiple traits and showing recombination rates could be predicted additively, as one would expect given a linear structure. Thus, the first chromosomal linkage map was made which measured the distance between two loci in genetic map units (m.u.; also called centimorgan, cM) where 1 product of 100 meiosis events is recombinant

[373]. These experiments also compellingly synthesised Mendel's laws of inheritance and the chromosomal theory of inheritance.

Following Morgan and Sturtevant, the next years saw a plethora of revolutionary discoveries including conceptually connecting genes to enzymes via George Beadle and Edward Tatum's "one gene-one enzyme" hypothesis in 1941 (bringing together genetics and biochemistry) [28],[5] determining the structure of deoxyribonucleic acid (DNA) by Francis Crick and James Watson in 1953 [412], and the invention of technology to sequence DNA by Frederick Sanger in 1977 [334]. These developments, along with many others (reviewed in [374, 124]), enabled the study of human genetics in ways that were previously unimaginable.

### 1.2.5 Human linkage studies

Similar to the phenotypic markers, like body or eye colour, that Morgan and Sturtevant used to build linkage maps in fruit flies, molecular markers can also be used for linkage mapping, and through the development of molecular biology, many molecular markers emerged. Initially, restriction fragment length polymorphisms (RFLPs) were the primary marker sets [86, 40], but with the advent of polymerase chain reaction (PCR) microsatellites (tandemly repeated DNA sequences that produced length polymorphisms) became widely used [414, 276, 417]. Further advances in genome sequencing and microarray technology (reviewed in Section 1.4.1.2) now provide the ability to directly assay the genotype of a sample with dense single nucleotide polymorphisms (SNPs), and these are the primary markers used in modern studies.

To be more specific, early human genetic studies in the 1980s applied RFLP linkage maps of molecular markers to disease pedigrees across multiple families to locate a disease locus within the human genome [40, 198]. RFLPs were identified by digesting DNA with restriction enzymes and separating the resulting DNA fragments by length through gel electrophoresis. The fragment patterns were imaged as Southern blots, where bands were representative of different alleles. By analysing RFLP patterns of a pedigree in conjunction with RFLP linkage maps that oriented RFLPs relative to each other [86], it was possible to use RFLPs as milestone markers to close in on a specific disease locus. The closer an RFLP was to

[5] Although this connection was already made in 1902 by Archibald Garrod studying alkaptonuria [122]. Like Mendel's laws, the idea that genes were connected to enzymes was largely unappreciated by the scientific community.

the causal gene, the less likely that recombination events would occur, and therefore strong cosegregation patterns will be observed among the presenting families.

Once a region was identified, it could be narrowed further by using additional markers of increased density in the specific region (fine mapping). If a researcher could narrow down to a small enough region, it could be sequenced and candidate genes identified. But prior to the development of complete physical maps and high throughput DNA sequencing, this kind of positional cloning was a long and painstaking process [334]. Until 1989, the only disease genes identified by this approach were those in which rare chromosomal rearrangements pointed to the precise location of the responsible gene, such as Duchenne muscular dystrophy [269]. The successful identification of the cystic fibrosis gene in 1989 by pure positional cloning [320, 313, 183] proved that this technique could be successfully used in the absence of such chromosomal rearrangements, though the work took many years.

With the success of the Human Genome Project [199, 396] and the advent of modern sequencing technologies (reviewed in [131]), linkage studies are now far less laborious and can cheaply be performed at the level of DNA sequence. Linkage studies continue to be extremely useful for *Mendelian traits*, a subset of traits caused by, in its simplest form, a single mutation of high penetrance (meaning a high proportion of the cases with the mutation also exhibit the trait of interest), or allelic heterogeneity, where multiple rare mutations occur at the same locus giving rise to the phenotype [280]. At the time of writing, more than 4,000 Mendelian disorders have had their precise molecular cause uncovered by this approach—most of those in the last 15 years [64].

### 1.2.6 Genome wide association studies (GWASs)

Despite early successes for rare, monogenic diseases like cystic fibrosis, little headway was made in unraveling the genetics of common diseases and traits using linkage studies. Recognising such results may signal a genetic architecture similar to Fisher's 1918 paper [99], defined by common genetic variants of small effect sizes, visionary calls [314, 197] were made for common variant association studies or genome wide association studies (GWASs), a study design that would be better powered to detect polygenic effects where allele frequencies are compared between cases and controls or associated with continuous traits. These visionary calls stimulated the development and application of technologies to cheaply assay the genetic diversity of human populations (reviewed in [194]).

The pivotal technological advancements that enabled GWA studies were the development of microarray technology (described in Section 1.4.1.2), combined with growing catalogues of common human SNPs [327, 162, 163], and crucially the ability to pinpoint these SNPs precisely along the sequence of the human genome [199, 396].

Human genetic variation is rare and distributed across the genome: one person has on average of ~10 million common SNPs out of the 3.2 billion base pairs (bp) in the human genome [193]. Furthermore, variants are inherited non-randomly in linkage disequilibrium (LD) blocks, neighbourhoods of 1 to ~100s of kilobases (kb), bounded by recombination hotspots [181, 134]. Because *Homo sapiens* has a relatively recent origin and recombination is rare outside of hotspots, specific combinations of alleles tend to travel together on a particular chromosome, known as haplotypes. This feature of the human genome provides a significant advantage for GWA studies. By cataloguing the common SNP pool through large scale SNP discovery and haplotype mapping [327, 162, 163], researchers did not need to assay every SNP to capture genome wide information (which contains redundant information due to LD), but rather could scale down to a subset of SNPs that tagged each LD block and would fit on a single microarray, typically consisting of ~200 thousand to 2 million probes [401]. Subsequently, after genotyping a participant, the individual's patterns of genetic variation could be matched against a database of more complete haplotypes and the missing, unmeasured genetic variation could be accurately recovered through statistical imputation [239, 45, 220]. Of course, such methods only work for common genetic variation with a minor allele frequency (MAF) > ~1%; nonetheless, for common variants, these innovations enabled high throughput and cost effective genotyping across thousands of participants, leading to successful mapping of polygenic traits.

In the early 2000s, the first GWA studies were launched and published results began to appear in as early as 2005 [187]. By 2007, the feasibility of GWA studies across a variety of traits was firmly established with publication of the Wellcome Trust Case Control Consortium (WTCCC) papers, which encompassed 7 common diseases of major importance to public health (including T2D as described in Section 1.3.2.2) [419]. Since then, the field has expanded at a phenomenal rate, encompassing a host of common human traits and yielding many insights, especially in the area of complex disease (reviewed in [400, 401]). Overall, two clear trends have emerged [400]. First, it is clear that many complex traits are highly polygenic, influenced by the combined small effects of hundreds to thousands of common variants, perhaps even more than predicted by the early biometricians. Second, pleiotropy, where a variant influences multiple apparently unrelated phenotypes, is the rule for complex traits rather than the exception (reviewed in [399, 401]).

### 1.2.7 Interpreting GWAS loci

Despite the routine success of GWA studies, only a small fraction of the ~10,000 [231] independent variant-trait associations from GWA studies have led to the identification of specific genes or molecular mechanisms underlying complex diseases and traits. Increasing our knowledge of the effect of trait-associated genetic variation on specific genes and molecular mechanisms would enable targeted development of efficacious treatments and interventions. The knowledge gap is due to the fact that the vast majority of the GWAS loci for complex traits lie in non-coding portions of the genome [243, 409]. Because these loci are often surrounded by several genes or no genes at all (i.e., gene deserts), it is far from straightforward to identify target genes and molecular pathways through which they exert their effects. Furthermore, sets of variants are commonly inherited in tandem, due to LD, thereby obscuring the actual causal variant (or series of causal variants) identified in the region of a GWAS association.

Driven by the desire to hone in on the specific causal variants and genes underlying disease, and enabled by the development of sequencing-based high throughput molecular assays, the biomedical research community launched efforts to understand the effects of genetic variation in the context of molecular traits across human cell and tissue types. Through global projects like the Encyclopedia of DNA Elements (ENCODE) [90], NIH Roadmap Epigenomics [316], and BLUEPRINT [2], the focus has been to chart a map of the regulatory landscape for key cell and tissue types, generating public databases of gene expression and epigenomic signatures such as transcription factor binding, histone modifications, and chromatin interactions. These resources have enabled researchers to (1) identify key tissues underlying disease based on enrichment of GWAS loci in regulatory features, and (2) narrow down a GWAS locus to a smaller subset of variants that are likely to perturb key regulatory features. Building on the success of earlier studies (described in Section 1.4.3), the Genotype-Tissue Expression (GTEx) project [137] was also launched to further enrich these databases with *in vivo* information on gene expression in multiple tissues from large numbers of deeply genotyped individuals. By testing for associations between genetic variation and a molecular trait, molecular *quantitative trait loci* (QTLs) can be identified. Such studies are extremely powerful, as one can intersect loci from GWA studies with loci from QTL studies to identify candidate molecular traits underlying a GWAS locus.

This thesis is a QTL study and is focused on two molecular traits, gene expression and DNAme, in skeletal muscle. This study is motivated by the desire to fill in the knowledge gap particularly for T2D GWAS loci, as skeletal muscle is a relevant T2D tissue. Before describing how this study fits in the context of other QTL studies, I first give an overview of T2D aetiology and genetics, as T2D constitutes the underlying motivation for the study.

## 1.3 Type 2 diabetes (T2D)

### 1.3.1 T2D background and aetiology

One of the most studied polygenic diseases is non-insulin-dependent diabetes mellitus (NIDDM), also known as type 2 diabetes (T2D). Over the last decade, GWA studies have identified > 100 independent risk loci with high confidence [340, 82, 419, 439, 88, 402, 264, 83, 111, 342]. A sense of urgency surrounds research on this disease as T2D accounts for ~90% of diabetes cases, and diabetes affects ~415 million people worldwide and is projected to increase 55% by 2040 [161]. Moreover, T2D is the 6th leading cause of death in the world, costs an estimated $673 billion globally, and disproportionately affects individuals in lower socioeconomic segments of society (2015 estimates [161, 426]).

Biologically, T2D is characterised by insulin resistance and dysfunction in insulin secretion. Insulin, produced by the pancreatic islet beta cells, plays a central role in regulating blood glucose, keeping glucose confined to a narrow range [176] through complex interactions with a variety of tissues. Increased blood glucose levels trigger insulin release spikes that promote glucose uptake in muscle, halt fat breakdown in adipose, stimulate glucose storage in adipose, block increased glucose production in the liver, and trigger the hypothalamus to regulate appetite (reviewed in [372, 177]). When this homeostatic equilibrium is disrupted, blood glucose levels become unregulated leading to the onset of T2D and the accompanying health difficulties. Severe elevation of glucose (hyperglycemia) can lead to diabetic coma. But even modest hyperglycemia over many years produces long term health complications, such as cardiovascular disease, neuropathy, nephropathy, retinopathy, diabetic foot syndrome, and periodontitis [161].

T2D has its origins in nature, nurture, and complex interactions between the two. A combination of genetic ($h^2$~26-69% from twin studies [300, 7]) and environmental factors contribute to T2D risk. While it is clear that obesity, lack of physical exercise, a sedentary lifestyle, and increased consumption of energy-dense foods have contributed to the rapid increase

| Implied Mechanism | Exposure |
|---|---|
| Energy expenditure | Basal metabolism, exercise (or lack thereof), ambient temperature |
| Diet | High energy content foods, vitamins (e.g., vitamin D), macronutrients, micronutrients |
| Microbiome | Diet (processed foods), antibiotic usage, bariatric surgery, fecal transplant |
| Early life influences | Maternal disease, maternal nutrition, postnatal growth |
| Other | Sleep debt, environmental chemicals, chronic inflammation |

**Table 1.1** Examples of T2D exposures and implied mechanisms (these mechanisms have not necessarily been shown to be causal). Table is not comprehensive and is derived from Franks and McCarthy [108, Figure 1].

in T2D incidence [279], epidemiological studies have identified additional exposures that indicate a variety of mechanisms that could potentially contribute to T2D development. These exposures include the microbiome, early life influences (in utero and postnatal), environmental chemicals, sleep deprivation, and inflammation (reviewed in [108]; Table 1.1). Such observations are critically important in characterising T2D for the development of efficacious therapies and lifestyle interventions; however, their interpretation is extremely complex, especially since causality cannot be inferred due to the possibility of reverse causation or other sources of confounding [360].

The need to unravel causality highlights one of the most promising and perhaps unforeseen applications of human genetics—Mendelian randomisation (MR; summarised in [359]). In cases where genetic studies have identified a strong genetic proxy for an exposure, the genetic association can be used as an instrument variable to test for causality between the exposure and outcome, since variants are set by Mendelian inheritance at conception (i.e., not subject to reverse causality or confounding), and randomised at fertilisation (i.e., participants are randomly assigned to the "risk allele" group). Thus, an MR study is analogous to randomised controlled trials (RCTs) where the variant(s) substitutes for the drug perturbation and divides the population into those who receive the exposure (i.e., drug or treatment) and those who do not receive the exposure over a course of a lifetime.

Retrospective MR studies have mirrored the results of RCTs [148]. For instance, a recent MR analysis focused on genetic variants that affect vitamin D levels finds no evidence for a causal relationship between 25(OH)D and T2D [431], which is consistent with the reports of a RCT [195]. However, proper application and interpretation of MR requires deep, careful, and critical thought on the overall analysis, especially in regards to the validity of the instrument variable(s) [49, 50, 148]. Nonetheless, the rapidly growing MR field promises to tease apart many of the causal mechanisms behind a myriad of intermediate and disease outcomes including T2D [379], enabling the development of more efficacious therapies and lifestyle interventions.

In addition, the past decade of epidemiological and genetic T2D studies have led to refined understanding of T2D disease architecture [244]. Traditionally, T2D diagnosis has operated by categorising individuals into rigidly defined classifications with standardised treatments and protocols. But such categories can obscure important differences between individuals. With the rise of precision medicine, the plethora of omics and imaging technologies are thought to enable more precise categorisation. Indeed, for disorders like monogenic diseases or cancer, where a small number of possible pathways can have a large phenotypic impact, these technologies have enabled clinicians to more accurately assign patients to their appropriate classification (reviewed in [13]). However, hopes that a similar approach could classify T2D patients into distinct subsets (T2D-A, T2D-B, etc.) with differing natural histories and response to therapy have not yet been realised. Due to the multifaceted nature of the disease, many patients occupy a grey space where assignment to a specific diagnostic bin simply does not fit, and to insist on doing so would result in an overfocus of classifying rather than treating.

Synthesising insights from the past decade of research, a new approach to understanding T2D has been proposed that implicitly recognises the heterogeneity of the disease [244]. We know islet function, islet regeneration (beta cell number), islet autoimmunity, obesity, fat distribution, and insulin resistance are all critical components of T2D pathophysiology (list drawn directly from McCarthy [244]). Instead of insisting on rigid cutoffs to define precise disease boundaries across what arguably constitutes a continuous, multidimensional space, the “palette model” inherently recognises the multifaceted disease space [244]. The analogy is drawn from painting, where key disease pathways (i.e., the multidimensional vectors) constitute base colours. The saturation of the colour for an individual, or their placement along the continuous dimension, is determined by a combination of genetic and exposure influences. As an individual moves through life, their placement in the multidimensional space (i.e., colour) changes. Gradually, one may drift away from areas of homeostasis towards diabetes across a combination of dimensions. Of course, some individuals will occupy extremes of one or two vectors (for instance monogenic diabetes), and could be "classified" to some extent or alternatively, thought of as archetypes. However, the vast majority of patients will fall somewhere within a continuous, multi-faceted palette spectrum, where many individually modest risk factors collectively contribute to disease, such that by themselves they are relatively unimportant.[6]

Recognising the inherent complexity of diabetes within the clinical model transforms the mindset in which diabetes is considered and handled by patients and practitioners, bringing the concept of diabetes closer to the aetiology. For instance, thinking of the disease as a spectrum lends itself to a mindset that promotes preventative measures through healthy lifestyle decisions, rather than deeming oneself as "healthy" until long term poor lifestyle decisions suddenly manifest themselves as T2D. In addition, this refined understanding suggests research on identifying and characterising the archetypes—the extremes of the multidimensional space—should be particularly fruitful, as the pathophysiological mechanisms underlying these rare cases are limited compared to general disease cases.

## 1.3.2 History of T2D genetics and FUSION

### 1.3.2.1 T2D linkage studies

This refined understanding of T2D disease architecture has been the result of decades of T2D genetics research. Like many early studies of disease genetics, early T2D studies were linkage studies. Using a collection of genetic markers, these studies analysed a marker's segregation, or transmission patterns according to individuals with and without the disease in family pedigrees.

One of the first T2D linkage studies was the Finland-United States Investigation of NIDDM (FUSION) study (https://fusion.sph.umich.edu). This international collaboration between Finland and the United States began under the leadership of Michael Boehnke (USA), Francis Collins (USA), and Jaakko Tuomilehto (Finland). Subsequently over the years, the collaboration has grown to include over 8 laboratories located in the United States, Finland, and Europe. The study is exclusively focused on diabetes in Finland, for three reasons (1) the excellent health records system; (2) the strong tradition of participation in medical research; and (3) the relatively homogeneous Finnish population resulting from centuries of isolation,

[6] In some sense, our biological understanding of T2D has pushed us back towards thinking along the lines of a liability distribution, proposed by Pearson and Lee in 1900 [291].

amplifying the potential of unraveling hereditary factors compared to more outbred human populations.

As one would expect, given the now known polygenicity of T2D consisting of many common variants with small effect sizes, these early linkage studies, based on affected sibling pairs, were underpowered and had limited success [391, 125, 126, 351]. However, in the case of FUSION, these studies forged international relationships, led to a remarkably large and well characterised clinical data set, and worked out the logistical framework for high throughput genotyping and sophisticated data analysis. Thus, as technology matured and GWA studies became feasible, FUSION was at the forefront of the field.

#### 1.3.2.2 T2D GWASs

The stage was set, and by the late 2000s decades of genetics research enabled the first wave of T2D GWA studies, which included FUSION [340], WTCCC [419, 438], and the Broad Institute Diabetes Genetics Initiative (DGI) [82]. While these initial studies were modest in success, bringing the total number of T2D loci to 10, it became apparent that, due to the large penalty for multiple hypothesis testing, much larger sample sizes would be needed to unlock the genetics of T2D. The most straightforward and low cost way to expand sample sizes was through collaboration, and therefore, large consortia, consisting of many studies, were formed—including DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) and the Meta-Analysis of Glucose and Insulin-related traits Consortium (MAGIC).

These studies rapidly expanded the list of genetic risk loci. The first DIAGRAM study identified 6 novel T2D loci [439]. Shortly thereafter, MAGIC, which had the goal of investigating the effects putative T2D loci on glycemic and other T2D-relevant traits, identified 16 loci associated with fasting glucose related traits and 2 loci with fasting insulin related traits [88]. In addition to greatly expanding the number of loci associated with T2D and T2D-related traits, these studies stood out as examples to the broader human genetics community of the importance of data sharing and the tremendous insights to be gleaned through collaboration.

Since these seminal studies, the T2D genetics community has continued to conduct even larger GWA studies, collectively identifying > 100 T2D loci [340, 82, 419, 439, 88, 402, 264, 83, 111, 342]. However, despite these successes, these loci only explain a small portion (~10%) of the overall heritability of T2D estimated from family and twin studies [264]. One hypothesis is that this “missing heritability” resides in rare variants (MAF < 1%) of large effect sizes, primarily isolated to a family [237]. According to this hypothesis, while each individual case is rare, their effects collectively add up to a significant heritability component. Capturing these variants would require large sample sizes combined with whole genome sequencing (WGS) or exome sequencing, since by definition these polymorphisms are not common and therefore excluded from most genotyping arrays (minimum MAF ~1%). To test the rare variant hypothesis, two studies were launched: T2D Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) and Genetics of T2D (GoT2D). With the publication of the first results from these studies [111], the emerging picture suggests that it is unlikely that the population-level missing heritability resides in rare variants.

#### 1.3.2.3 A decade of T2D GWASs: insights and future prospects

With few exceptions [258], the overwhelming message of the past decade of GWA studies points to a model where T2D is simply a very polygenic disease and consists of many variants of small effect sizes that will require an ever growing number of samples to detect.

Indeed, the genetic architecture of T2D and other complex traits appears so polygenic that Boyle et al. [43] recently proposed an omnigenic model, whereby genetic effects on all genes expressed in disease-relevant cell types contribute to disease risk. First, this model posits that for any complex disease or trait, there are a handful of “core” genes with a direct and biologically interpretable role on disease aetiology (likely to operate in a tissue or cell type specific manner). Second, given a highly interconnected gene regulatory network where genes are separated from each other by only a few degrees, the model suggests that perturbations to almost any gene—even a peripheral, housekeeping gene expressed in all tissues—are liable to affect the regulation or function of a core gene. Thus, through signal propagation in a highly interconnected cellular regulatory network (e.g., transcriptional regulation, protein-protein inactions), genetic effects on any gene in a disease-relevant cell type are likely to contribute to disease risk, mediated by core genes. If such a model is true, one would expect that the ever increasing number of loci associated with T2D and other complex traits will spread uniformly throughout the genome (implying more and more genes across diverse molecular pathways), rather than clumping around loci that imply a few, common disease-relevant pathways. As described by the authors, initial evidence suggests an omnigenic model may be accurate; however, a more robust evaluation of an omnigenic model of complex trait genetics will be possible as more complete maps of the genetic risk factors for T2D and other complex traits are generated. Regardless of the outcome, the omnigenic model is an important synthesis of existing information and will help shape further research into the genetics of complex traits.

In addition, focusing on T2D, this high degree of polygenicity suggests, evolutionarily speaking, little selection for or against the risk alleles that compose the population burden of T2D, as there would be very little pressure on variants of small effect manifesting themselves later in adult life, past the prime years of reproduction. The neutrality of T2D risk alleles in human prehistory is further supported by formal adaptation studies that find little evidence for selection [14, 98] and the fact that most T2D associations replicate across human populations [411, 83]. Collectively, these results imply that the default model when considering T2D risk allele is not a “thrifty model” where a T2D risk allele rose to frequency because it was advantageous in the distant hunter-gatherer past; instead, most T2D risk alleles appear to be common simply due to genetic drift [14, 111]. As with most models, however, there may be exceptions and some cases likely caused by local adaptations to unique environments, for instance the *CREBRF* mutation found for BMI in Samoans [257], but such cases are outliers, not the majority of the population burden.

The future of T2D genetics will include GWA studies of ever larger sample sizes, gradually transitioning from genotyping arrays to sequencing approaches as whole genome sequencing prices continue to plummet [131]. Ongoing efforts to study T2D genetics in diverse human populations will also be important, as such studies have the potential to identify variants with particularly strong effects due to varying allele frequencies in isolated populations, as well as potentially unique selective pressures of the local environment [258, 102, 350, 111]. In addition, the value of studying the genetics of rare, monogenic, T2D-related diseases, like Maturity Onset Diabetes of the Young (MODY), congenital hyperinsulinemia, and neonatal diabetes, should not be overlooked. Indeed, the past decade of genetics research suggests the biological divide between monogenic disorders and common disease is not binary, but rather exists on a spectrum [245]. In many instances, genes in which certain mutations are causative for monogenic conditions also harbor other alleles that represent risk factors for T2D. And variable penetrance is regularly observed too: many of the specific alleles central to MODY have also been observed in individuals clinically presenting as T2D, or even as completely normal [101], suggesting penetrance estimates may have been overestimated perhaps in part due to genetic background and environment. Finally, as mentioned to earlier, more complete genetic maps for T2D will help inform whether an omnigenic model accurately represents the genetic architecture of T2D.

Filling in the “missing heritability” and generating a comprehensive catalogue of the genetics of T2D is only part of the puzzle. Like most loci identified through GWA studies, the vast majority of T2D loci are found in non-coding portions of the genome [341], obscuring the underlying genes and molecular pathways. The T2D research community thus faces a significant challenge of linking non-coding variation to specific genes and other molecular traits, in order to understand the molecular mechanisms of disease. Since this is a problem not only for T2D but for many complex diseases, widespread efforts like GTEx (see Section 1.2.7) have been launched to map expression QTLs in a variety of tissues [137]. The T2D genetics community is further supplementing these efforts through additional QTL studies in key T2D tissues such as pancreatic islets, which are difficult to obtain and not well assayed by GTEx [95, 393, 395]. Already these efforts are beginning to bear fruit as nearly a third of T2D GWAS loci now have a plausible mechanism, many of which are linked to pancreatic islet biology [401].

However, the majority of T2D GWAS signals still lack a clear mechanism, highlighting the need for ongoing research. FUSION is actively contributing to these efforts through the FUSION tissue biopsy study, focused on the genetic regulators of gene expression and DNAme in skeletal muscle and adipose tissue. Before describing the study, I provide a brief overview of molecular traits and molecular QTLs.

## 1.4 Molecular traits

In this section, I describe and provide a brief overview of the two molecular traits most relevant to this thesis—gene expression and DNA methylation. Subsequently, I provide an overview of molecular quantitative trait loci—i.e., genetic associations with molecular traits. This overview is particularly focused on expression as a quantitative trait, since the field developed around expression before expanding into other traits.

## 1.4.1 Gene expression

### 1.4.1.1 Gene expression biology

DNA is the stable, long term information storage molecule of the cell. By contrast, RNA is more dynamic and serves both as information transfer molecule for protein synthesis, as well as a functional molecule occupying a variety of regulatory roles (reviewed in [265]). There are many types of RNA including transfer RNA (tRNA), ribosomal RNA (rRNA), messenger RNA (mRNA), microRNA (miRNA), and a plethora of poorly characterised, non-coding RNA (ncRNA; reviewed in [265]). Constituting a small fraction (1-2%) of the total RNA in a cell [68], mRNA is one of the most studied forms of RNA, as it encodes the information for the proteins of a cell.

RNA polymerase II (Pol II) transcribes mRNA from DNA sequence by binding at the promoter of a gene, *initiating* transcription at the transcription start site (TSS), *elongating* the growing mRNA molecule by moving along the DNA from the 5' end towards the 3' end, and finally *terminating* transcription, which in humans generally means simply falling off the transcribed DNA at an ill defined region beyond the final exon (though occasionally human genes do have a structured termination site). As Pol II transcribes the growing mRNA molecule, a variety of mRNA processing steps occur including the addition of a 5' cap, splicing events, and the addition of ~200 adenosine residues, termed a poly(A) tail, to the 3' end. The 5' cap involves adding a modified guanosine triphosphate to the 5' end and signals to the cell the mRNA identity of the molecule—preventing its degradation by exonucleases and facilitating both nuclear export and translation. Splicing occurs co-transcriptionally [251] and describes a step where introns are excised from the mRNA molecule while exons are maintained, forming an elegant system where the same gene can code for different proteins across cell types through alternative splicing of gene isoforms. Other modifications such as the addition of the 3' poly(A) tail are thought to increase the stability of the molecule and regulate degradation (reviewed in [304, 260]). Finally, as a quality control mechanism and to clean up extra transcriptional byproducts, aberrant mRNA molecules are targeted and degraded by various mechanisms including miRNAs (reviewed in [286, 166]), so that only high quality mRNA is exported to the cytoplasm (reviewed in [375]), destined to be translated to proteins.

#### 1.4.1.2 Gene expression quantification: from microarrays to sequencing

Clearly, gene expression is critical for the function of a cell and therefore expression quantification has been the topic of extensive research (reviewed in [230]). Early quantification techniques, such as Northern blots [8] or reverse transcriptase quantitative PCR (RT-qPCR; [30]; reviewed in [109]), were low throughput, laborious, and consequently expensive. A fundamental breakthrough came with the development of microarrays [336].

Microarrays consist of oligonucleotide probes that are arrayed onto glass, silicon, or plastic substrates. During measurement, fluorescently labeled DNA is washed over the array and the probes hybridise to their complementary sequence. By knowing the sequence behind each probe and imaging the fluorescence intensity at each probe, a quantitative readout is obtained [230]. Various forms of microarray technologies have been used for a wide variety of applications including the measurement of genotypes, gene expression, and DNAme (reviewed in [150]). In the context of gene expression, protocols generally involve isolating total RNA, optionally selecting for specific types of RNA (e.g., mRNA), reverse-transcribing RNA into complementary DNA (cDNA), labelling the cDNA with fluorescent dyes or biotin, and hybridising the labeled cDNA products to the array [339, 230].

Microarrays enabled the first wave of expression studies, but they also had several limitations including (1) only known sequences could be assayed (preventing novel transcript detection) and (2) the dynamic range of detectable signal was limited due to high background levels of noise [408]. These limitations were overcome with the advent of low-cost next generation sequencing (NGS) technologies (reviewed in [131]).

These NGS technologies revolutionised molecular biology. With sequencing, nearly any molecular phenomena that involves DNA can be scaled to a genome-wide level, as long as the phenomena is reducible to the level of DNA sequence. For instance, one of the most recent “seq” applications is “assay for transposase-accessible chromatin with high throughput sequencing” (ATAC-seq), which uses Tn5 transposase to integrate sequencing adaptors into regions of open chromatin. Thus, by enabling sequencing, the reads obtained from these adaptors profile chromatin accessibility [46]. An overview of popular “seq” technologies is provided by Wold and Myers [423] and Soon et al. [362]; however, it should be noted that creative, new applications like ATAC-seq are always being devised and therefore a review is quickly outdated.

In the context of gene expression, NGS enabled the direct sequencing of cDNA libraries, termed RNA-seq (reviewed in [408, 230]). RNA-seq overcame many of the microarray limitations as RNA-seq (1) requires no prior knowledge of specific gene sequences to design probes and (2) generates an extremely dynamic and digital signal (mapped reads) with minimal background noise. Given these benefits, combined with the ever plummeting costs of sequencing [131], RNA-seq has emerged as the standard assay for transcriptomics.

In a basic RNA-seq experiment, RNA is isolated from the cell, optionally filtered for specific types of RNA, optionally fragmented to ~200-300 bp (required for most NGS technologies), reverse-transcribed to cDNA, ligated to sequencing adaptors, optionally PCR amplified, and sequenced (reviewed in [240, 230]). Following sequencing, digital reads that pass quality control filters are assembled, or stitched back together. A variety of methods exist for transcriptome assembly, including mapping back to a reference genome (or transcriptome) and de novo assembly (reviewed in [240, 68]). Generally following mapping, gene level expression is quantified by counting the number of uniquely mapped reads (as opposed to read that map to multiple locations) overlapping the exons of a gene. Read counts are subsequently normalised to account for various biases, such as gene-length and sequencing depth, to enable within sample and between sample comparison of genes [68] (see Section 2.6.2). In addition, RNA-seq enables quantification of other molecular features such as gene isoforms [386, 132], exon fragments [9], splicing events [218], and (if variants are present in the transcribed region) allele specific expression (ASE) [56].

The methods and applications for RNA-seq are rapidly evolving as new technologies continue to emerge. One particularly exciting expansion is the development of single cell transcriptomics, where the transcriptome of a single cell is assayed. The development and refinement of single cell methods (reviewed in [329]) promises to revolutionise our understanding of human biology, for instance by cataloguing all cell types in the human body [308].

In addition, new long-read sequencing technologies, such as nanopore sequencing, promise to overcome limitations in accurately quantifying the abundance of specific transcripts by sequencing entire molecules of RNA that have been reverse transcribed to cDNA [36, 52]. This technological development is particularly important, because many disease-associated variants are thought to exert their effects through alternative splicing [219]. Current NGS technologies are based on short reads (~200-300 bp). Therefore, transcripts are sampled in fragmented form and must be computationally estimated through isoform deconvolution [386, 132] or alternative measures like quantification of exons [9] or splicing events [218]. While such methods are useful, they are far from perfect, especially for highly spliced genes

[36]. By essentially eliminating a fragmentation step, long accurate sequencing technology would constitute a real advancement for transcriptome sequencing by directly measuring the splice isoforms that all of these techniques attempt to reconstruct from short read data.

### 1.4.2 DNA methylation

#### 1.4.2.1 Epigenetics

The word *epigenetics* was coined by Conrad Waddington in 1942 [403, 282]. The term vaguely described the causal mechanisms by which genetic variation affects phenotype, starting with the fertilised zygote and moving through developmental stages, ultimately resulting in a complex organism with varied phenotypes (reviewed in [97]). Waddington later clarified this concept of what would now be called developmental genetics, in his famous diagram of the "epigenetic landscape" where the cell is envisioned as a ball, rolling down a hill through valleys. As the ball encounters various ridges, it decides a path which places it on a trajectory towards a specific valley or cell type [404]. Waddington's term harkens back to a late 19th century embryologist debate between those who envisaged development as the enlargement of preformed elements (preformationism) and those who viewed development as a series of chemical reactions, executing a complex developmental plan (epigenesis) [97]. Thus, in its roots, one can see anticipations of the modern use of epigenetics to mean "mitotically and/or meiotically heritable changes in gene function that cannot be explained in DNA sequence" [326]. Based on this definition, epigenetics describes a variety of cellular phenomena including DNA methylation (DNAme), histone modifications, and to some extent ncRNAs (reviewed in [128, 97]).

#### 1.4.2.2 DNA methylation mechanisms

Perhaps the most studied and characterised epigenetic modification is DNAme, proposed as a mechanism of cell memory in 1975 [312, 155]. DNAme describes an additional chemical modification to DNA where DNA methyltransferases (DNMTs) attach a methyl group typically to the 5th carbon atom of a cytosine base, chemically changing it to 5-methylcytosine (5mC).

Generally in mammals, methylation occurs at cytosine residues adjacent to guanine residues connected by the DNA phosphate backbone (CpG site), producing diagonally symmetric methylation patterns across the DNA strands. Though uncommon, non-CpG methylation such as CHG or CHH, where H is a non-G nucleotide, have been observed across a variety of species and cell types, including human (reviewed in [289, 445, 168]). Depending on the surrounding bases, these methylation patterns may not be symmetrical, suggesting that the mechanisms which preserve CpG methylation through cell division (described below) do not apply to non-CpG methylation [289]. Non-CpG methylation is generally rare and is more rare in differentiated tissues compared to pluripotent cells or neuronal tissue [289, 168].

Early studies suggest non-CpG methylation sites may have functional importance (reviewed in [289, 445]). For instance, compared to somatic tissues, induced Pluripotent Stem (iPS) cells acquire non-CpG methylation [446, 224]. In neurons, CpH methylation is enriched around methyl-CpG-binding protein 2 (MeCP2) binding [140], and is positionally conserved between independent samples (i.e., not stochastic) [225]. Non-CpG methylation may also be linked to response to environmental factors. For example, skeletal muscle non-CpG hypermethylation patterns in the promoter of *PPARGC1A* are associated with decreased *PPARGC1A* expression, T2D (T2D participants show hypermethylation), and are inducible by tumor necrosis factor alpha as well as free fatty acid exposure [22]. These hypermethylation signals are also associated with obesity and return to non-obese levels after gastric bypass surgery and weight loss [24]. Finally, in healthy individuals, *PPARGC1A* promoter methylation decreases shortly after exercise and is subsequently followed by increased *PPARGC1A* expression hours later. Notably, by the time *PPARGC1A* expression changes are observed, methylation changes were no longer apparent [23]. Across these several *PPARGC1A* studies, the key changes were observed in non-CpG methylation, not CpG methylation. Collectively these findings suggest an important role of non-CpG methylation; however, at this time it is difficult to draw general conclusions as non-CpG methylation is still poorly characterised.

When a mammalian zygote is formed, DNAme is rapidly erased, or demethylated, throughout the genome (reviewed in [212]). As the embryo develops, a series of de-novo methylation events take place. These orchestrated DNAme events are performed by DNMTs, which operate in either a de-novo or maintenance fashion. Currently, there are three DNMTs of significant importance in mammals: DNMT1, DNMT3A, and DNMT3B (reviewed in [174, 212]). DNMT3A and DNMT3B are essential to laying down de-novo DNAme patterns during development. Though DNMT3A and DNMT3B may be involved in specific contexts [174], DNMT1 primarily functions in maintaining DNAme patterns, by scanning the genome and methylating cases of hemimethylated DNA, where a cytosine residue is methylated on only one strand. Such scenarios often occur during DNA replication where the newly synthesized strand of DNA lacks chemical modifications. Thus, via this mechanism DNAme can propagate through cell division, fulfilling the criteria of an epigenetic modification.

In contrast to the establishment of DNAme patterns, DNAme can be erased through passive or active mechanisms during development or in response to environmental cues (reviewed in [212]). Passive methylation involves the inhibition of DNMT1 such that methylated sites are not maintained over DNA replication events. Active demethylation involves the direct removal of the methyl group from 5mC and conversion back to an unmethylated cytosine residue.

Recently, it was discovered that additional, oxidised 5mC derivatives exist in the mammalian genome [382, 191], catalysed by the Ten-Eleven Translocation (TET) family of proteins (reviewed in [212, 428]). These enzymes can convert 5mC to 5-hydroxymethylcytosine (5hmC), 5hmC to 5-formylcytosine (5fC), and 5fC to 5-carboxylcytosine (5caC). 5fC and 5caC can in turn be actively converted to an unmethylated cytosine residue by thymine DNA glycosylase and the base excision repair pathway. Alternatively, DNMT1 prefers hemimethylated 5mC as a substrate, so 5mC derivatives may also be passively demethylated [428].

To date, whether or not these oxidised 5mC derivatives have a functional role or are simply intermediate products in a demethylation process is poorly understood (reviewed in [348, 428]). Across all CpGs, early experiments show oxidised 5mC derivatives are a minority with 5hmC constituting ~1-30% of CpGs and 5fC/5caC constituting ~8-10% of CpGs, depending on cell type [445]. Such quantities are quite small compared to 5mC which can be found at about ~70-90% of CpGs [445, 212]. Although scant, initial evidence suggests a functional role worth further exploration. For instance, 5hmC is enriched around enhancers and DNase1 hypersensitivity sites (DHS), 5fC around poised enhancers and exons, and 5caC around satellite repeats [445].

Unless special protocols are used [434, 38], 5hmC is indistinguishable from 5mC in commonly used DNAme readouts based on bisulfite conversion [156, 37] (see Section 1.4.2.4). Therefore, due to the popularity of bisulfite treatment protocols, unless otherwise stated, many studies, including this study, will describe DNAme patterns that do not truly represent 5mC, but rather the sum of 5mC and 5hmC. In this thesis, when I refer to DNAme, I am not distinguishing between 5mC or 5hmC.

#### 1.4.2.3 DNA methylation function

As proposed in 1975 [312, 155], DNAme has been generally thought of as linked to repression, silencing, and general inactivation of the genome. Indeed, many lines of evidence support this notion. DNAme is essential to genome stability and is found around centromeres, microsatellites, transposable elements, and other repetitive elements in mammalian genomes (reviewed in [305]). DNAme is necessary for X-chromosome inactivation in females, where one of the two female copies of the X-chromosome is silenced so that transcription occurs on only one copy for the majority of X chromosome genes [212]. DNAme also serves in imprinting where one parental copy of a gene is transcriptionally silenced (reviewed in [20]). Finally, there has been incontrovertible evidence for some time that DNAme in promoters silences gene expression (reviewed in [212]).

However, despite all of these observations, there is no established, comprehensive model for how DNAme mechanistically functions in mammalian genomes. One model envisages DNAme as a “locking” mechanism, where it aids in maintaining chromatin states rather than initiating chromatin remodelling [173]. For instance in mouse, methylation of the *Hprt* gene occurs after X-chromosome inactivation and *Hprt* silencing, suggesting DNAme is not the primary mechanism in X-chromosome transcriptional silencing [227]. However, DNAme can also attract TFs (reviewed in [445]), including methyl-CpG binding domain proteins, like MeCP2, which associate with repressor complexes that alter the surrounding chromatin structure, pointing to a role as a chromatin remodelling pioneer [212]. Still other data suggest that nucleosome histone modifications may make DNA differentially susceptible to methylation [174]. Through all of these lines of evidence, a picture emerges of a very complex and intertwined relationship between DNAme, histone modifications, and other regulatory mechanisms that is highly context dependent.

Enabled by high throughput technologies to assay methylomes genome wide, the past decade of epigenetics research has helped establish and contextualize the diverse roles of DNAme. In somatic mammalian tissues, the majority (~70-90%) of CpG sites are methylated [212, 445]. CpG dinucleotides are globally depleted (~5x) from the human genome [34], likely stemming from the fact that 5mC is prone to spontaneously deaminate from a cytosine to a thymine residue [212].

An important exception to this global CpG methylation pattern is the presence of specific ~1 kb stretches of mostly unmethylated CG-dense regions called CpG islands (CGIs; reviewed in [77]). CGIs generally mark TSSs. In humans and in mice, ~50% occur in canonical promoters. The remaining "orphan CGIs" are split with ~25% occurring in intragenic regions and ~25% occurring in intergenic regions. Despite the various genomic contexts of CGIs, nearly all show evidence of transcriptional initiation, perhaps because CGIs mark open, nucleosome deficient chromatin and therefore do not require additional ATP-dependent chromatin remodelling complexes for nucleosome displacement [77]. Many of the orphan CGIs show transcription of ncRNA, and exhibit tissue specific activity. Of all CGIs, the intragenic CGIs show the greatest number of differences in DNAme across somatic tissues, which may be linked to alternative splicing [77].

In addition, the high CpG content of CGIs can recruit proteins that promote H3K4me3, an activating chromatin mark [385]. Conversely, the G+C richness of CpGs also attracts proteins associated with H3K27me3, a repressive mark [250]. Therefore, in specific cellular contexts like embryonic stem (ES) cells, many CGIs lie in a bivalent chromatin states. As differentiation occurs, these states flip, like a switch, into active or repressed [77, 205]. Because of the peculiar fact that ~70% of all promoters have a CGI [77], CGI promoters have been intensely studied. In this specific context, it is clear that DNAme of CGI promoters blocks TF binding and gene transcription is inhibited [173, 77]. Because many early studies focused on this specific context (promoter CGIs), this observation has shaped the general perception that DNAme decreases gene expression [173].

However, as technologies have enabled the study of DNAme in other contexts, this intuition that DNAme always decreases gene expression has been clearly refuted [173]. For instance, DNAme in the context of gene bodies can be associated with increased levels of transcription, possibly even stimulating the elongation phase of transcription [173]. Building on this observation, there is a growing body of evidence that supports a regulatory role of gene body DNAme in transcript splicing (reviewed in [210]). Compared to introns, DNAme is more abundant in exons [210], and has been shown to be capable of directly causing alternative splicing [432]. However, the effects of DNAme appear to be context specific—sometimes promoting exon inclusion and other times promoting exclusion. In cases of strong splicing programs for constitutive exons, perhaps due to a strong splice motif, weaker DNAme effects on splicing are often suppressed. These observations have led to a model where DNAme functions as a "fine-tuning" mechanism for alternative splicing. Such a model is consistent with the fact that DNAme cannot be required for splicing since other organisms that lack DNAme, like *Drosophila melanogaster* and *Saccharomyces cerevisiae*, have spliced genes [210]. Mechanistically, DNAme has been shown to affect splicing by altering the kinetics of Pol II elongation, for instance by creating "roadblocks" via CTCF or MeCP2 recruitment, as well as by attracting proteins which associate with splicing factors, such as HP1 [210]. Despite these clear cases, given the growing catalogue of TFs that read and write DNAme (reviewed in [445]), there are likely many more key TFs involved in DNAme-linked splicing that have yet to be discovered.

Even less characterised are intergenic regions, although initial evidence suggests that the methylation patterns in these regions is very important. In mouse, Stadler et al. [364] describe intergenic lowly methylated regions (LMRs) that are cell type specific and overlap DHSs and enhancers. These regions were also strongly correlated with increased expression of nearby genes (i.e., methylation of LMR reduced expression). In human cell lines, Charlet et al. [58] found some H3K27ac peaks, a hallmark of active enhancers, coexist with DNAme in enhancers, but not promoters. In cases where TCF4, a TF associated with enhancers and H3K27ac peaks, was bound within a H3K27ac peak, an abrupt decrease DNAme was observed. Furthermore, genetic or pharmacological reduction of DNAme decreased H3K27ac, suggesting that DNAme is important to broader enhancer integrity, but lack of DNAme is linked to TF binding within enhancers. It should be noted, however, that TF binding does not always coincide with decreased DNAme, and indeed some TFs may prefer DNAme when binding (reviewed in [445]). The important message from these studies is DNAme shows complex patterns of functional importance in intergenic regions such as enhancers. Given the importance of enhancers in regulating tissue specific gene expression and the general enrichment of GWAS loci in disease relevant, tissue specific enhancer states [287], such results may have important implications for disease and motivate the further study of DNAme in this context.

Collectively, these observations demonstrate the role of DNAme is far more complex and nuanced than previously appreciated. Understanding the functional importance of DNAme will require the integration of multiple molecular traits, including gene expression, DNAme, and histone marks, across multiple tissues.

#### 1.4.2.4 DNA methylation quantification

Like gene expression, both array and sequencing-based approaches have been developed to measure DNAme (reviewed in [433]). Most array based studies (including this study) use Illumina microarrays, which were recently upgraded to the MethylationEPIC array from the Methylation450 array. Both platforms use the same BeadArray technology (reviewed in [150]), except the EPIC array deploys more probes (850k vs 450k) that capture a greater portion of intergenic genomic regions, like enhancers and DNase hypersensitive sites, and also includes several non-CpG sites identified in human stem cells [261]. The gold standard, however, is whole genome bisulfite sequencing (WGBS), which assays genome wide methylation, unrestricted by probe sites. Both of these methods use a bisulfite conversion protocol.

Sodium bisulfite converts unmethylated cytosine to uracil, which after PCR, becomes thymine. Thus, with this reaction, methylation can be measured by quantifying C→T changes, which essentially enables the use of genotyping technologies to measure DNAme [35]. As mentioned earlier, both 5mC and 5hmC protect cytosine during a standard sodium bisulfite protocol [156, 37], while cytosine, 5fC, and 5caC are converted to uracil [272]. Therefore, unless otherwise stated, the DNAme readout from commonly used protocols will be both 5mC and 5hmC.

In contrast to the prolific use of RNA-seq to assay gene expression, WGBS is not yet the current standard for methylation studies. The main reason is the prohibitive cost, as most sequencing reads are wasted assaying non-methylated sites. For this reason, many studies use arrays which are cheaper, but are restricted to predefined regions, and thus may miss key methylation events (see Section 4.3.2). Alternative sequencing-based approaches have been developed that reduce costs by enriching for methylated sites (reviewed in [433]), but are still more expensive than array based technologies.

Looking to the future, new DNAme technologies stand on the horizon that promise to overcome many of the current limitations. To date most methylation studies have been conducted on cultured cell lines or bulk tissue samples, as existing protocols require relatively large amounts of input DNA. Thus, the final methylation signal, derived from the aggregate signal across all cells in the sample, ranges from 0 (unmethylated) to 1 (methylated). However, new single cell technologies [139, 65] are beginning to shed light on the precise methylome which, at the level of individual cells, is essentially binary apart from hemimethylated sites. As technologies mature, these techniques will provide critical insights into the methylome across a cell's lifespan, cell types, and in relationship to disease. In addition, nanopore sequencing has been shown to be able to directly detect 5mC [354]. New sequencing technologies, combined with ever decreasing sequencing costs [131], promise to further drive down the currently prohibitive price of WGBS, enabling widespread, unbiased methylome surveys.

### 1.4.3 Molecular quantitative trait loci (QTLs)

By 2001, it was clear that rapid technological advances could in theory enable genome wide mapping of QTLs for molecular traits, and since mRNA could be easily captured and quantified, it was the most obvious target [170]. One year after the initial proposal by Jansen and Nap [170], Brem et al. [44] generated the first genome wide map of expression quantitative trait loci (eQTLs) using yeast. Enabled by technology that allows for the transformation of B lymphocytes into lymphoblastoid cell lines (LCLs; reviewed in [349]), an immortal cell line essentially providing an unlimited resource of cellular material with many applications, this initial success in yeast was quickly followed by studies in human [61, 259, 263, 370, 371]. Since then, eQTL studies have been conducted across many organisms, developmental stages, tissues, and molecular traits (reviewed in [112, 4]).

Mechanistically, a variant that affects gene expression can operate through *cis* regulation, directly influencing expression from the same DNA molecule as the target gene in an allele specific manner, or through *trans* regulation, indirectly influencing expression perhaps from a different DNA molecule [317, 112, 4]. In general, regulatory variation that is proximal to the target feature (the TSS of gene in the case of eQTLs) have been shown to operate in *cis* [62, 294], and therefore proximal is often equated to *cis* regulation. However, that is not always the case. For instance in yeast, the AMN1 protein indirectly regulates expression of itself through interactions with other TFs. A variant in the *AMN1* gene causes an amino acid change that alters the regulatory interactions of the regulatory feedback loop, ultimately perturbing the *AMN1* expression. Notably, the ratio of expression of the amino acid changing alleles in the diploid hybrid are not significantly different. Therefore, in this case, the amino acid altering mutation is a proximal eQTL; however, the regulatory effect on expression is *trans* as it affects both alleles in the heterozygous diploid [321]. Using sequencing data, cases of true *cis* regulation can be detected by allele specific expression (ASE), where preferential binding to one parental regulatory allele results in increased expression of a transcribed allele in phase with the regulatory allele. In contrast, *trans* alleles may occur anywhere throughout the genome, proximal or distal. Due to the severe statistical penalty that must be paid when mapping distal QTLs, many studies only consider proximal regulation—often defined as +/- 1 Mb from the genomic feature (e.g., TSS in the case of genes) [275].

Since mRNA is critical to information transfer out of the nucleus and is easily isolated, the majority of QTL studies to date focus on gene expression (eQTLs). Studies show that the majority of eQTLs are shared across tissues [136, 104, 138], and that eQTLs are generally, but not always, located upstream of the genes they regulate [138]. In addition, Kilpinen et al. [185] recently demonstrated that iPS cells contain a host of eQTLs not found in somatic tissues. This supports a growing notion that many eQTL effects are specific to a cellular context, such as a developmental time point [185, 54] or a specific environmental exposure [356, 319, 135, 21, 238, 159, 235, 306, 430, 204, 96, 47, 267, 444, 188]. Identifying the correct cellular context to observe an effect may make ascertaining the causal effect for some GWAS loci difficult; however, a preliminary report suggests such response eQTLs may exhibit effects on other molecular traits, like chromatin accessibility (caQTLs), in a non-induced state, such that the cell is "primed" to respond to a specific context [3].

Finally, alongside eQTL studies, a growing number of studies have mapped genetic regulators of other molecular traits (reviewed in [112, 4]) including DNA methylation (mQTLs), histone marks (hQTLs), DNaseI sensitivity (dsQTLs), and more recently CTCF binding (CTCF-QTLs) [84] as well as ATAC-seq (caQTLs) [196, 3]. In aggregate, these studies suggest a substantial proportion of variation across many layers of molecular traits is driven by genetic variation. To date, the extent to which genetic variants directly affect molecular traits or are mediated through one trait or another is currently unknown—although initial studies suggest that genetic effects across layers of molecular traits are often independent [141, 273]. It should be stressed, however, that these studies span a limited number of molecular traits and are by no means comprehensive.

### 1.4.4 QTL mapping

In order to model genetic effects, trait association studies commonly make two simplifications [51]. First, variants that are not biallelic and contain more than two alleles are removed. Second, only additive effects are modelled, meaning each minor allele copy results in a similar proportional change in the phenotype. Dominant or recessive effects are not typically modelled explicitly; however, in addition to capturing additive effects, additive models have reasonable power to detect dominant effects [209].

Similar to the development of specialised tools like PLINK [57] during the maturing stages of GWA studies, a variety of tools have been developed for QTL mapping for molecular traits including matrix eQTL [345], LIMIX [222], RASQUAL [196], and QTLtools [81]. Each tool provides slightly different options. For instance, LIMIX developed by Oliver Stegle's group at the EMBL-EBI is an extremely efficient, one stop toolkit for a variety of linear mixed model applications ranging from simple linear models to multivariate models. By contrast, RASQUAL, which was released during my PhD studies, uses innovative models that capture molecular trait biology to map truly *cis* QTLs by jointly modelling feature level signal (e.g., gene expression) along with allele specific signal (e.g., allele specific expression) in sequencing data. By utilising two orthogonal sources of information, RASQUAL boosts power and identifies robust *cis* signals. Building on previous collaborations in the Birney laboratory [54], I developed a QTL pipeline around LIMIX at the start of my PhD research; however, for sequencing-based *cis* QTL mapping, RASQUAL is an excellent alternative approach.

## 1.5 FUSION tissue biopsy study

Having helped lead the development of complex trait genetics from early linkage studies to the now mature field of GWA studies, the collective FUSION laboratories have a record of looking forward, envisioning the hurdles that need to be surmounted, with the overarching goal of translating genetic findings into clinical applications. Recognising the need to link non-coding GWAS loci to molecular traits, FUSION began designing a molecular trait biopsy study in the 2000s with the goal of mapping molecular QTLs. Skeletal muscle and adipose were selected as the primary target tissues since they are the only tissues related to insulin response and glucose homeostasis that could be sampled from living participants who could undergo extensive phenotyping (see Section 1.3.1 for a description of tissues critical for glucose regulation). In its entirety, this study encompasses tissue samples of *vastus lateralis* skeletal muscle, abdominal subcutaneous adipose, and skin from ~318 Finnish individuals with Normal Glucose Tolerance (NGT), Impaired Glucose Tolerance (IGT), Impaired Fasting Glucose (IFG), or recently diagnosed T2D before the onset of treatment (Table 2.5). In most cases, all three tissues were collected from each participant, who had previously undergone extensive phenotyping and genotyping. Focusing on T2D-related tissues (muscle and adipose), each tissue sample has been subjected to deep RNA-seq and DNAme analysis (Illumina Infinium MethylationEPIC BeadChip).

The primary goal of the FUSION biopsy study is to identify genes and DNAme sites linked to T2D and T2D-related traits through direct associations and by integrating GWAS loci with molecular QTLs. Molecular traits identified from this study, as well as others, can then be followed up with functional studies to further clarify and establish the relationship with T2D and T2D-related traits. As noted, skin biopsies have also been acquired with the long term goal of using these cells and iPS technology to generate islet cells, essential to T2D. In fact, at the time of writing, 50 skin fibroblast cultures have already been transformed into iPS cells.

## 1.6 The scope of this thesis

In this thesis, I present an analysis of genetic effects on expression and DNAme in skeletal muscle, which was prioritised over adipose tissue in data generation.

> In Chapter 2, I describe the measures taken to ensure good quality of the genotype, expression, and DNAme data. For expression, this section focuses on skeletal muscle RNA-seq data, as muscle RNA-seq was prioritised over adipose RNA-seq. For DNAme, I describe quality control steps across all FUSION biopsy samples including the skeletal muscle and adipose, as well as several additional pancreatic islet samples derived from cadaveric donors. Finally, I also describe a measure of skeletal muscle specificity developed for both gene expression and DNAme.
>
> In Chapter 3, I characterise the relationship between expression and DNAme. In this analysis, I find evidence of latent and potentially confounding sources of correlation between expression and DNAme. I show that some of this correlation may be due to differences in tissue heterogeneity across samples, which has been previously underappreciated in similar studies. Accounting for this correlation, I chart associations between expression and DNAme and characterise the genomic context of these associations.
>
> In Chapter 4, I generate maps of genetic regulators of gene expression and DNAme. I analyse these maps in the context of a panel of chromatin states across several cell/tissue types, identifying patterns of enrichment. In addition, I integrate QTL maps with TF binding predictions in order to identify skeletal muscle activators and repressors.
>
> In Chapter 5, I integrate QTL maps with variation linked to T2D and T2D-related traits. I find QTLs for many GWAS loci and summarise the top results, showing how DNAme can potentially inform the location of key regulatory events (TF

binding at the canonical promoter, alternative splicing, distal regulation, etc.). I also describe QTLs for molecular traits with signal patterns highly specific to skeletal muscle, as these loci have the greatest potential to inform skeletal muscle T2D pathophysiology. I replicate an association with a T2D GWAS locus and a highly muscle specific *ANK1* isoform, previously identified using an earlier version of this dataset [341]. I also analyse mQTLs at the same T2D GWAS locus for highly muscle specific DNAme sites around *ANK1*, showing that the strong genetic effects on *ANK1* expression and DNAme appear to be statistically independent. Inspired by this observation, I characterise the relationship between expression and DNAme genome wide, through a mediation analysis using QTLs. I show that in the majority of cases, associations between gene expression and DNAme appear to be independent.

In Chapter 6, I map genetic effects on gene expression that are specific to an environmental context, treating the rich phenotyping data on FUSION participants as potential environments. Though this study was underpowered, I highlight a candidate context specific effect between *FHOD3* and low-density lipoprotein cholesterol as well as systolic blood pressure.

Finally, in Chapter 7, I summarise the key findings of each chapter and outline steps for further research motivated by the results presented in this thesis.

Altogether, these analyses add new insight into the genetic regulators of skeletal muscle expression and DNA methylation, and contextualise how these genetic regulators affect fundamental muscle biology in normal individuals and those with T2D.

# Chapter 2

# Data generation and quality control

## 2.1 Introduction

In this chapter, I describe how my colleagues and I generated the data I analysed in this thesis and the quality control procedures I employed for the DNAme data. To provide the reader with the context of these data, I will briefly outline the key characteristics of the cohort, sample collection, and the molecular traits generated on these samples. My collaborators, primarily in the Collins laboratory, generated all of the data analysed (see Acknowledgments and Scott et al. [341]). Data analysis was conducted by the FUSION tissue analysis team, mainly composed of University of Michigan and NIH analysts, including myself. To accomplish the many tasks involved in data analysis and quality control, we subdivided the tissue analysis team into project oriented groups—phenotype, genotype, RNA-seq, and DNAme. Within each task group, consisting of 3-4 analysts, decisions were made and subsequently presented to the full analysis team for collective approval or modification.

My role in this process was as follows: in collaboration with a clinician, I processed the medication phenotype information. I also led the genotype and RNA-seq data generation and quality control—overseeing analysis steps, decisions about sample exclusions, and performing specific analyses. Finally, I directed and executed the bulk of the DNAme analysis and quality control.

## 2.2 Participant recruitment

The FUSION tissue biopsy study is part of a long term epidemiological cohort, published extensively on over the last 10 years (https://fusion.sph.umich.edu). The full details of participant recruitment for the tissue biopsy study are described in detail in Scott et al. [341]. For completeness sake, I have reproduced part of that text to provide relevant background information. We attempted to contact still-living FUSION spouses and offspring who participated in FUSION study visits between 1994 and 1998 [391], individuals who had participated in the population-based Savitaipale Prospective Diabetes Study [390], the FINRISK 2007 survey, the Dose Responses to Exercise Training (DR's EXTRA) study [190] and the Metabolic Syndrome in Men (METSIM) study [365]. Additional subjects were recruited by newspaper advertisements. We excluded individuals: (1) with drug treatment for diabetes, (2) with diseases that might be expected to confound the analyses (for example, cancer, skeletal muscle diseases, acute or chronic inflammatory diseases), (3) with diseases that increase haemorrhage risk during biopsy (for example, von Willebrand's disease, haemophilia, severe liver diseases), (4) taking medications that need to be taken daily and increase haemorrhage risk in the biopsies including warfarin (patients on acetylsalicylic acid were instructed to stop for 7 days prior to biopsy), (5) taking medications that could confound the analyses (for example, oral corticosteroids, other anti-inflammatory drugs such as 5-ASA, infliximab or methotrexate), and (6) of age < 18 years. The study was approved by the coordinating ethics committee of the Hospital District of Helsinki and Uusimaa. A written informed consent was obtained from all the subjects.

## 2.3 Clinical visit and phenotyping

Full clinical procedures are described in Scott et al. [341] and are copied below. Clinical visits were performed in Helsinki, Savitaipale and Kuopio on average 14 days prior to biopsies. The clinical visit followed a 12-hour overnight fast and centered around a 4-point (0, 30, 60, 120 min) 75 g oral glucose tolerance test (OGTT). We defined glucose tolerance categories of normal glucose tolerance (NGT), impaired glucose tolerance (IGT), impaired fasting glucose (IFG) and T2D using World Health Organization (WHO) criteria [427]. We determined OGTT plasma glucose (fluoride citrate plasma) concentrations by hexokinase assay (Abbott Architect analyser, Abbott Laboratories, Abbott Park, IL, USA).

Several metabolites were also measured at the first time point, during the fasting state. Serum insulin and serum C peptide concentrations were assayed by chemiluminescent microparticle immunoassays using the Abbott Architect analyser (Abbott Laboratories, Abbott Park, IL, USA). Serum triglycerides, total and HDL cholesterol were measured by enzymatic methods with the Architect analyser. LDL cholesterol concentration was calculated using the Friedewald formula [110]. All laboratory analyses were performed at a certified core laboratory at the National Institute for Health and Welfare, Helsinki, Finland. In addition, during the clinical visit, anthropomorphic information, health history, medication, and lifestyle questionnaires were collected. Height was measured to the nearest 0.1 cm. Height and weight were measured in light clothing. Waist circumference was measured midway between the lower rib margin and the iliac crest. Hip circumference was measured at the level of the trochanters. Body mass index (BMI) was calculated as weight (kg) divided by the square of height (m).

In collaboration with a physician (Andrea Ramirez, NIH), I processed the medication information. This involved (1) parsing the medical information which included Anatomical Therapeutic Chemical (ATC) codes [425], (2) physician review and diagnosis, and (3) validation of physician diagnosis by cross referencing with the MEDication Indication (MEDI) database [416]. I performed the computational steps of items 1 and 3. With the help of the database managers (Heather Stringham, University of Michigan and Leena Kinnunen, Finland National Institute for Health and Welfare), I verified that medication information was missing from one participant who was excluded from medication analyses. This participant was also the non-Finnish participant (see Section 2.5.2).

## 2.4 Biopsy visit

The following text is reproduced from Scott et al. [341] and describes biopsy procedures. Biopsies were performed using a standardised protocol and one physician trained all doctors performing biopsies. We instructed participants to avoid strenuous exercise for at least 24 hours prior to biopsy. Following overnight fast, we obtained ~250 mg vastus lateralis skeletal muscle using a conchotome, under local anaesthesia with 20 mg ml $^{-1}$ lidocaine hydrochloride without epinephrine. Altogether 9 experienced and well-trained physicians collected 331 muscle biopsies in 2009-2013 in 3 different study sites (Helsinki, Kuopio and Savitaipale). Three physicians, one in each site, performed most of the biopsies (237 biopsies). The muscle samples were cleaned of blood, fat and other non-muscle tissue by scalpel and forceps, rinsed with NaCl 0.9% solution, and frozen in liquid nitrogen. Samples were frozen within 30 seconds after sampling. Muscle samples were then stored at -80 °C for a duration of 0-4 years before analysis. Overall, the biopsy procedure was well tolerated. Apart from a few expected cases of bruising, numbness at the biopsy site and vasovagal reactions, there were no clinically significant adverse sequelae. All biopsies were shipped to the Collins lab at the NIH where they were processed to generate genotype and molecular trait information.

## 2.5 Genotyping procedures and quality control

In the genotype quality control steps outlined below, I helped oversee the analysis carried out by Narisu Narisu (NIH) and Anne Jackson (University of Michigan). I helped choose the final filtering parameters for imputation, flagged the non-Finnish sample by analysing genotype principal components (PCs), and set the final number of PCs we included in later analyses as covariates. Below, I outline the genotyping steps. This text is partially drawn from Scott et al. [341], but altered to account for additional genotypes.

We extracted DNA from blood. DNA samples were genotyped at the Genetic Resources Core Facility of the Johns Hopkins Institute of Genetic Medicine. 327 samples were genotyped on the HumanOmni2.5-4v1_H BeadChip array, while 4 were genotyped on the InfiniumOmni2-5Exome-8v1-3 BeadChip array (Illumina, San Diego, CA, USA). We mapped the Illumina array probe sequences to the GRCh37/hg19 genome assembly using the Burrows-Wheeler Aligner [213]. We excluded SNPs with probe alignment problems, known variants in the 3' end of probes, and reduced the original set to 2,277,032 common markers between arrays. We further filtered out markers with call rates < 95%, minor allele count (MAC) < 1, or Hardy-Weinberg equilibrium p-value < $10^{-4}$, leaving 1,571,557 SNPs for subsequent analysis (including 33 chrY and chrM markers). All alleles were oriented relative to the reference.

### 2.5.1 Genotype imputation

In order to reduce the effect of ambiguous SNPs with respect to pre-phasing and subsequent imputation, we removed array markers exhibiting an alternate allele frequency difference of

> 20% with phase 3 1000 Genomes European data, palindromic SNPs with a minor allele frequency > 40%, genotype missingness > 2.5%, or Hardy-Weinberg p-value < $10^{-4}$. A total of 1,556,249 markers were used in pre-phasing and imputation.

We performed pre-phasing and imputation separately on autosomal and chrX markers using the Michigan Imputation Server [74]. We used eagle v2.3 [228] for autosomal marker pre-phasing and shapeit v2.r790 [80] for chrX markers. We subsequently used minimac3 [74] for imputation of missing genotypes using the Haplotype Reference Consortium (HRC) panel (hrc.r1.1.2016) [246]. At the time of imputation, the HRC panel only supported SNP imputation.

### 2.5.2 Sample quality control

We analysed sample relatedness and identified two unexpected pairs of first-degree relatives using KING [236]. Each was an NGT-IGT pair; from each pair we excluded the NGT participant. We performed PCA, merging the FUSION samples with a population reference panel to verify Finnish ancestry using LASER [407]. We identified and removed one non-Finnish participant (Figure 2.1, Table 2.1). Within the remaining FUSION samples we performed PCA using eigenstrat [301] on 437,182 genotyped, autosomal SNPs with MAF > 1% and Hardy-Weinberg equilibrium p-value < $10^{-4}$, after excluding SNPs from regions of high LD and LD pruning SNPs to a pairwise $r^2$ threshold of 0.5 [302, 413]. We found the first 4 PCs to be significant (p-value < 0.1) and included them in later analyses to account for sample relatedness (Figure 2.2). One sample swap identified in the RNA-seq and DNAme data was determined to be a genotype swap based on sib-pairs from previous FUSION studies. This swap was corrected in the raw data before PCs were generated and after imputation.

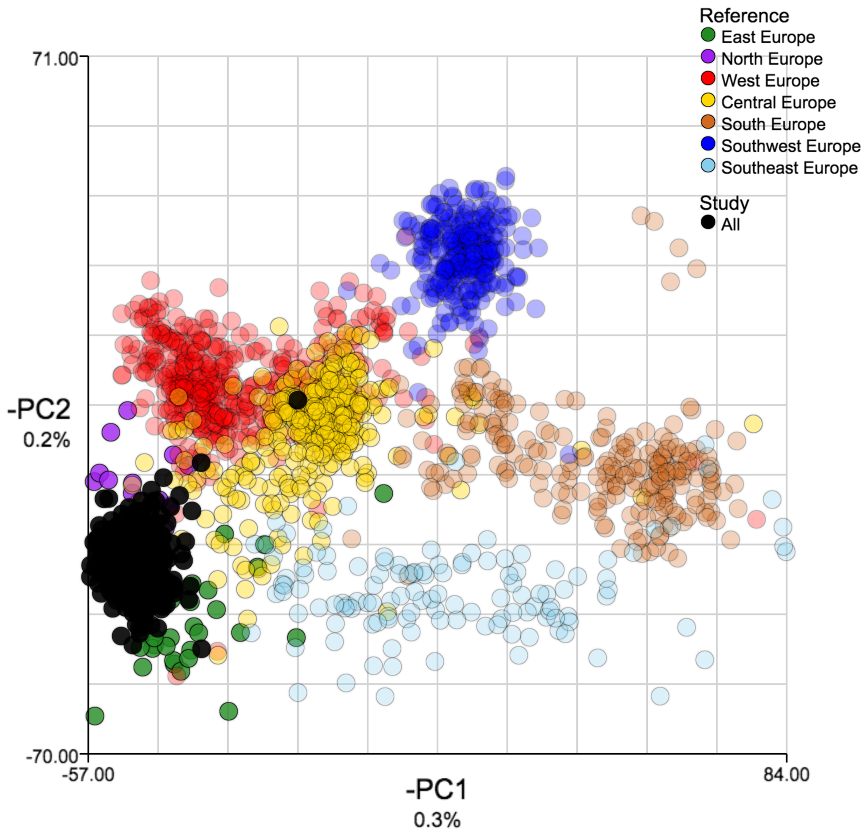


**Figure 2.1** FUSION genotype PCs (black) projected onto European populations. We identified and removed one non-Finnish participant that clustered with central Europeans.

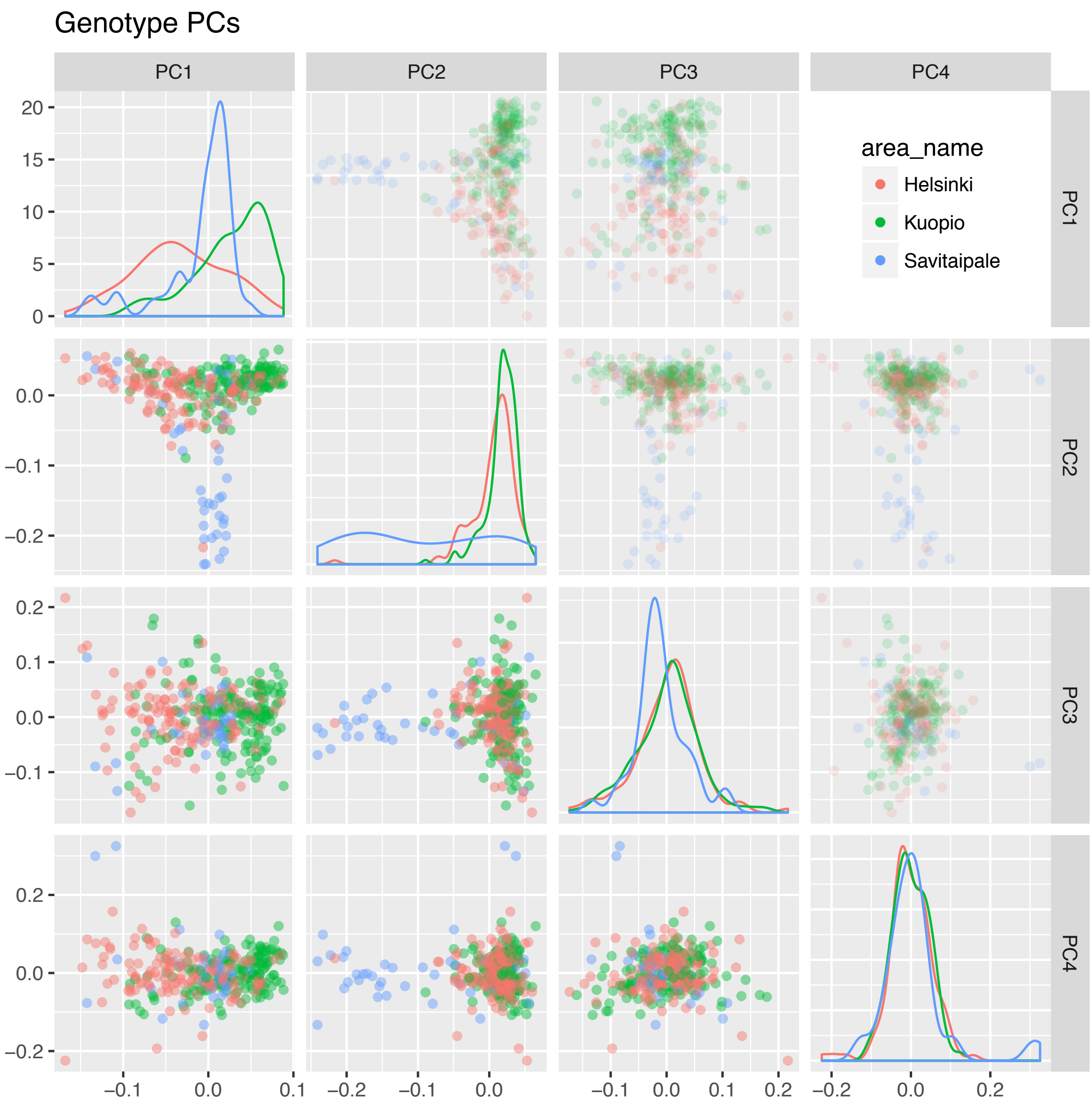


**Figure 2.2** FUSION genotype PCs. PCs generated after all genotype sample exclusions, coloured by sample collection site.

## 2.6 Measuring gene expression and quality control

We previously published the FUSION skeletal muscle gene expression data [341]. For the initial publication, NIH colleagues, Michigan colleagues, and I developed the code used to map reads and perform QC. Since completing the work for that publication, we sequenced additional skeletal muscle biopsies and made modifications to the analysis pipeline. Here, I briefly outline the major points of the RNA-seq analysis pipeline used in the bulk of this thesis. I note differences between this section and Scott et al. [341] at the end of this chapter (Section 2.8). When appropriate, parts of the text below are reproduced from Scott et al. [341] and slightly updated to reflect additional samples and slight alterations to the data processing pipeline. In regards to the pipeline, Peter Chines (NIH) executed the bulk of the RNA-seq pipeline, and I helped decide on sample and gene filters, which were presented and approved by the larger FUSION tissue analysis group. In addition to developing code that was deployed in many sections of the RNA-seq analyses, I specifically ran the PCA outlier analysis as well as the tissue deconvolution analysis, described below.

### 2.6.1 RNA isolation and sequencing

We visually dissected 30-50 mg of each frozen muscle biopsy sample (323 biopsies) to avoid adipose tissue. Total RNA was extracted and purified with Trizol (Invitrogen, Carlsbad, CA). RNA integrity numbers (RIN) ranged from 6.6 to 9.4 (median 8.4). RIN information was missing from one sample due to a technical error in the Bioanalyzer machine (Agilent). All other samples processed in the batch with the missing sample had an average RIN of 8.5, suggesting high quality. Using all muscle RNA-seq samples, we mean imputed the missing RIN for downstream analysis, estimating it to be 8.4.

To minimize and quantify batch effects, we randomly queued samples for sequencing using a 24-sample barcode-pooling approach and targeted proportional representation of the OGTT states (NGT, IGT, IFG and T2D) in each sequencing batch. External RNA Controls Consortium (ERCC) RNA controls [171] were spiked prior to barcoding to facilitate library QC. In total, we submitted 323 Poly(A)-selected RNA samples for sequencing at the NIH Intramural Sequencing Center (NISC) using the Illumina TruSeq directional mRNA-seq library protocol to a targeted depth of 80 million 100 bp paired-end reads per sample.

### 2.6.2 RNA-seq processing and quality control

We retained RNA-seq reads passing the Illumina chastity filter and mapped reads to a reference sequence composed of ERCC control fragments and all chromosomes and contigs from GRCh37/hg19, excluding alternate haplotypes, replacing chromosome M with the Cambridge Reference Sequence and masking the pseudoautosomal region on chromosome Y. We aligned reads using STAR v2.3.1y [85] with default parameters and a splice junction catalogue based on GENCODE v19 [146]. Duplicate read pairs were retained. Non-uniquely mapping reads and read pairs with unpaired alignments were discarded.

We performed RNA-seq QC at the level of read groups, defined as a library on a lane, using QoRTs v1.1.18 [147]. Seeing little variation from lane to lane, we summarised the QoRTs measurements by taking the mean for each sample. We looked for outliers using a variety of measures including GC content, transcriptional diversity, and gene body coverage. There were no outliers for GC content. For transcriptional diversity, we calculated the distribution of the fraction of total transcription in 500 roughly equal count bins, according to the median counts for each gene, then compared each sample to the median of all samples using the Kolmogorov-Smirnov test (ks.test function in R), dropping 7 outlier samples (p-value < 0.01; Figure 2.3). Many of the samples removed as outliers in the transcriptional diversity analysis also had a decreased fraction of skeletal muscle when included in tissue heterogeneity estimates. Most notably, all 4 samples with an estimated skeletal muscle fraction < 90% were dropped based on transcriptional diversity measures. For gene body coverage, we compared samples based on the fraction of reads in 40 bins along the normalised length of all genes; we dropped four samples as outliers based on their coverage the 3' end in the (0.9,0.925] bin, possibly indicating RNA degradation.

We used verifyBamID v1.1.1 [175] with the following parameters "--ignoreRG --precise --best --maxDepth 100" to remove RNA-seq samples comprised of reads derived from more than one individual and and identify sample swaps by comparing transcribed SNPs to SNP chip genotype data. We identified two pairs of sample swaps and removed one sample that showed high levels of contamination (~8%). In addition, we removed six intentional replicates, one unintentional replicate, one participant of non-Finnish ancestry, and one of 2 pairs of first-degree relatives. After all exclusions, there were 301 muscle RNA-seq samples available for analysis (Table 2.2).

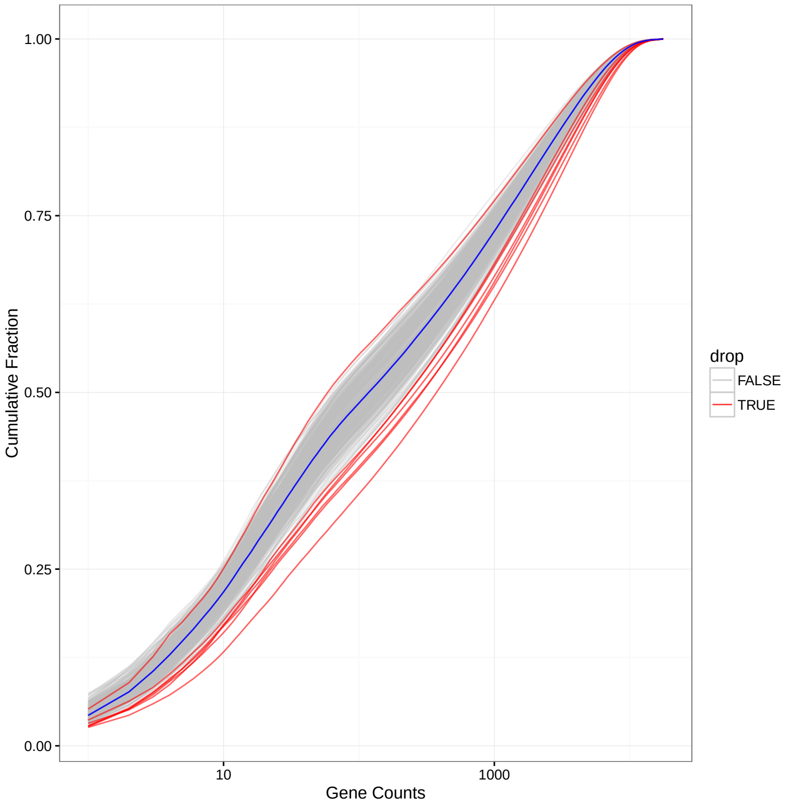


**Figure 2.3** Transcriptional diversity of each sample. Red samples exhibit increased transcriptional diversity and were dropped. The blue line shows the median across all samples.

We previously [341] quantified gene expression as fragments per kilobase per million reads (FPKM) [266, 386]. This unit of measure accounts for bias in transcript length, where even if expressed at the same level, longer transcripts have more reads because they produce more molecules in the fragmentation step of Illumina sequencing. In addition, the FPKM unit of measure controls for variable sequencing depth (total number of reads obtained in one sequencing run), which if not accounted for would make genes equally expressed in two samples appear more expressed in the sample with greater sequencing depth. Within a sample, for a gene, $g$, the FPKM is calculated as:

$$FPKM_g = \frac{c_g 10^9}{l_g N} \tag{2.1}$$

where $c$ is the number fragments mapping to a gene's exons, $l$ is the length of the gene (sum of exons—number of possible start positions for a fragment), and $N$ is the sequencing depth of a sample (number of mapped reads).

Transcripts per million (TPM) is another expression measurement unit, slightly different from FPKM, and is thought to be a more accurate measurement of relative molar RNA concentration [405, 211]. Within a sample, for gene, $g$, among $n$ total genes, the TPM is calculated as:

$$TPM_g = \frac{c_g}{s_g} \times \frac{1}{\sum_{j=1}^{n} c_j s_j} \times 10^6 \tag{2.2}$$

where $c$ is the number fragments mapping to a gene's exons and $s$ the effective gene length defined as the length of unioned exons for a gene minus the median insert length of a sample.

Given these reports on TPM [405, 211], we decided to change the unit of measure for gene expression from FPKM to TPM. However, for exon expression, it was unclear how to calculate effective gene length, because, unlike for genes, the exon fragment length is often smaller than the insert size. In such cases, the effective gene length, $s$, would be negative. Therefore, we used FPKM as a unit of measure for exon fragments.

Definitions for all transcriptome features were based on GENCODE v19 [146]. We counted fragments mapping to genes using htseq-count v0.5.4 [10], and used QoRTs to parse GENCODE v19 exon annotations into non-overlapping fragments and count exon reads. To reduce the number of transcripts per gene, to avoid identifiability issues, and to restrict analysis to high-confidence transcripts, we estimated transcript expression values for the subset of GENCODE transcripts with the tag “basic” in the GTF file.

After the above exclusions and swaps, we adjusted the gene expression TPMs for age, sex, batch, and RIN, and performed PCA on the residuals to look for additional outliers. We selected the first 2 PCs, which accounted for > 40% of the cumulative variance explained and transformed the PCs to z-scores. We found no striking outliers that warranted removal, defined as |z-score| > 5 (Figure 2.4). For subsequent linear models, we filtered for genes with $\geq$ 5 counts in > 25% of samples and inverse normalised the TPMs. Additionally, using the filtered expression data did not affect the PCA outlier decisions.

**(a)** Cumulative variance explained **(b)** PCs

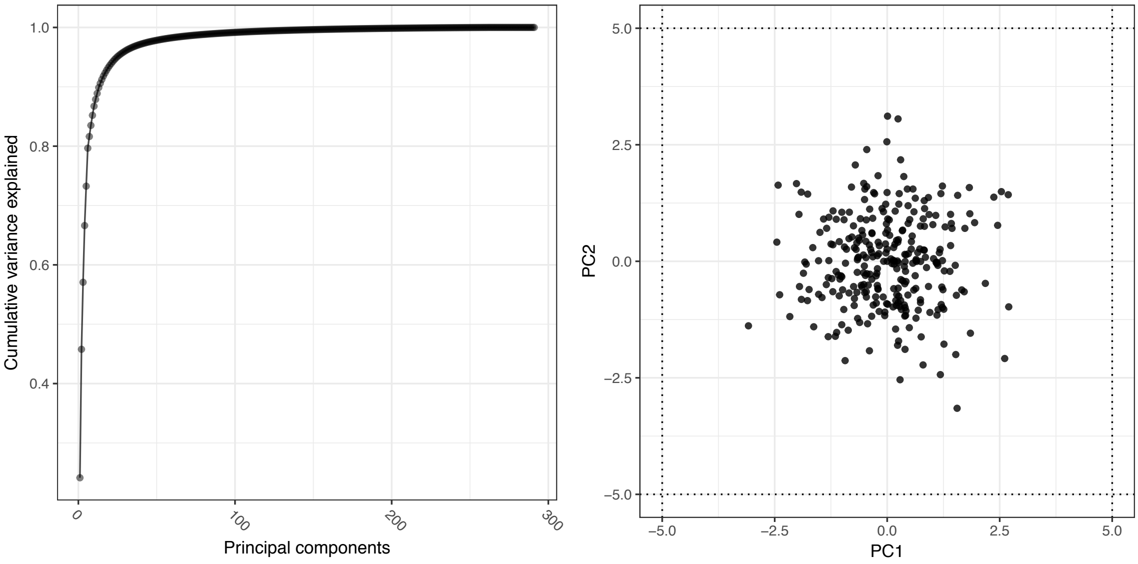


**Figure 2.4** Expression PCA. (a) Cumulative variance explained by each PC. (b) No outliers were identified in the first 2 PCs.

### 2.6.3 Gene expression tissue specificity index

My collaborators at the University of Michigan previously developed a method to measure the cell/tissue type specificity of gene expression, termed the expression specificity index (ESI) [341, 395]. Genes with a large ESI are highly and specifically expressed in a single cell/tissue type based on the reference panel used to generate the index. For instance, we previously used this method to identify genes with muscle specific expression patterns based on a muscle expression specificity index (mESI) that was generated using 16 tissues from Illumina Human Body Map 2.0 [341].

In order to calculate mESI values over a more comprehensive reference panel, I applied this method to 49 tissues from GTEx (v6p), removing tissues with < 25 samples (bladder, ectocervix, endocervix, and fallopian tube). Using the raw read counts, Peter Chines and I quantified GTEx gene expression as TPMs as opposed to RPKMs in order to be consistent with the FUSION data. For each gene in each tissue type, I calculated the average expression across samples to build a reference transcriptome panel. With this reference transcriptome panel, I calculated muscle specificity as previously described [341, 395] and reproduced here, slightly modified to fit the GTEx data.

We calculated the relative expression of each gene (*g*) in skeletal muscle compared with all 49 tissues (*t*) as *p*:

$$p_{g,muscle} = \frac{x_{g,muscle}}{\sum_{t=1}^{49} x_{g,t}} \tag{2.3}$$

We next calculated the entropy for expression of each gene across all 49 tissues as *H*:

$$H_g = -\sum_{t=1}^{49} p_{g,t} log_2(p_{g,t}) \tag{2.4}$$

Following previous studies [149, 338] we defined muscle tissue expression specificity (*Q*) for each gene as:

$$Q_{g,muscle} = H_g - log_2(p_{g,muscle}) \tag{2.5}$$

To aid in interpretability, we divided $Q$ for each gene by the maximum observed $Q$ and subtracted this value from 1 and refer to this new score as the mESI:

$$mESI_g = 1 - \frac{Q_{g,muscle}}{max(Q_{muscle})} \tag{2.6}$$

The final mESI scores near zero represent low and/or ubiquitously expressed genes, and scores near one represent genes that are highly and specifically expressed in skeletal muscle.

It should be stressed that the actual cell/tissue type specificity of this measurement depends on the quality of the reference panel used. In this case, I used GTEx as a reference panel since it is the most comprehensive multi-tissue gene expression dataset to date. However, like the FUSION data described in Section 2.6.4, the GTEx data are derived from tissue biopsies that are composed of a population of heterogeneous cell/tissue types, which could obscure the true cell/tissue type expression signature of a gene. In the future, such issues could be mitigated by using a multi-tissue, single cell expression dataset as a reference panel [308].

### 2.6.4 Gene expression tissue deconvolution

To estimate tissue heterogeneity in the FUSION tissue biopsies, I compared FUSION TPMs across all protein coding genes, not filtering for genes with $\geq 5$ counts in $> 25\%$ of samples, to the average TPM in the GTEx data using DeconRNASeq v1.16.0 [129]. I did not filter genes for counts because genes that are not expressed may be useful in tissue deconvolution. For instance, if a highly expressed gene specific to adipose was not expressed in our data, it would indicate adipose contamination is unlikely. In order to limit to the most relevant genes per tissue, I selected the top 500 tissue specific genes for each considered tissue in the GTEx reference panel. For skeletal muscle, I selected “skin not sun exposed suprapubic”, “whole blood”, “adipose subcutaneous”, “muscle skeletal”, and “EBV transformed lymphocytes” as a reference panel from GTEx. Across the skeletal muscle samples after QC, I estimated

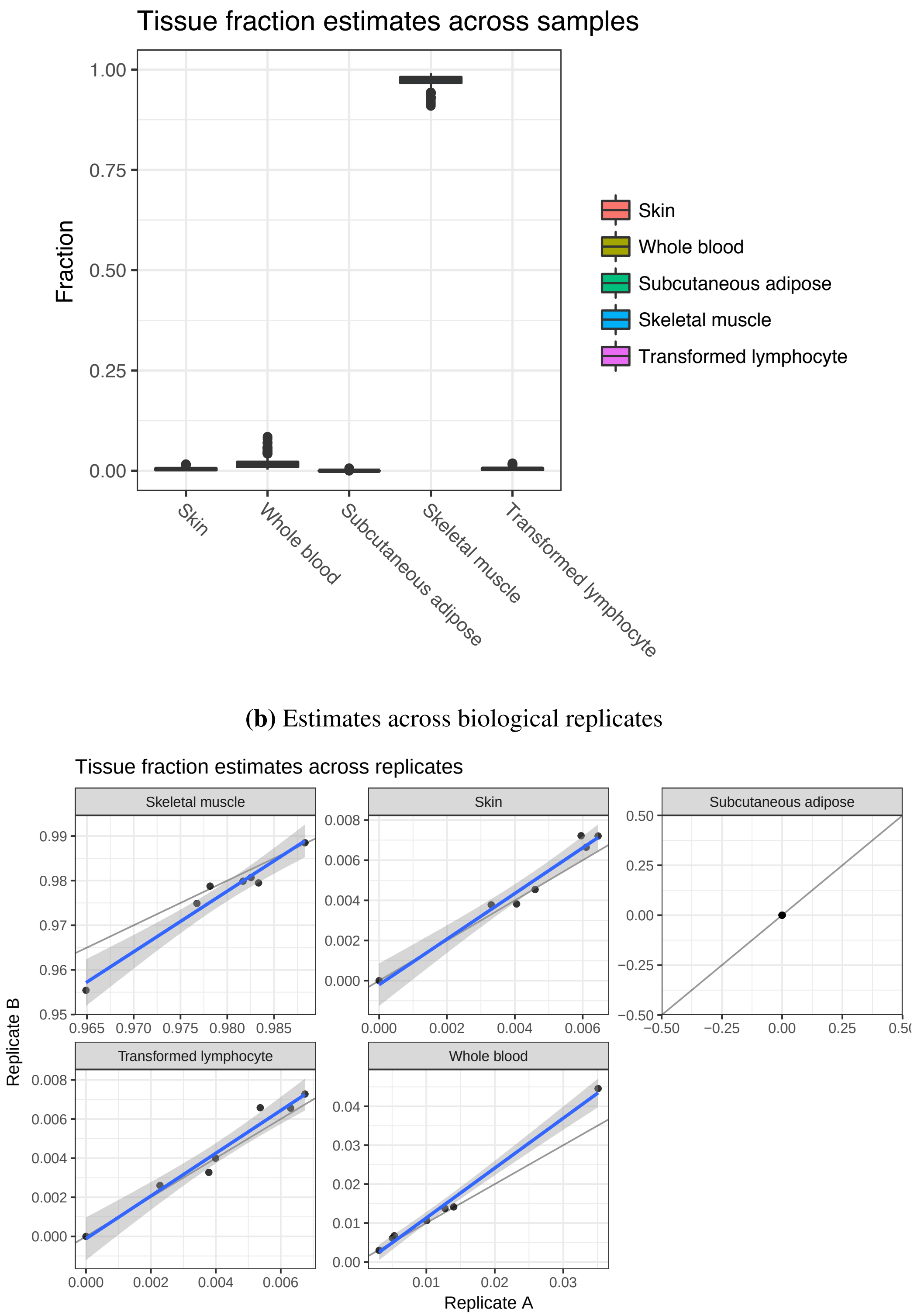
$< 0.1\%$ adipose, 0-2% skin, 0-2% lymphocytes, 0-8% whole blood, and 91-99% skeletal muscle tissue heterogeneity (Figure 2.5a).

In order to quantify the reproducibility of our tissue estimates, I compared the tissue fraction estimates of 6 replicates (obtained by separately processing the same source tissue) and one unintentional replicate due to a sample swap, confirmed by genotypes. I found the tissue estimates between these samples to be remarkably similar (Figure 2.5b).

Because these replicates were a separate processing of the same tissue source, such replicates are similar to how we generated the DNAme data—by separately processing the same frozen tissue stock. Since the expression-based tissue heterogeneity estimates between these replicates is so similar, I used them for the DNAme data, as no appropriate reference methylation panel could be found without integrating data from many studies, which would introduce confounding batch effects.

Finally, I note that these estimates are not foolproof as (1) the reference panel is also composed of heterogeneous tissue (e.g., no "pure" muscle) and (2) the reference panel may not encompass all relevant cell types, for instance specific blood cell types instead of simply "whole blood". Therefore, in subsequent analysis, I treated the tissue heterogeneity estimates as an approximation of the true tissue heterogeneity.

**(a)** Estimates across samples



**(b)** Estimates across biological replicates

**Figure 2.5** Tissue heterogeneity estimates. (a) Estimates across samples. (b) Estimates across biological replicates.

## 2.7 Measuring DNA methylation and quality control

I oversaw the DNA methylation QC and directly ran most of the steps, outlined below. The muscle samples used in this thesis were submitted along with several other samples for DNAme measurement including adipose, blood, islet, and EndoC betaH1 (pancreatic beta cell line). The adipose and blood samples were also collected as part of the FUSION biopsy study and are matched to the skeletal muscle biopsy participants. I processed all of these samples together and therefore describe the QC steps across all samples, even though this thesis focuses on the muscle biopsy samples.

### 2.7.1 DNA isolation and methylation quantification

We visually dissected ~25 mg of each frozen muscle and ~100-150 mg of frozen adipose biopsy sample, taking care to avoid adipose tissue in muscle and particularly bloody tissue sections in adipose. Genomic DNA was obtained from ~2,000 islet equivalents of islet tissue, cultured in various glucose stimulation states. Finally, for the pilot plate, DNA was extracted from ~6 ml of whole blood.

For each tissue and blood sample, genomic DNA was extracted using DNeasy Blood & Tissue Kits (QIAGEN), according to the manufacturer's recommendations. Genomic DNA extraction from pancreatic islets and EndoC was performed using the Gentra Puregene Cell kit (QIAGEN), according to the manufacturer's protocol. We submitted 200 ng of genomic DNA for 736 samples (41 islets, EndoC, 333 adipose, 337 muscle, and 24 blood samples) to the Center for Inherited Disease Research (CIDR) at the Johns Hopkins University, where they were bisulfite-converted using EZ DNA methylation Kits (ZYMO research), as part of the TruSeq DNA Methylation protocol (Illumina). Following bisulfite-conversion, CIDR measured DNAme using the Illumina Infinium HD Methylation Assay with Infinium MethylationEPIC BeadChips according to manufacturer's instructions.

With the addition of 4 controls (ZYMO research) per plate (two 100% methylation, one 0% methylation, and one 50% methylation generated by mixing 100% and 0% in an equimolar mixture), we filled the 768 total possible samples spanning 96 sentrix slides, each containing 8 arrays, run in batches of 12 across 8 plates. In order to test the CIDR pipeline before submitting the entire FUSION sample set, one plate (WG3000808) was submitted as an earlier pilot plate which contained all of the blood samples as well as 21 samples that were technical replicates of samples located on later plates. The 644 remaining adipose, muscle, and islet samples were randomised across the remaining plates.

### 2.7.2 Technical sample filters and quality control

I processed the idat files, which contain the raw methylation signal, using minfi v1.20.2 [12]. I dropped 15 failed samples (5 islet, 3 adipose, and 7 muscle) where > 1% of probes had a detection p-value > 0.05, as performed by Hannon et al. [144]. Similar to Aryee et al. [12], I compared the median signal intensity of the raw methylated (Meth) and unmethylated (Unmeth) channels across samples. I dropped one adipose sample where the difference between the median Meth and Unmeth signal was > 1 as well as 5 samples (1 adipose, 4 muscle) where the either the Meth and Unmeth signal was < 10 (Figure 2.6).

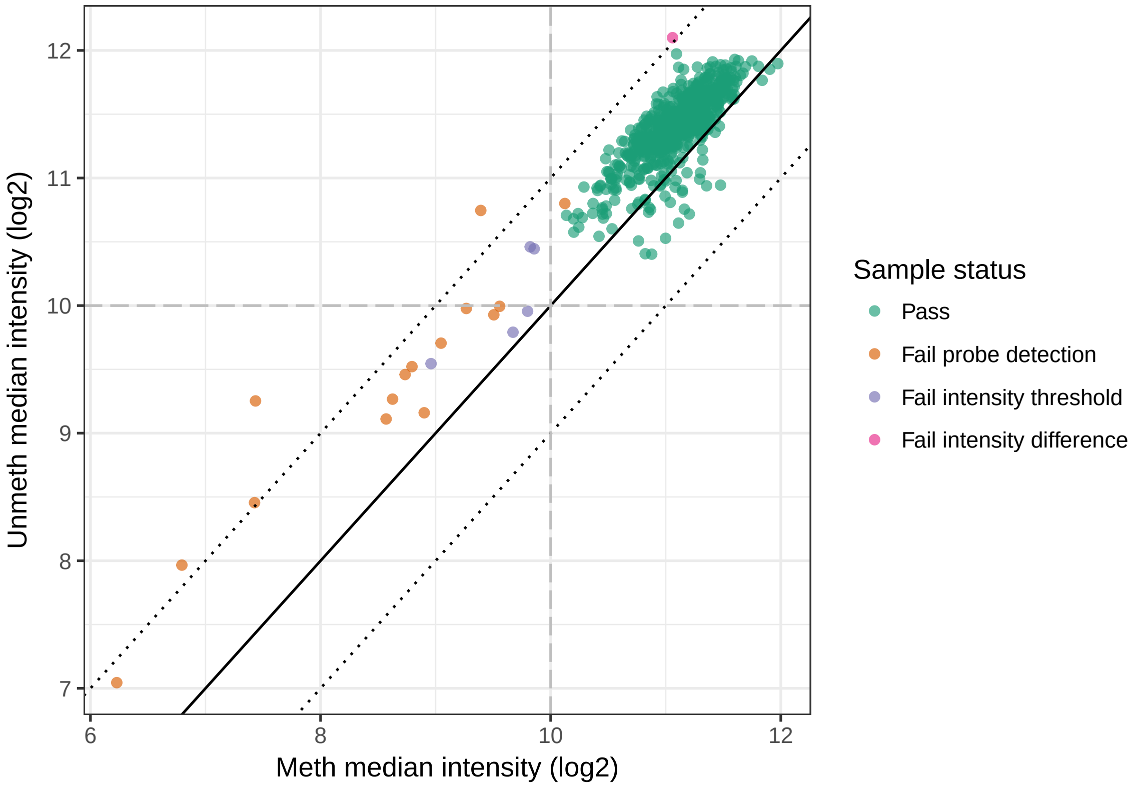


**Figure 2.6** Median signal intensity of the methylated (x axis) and unmethylated (y axis) channels across samples, coloured by QC status. The solid black line is the identity line. Small dot black lines depict boundaries where the deviation from the identity line is greater than 1. Grey lines show the intensity cutoff at 10.

Next, using the returnControlStat function from shinyMethyl v1.10.0 [107], I analysed the bisulfite conversion, extension, hybridisation, negative, non-polymorphic, specificity, staining, and target removal control probes. Many of the samples flagged in the previous steps were also outliers in the control probes (Figure 2.7). Of the samples that passed previous QC measures, I removed 1 islet, 8 adipose, and 3 muscle samples where the absolute value of the signal intensity z-score was > 3. The samples that were slight outliers, yet passed this filter, were not obvious outliers when looking at final tissue specific PCs after removing sample plate, sentrix position, plate position, age, and sex effects.

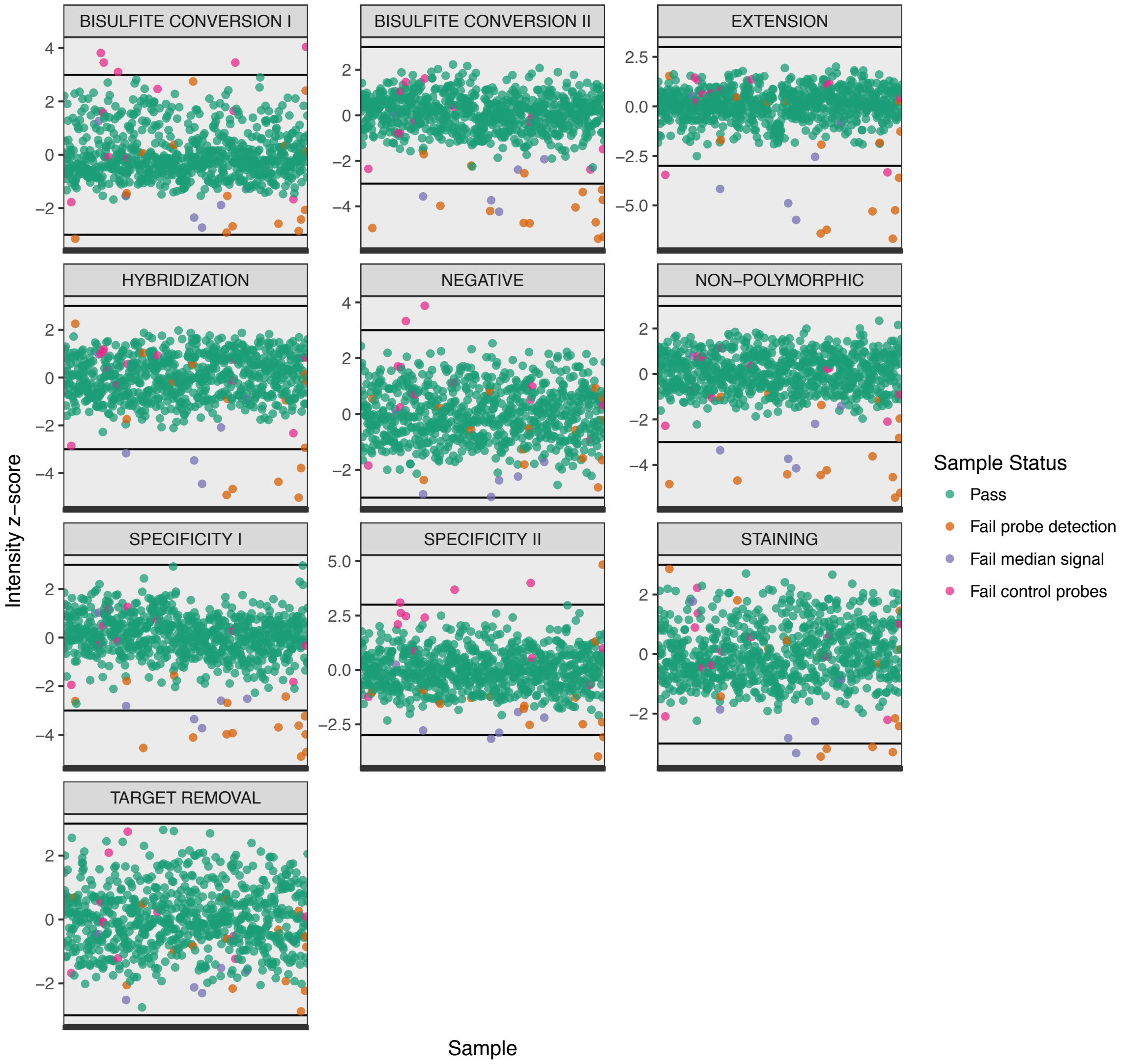


**Figure 2.7** Control probe QC. Mean control probe signal, transformed into z-scores across various classes of control probes, generated using returnControlStat from shinyMethyl. Samples coloured by QC status up to this filter.

### 2.7.3 Sample swap identification using genotypes

On the EPIC array, there are 59 SNP probes, designed to enable users to verify sample identity [295]. Peter Chines and I verified sample identity by comparing EPIC genotype calls from beta values to imputed genotypes. From the 59 EPIC SNP probes, we dropped (1) rs11249206 and rs2857639 because a variant in the HRC panel overlapped the last 10 bp of the probe (Section 2.7.5), (2) 6 markers that fail Hardy-Weinberg (p-value < $10^{-20}$), (3) rs6471533 because it had many beta values in-between genotype clusters, (4) rs939290 because it is tri-allelic in the HRC reference panel, and (5) two additional SNPs with more than 10 mismatches across samples. In total we compared 47 markers to imputed genotypes.

We dropped samples with a Manhattan distance between dosage vectors > 3 using the EPIC genotypes and the expected genotypes. In cases where we identified a perfect genotype match with a different sample, we changed the DNAme sample identifier to match the genotype sample identifier. Other cases with no match may indicate cases of contamination by another sample. In total, we dropped 15 samples that did not match the imputed genotypes (4 islet, 3 adipose, and 8 muscle). Of these failed samples, 10 were already removed due to a high fraction of failed probes. The additional 5 failed samples included 1 adipose and 4 muscle samples. We dropped an additional islet sample, despite a perfect genotype match, because we could not determine the glucose stimulation state due to the sample swap. Finally, we identified one sample swap previously found in the RNA-seq data, as well as three mislabeled adipose samples, two of which created sample replicates which were used in the later replicate analysis (however one replicate failed QC). After correcting sample swaps using genotypes, we further verified the identity of samples by comparing the recorded sex to the predicted sex based on methylation using the minfi getSex function with default parameters (Figure 2.8).

### 2.7.4 Plate quality control

I analysed bulk signal trends across arrays. I found that the pilot array, WG3000808, showed different trends in the overall distribution of beta values, M-values, Meth, and Unmeth signal intensities (Figure 2.9). In order to ensure these trends were not due to a different tissue being included in the pilot plate (blood), I removed all blood samples and found the batch effect persisted. I further analysed the median Meth and Unmeth signal of each sample and found a consistent, different profile for the WG3000808 plate (Figure 2.10). As per design, the muscle and adipose tissue samples on the WG3000808 array are technical replicates. I

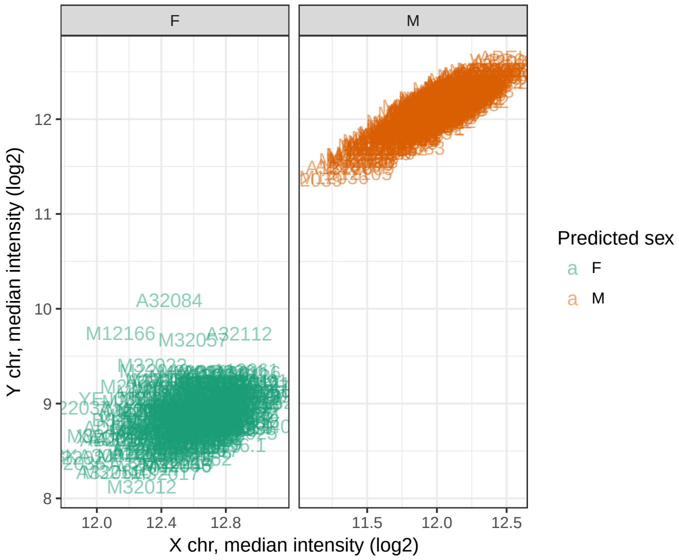
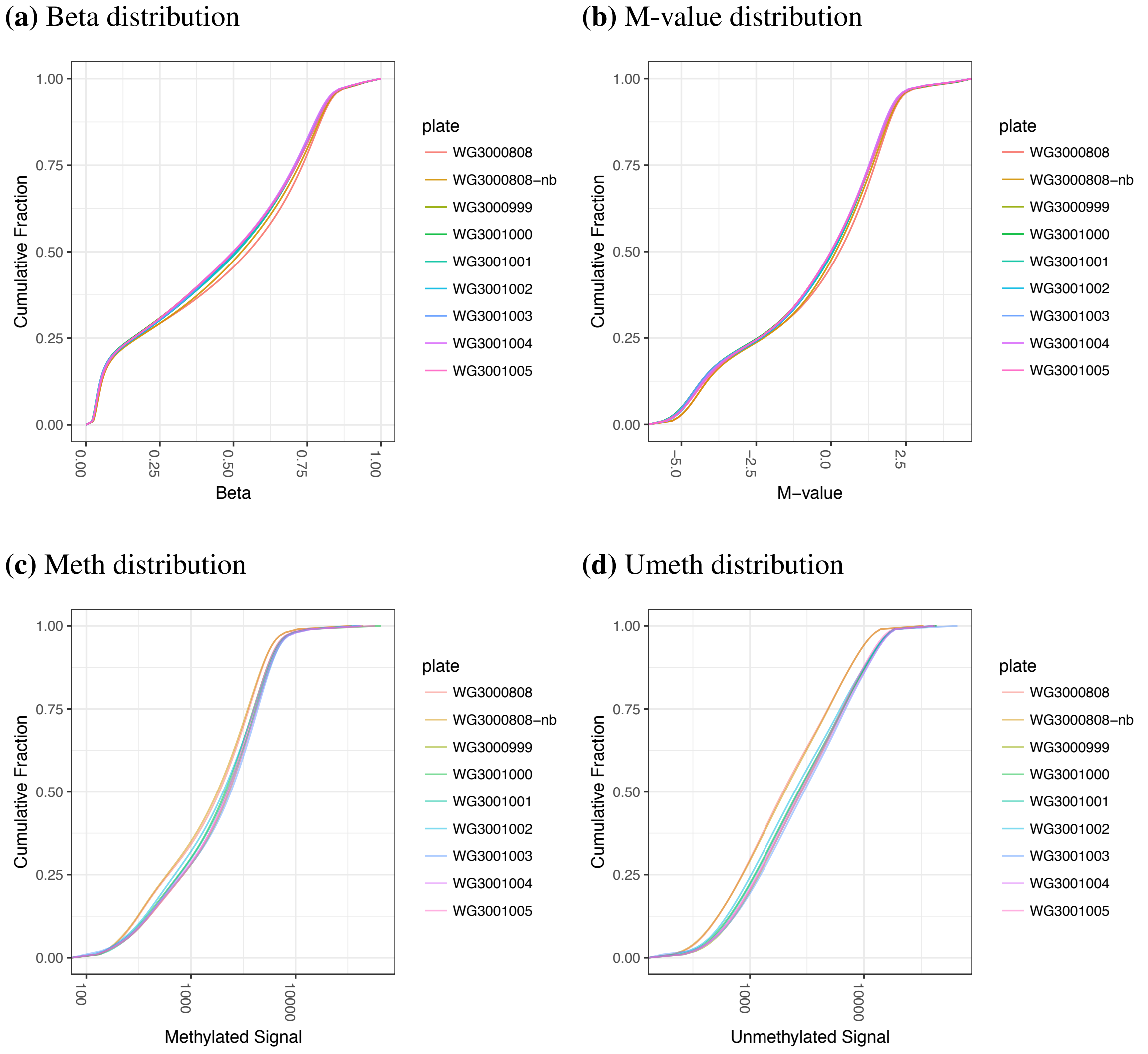


**Figure 2.8** Recorded sex validation using methylation. Facets show the recorded sex and colours show the predicted sex.

compared the raw beta signal of these samples to replicate pairs scattered across other plates and found a general trend in direction of the skewing of DNAme estimates across samples (Figure 2.11).

Unless specified otherwise, I removed the WG3000808 plate from further analyses, described after this section. However, before doing so, I included WG3000808 in the sample QC steps (see Section 2.7.7) before re-running these steps without this plate. Including or excluding WG3000808 did not affect the final result of any filters. Compared to other sample QC steps, excluding WG3000808 had the largest effect on the tissue specific PCA analysis, although this effect was still only marginal and did not affect the final outlier samples. Of the additional QC steps, when analysed with the other plates, no WG3000808 samples were flagged as outliers in the tissue methylation distribution analysis, the multi-tissue PCA analysis, or tissue specific PCA analysis on residuals after accounting for sample plate, sentrix position, plate position, age, and sex effects. Three adipose samples that failed on the other plates could have been rescued by drawing from WG3000808, but I rejected that option based on the large batch effect.



**(a)** Beta distribution

**(b)** M-value distribution

**(c)** Meth distribution

**(d)** Umeth distribution

**Figure 2.9** Plate methylation distribution. WG3000808-nb depicts WG3000808-“no blood” where blood samples on WG3000808 are dropped. (a) Beta values. (b) M-values. (c) Methylated signal intensity. (d) Unmethylated signal intensity.

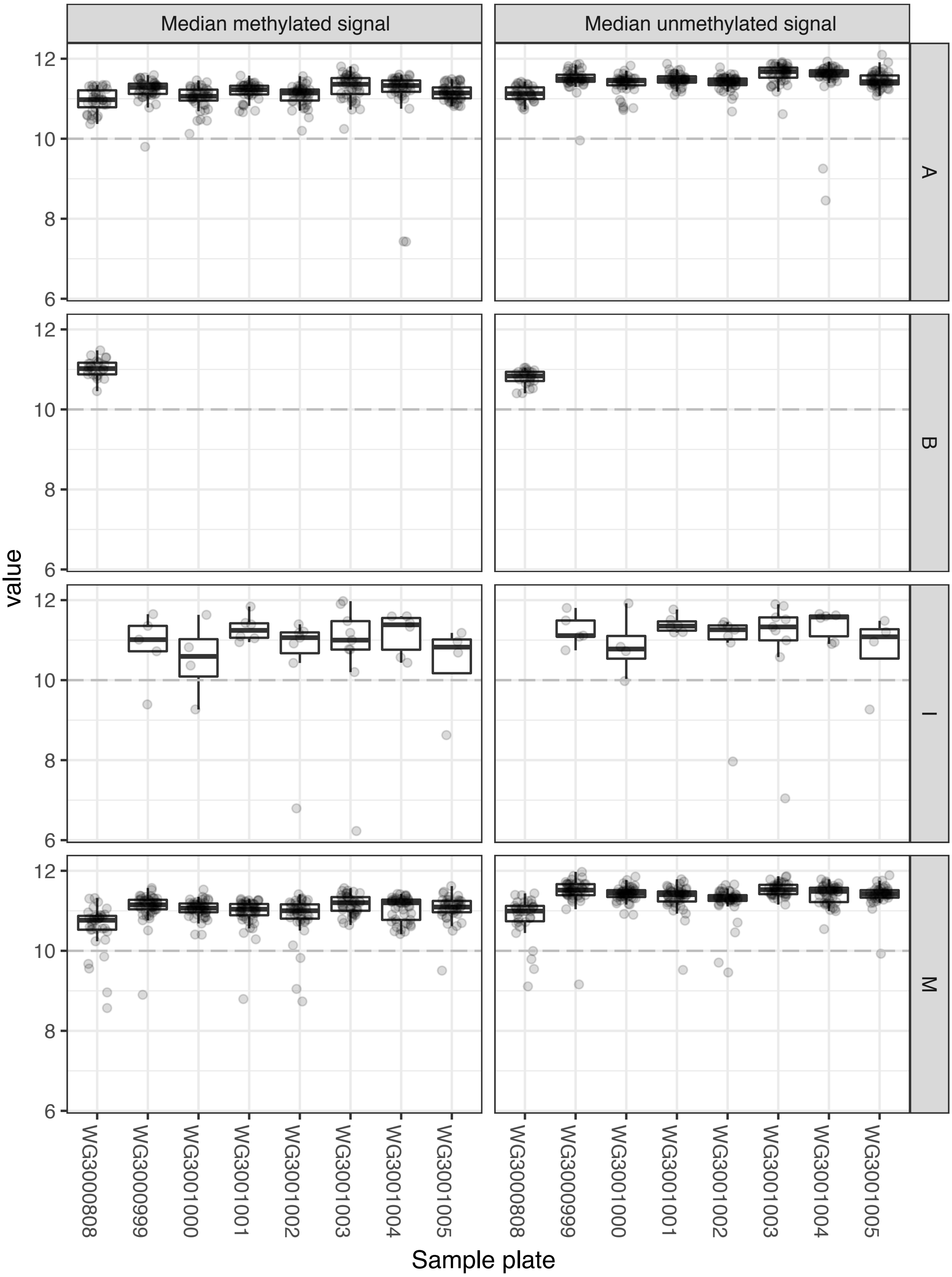
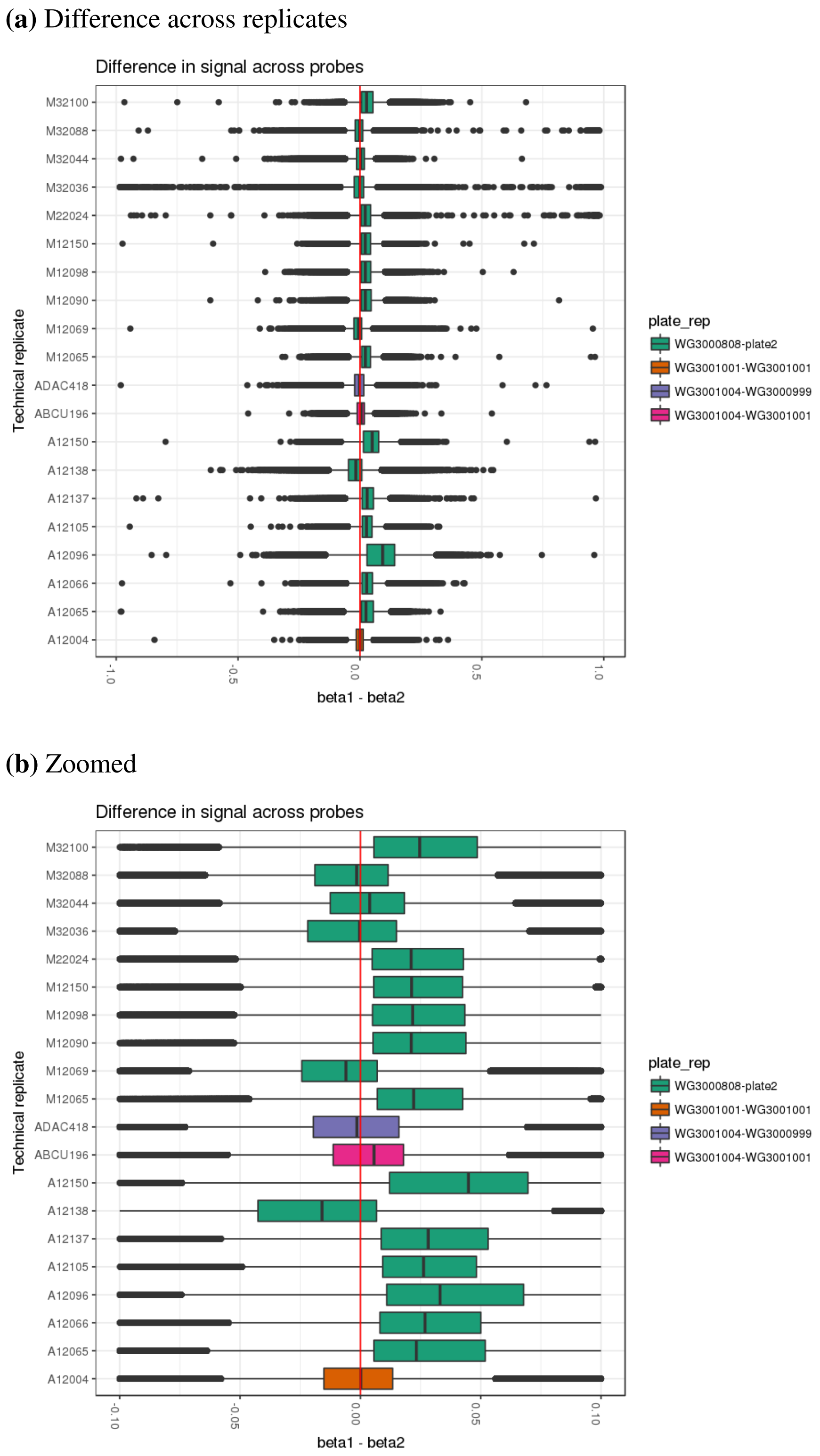


**Figure 2.10** Distribution of median methylated or unmethylated signal of each sample per plate. Dashed line at 10 shows the intensity cutoff point. The row facets split tissues: (A)dipose, (B)lood, (I)slets, and (M)uscle.



(a) Difference across replicates

(b) Zoomed

**Figure 2.11** Difference in raw methylation across replicates. Colors indicate the order of subtraction for comparison. For all green plots, methylation differences were calculated by taking the WG3000808 replicate, the non-WG3000808 replicate. Note that the orange replicate occurs on the same plate. (a) Full difference scale. (b) Zoomed in differences.

### 2.7.5 Probe filters and quality control

Peter Chines and I also worked together to remove potentially bad probes from the EPIC array. Some probes have been reported to be cross-reactive, mapping to more than one genomic location [60, 303, 247, 442]. Measurements from such probes are unreliable as they likely represent aggregate DNAme signal across multiple sites. To identify cross-reactive probes on the EPIC chip, we mapped non-control probes back to the entire bisulfite-converted genome (leaving out alternative haplotypes, and ignoring a single hit to a random contig when there is a single corresponding hit to a primary chromosome), using Novoalign's -b4 option, with allowance for up to three mismatches in the 50 bp probe alignment beyond the best alignment seen (-R120 option). We kept only uniquely mapping probes, removing 49,495 probes (Figure 2.12).

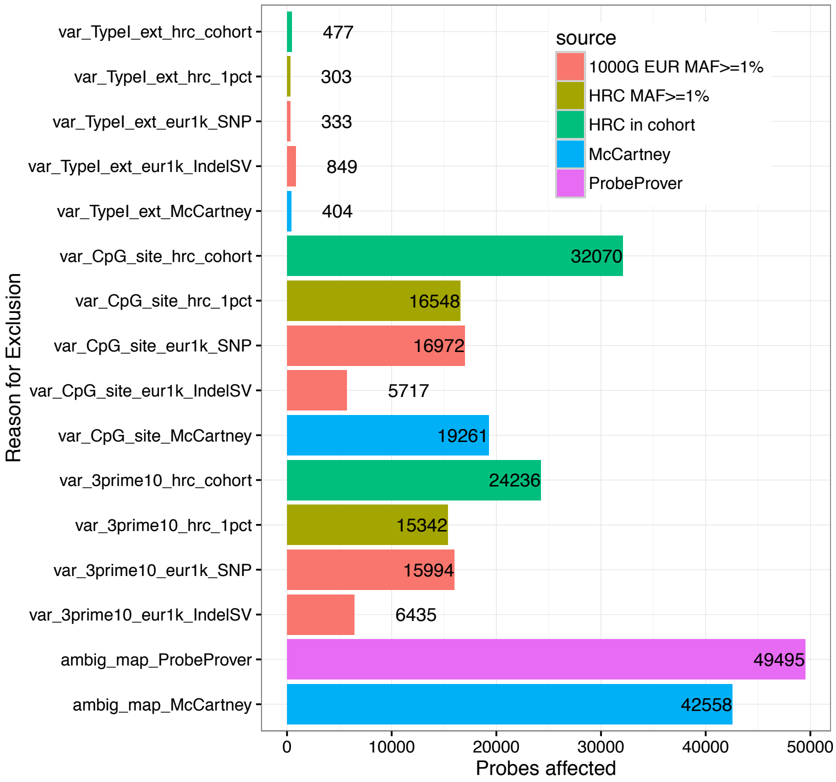
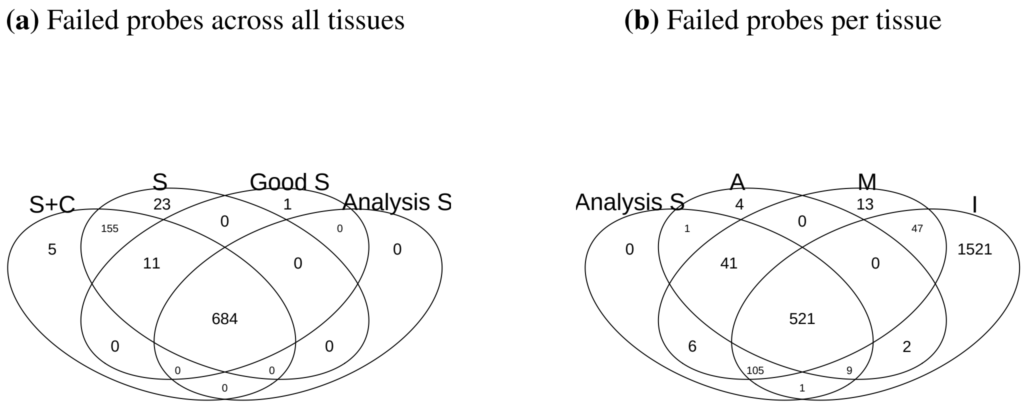


**Figure 2.12** Summary of blacklist probes excluded from analysis. ProbeProver is the term used to describe the method developed and used by FUSION for ambiguous probe mapping.

In addition, probes may also contain SNPs, which if common to the population of interest, could lead to biases in inter-individual studies. For example, "methylation" signals at polymorphic CpGs merely reflect the underlying genetic polymorphism [60] as well as exhibit significantly increased variation compared to all other probes [303]. In order to avoid such biases, we removed probes with a SNP within 10 bp of the 3' end of the probe, within the target CpG itself, and finally, in the case of type I probes, if the variant overlaps the single base extension site. We used 10 bp as a cutoff because it is consistent with previous studies [303].

For variants we used common (MAF $\geq 1\%$) SNPs, indels or structural variants in the phase 3 1000 Genomes European dataset, common (MAF $\geq 1\%$) SNPs in the HRC reference panel r1.1, and SNPs appearing at all in our own samples, even at low frequency, after imputation to the HRC reference panel. We chose to filter probes overlapping a SNP at any frequency in our imputed HRC genotypes, because we will likely use different sample subsets for future integrated muscle and adipose studies. We wanted a consistent analysis data frame across all studies, instead of applying a different MAF filter for only adipose or only muscle samples. In total we removed 63,840 probes due to SNP overlaps. As a final step, we combined our blacklist with a previously published EPIC probe blacklist from McCartney et al. [247] for a total of 120,627 unique probes which were removed from subsequent analysis (Figure 2.12).

After removing blacklist probes, I flagged probes with a high detection p-value, defined as p-value > 0.05 in $\geq 5\%$ of samples, for removal before later analyses. The probe detection p-value quantifies the probability that the combined Meth and Unmeth signal is above the background signal, estimated using negative control probes. One potential cause of such low quality signal could be due to spatial artefacts on the array [79]. I evaluated various methods to remove low quality probe filters. First, I considered across all tissues using four samples sets: (1) all samples and controls, (2) dropping controls, (3) only samples that passed QC, and (4) only the final, analysis samples (after dropping samples removed in genotype QC step and selecting one of each replicate pair). Overall, I found the different sample subsets affected only a small number of probes, relative to the whole dataset (Figure 2.13; note WG3000808 was dropped for this analysis). Second, using the final analysis samples (after tissue specific filters), I evaluated a per tissue probe filter. I found an increased number of probes that failed in islet samples, likely due to the fact that fewer islet samples were assayed. I decided to use a conservative approach, removing probes that failed $\geq 5\%$ of the final analysis samples per tissue type. After blacklist filters, I removed 578 adipose probes, 733 muscle probes, and 2,206 islet probes.


(a) Failed probes across all tissues

(b) Failed probes per tissue

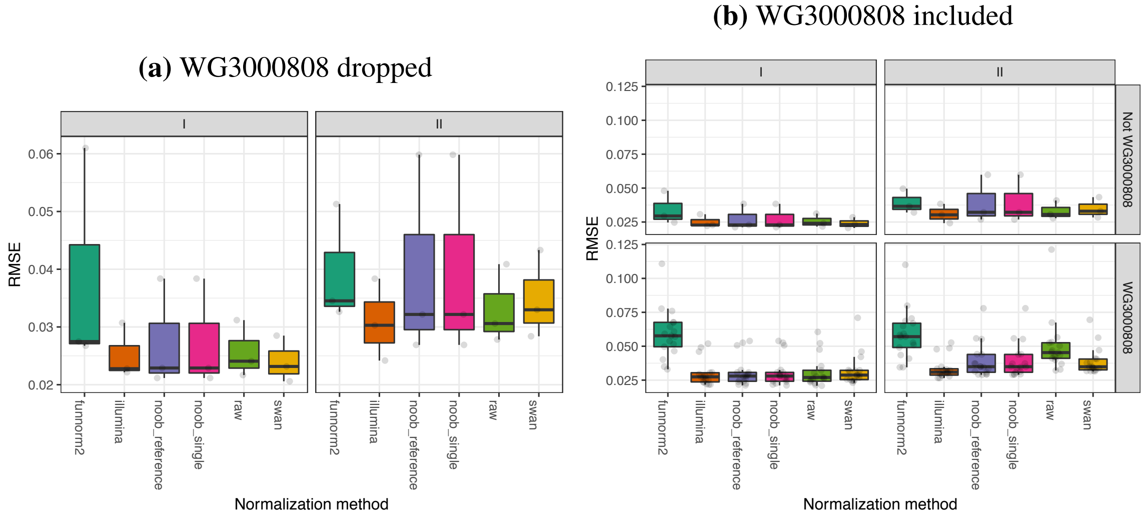
**Figure 2.13** Overlap of low quality probes with a high detection p-value across different sample sets. (a) S+C indicates that the probe failure rate was calculated across tissue samples and controls. S excludes the controls from the probe failure rate calculations. Good S excludes controls and samples that did not pass QC. Analysis S only uses samples that are included in the final analysis dataframe, after dropping samples removed in genotype QC step and selecting one of each replicate pair. (b) Comparison of failed probes across each tissue in the Analysis S set.

### 2.7.6 Normalisation

I processed the raw methylation data and compared potential normalisation techniques implemented in minfi: preprocessRaw (no normalisation), preprocessIllumina, preprocessSWAN, preprocessNoob, and preprocessFunnorm. For all methods I used default settings, except for preprocessFunnorm where I also included sex information. As a QC metric, I used the 3 replicates pairs that passed sample QC (2 islet and 1 adipose), calculating the root mean squared error (RMSE) between replicates across normalisation techniques, after removing blacklist and failed probes. One of the three replicates was created by a sample swap and I therefore could not determine if it was technical or biological replicate. This unintentional replicate also occurs on the same plate. The two remaining replicates are technical replicates.

I found the Illumina normalisation method implemented in minfi minimised RMSE (Figure 2.14). I also observe increased variability in type II probes, consistent with previous studies [78]. These trends were also consistent when drawing from the 16 additional technical replicates on the WG3000808 plate that passed QC. Note that I excluded the WG3000808 plate before the normalisation step in Figure 2.14a, while in Figure 2.14b, I processed

WG3000808 with the other plates in the normalisation step before splitting the two replicate sets in the final plot.


(a) WG3000808 dropped

(b) WG3000808 included

**Figure 2.14** Comparison of difference in methylation across technical replicates. (a) Comparison of non-WG3000808 replicates, where WG3000808 was excluded before normalisation. (b) Comparison including WG3000808 replicates, where WG3000808 was kept during the normalisation step. Replicates facetted according to if one replicate pair was on WG3000808.

To generate the final dataset, I jointly processed the entire dataset using the Illumina normalisation method, so that complete dataset is available for future analyses. This joint processing included the WG3000808 plate, as it makes no difference for Illumina normalisation because the "reference" control sample was not on WG3000808. I used the default minfi settings (bg.correct = TRUE, normalize = controls, reference = 1), which involves a two step procedure. Step 1 normalises across samples using the control probes: (1.1) calculate the average red and green signal intensity separately across control probes for each sample, (1.2) average the red and green per sample signal intensity, (1.3) select one sample as a reference sample (by default, the function selects the 1st sample, which for FUSION is A12001—a high quality sample not flagged in any QC steps or in PCs and also not on the WG3000808 plate), (1.4) scale the red and green signal intensity across all probes by a scaling factor such that the average green and red signal intensity of control probes across samples and color channels is identical to the reference sample (i.e., for the control probes the red signal is equivalent to the green signal and is constant across all samples). Step 2 normalises the background signal within each sample: (2.1) for each sample, sort the 411 negative EPIC control probes in increasing order, selecting the 31st probe (i.e., select the probe with the 31st lowest intensity), and (2.2) subtract signal of this probe from all of the probes for each sample when the signal goes negative, set the value to 0.

### 2.7.7 Tissue specific sample filters and quality control

For the analyses in this section, I used Illumina normalised M-values, as opposed to beta values because M-values are more homoscedastic [87] and are therefore better suited for standard statistical models that assume homoscedastic data (constant variance across the data independent of the mean). In order to minimize noise, I removed blacklist probes.

From the remaining samples, I identified samples with abnormal methylation distributions by calculating percentiles across probe types per sample and comparing this to the median distribution per tissue using the Kolmogorov-Smirnov test (ks.test function in R). I identified and dropped 3 samples (2 adipose and 1 muscle) with p-value < 0.01. The samples that failed these filters show a significant deviation from the methylation distribution of samples that passed technical QC filters (Figure 2.15, Figure 2.16, Figure 2.17).

I performed PCA across all samples and found the first two PCs separated tissue types. In order to have a larger representation of tissue types, I included samples from the WG3000808 plate, which expanded our total tissues to blood, islet, adipose, and muscle. Based on these clusters, I dropped 6 samples (1 islet, 2 adipose, and 3 muscle) that clustered with an unexpected tissue (Figure 2.18).

To further remove potential outliers that would affect later analyses, I calculated PCs on a per tissue basis after removing sample plate, sentrix position, plate position, age, and sex effects from the Illumina normalised M-values values and dropping sex chromosome probes. For each tissue, I selected the minimum number of PCs to explain 20% of the variance. I transformed the PCs to z-scores and dropped samples where |z-score| > 5, removing one muscle (outlier in PC2) and one adipose (outlier in PC5) sample. After a first pass of sample exclusions, I repeated this process and found no additional strong outliers (Figure 2.19, Figure 2.20, Figure 2.21).

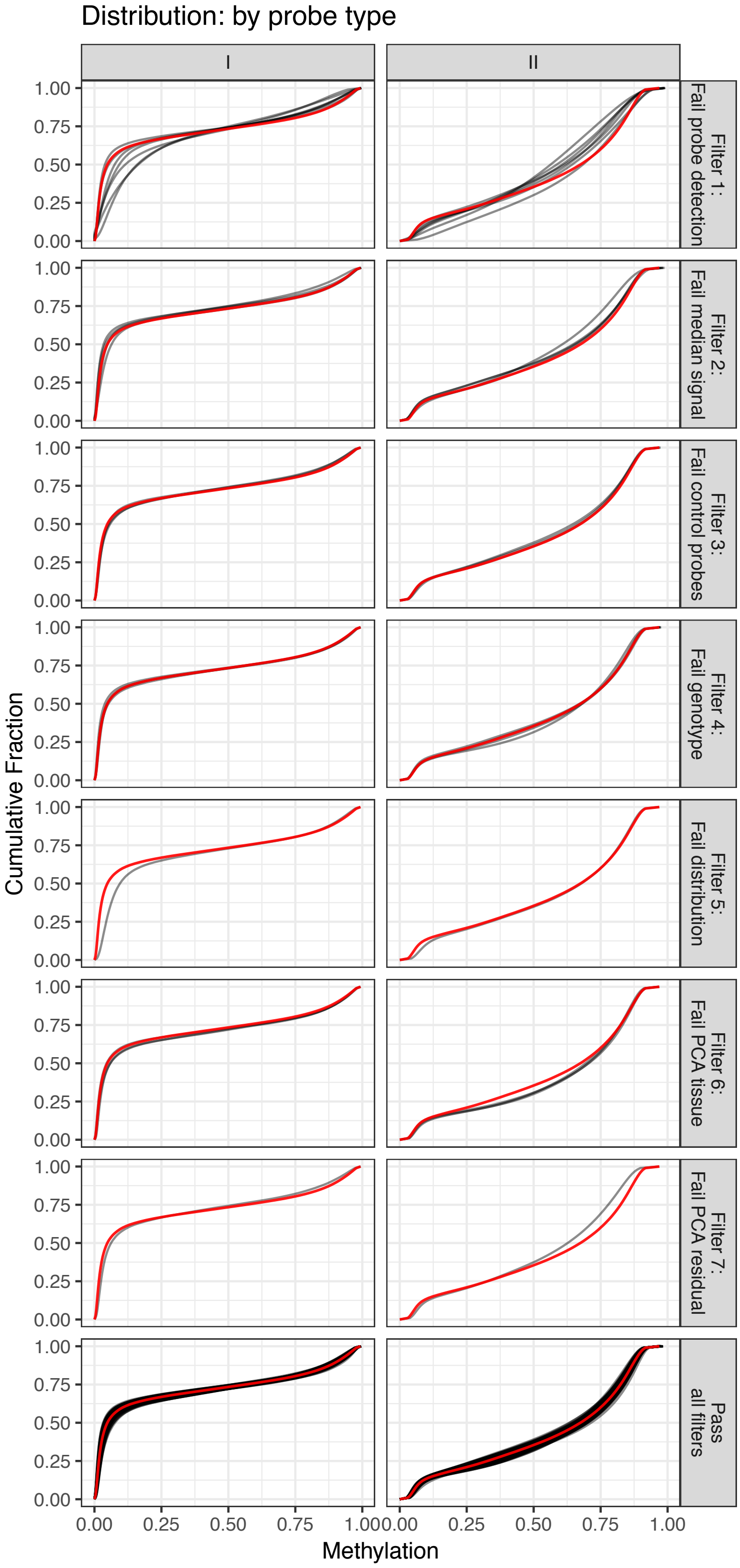


**Figure 2.15** Muscle QC summary: methylation distribution. Distribution of Illumina normalised beta values faceted by probe type and QC status of sample. Red line depicts the median across passed samples.

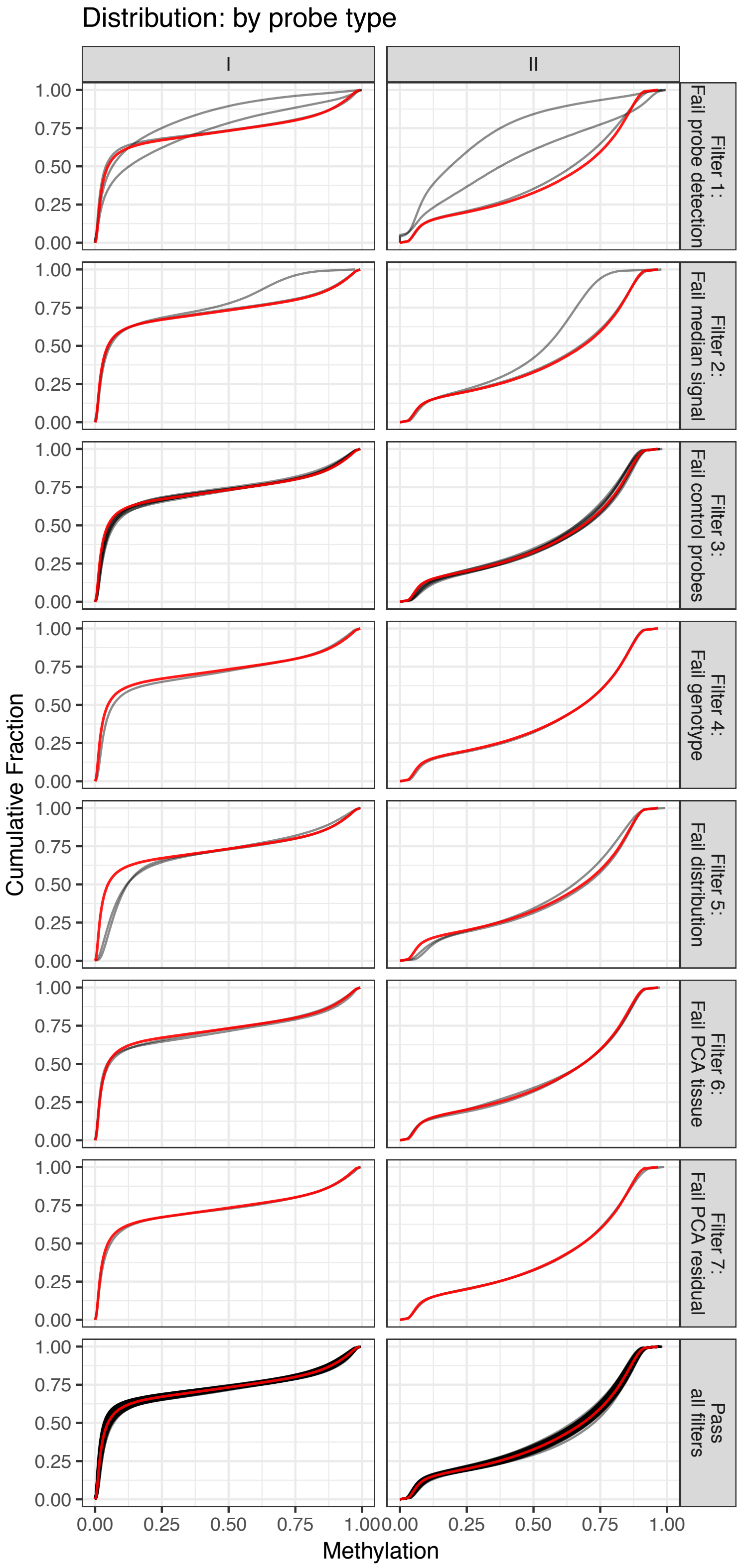


**Figure 2.16** Adipose QC summary: methylation distribution. Distribution of Illumina normalised beta values faceted by probe type and QC status of sample. Red line depicts the median across passed samples.

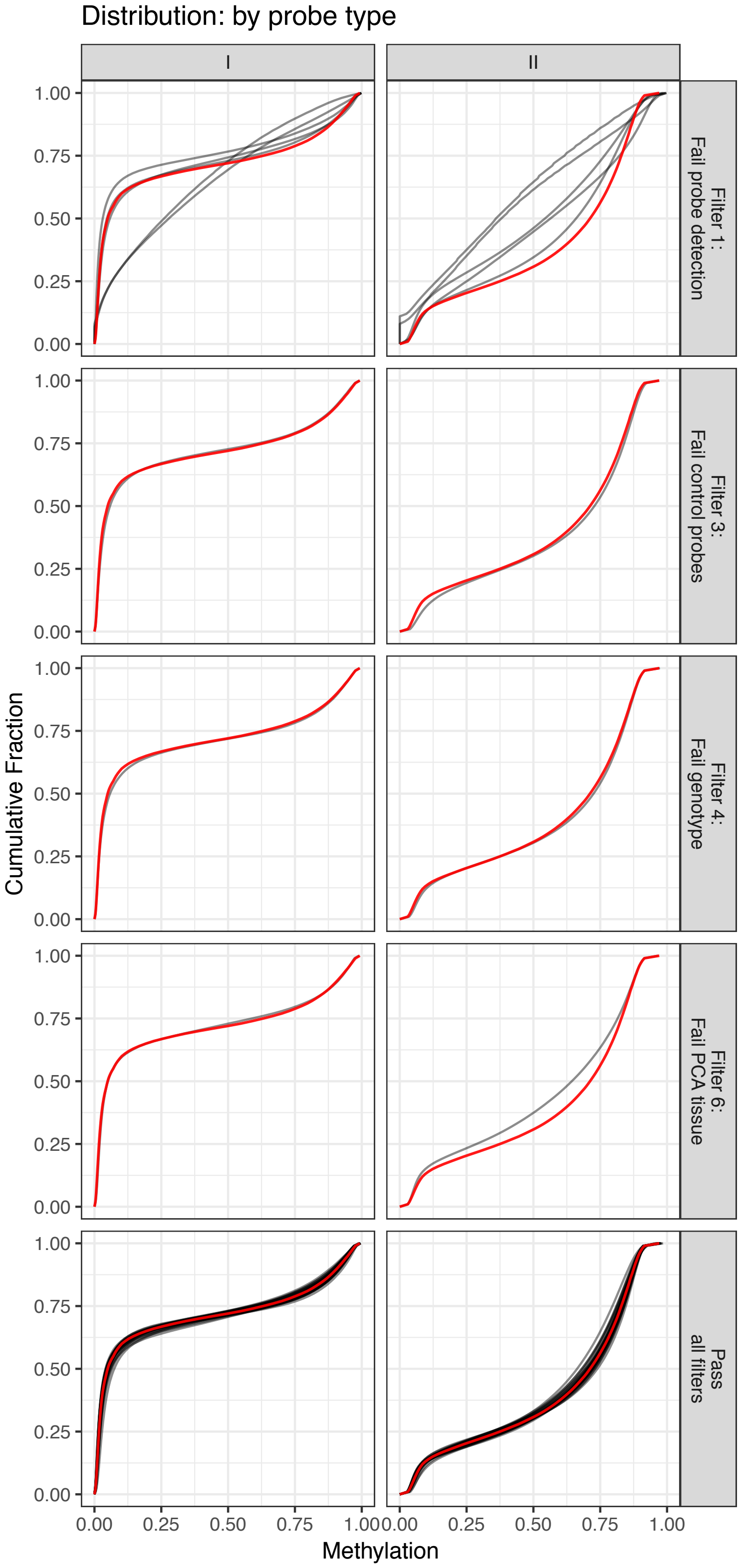


**Figure 2.17** Islet QC summary: methylation distribution. Distribution of Illumina normalised beta values faceted by probe type and QC status of sample. Red line depicts the median across passed samples.

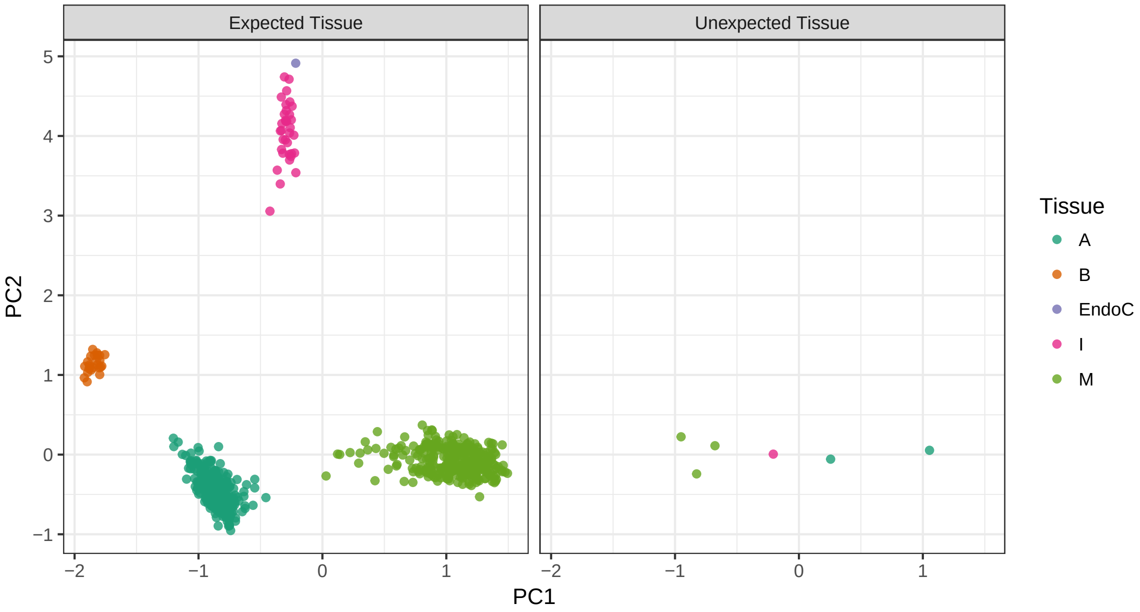
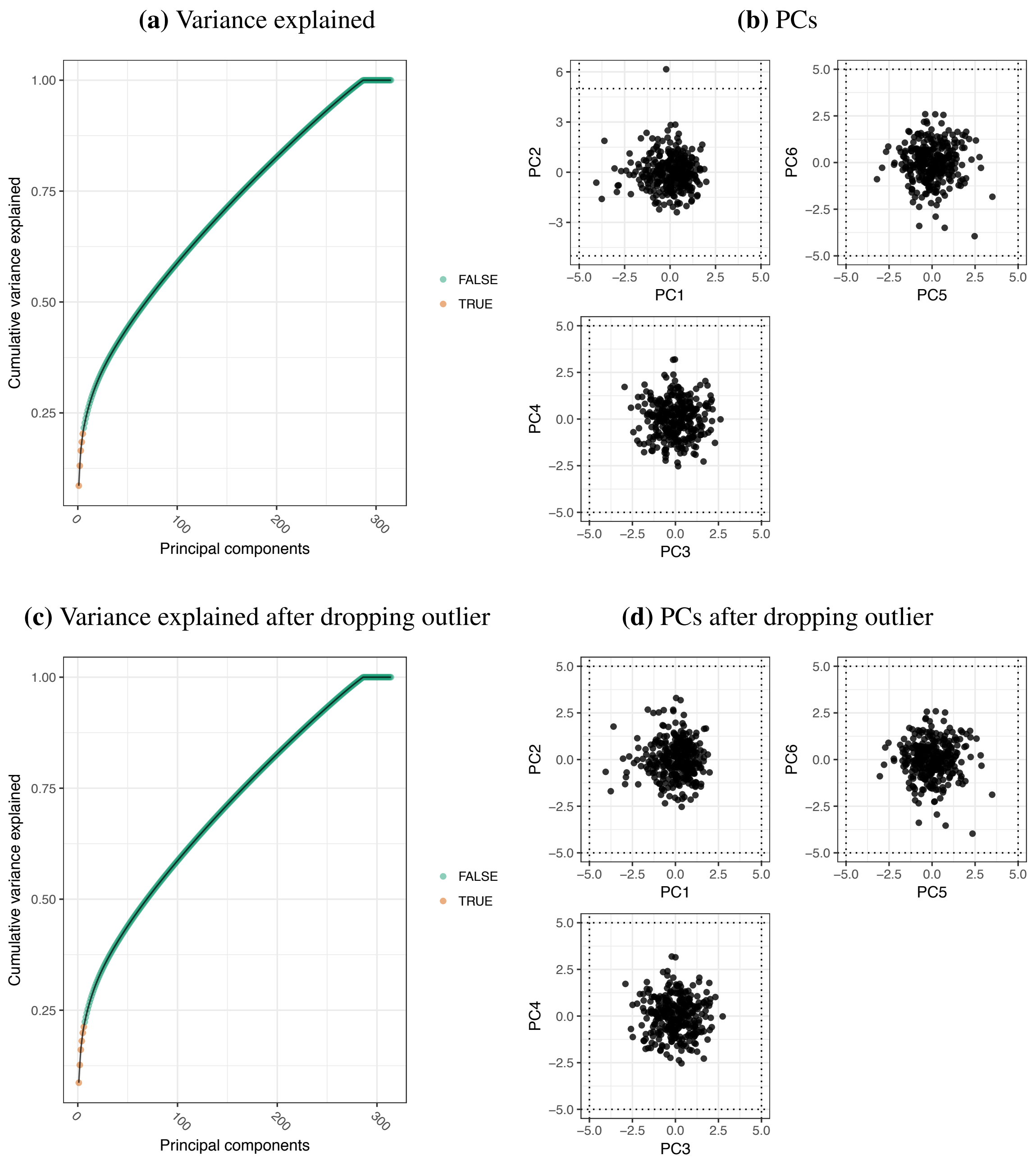


**Figure 2.18** Multi-tissue methylation PCA. Samples that clustered with unexpected tissues were removed from the analysis.



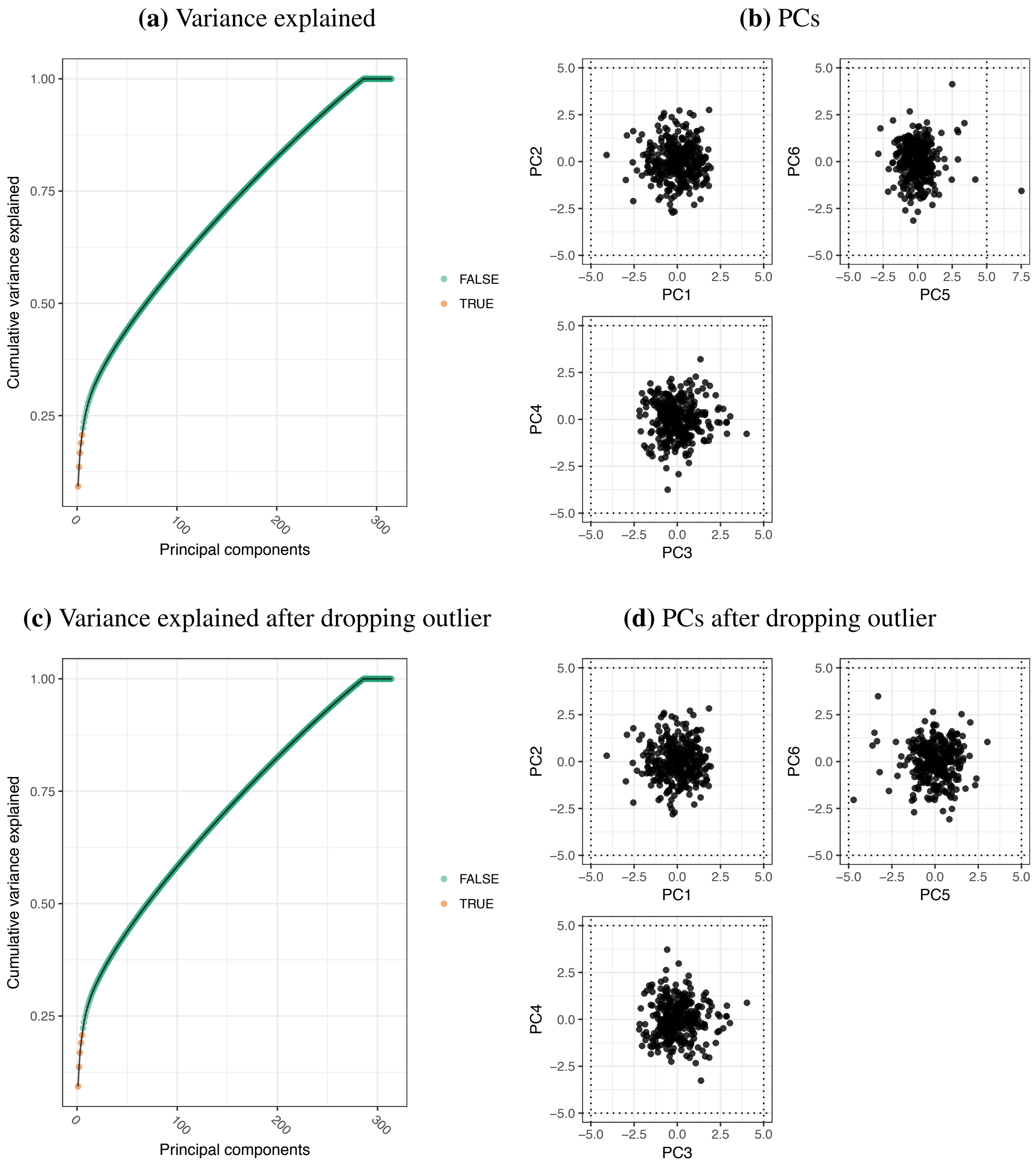
**(a)** Variance explained

**(b)** PCs

**(c)** Variance explained after dropping outlier

**(d)** PCs after dropping outlier

**Figure 2.19** Muscle methylation PCA outliers. (a) Cumulative variance explained by each PC. Orange PCs are the minimum number of PCs to explain 20% of the variance. (b) PCs that explain 20% of the variance, transformed into z-scores. Dashed lines indicate cutoffs. (c) Cumulative variance explained after dropping outlier samples. (d) PCs after dropping outlier samples.



**(a)** Variance explained

**(b)** PCs

**(c)** Variance explained after dropping outlier

**(d)** PCs after dropping outlier

**Figure 2.20** Adipose methylation PCA outliers. (a) Cumulative variance explained by each PC. Orange PCs are the minimum number of PCs to explain 20% of the variance. (b) PCs that explain 20% of the variance, transformed into z-scores. Dashed lines indicate cutoffs. (c) Cumulative variance explained after dropping outlier samples. (d) PCs after dropping outlier samples.

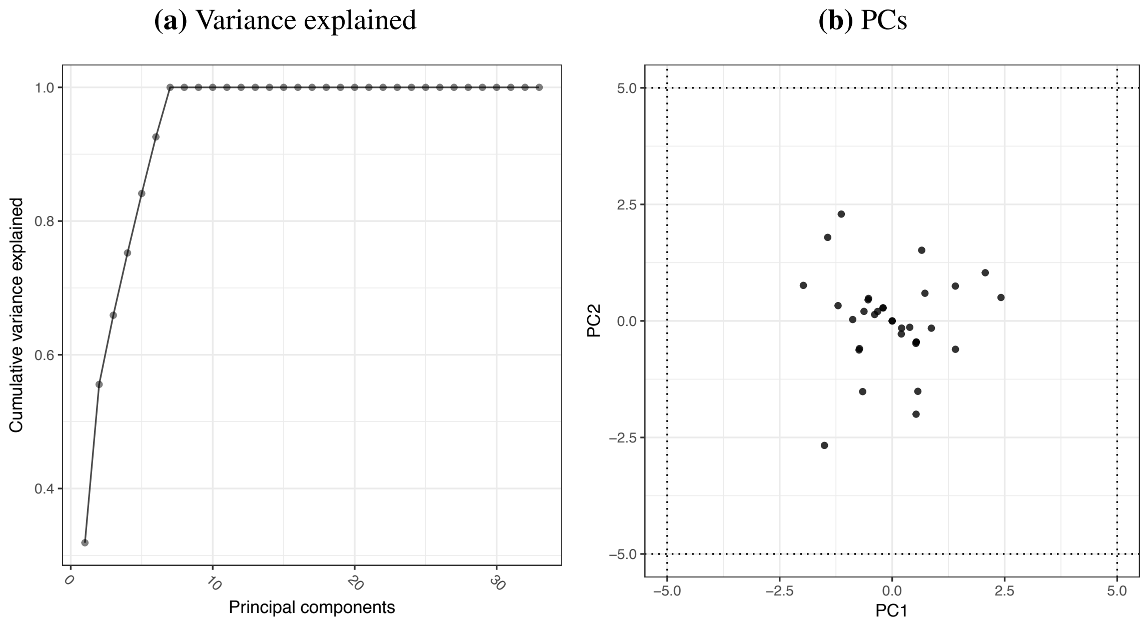


**Figure 2.21** Islet methylation PCA outliers. (a) Cumulative variance explained by each PC. Orange PCs are the minimum number of PCs to explain 20% of the variance. (b) PCs that explain 20% of the variance, transformed into z-scores. Dashed lines indicate cutoffs.

Finally, in order to choose between technical replicates, I selected the sample with the largest number of detected probes (p-value $\leq 0.05$). There were 4 replicate pairs not on the WG3000808 plate. Of those pairs, one pair failed QC and one pair came from a non-Finnish participant, leaving 2 remaining pairs. In addition to the adipose replicate, I dropped the muscle sample from the non-Finnish participant. Finally, I dropped the adipose and muscle samples from two participants that had a first degree relative (removed 4 total samples). In total, after excluding WG3000808 samples and the other filters described above, there were 31 islet, EndoC, 276 adipose, and 282 muscle samples that passed all QC filters (Table 2.3).

### 2.7.8 Methylation specificity index

Similar to the gene expression specificity index described earlier, John Didion (NIH) generated a methylation specificity score (MeSS). This measure quantifies the tissue specificity of the methylation patterns such that a high specificity score indicates instances where other tissues are highly methylated and the target tissue is unmethylated, or vice versa. I helped select the input tissue reference panel and performed a validation experiment using gene expression. The input data sources and analysis pipeline are described below.

As a methylation reference panel, we used 21 human tissues (Table 2.6) from the Roadmap Epigenomics project [316] that were processed using a standardised pipeline as part of the MethBase repository [361]. Because of variable and sometimes low genome coverage across tissues in the reference panel, we used BSmooth v1.8.2 [145] which utilizes local likelihood smoothing techniques to smooth DNAme values and thereby increase the precision of low coverage WGBS. As parameters for BSmooth, we set a minimum window size of 500 bp and 50 CpGs, and maximum gap between consecutive CpGs of 1000 bp.

To calculate methylation specificity, we used a method recently developed by Liu et al. [226]. Briefly, using the raw methylation Beta values (*rm*), we calculate a weighted mean (*TB*) for each CpG (*r*) across 21 cell/tissue types (*s*) using a one-step Tukey biweight:

$$TB_r = \frac{\sum_{s=1}^{21} w_{r,s} rm_{r,s}}{\sum_{s=1}^{21} w_{r,s}} \tag{2.7}$$

where *w* is a weight parameter calculated as described in Zhang et al. [443]. Beta values are transformed by taking absolute difference from the mean, setting the minimum allowed absolute difference to 3:

$$m_{r,s} = max(|rm_{r,s} - TB_r|, 3) \tag{2.8}$$

The transformed beta values (*m*) are converted to probability (*p*) by dividing the absolute difference from the mean by the sum of absolute differences across all cell/tissue types:

$$p_{r,s} = \frac{m_{r,s}}{\sum_{s=1}^{21} m_{r,s}} \tag{2.9}$$

Using these probabilities, we calculate Shannon entropy (*H*), noting that the choice of the logarithm base has no effect on the relative relationships of specificity values, so we follow the practice of Liu et al. [226]:

$$H_r = -\sum_{s=1}^{21} p_{r,s} log_{21}(p_{r,s}) \tag{2.10}$$

We weight entropy (*HW*) based on the observed range of beta values:

$$HW_r = H_r \times \frac{1 - D_r}{100} \tag{2.11}$$

where *D* is the difference between the minimum and maximum methylation values at CpG *r*. Finally, the weighted entropy is converted to a Methylation Specificity Score (MeSS) following the methodology described earlier for gene expression (Section 2.6.3).

$$Q_{r,muscle} = H_r - log_{21}(p_{r,muscle}) \tag{2.12}$$

$$MeSS_r = 1 - \frac{Q_{r,muscle}}{max(Q_{muscle})} \tag{2.13}$$

Similar to the gene expression specificity index (Section 2.6.3), the methylation specificity score is dependent on the reference panel used. To date, the most comprehensive multi-tissue panel is WGBS data from the Roadmap Epigenomics project; however, as sequencing costs decrease and additional WGBS datasets are generated (possibly at the level of single cells), it will be possible to obtain more accurate cell/tissue type methylation specificity scores.

Using the final expression-methylation associations (expression quantitative trait methylation; eQTMs) described in Chapter 3, I validated the muscle methylation specificity score by comparing the muscle specificity indices of associated expression and methylation sites. I found that for nearby eQTMs, the gene expression and methylation specificity scores were correlated and that the correlation diminished according to the eQTM distance (Figure 2.22).

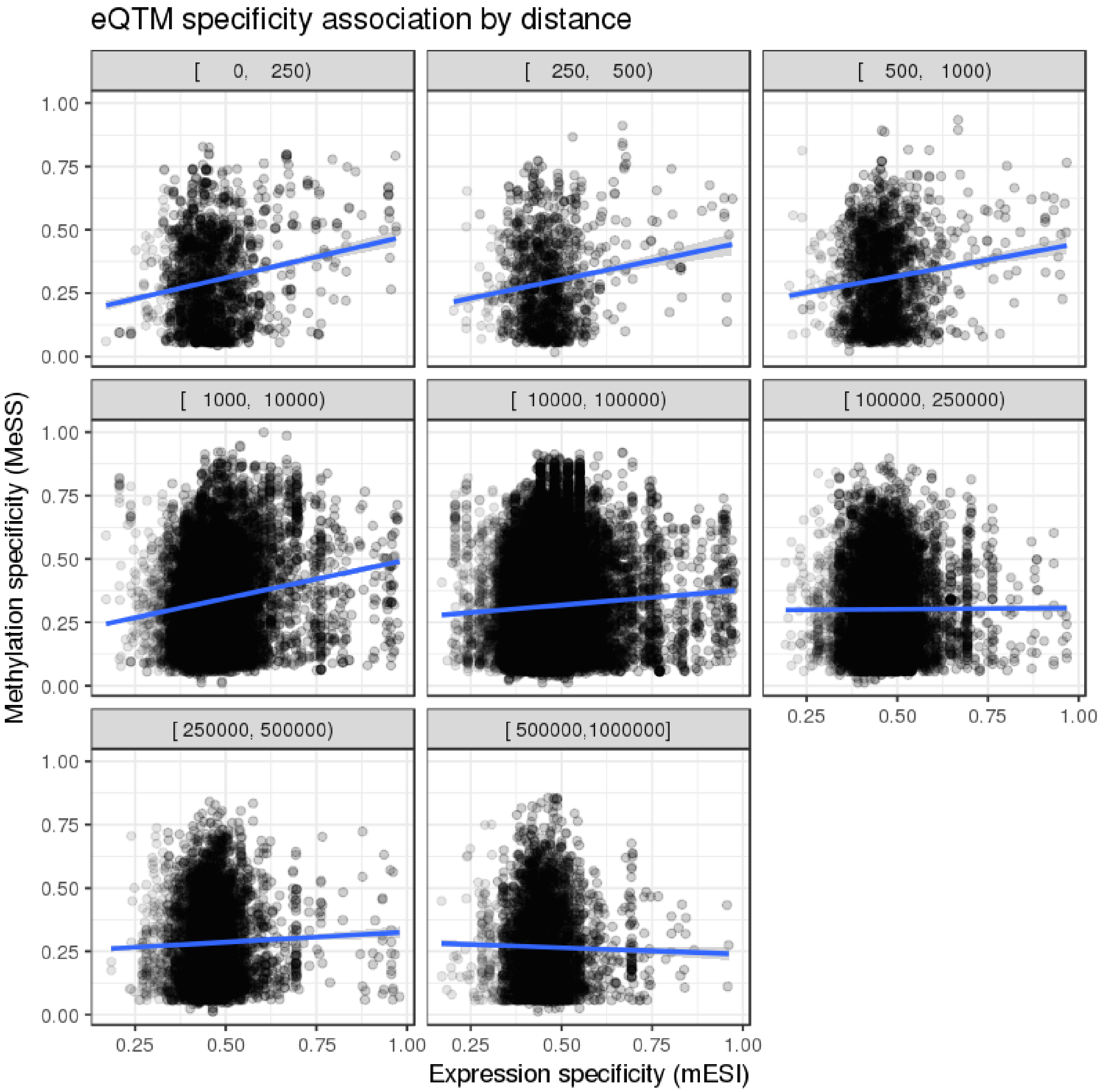


**Figure 2.22** Gene-methylation specificity index correlation. Correlation facetted by eQTM distance.

## 2.8 Intermediate data freeze note

The preliminary dataset published in Scott et al. [341] was used for Chapter 6 of this thesis. Since that publication, we generated RNA-seq and genotype information on additional samples and adjusted the analysis pipeline. Here, I describe the major differences between Scott et al. [341] and the analysis described in this thesis chapter.

In Scott et al. [341], we analysed gene expression from 271 skeletal muscle biopsies, 267 of which had array genotype information. Since then, we generated RNA-seq data from additional skeletal muscle biopsies from the same study, for a total of 301 samples after QC—all genotyped on arrays. Previously, we dropped one library that was a slight outlier in insert size. After evaluating this library in the context of the full dataset, we decided to reinstate it. In addition, we previously quantified gene expression as fragments per kilobase per million reads (FPKM), whereas in the revised pipeline we use TPM, which is better measurement of relative molar RNA concentration [405, 211]. We also now consider genes of all biotypes after count filters, instead of only protein_coding, pseudogene, antisense, lincRNA, processed_transcript, sense_intronic, sense_overlapping gene biotypes from Scott et al. [341] Additionally, we use GTEx as a reference panel for tissue deconvolution, as opposed to the Illumina Human Body Map 2.0 dataset, which has the advantage of using the average gene expression over many samples to generate tissue signatures, rather than a single biopsy. Finally, we use a different genotype imputation pipeline, capitalising on the newly developed Michigan Imputation Server [74].

## 2.9 Additional data sources

In addition to the data described above, I use other resources throughout this thesis. The first resource is genome wide chromatin state maps generated by collaborators at the University of Michigan [395]. These chromatin states were learned jointly by applying the ChromHMM (v1.10) algorithm [92, 94, 93] at 200 bp resolution to six data tracks (Input, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3) from publicly available chromatin immunoprecipitation followed by sequencing (ChIP-seq) data [94, 287, 316, 253]. Descriptions of each cell/tissue type and the source study can be found in Table A.1. I do not include these abbreviations in the Nomenclature of this thesis. As is described in Varshney et al. [395], my collaborators selected a 13-state model and mapped the biological function names to match the Roadmap Epigenomics “extended” 18-state model [316]. In addition, to identify open chromatin regions I used previously published skeletal muscle [341], islet [395], adipose [6], and GM12878 [46] ATAC-seq data. All ATAC-seq data was uniformly processed as described in Scott et al. [341], using the same read trimming, alignment, filtering and peak calling pipeline.

| Samples | Description |
|---|---|
| 2 | One sample from each of 2 first degree relative pairs (drop NGT and keep IGT in both cases) |
| 1 | Non-Finnish participant |
| *328 / 331* | *Total samples passed / total samples submitted* |

**Table 2.1** Genotype QC summary of all FUSION biopsy samples.

| Samples | Description |
|---|---|
| 1 | Contaminated with a different sample ~8% |
| 7 | Outlier in transcriptional diversity |
| 4 | Extreme 3' bias in gene body coverage |
| 2 | Genotype QC: drop one of 2 pairs of first degree relative |
| 1 | Genotype QC: drop non-Finnish participant |
| 7 | One sample from each of 7 replicate pairs |
| *301 / 323* | *Total samples passed / total samples submitted* |

**Table 2.2** RNA-seq QC summary of skeletal muscle biopsies.

| All Samples | Muscle Samples | Description |
|---|---|---|
| 13 | 5 | Failed low quality probe filter |
| 3 | 1 | Outlier in median methylated and unmethylated plot |
| 12 | 3 | Outlier in control probes |
| 6 | 4 | No clear genotype match |
| 3 | 1 | Outlier in methylation distribution |
| 6 | 3 | Clustered outside of expected tissue in PCA |
| 2 | 1 | Failed residual PCA filter |
| 4 | 2 | Genotype QC: drop one of 2 pairs of first degree relative |
| 3 | 1 | Genotype QC: drop non-Finnish participant (also happened to be an adipose sample replicate) |
| 2 | 0 | One sample from each of 2 replicate pairs that passed all previous steps |
| *590 / 644* | *282 / 303* | *Total samples passed / total samples submitted* |

**Table 2.3** DNAme QC summary. WG3000808 plate not included.

| Samples | Description |
|---|---|
| 301 | Muscle RNA-seq after QC |
| 282 | Muscle EPIC DNAme after QC |
| 265 | Common samples between RNA-seq and DNAme |
| 318 | Union of all samples with RNA-seq and DNAme |

**Table 2.4** Summary of molecular trait overlaps in skeletal muscle biopsies.

| | RNA-seq Samples | DNAme Samples | Common RNA-seq & DNAme Samples | Union RNA-seq & DNAme Samples |
|---|---|---|---|---|
| N | 301 | 282 | 265 | 318 |
| Sex=M(%) | 174 (57.8) | 159 (56.4) | 152 (57.4) | 181 (56.9) |
| Age (mean (sd)) | 59.91 (7.66) | 59.98 (7.92) | 59.87 (7.72) | 60.01 (7.84) |
| BMI (kg/m$^2$; mean (sd)) | 27.45 (4.13) | 27.61 (4.27) | 27.45 (4.14) | 27.59 (4.25) |
| Fasting High-density Lipoprotein (mmol/l; mean (sd)) | 1.44 (0.36) | 1.44 (0.34) | 1.43 (0.35) | 1.44 (0.35) |
| Fasting Low-density Lipoprotein (mmol/l; mean (sd)) | 3.41 (0.89) | 3.39 (0.88) | 3.40 (0.89) | 3.40 (0.88) |
| Fasting Triglycerides (mmol/l; mean (sd)) | 1.42 (0.90) | 1.44 (0.93) | 1.45 (0.94) | 1.42 (0.89) |
| Fasting Total Cholesterol (mmol/l; mean (sd)) | 5.49 (1.03) | 5.48 (1.03) | 5.50 (1.02) | 5.48 (1.03) |
| Systolic Blood Pressure (mmHg; mean (sd)) | 135.15 (16.09) | 135.59 (16.43) | 135.09 (15.89) | 135.60 (16.56) |
| Diastolic Blood Pressure (mmHg; mean (sd)) | 83.22 (9.19) | 83.42 (9.56) | 83.25 (9.23) | 83.37 (9.49) |
| Fasting Serum Insulin (mU/l; mean (sd)) | 8.59 (5.20) | 8.83 (5.43) | 8.66 (5.31) | 8.74 (5.32) |
| Fasting Serum C-peptide (pmol/l; mean (sd)) | 703.78 (282.20) | 714.44 (294.21) | 703.85 (280.78) | 713.18 (294.03) |
| Fasting Plasma Glucose (mmol/l; mean (sd)) | 6.27 (0.78) | 6.26 (0.78) | 6.27 (0.79) | 6.26 (0.78) |
| Ever Smoker = Y (%) | 43 (14.3) | 40 (14.2) | 39 (14.7) | 44 (13.8) |
| Antihypertensive = Y (%) | 87 (28.9) | 82 (29.1) | 76 (28.7) | 93 (29.2) |
| Statin = Y (%) | 52 (17.3) | 46 (16.3) | 45 (17.0) | 53 (16.7) |
| Synthetic Thyroid Hormone = Y (%) | 21 (7.0) | 21 (7.4) | 19 (7.2) | 23 (7.2) |
| Oral Glucose Tolerance Test Status (%) | | | | |
| Normal Glucose Tolerance (NGT) | 108 (35.9) | 98 (34.8) | 90 (34.0) | 116 (36.5) |
| Impaired Fasting Glucose (IFG) | 43 (14.3) | 39 (13.8) | 38 (14.3) | 44 (13.8) |
| Impaired Glucose Tolerance (IGT) | 73 (24.3) | 73 (25.9) | 71 (26.8) | 75 (23.6) |
| Type 2 Diabetes (T2D) | 77 (25.6) | 72 (25.5) | 66 (24.9) | 83 (26.1) |

**Table 2.5** Clinical traits. Phenotypic characterisation of participants in the FUSION tissue biopsy study.

| Tissue | Mean coverage | Median coverage |
|---|---|---|
| Adipose | 129.04 | 133 |
| Adrenal gland | 72.65 | 70 |
| Bladder | 55.07 | 54 |
| Blood (Macrophages) | 36.43 | 30 |
| Blood (Natural Killer cells) | 26.97 | 24 |
| Blood (T cells) | 34.39 | 28 |
| ES-derived Ectoderm | 86.76 | 79 |
| ES-derived Endoderm | 37.77 | 36 |
| ES-derived Mesoderm | 64.82 | 61 |
| Oesophagus | 69.95 | 72 |
| Gastric | 23.45 | 23 |
| Heart (Left ventricle) | 109.57 | 111 |
| Liver | 50.1 | 51 |
| Lung | 79.260 | 80 |
| Ovary | 51.45 | 51 |
| Placenta | 27.47 | 25 |
| Psoas muscle | 42.24 | 40 |
| Sigmoid colon | 118.2 | 122 |
| Small intestine | 109.4 | 108 |
| Spleen | 99.44 | 100 |
| Thymus | 72.05 | 75 |

**Table 2.6** Methylation specificity score (MeSS) reference panel. Psoas muscle was used to calculate muscle specific methylation patterns.

# Chapter 3

# Relationships between molecular traits

## 3.1 Introduction

The relationship between gene expression and DNAme has been of great scientific interest since DNAme was proposed as a mechanism of cell memory in 1975 [312, 155]. These initial studies established the notion that DNAme is a repressive mark (i.e., negatively associated with expression); however, more recent studies have shown DNAme can be positively or negatively associated with gene expression (reviewed in [173] and Section 1.4.2.3). Collectively, these findings paint a complex and poorly understood relationship between expression and DNAme that varies greatly based on the genomic context of the DNAme site (e.g., location in relationship to a gene body, CGI, or surrounding histone modifications).

In this chapter, I build on these studies by (1) mapping associations between gene expression and DNAme in skeletal muscle and (2) analysing the relationship between expression and DNAme (i.e., positive or negative) across a variety of genomic contexts (based on overlaps with DNAme sites) including CGIs, muscle chromatin states, muscle ATAC-seq peaks, as well as the distance of the DNAme site to the gene TSS. Drawing from Gutierrez-Arcelus et al. [141], I call these associations expression quantitative trait methylation (eQTM). To date, this is one of the first studies considering eQTMs using the EPIC array, which assays a greater portion of intergenic genomic regions, such as enhancers [261], than the previous 450k arrays that focused on genic regions like promoters [333, 33]. The use of the EPIC array makes this study one of the most comprehensive charts of the regulatory network of gene expression and DNAme to date.

In addition, prior to eQTM mapping and characterisation, I describe how I analysed and largely removed possible confounding effects through latent factor analysis. I show that a potential confounder for eQTM studies is variable tissue heterogeneity across samples and that by learning latent factors, it is possible to control for this correlation structure.

As a final note, I called eQTMs both at the level of gene expression and at the level of exon expression. To distinguish between the two I refer to gene level associations as eQTMs and exon level associations as exQTMs. However, for the bulk of this chapter, I focus on eQTMs to avoid redundancy as the eQTM trends hold true for exQTMs.

## 3.2 Controlling for unwanted variation in molecular traits

Before carrying out the analysis of any dataset, it is important to understand sources of variation, as accounting and correcting for unwanted variation boosts power and safeguards against drawing incorrect conclusions due to confounding signals. Many studies document potentially hidden (i.e., latent) sources of variation in high throughput molecular technologies (reviewed in [207]). These sources of variation can be (1) technical or (2) biological in origin and are often masked by some level of background, stochastic noise. Technical artefacts may be measured directly, indirectly, or in some instances may be completely unknown. For instance, ’t Hoen et al. [381] describe strong batch effects approximated by the specific sequencing laboratory site in RNA-seq data, even though special efforts were made to use identical protocols across laboratories. This is an example of an indirectly measured technical artefact, where sequencing site functions as a surrogate for underlying differences (e.g., library insert size). Likewise, biological mechanisms, both known and unknown, contribute to variability. Obviously the identification of certain biological signals is the goal of a study, but potential confounding biological effects must also be considered. For instance, using publicly available methylation data drawn from several studies, Jaffe and Irizarry [167] demonstrate that differences in blood cell type composition are a key, unmeasured confounder of age associated DNAme signals in whole blood. For the original studies that Jaffe and Irizarry re-analysed, cell type heterogeneity represents an unknown biological source of variability that had the potential to confound the detection of biological signal of interest—age associated DNAme. Finally, biological systems are inherently noisy and there will always be some degree of random variation or noise in high throughput molecular measurements.

The goal of latent factor analysis is to identify and control for unknown and unwanted technical or biological sources of variation, while preserving the signals of interest. These techniques [206, 366, 367, 315, 307], take advantage of the fact that high throughput molecular technologies assay many features (e.g., genes or DNAme sites) and use these many observations to derive factors that affect many features. However, these techniques cannot in themselves distinguish the type of confounding—from technical to biological—and therefore careful thought must be given to the particular method employed. For instance, when analysing proximal genetic effects on molecular traits, the goal of the study is to identify genetic effectors that lie near their gene targets in genomic space, which inherently means the signals of interest will not be widespread and will not affect a large number of features. Rather, the signal of interest should be limited to a handful of features. In such a scenario, it would be appropriate to use latent factor techniques liberally, since factor analysis techniques do not learn factors that only affect a few features. Such insight informs the design and use of techniques like PEER [366, 367].

On the other hand, if one is seeking to detect distal genetic effects, which are assumed to function by perturbing a *trans* acting factor that regulates many other genes, strategies that identify and remove many latent factors would not be appropriate. Instead, alternative approaches that limit the aggressive identification of latent factors should be taken, such as GNet-LMM [307]

As a final example, consider the case of a differential expression analysis or differential methylation analysis where the outcome signal of interest is known beforehand and may be associated with other measured features, such as a participant's BMI. In such cases, since the outcome of interest is known beforehand, it makes sense to use the outcome signal of interest to ensure that the signal of interest is not washed away when identifying latent factors. Such is the approach used by SVA [206].

### 3.2.1 PEER for latent factor analysis

To account for known and unknown technical and biological effects for all subsequent molecular trait analyses (eQTL, exQTL, mQTL, eQTM, exQTM), I used PEER v1.0 [366, 367]. For both gene expression and DNAme, I excluded sex chromosome features as in previous mQTL studies [19, 406, 144]. Second, I transformed the gene expression (TPM), exon expression (FPKM), and methylation (M-values) signal across all samples that passed QC for each molecular trait (301 samples for gene expression and 282 samples for DNAme) using rank-based inverse normalisation by ranking the data and then fitting it to a normal distribution. Finally, I ran PEER with up to 110 latent factors and inverse normalised the PEER residuals.

When running PEER, I included age, sex, OGTT status, and the top 4 genotype PCs as covariates. These variables are standard, known sources of variation that could confound analysis. In addition, I included OGTT status because I wanted to identify general QTLs and QTMs that are not dependent on a particular disease context. Therefore, I needed to account for the intentionally biased design of the FUSION study to sample each OGTT state, even though we do not find many T2D associated signals in gene expression [341]. Finally, I included additional measurements to capture known sources of technical variation. For gene and exon expression, I included RIN and sequencing batch, two established sources of bias in RNA-seq data [381, 115, 343]. For DNAme, I included sample plate, sentrix position, and plate position as covariates, to account for well known batch and spatial effects of Illumina methylation arrays [79].

### 3.2.2 Correlation between gene expression and methylation factors

As described previously, PEER does not have knowledge of technical or biological sources of variation. If the effect of a particular source of variation (e.g., insulin response or cell type heterogeneity) was extremely strong, PEER would likely identify factors that correlate to that source of variation. Alternatively, in the cases of weaker effects, PEER may identify factors that describe a combination of effects that may be technical or biological in nature. In order to understand potential unknown biological sources of variation, I compared the PEER factors discovered separately in gene level expression and DNAme to each other and to other phenotypic traits of the FUSION tissue biopsy participants (see Table 2.5).

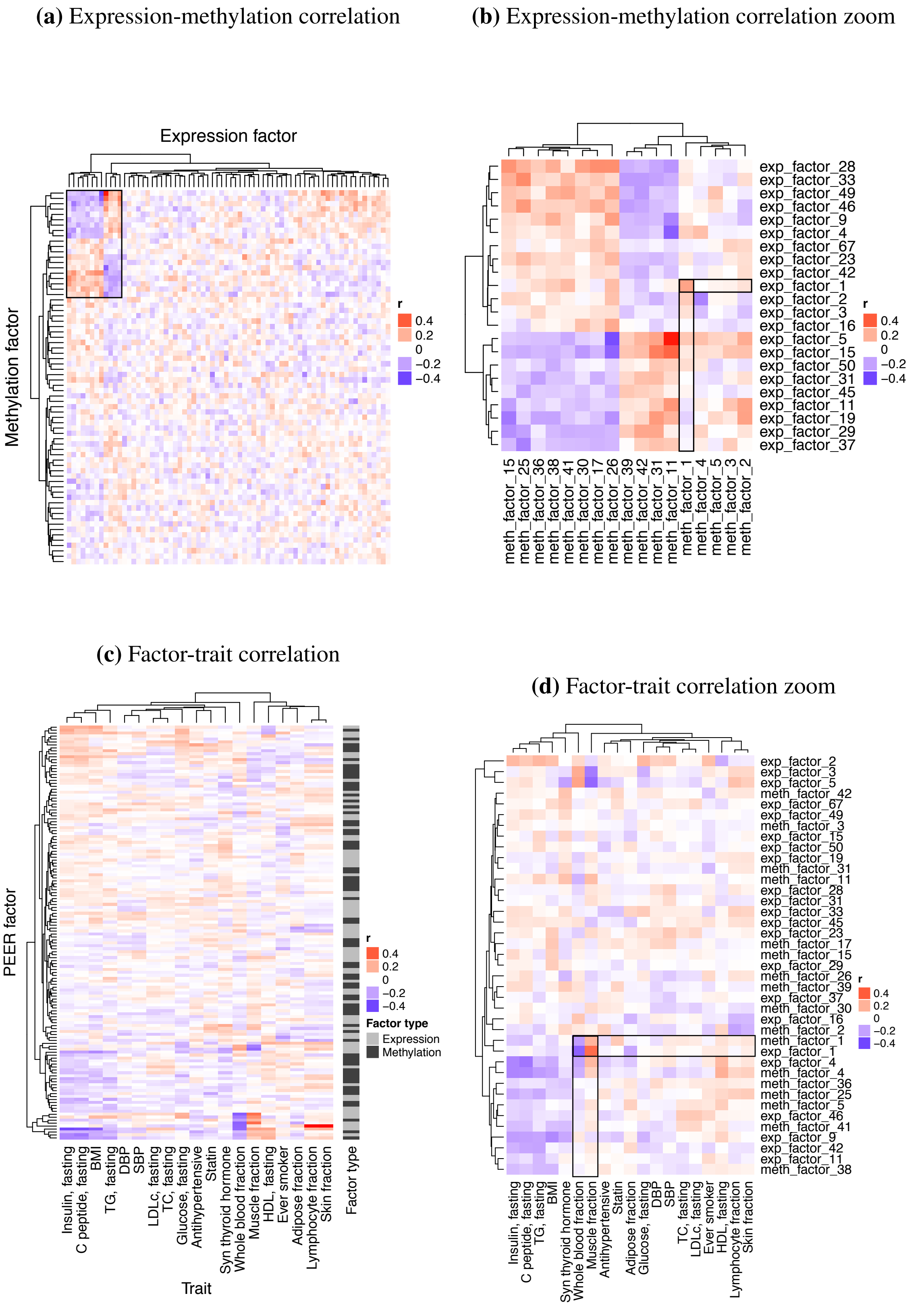
Because gene expression and DNAme are measured using two very different technologies, and represent two completely different biological phenomena, one would expect any correlation between PEER factors to relate to biological variation rather than technical variation. The only potential source of correlated technical variation could be biases that occurred during tissue collection and processing that are subsequently apparent in all assays that use the same tissue source. Comparing the first 70 PEER factors for gene expression and DNAme,[1] I find structured correlation, suggesting potential confounding effects (Figure 3.1).

Since this correlation structure is likely biological, I calculated the correlation between each PEER factor and phenotypic traits to identify potential biological sources of factors. Prior to associations, I removed sex and age effects from continuous traits by fitting a linear model, and inverse normalised the residuals. For both gene expression and DNAme, the first PEER factor showed some correlation with skeletal muscle fraction and whole blood tissue estimates (derived from gene expression). In addition, I found the first PEER factor from both expression and DNAme also clustered together and were most associated with each other compared to all other factors ($r = 0.24$). Together, these results suggest that a significant component of variation captured by the first PEER factor corresponds to tissue heterogeneity. This would not be surprising, as I found tissue heterogeneity constituted the first PCs in the methylation QC across an array of tissues (muscle, adipose, blood, and islets; Section 2.7.7).

I also find a collection of factors that are associated with BMI, insulin, C peptides, and triglycerides (TG). These results indicate a strong molecular response to such signals, which is not surprising as (1) these traits are correlated, and (2) insulin is a key signalling hormone that initialises a signal cascade in skeletal muscle to uptake circulating blood glucose. This signalling cascade involves the relocation of many transporter proteins, so the fact that the PEER factors suggest large changes in the transcriptome and methylome related to these traits is expected. Indeed, as described in Scott et al. [341], we found many genes associated with insulin and BMI, and preliminary DNAme results also suggest similar methylome trends.

Despite the association of PEER factors and tissue heterogeneity estimates, I decided not to include the tissue heterogeneity estimates as covariates because these estimates are only as good as the reference panel, which in this case was composed of a collection of heterogeneous tissues, rather than profiles of specific cell types. Instead, for further analyses, I made the assumption that PEER will properly account for tissue heterogeneity, which I believe is a reasonable assumption given (1) the correlation of PEER factors and tissue heterogeneity estimates, and (2) previous studies which find similar latent factor techniques perform fairly well at capturing tissue heterogeneity [206, 248]. Thus, while tissue heterogeneity estimates

[1] I analysed up to 70 PEER factors as I found 70 PEER factors to be the optimal number of PEER factors for eQTM mapping as described in Section 3.3.1 below.



**(a)** Expression-methylation correlation

**(b)** Expression-methylation correlation zoom

**(c)** Factor-trait correlation

**(d)** Factor-trait correlation zoom

**Figure 3.1** PEER factor correlations. (a) Association between 70 expression and 70 methylation factors. (b) Associations between expression and methylation factors zoomed in on key correlation structure in upper corner of panel a. (c) Association between all factors (70 expression, 70 methylation) and traits. (d) Association between factors from panel b and traits.

were a crucial QC step that gave me confidence in the "purity" of our skeletal muscle samples (all samples in analysis had > 90% skeletal muscle), I did not use them in further analyses.

## 3.3 eQTM mapping

I mapped eQTMs, associations between gene expression and DNAme, across 265 samples that passed QC using LIMIX v0.7.74 [222]. Prior to associations I used PEER to control for known (e.g., batch) and unknown confounding effects. Let $y_j$ be a vector of inverse normalised PEER expression residuals for gene or exon $j$ across individuals. I consider the following linear model to map eQTMs and exQTMs:

$$y_j = \underbrace{\alpha_j \mathbb{1}}_{\text{intercept}} + \underbrace{\beta_j m}_{\text{methylation eff.}} + \underbrace{\psi_j}_{\text{noise}}, \quad \psi_j \sim \mathcal{N}(0, \sigma_e^2 I) \tag{3.1}$$

where $\alpha_j$ is the intercept, $m$ denotes the inverse normalised PEER DNAme residuals of the probe in consideration across individuals, $\psi_j$ Gaussian noise, and $\beta_j$ the effect of methylation. I correct for the number of tests using Storey's FDR [369].

Previous eQTM studies use a wide range of window sizes, defined as the distance between the TSS (or exon) and the DNAme site, ranging from 50 kb [299, 141] to 1 Mb [59], as well as distances in between [406]. In order to evaluate the potential window sizes, I mapped eQTMs using all DNAme sites within 10 Mb from the TSS of the target gene. When only accounting for known covariates (and no PEER factors), I find a constant eQTM discovery rate at distances > 1 Mb (Figure 3.2). I reasoned this constant discovery rate was due to confounding correlation between gene expression and DNAme. To test this hypothesis, I randomly paired genes with DNAme sites on different chromosomes and calculated the discovery rate, finding it to be nearly identical to the eQTM discovery rate on the same chromosome at distances > 1 Mb. This result suggests there is indeed confounding correlation between gene expression and DNAme, which one may have predicted from the PEER factor association analysis (Section 3.2.2). Furthermore, drawing from the PEER factor association analysis, this correlation structure is likely due to a combination of tissue heterogeneity and perhaps insulin or BMI effects.

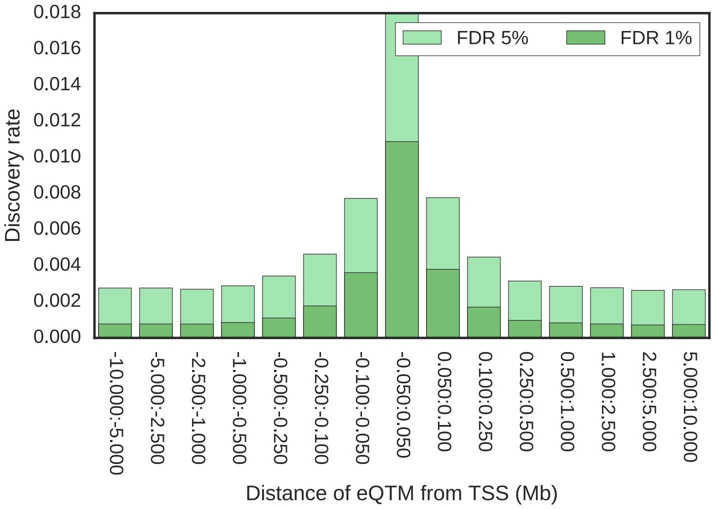
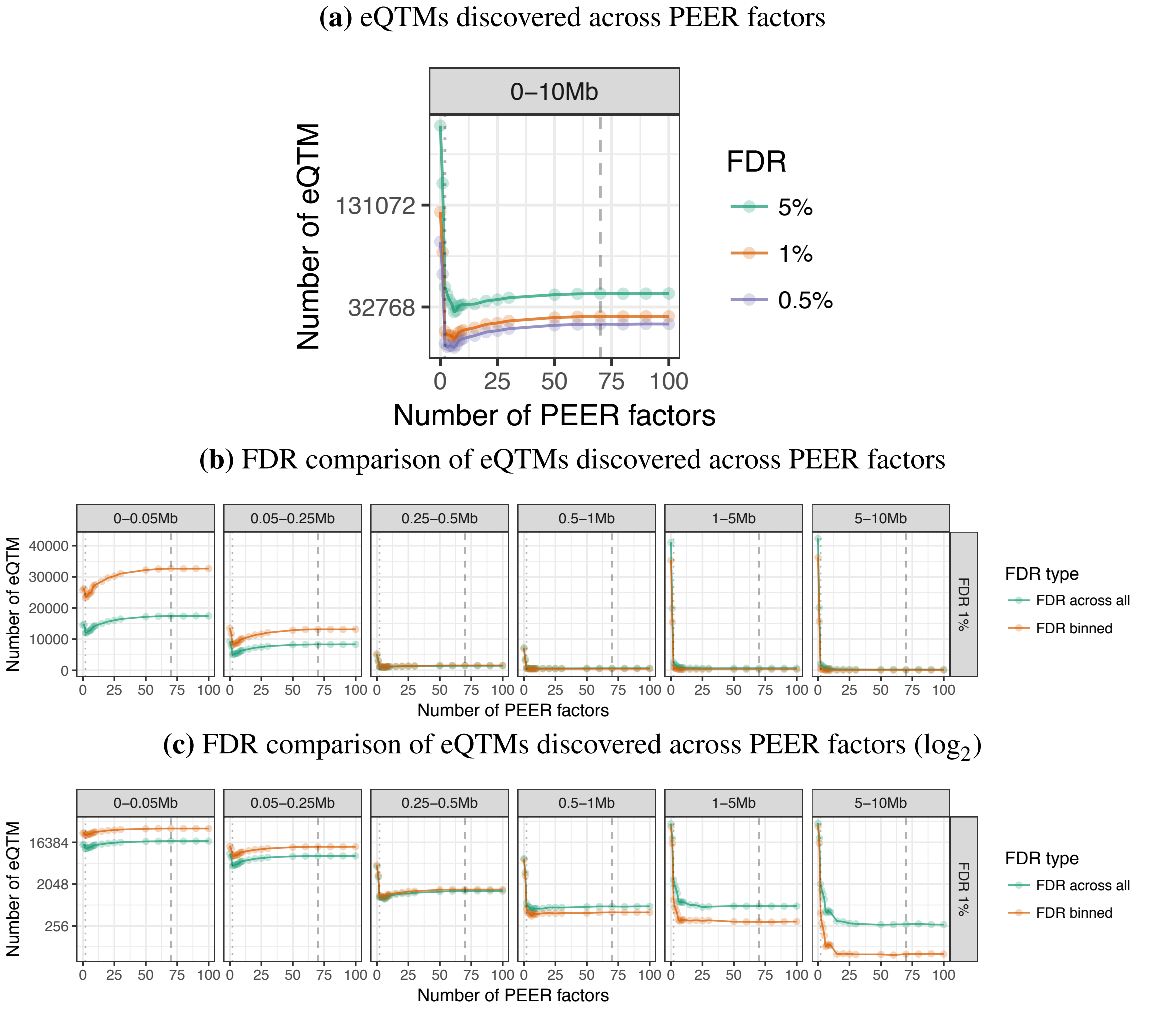


**Figure 3.2** eQTM discovery rate by distance: PEER 0. Discovery rate of eQTMs binned by distance from TSS, accounting for known covariates but not for PEER factors.

### 3.3.1 PEER factor optimisation

In order to account for the confounding correlation between gene expression and DNAme, I mapped eQTMs, iteratively increasing the number of PEER factors included in the expression and DNAme analysis (i.e., 1 PEER factor means PEER was run learning 1 PEER factor in expression and DNAme separately). Interestingly, I observed a striking drop in the number of eQTMs discovered when I accounted for the first 2 PEER factors from both gene expression and DNAme (Figure 3.3). Given the PEER factor-trait associations, this suggests that tissue heterogeneity is a key driver behind the confounding correlation.



**(a)** eQTMs discovered across PEER factors

**(b)** FDR comparison of eQTMs discovered across PEER factors

**(c)** FDR comparison of eQTMs discovered across PEER factors ($\log_2$)

**Figure 3.3** eQTM PEER optimisation. Dotted line at 2 PEER factors. Dashed line at 70 PEER factors. (a) Number of eQTMs discovered (y axis), iteratively accounting for more PEER factors (x axis). (b) Number of eQTMs discovered performing FDR over all eQTMs from 0-10 Mb in distance (green) and FDR over binned groups (orange). (c) Same data from panel b, but on $\log_2$ scale.

Recognising that the discovery of eQTMs > 1 Mb from the TSS may indicate confounding correlation, I binned the eQTMs across PEER factors by distance to see if there is evidence of nuanced trends missed when analysing 0-10 Mb eQTMs in bulk. In addition, I also re-ran the eQTM mapping process separately controlling for FDR within eQTM distance bins, rather than controlling for FDR at once across all eQTMs from 0-10 Mb. I performed this binning test to test if the "significant" distal associations might appear to have signal due to the strong signal of eQTMs near a TSS. Worded another way, I thought the distal eQTMs were being called significant by the Storey FDR procedure [369] primarily due to the signal at nearby eQTMs.

From this analysis, I discovered two trends. First, as I included additional PEER factors beyond 2, I found that the number of eQTMs near the TSS (< 250 kb) increased across PEER factors, while the number of eQTMs further away (> 1 Mb) were drastically reduced. The simultaneous increase in proximal eQTMs and decrease in distal eQTMs reached a steady state at 70 PEER factors. Second, I found far fewer distal eQTMs when performing the Storey FDR procedure across distance bins.

Together, these results suggest very few true eQTM signals exist in regions far away from the TSS. I therefore reduced the eQTM window to 1 Mb to be consistent with the window size used for QTL mapping. An alternative option would have been to use a recently developed method that performs multiple hypothesis correction while controlling for covariates, such as distance [160]. I did not use this method, however, since I wanted to be consistent with QTL methods and I was concerned about the scalability of using such an approach for QTLs—for instance, when analysing ~700,000 features in the case of mQTLs.

Before settling on using 70 PEER factors from both gene expression and DNAme, where the number of nearby and distal eQTMs stabilised, I wanted to understand the effect of including so many PEER factors, since such a high number of factors is likely to remove a combination of technical and biological effects. Therefore, I compared the p-values and the direction of effect between eQTMs called using 2 PEER factors and 70 PEER factors (Figure 3.4). I found that including 70 PEER factors increased power to detect the top, nearby eQTMs, and did not change the sign of effect for the vast majority of eQTMs. In addition, I observed a progressive downward shift in the of the best fit line towards the 2 PEER factor axis that increased across eQTM distance bins. Together, these results suggest that increasing PEER factors enhances the detection of the strongest, most robust eQTMs, while reducing eQTMs likely caused by other correlation sources. Based on these results, I chose 70 PEER factors to call eQTMs for subsequent analysis. Compared to the initial eQTM discovery rate, 70 PEER factors essentially eliminates the distal eQTM discovery rate (Figure 3.5)

Finally, I note that after mapping eQTMs across gene level expression data, I repeated this process for each exon, testing all DNAme sites within 1Mb, mapping eQTMs. For exQTMs, I found similar trends as those noted above, using total gene expression.

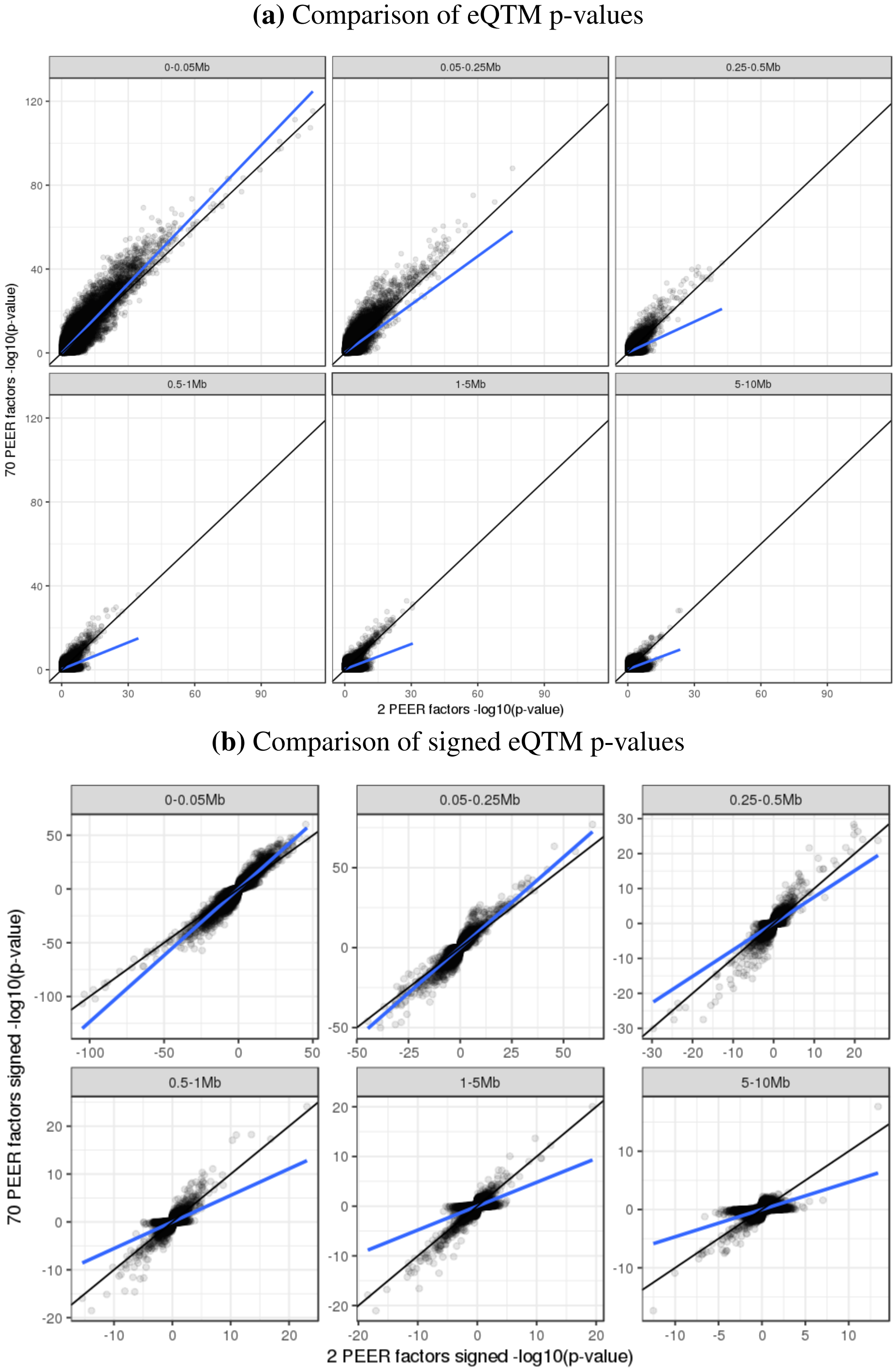


**Figure 3.4** PEER 2 vs PEER 70 eQTMs. Comparison of eQTMs discovered accounting for 2 PEER factors (plus covariates) and 70 PEER factors (plus covariates). Black line depicts the identity line. Blue line shows the line of best fit. (a) Comparison of $-\log_{10}$(p-values). (b) Comparison of signed $-\log_{10}$(p-values).

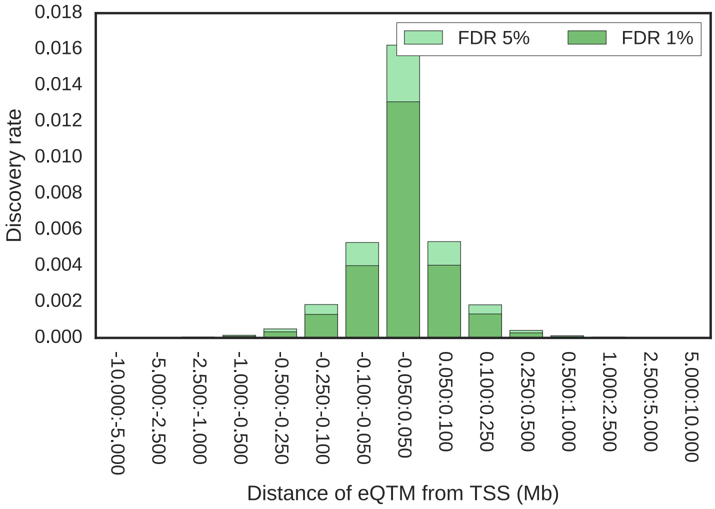


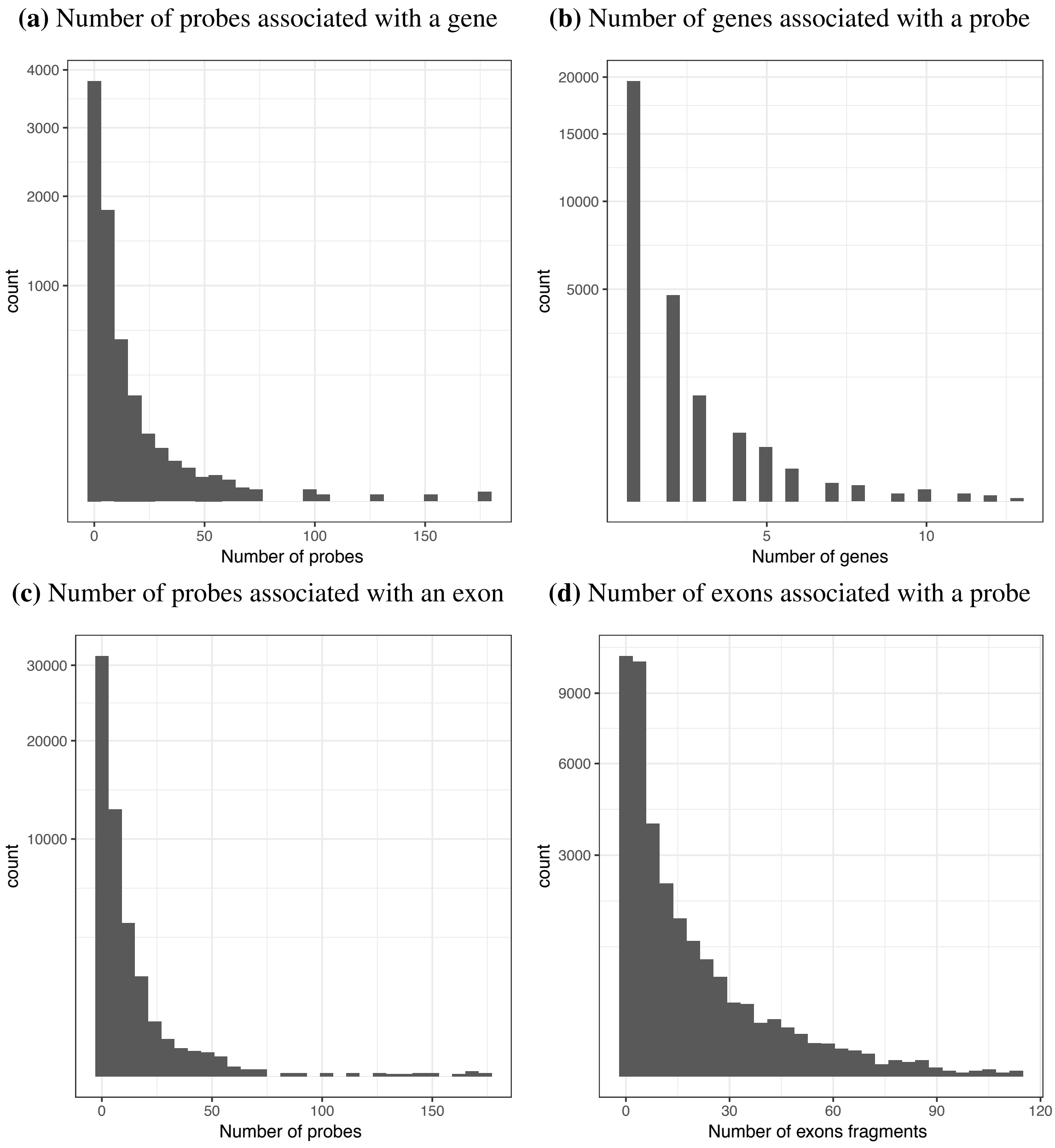
**Figure 3.5** eQTM discovery rate by distance: PEER 70. Discovery rate of eQTMs binned by distance from TSS, accounting for known covariates and 70 PEER factors.

## 3.4 eQTM results

I mapped eQTMs using a 1 Mb window and molecular trait residuals after removing known covariates and 70 PEER factors from each molecular trait separately. I found 38,115 eQTMs (FDR 1%; 56,131 FDR 5%) spanning 6,697 genes, with > 75% falling within 250 kb of the TSS. These trends were consistent across chromosomes, with the exception of chromosome 6, which was likely due to the complexity of the MHC region (chr6:28,477,797-33,448,354). Likewise, I used the same parameters for exQTMs and discovered 251,012 exQTMs (FDR 1%; 368,325 FDR 5%) spanning 51,289 exon fragments, with > 75% falling within 250 kb of the start of the exon.

### 3.4.1 eQTM overview

I found that genes on average were associated with 5.7 DNAme sites (Figure 3.6). With 178 eQTMs, *HOXD8* had the largest number of significant eQTMs. The genes with > 100 eQTMs were *HOXD3, HOXD9, HOXD8, HOXD-AS2*, *AC025183.1*—all of which are *HOX* related genes, except for *AC025183.1,* an antisense gene near *IRX4* on chromosome 2 (*IRX4* itself was associated with 95 eQTMs). DNAme sites were on average associated with 1.4 genes. With 13 gene associations, cg19051117 (chr19:21767566; within *RP11-678G14.3*) was associated with the most genes, including a variety of nearby zinc finger proteins as well as other non protein coding genes (lincRNA, antisense, unprocessed pseudogenes, etc.).


**(a)** Number of probes associated with a gene

**(b)** Number of genes associated with a probe

**(c)** Number of probes associated with an exon

**(d)** Number of exons associated with a probe

**Figure 3.6** General eQTM and exQTM characteristics. (a) Distribution of the number of probes associated with a gene. (b) Distribution of the number of genes associated with a probe. (c) Distribution of the number of probes associated with an exon. (d) Distribution of the number of exons associated with a probe.

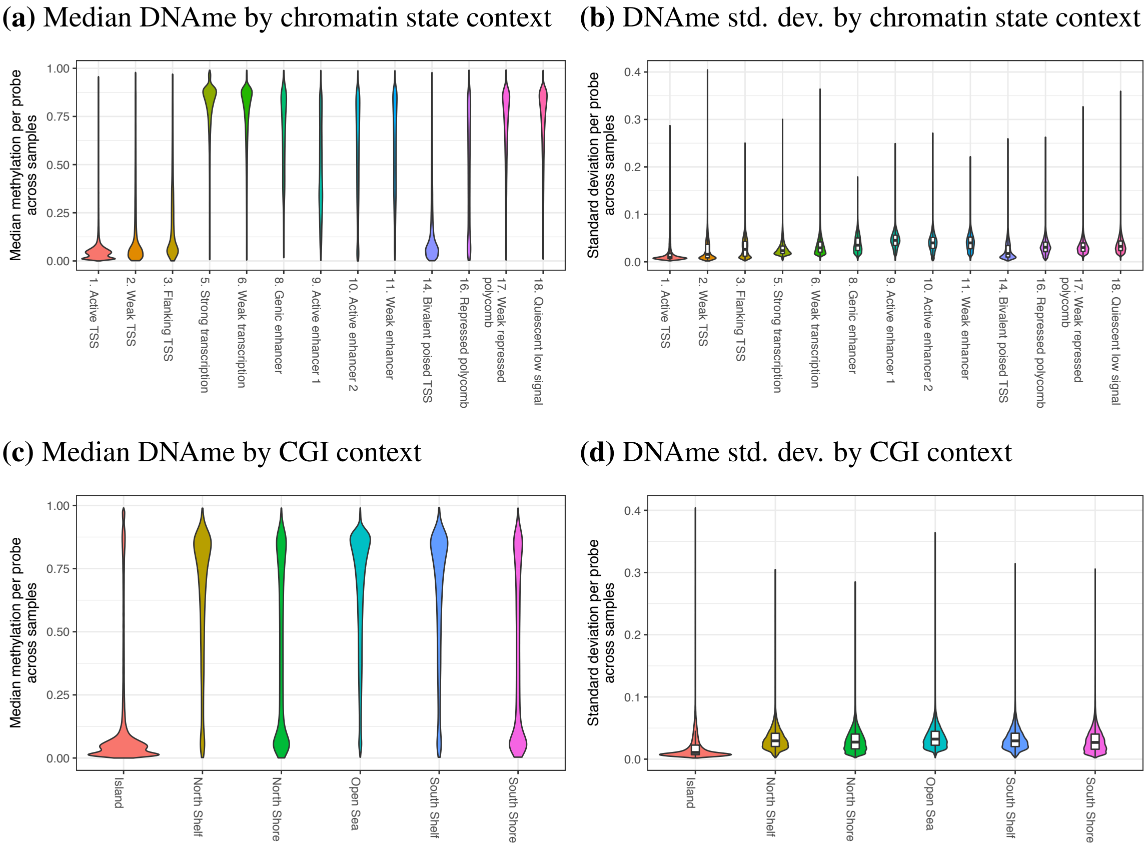
The exQTMs showed similar trends. I found that on average an exon fragment is associated with 4.9 DNAme sites. As with the eQTMs, exon fragment 12 in *HOXD8* was associated with the largest number of DNAme sites at 175 significant associations. On average, DNAme sites were associated with 7.5 exon fragments, which is more than for eQTMs, but unsurprising given the correlation structure of exon expression within isoforms. cg15710545 (chr6:32578114), located in the MHC region, was associated with 114 exon fragments, the largest number in the whole data set.

### 3.4.2 Relationship of eQTM effect and genomic context

Early studies linked increased DNAme with transcriptional inactivation [155, 312]; however, more recent studies show that DNAme can also be associated with increased expression, often depending on the genomic context [151, 223, 141]. These studies paint an increasingly complex and poorly understood picture of the relationship between gene expression and DNAme. I sought to characterise such trends within the skeletal muscle data. First, I summarised the overall trends across all DNAme sites (not just eQTMs) in relation to genomic context and found promoter related regions and CGIs (most which localise to promoters [335]), generally exhibit decreased methylation and variability in signal across all muscle samples (Figure 3.7). In addition, I also observed increased variability in transcription and enhancer related chromatin states. Both observations are consistent with the current understanding of DNAme patterns (reviewed in [173]).

I then focused on trends in eQTM effects. At an FDR 1%, I found that 59.64% of identified eQTMs were negatively associated with gene expression (i.e., repressive; 58.37% FDR 5%). When split by the relationship of the DNAme site to CGI, the majority of eQTMs had a negative effect across all CGI contexts (Figure 3.8a). I also evaluated the eQTM effect based on the skeletal muscle chromatin state context of the DNAme site (Figure 3.8b). I found that for TSS, enhancer, and other active transcription related states, the majority of eQTMs have a negative effect. However for bivalent/poised TSS, repressed, and quiescent/low signal states (states 14, 16, 17, and 18), the majority of eQTMs were positively associated with transcription. This shift in the number of negative eQTMs between these two groups of chromatin states was significant (chi-squared p-value < $2.2\text{x}10^{-16}$). These results are consistent with a previous study that finds a larger proportion of eQTMs with negative effects in promoter and enhancer states [141].

I also evaluated the chromatin state of DNAme sites that overlap a skeletal muscle ATAC-seq peak (Figure 3.8c). Of the few overlaps, I found the majority of ATAC-seq peak overlapping



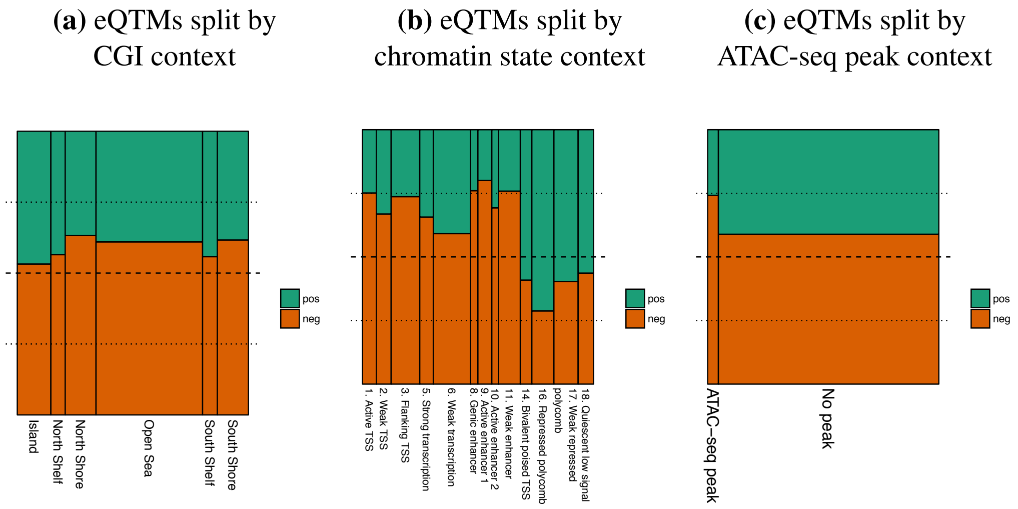
**(a)** Median DNAme by chromatin state context

**(b)** DNAme std. dev. by chromatin state context

**(c)** Median DNAme by CGI context

**(d)** DNAme std. dev. by CGI context

**Figure 3.7** Characterisation methylation by the probe genomic context. All calculations performed on methylation Beta values. (a) Median DNAme of probe across all samples divided by skeletal muscle chromatin state. (b) Standard deviation of DNAme across samples divided by skeletal muscle chromatin state. (c) Median DNAme of probe across all samples divided by CGI context. (d) Standard deviation of DNAme across samples divided by CGI context.


**(a)** eQTMs split by CGI context

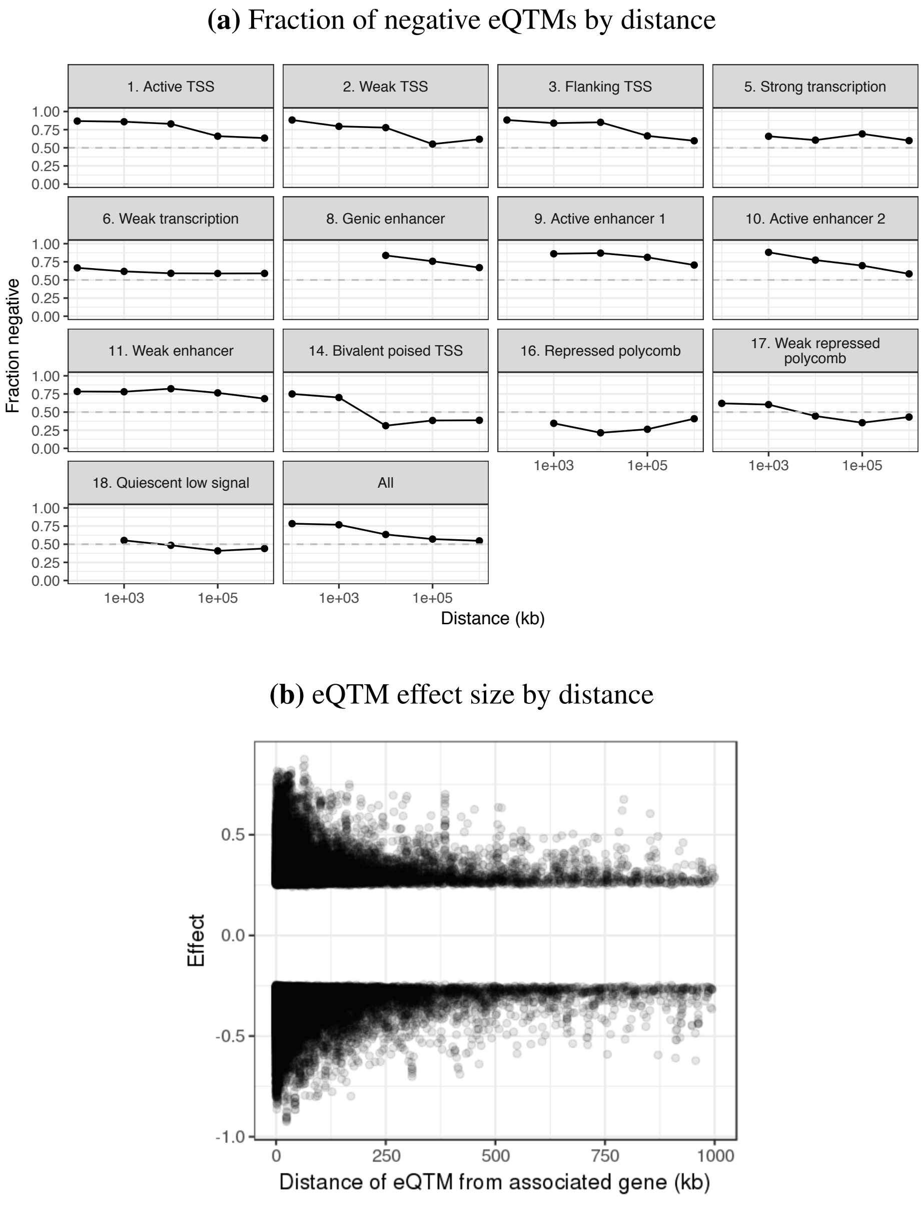
**(b)** eQTMs split by chromatin state context

**(c)** eQTMs split by ATAC-seq peak context

**Figure 3.8** Characterisation of eQTM effects by genomic context. Bins based on overlaps of the methylation sites and various genomic features. Across all plots, the size of the x and y axis is proportional to the fraction of eQTMs meeting that criteria. Dashed line at 50% and dotted lines at 25% and 75%. (a) Fraction of negative eQTMs based on CGI context. (b) Fraction of negative eQTMs based on skeletal muscle chromatin state context. (c) Fraction of negative eQTMs based on skeletal muscle ATAC-seq peak context.

eQTMs had a negative association. When only considering eQTMs that overlapped ATAC-seq peaks, the vast majority of ATAC-seq eQTMs (the DNAme sites) occurred within TSS related or active enhancer states. Otherwise, the trends in CGI and chromatin states remained largely the same.

Finally, I evaluated the fraction of negative associations with respect to distance. I found that across all chromatin states, DNAme sites closer to the target gene TSS tended to have a negative effect (Figure 3.9a). In addition, I found that eQTMs closer to the TSS also had larger effect (Figure 3.9b). Finally, I found very similar trends when considering exQTMs.



**(a)** Fraction of negative eQTMs by distance

**(b)** eQTM effect size by distance

**Figure 3.9** Characterisation of eQTM effects by distance. (a) Fraction of negative eQTMs by distance. Dashed line at 0.5. Bins range from 100 bp, 1 kb, 10 kb, 100 kb, to 1 Mb, and are non-overlapping (first point is 0 to 100 bp, second point is 101 bp to 1 kb, etc.). A minimum of 25 eQTMs required for bin. (b) eQTM effect by distance.

## 3.5 eQTM summary

In summary, these results highlight both the importance of latent factor analysis when analysing molecular traits and the context specific nature of expression-DNAme relationships. While mapping eQTMs, I found strong evidence for the existence of latent correlation structure between expression and DNAme. Using PEER, I was able to account for the correlation structure between gene expression and DNAme, enabling the identification of a confident set of eQTMs and exQTMs. I characterised the properties of these eQTMs (and exQTMs) and found them to be consistent with the current context specific understanding of expression-DNAme relationships (reviewed in [173, 212]). Consistent with earlier studies, which focused on DNAme in "active states" like TSS regions [173], I found eQTMs located in genomic regions linked to active gene regulation (including enhancers which are better captured by EPIC arrays than previous arrays) tended to be negatively associated with gene expression (i.e., increased DNAme is associated with repressed expression). By contrast, I found eQTMs in less active or repressed genomic regions, poorly analysed by earlier studies, tended to show positive associations with expression, consistent with a more recent notion that expression-DNAme trends are highly context specific [173]. Finally, I showed that eQTMs closer to the target gene TSS have stronger effects that tend to be negative.

# Chapter 4

# Molecular quantitative trait loci

## 4.1 Introduction

In this chapter, I describe how I mapped proximal eQTLs, exQTLs, and mQTLs, as well as characterise the general properties of these QTLs. I investigate the QTL enrichment properties in chromatin states across cell/tissue types. For simplicity, I focus primarily on eQTLs and mQTLs, since exQTLs follow nearly identical trends as eQTLs. In addition, I describe how I used QTLs to understand the regulatory effects of transcription factors in skeletal muscle. In the next chapter (Chapter 5), I specify how I integrated GWAS data in a complementary analysis to the one presented here. I do not consider distal QTLs, because the FUSION dataset is underpowered for such an analysis, both on theoretical grounds [275] and as confirmed by a preliminary analysis performed by colleagues at the University of Michigan.

## 4.2 QTL mapping

I mapped QTLs using LIMIX v0.7.74 [222]. As candidate SNPs, I considered all proximal SNPs within 1 Mb of the start of the feature (TSS for genes, exon start for exons, and DNAme sites). I used 1 Mb as it is a common distance threshold for proximal QTLs [275]. Using this window, I tested for genetic associations with rank-based inverse normalised PEER residuals, $y$, of feature $j$ across individuals using the linear model:

$$y_j = \underbrace{\alpha_j \mathbb{1}}_{\text{intercept}} + \underbrace{\beta_j g}_{\text{genetic eff.}} + \underbrace{\psi_j}_{\text{noise}}, \quad \psi_j \sim \mathcal{N}(0, \sigma_e^2 I) \tag{4.1}$$

where $\alpha_j$ is the intercept, $g$ the genotype vector of dosages across individuals, $\beta_j$ the genotype effect, and $\psi_j$ Gaussian noise. After running all associations, I corrected for the number of tests per QTL type using Storey's FDR [369].

### 4.2.1 Pipeline validation

Prior to analysing the complete FUSION dataset, I used the data freeze from Scott et al. [341], described previously (Section 2.8), to validate the LIMIX pipeline by comparing LIMIX eQTLs to the Scott et al. [341] eQTLs which were called using matrix eQTL [345]. I compared the p-values and effect sizes across all 5,344,655 autosomal SNP-gene pairs from the raw matrix eQTL output, as matrix eQTL thresholds the output based on a p-value cutoff so that not all SNP-gene pairs are saved (Figure 4.1). I found the results from the two pipelines to be nearly identical, with the exception of 17 SNP-gene pairs, spanning 11 total SNPs, where the MAF within FUSION samples was exactly 0.5. In these cases, the LIMIX effect sizes were oriented to reference genome allele while the matrix eQTL effect sizes were oriented to the alternate genome allele, flipping the orientation so that the effect alleles matched produced identical results. In addition, I found the LIMIX p-values were slightly smaller, possibly due to the likelihood-ratio test used by LIMIX compared to the t-statistic used by matrix eQTL.

## 4.3 QTL results

My collaborator, Narisu Narisu, prepared the imputed genotypes across the 318 samples that passed QC for QTL mapping, keeping only bi-allelic variants with $r^2 > 0.3$ and MAC > 10. I mapped QTLs across all available samples for each molecular trait: 301 for gene expression

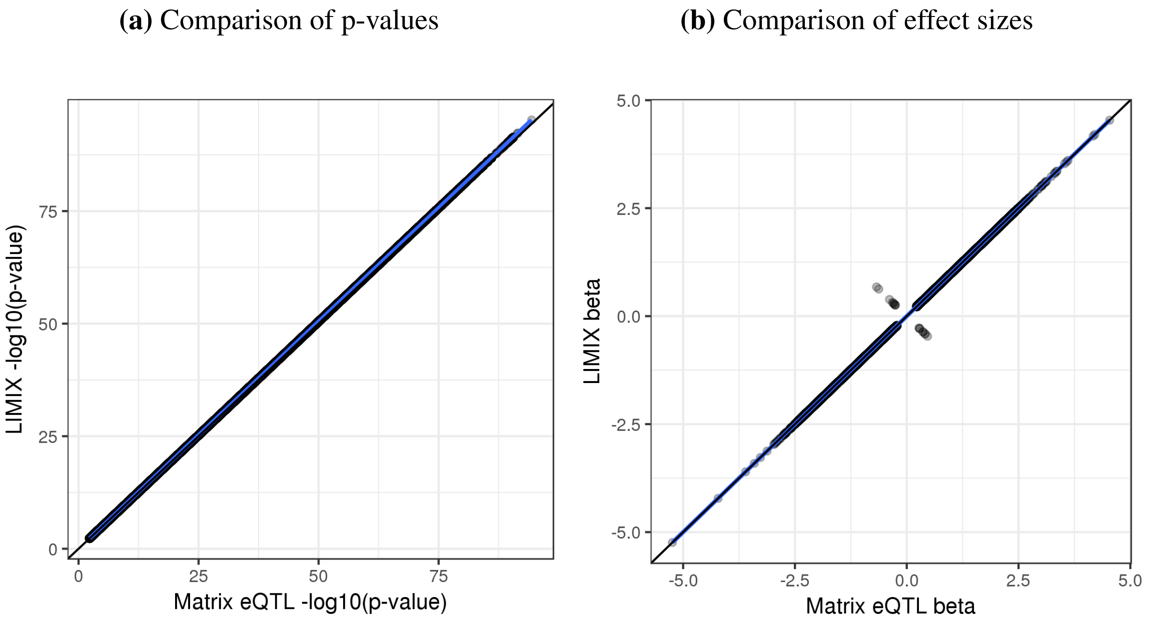
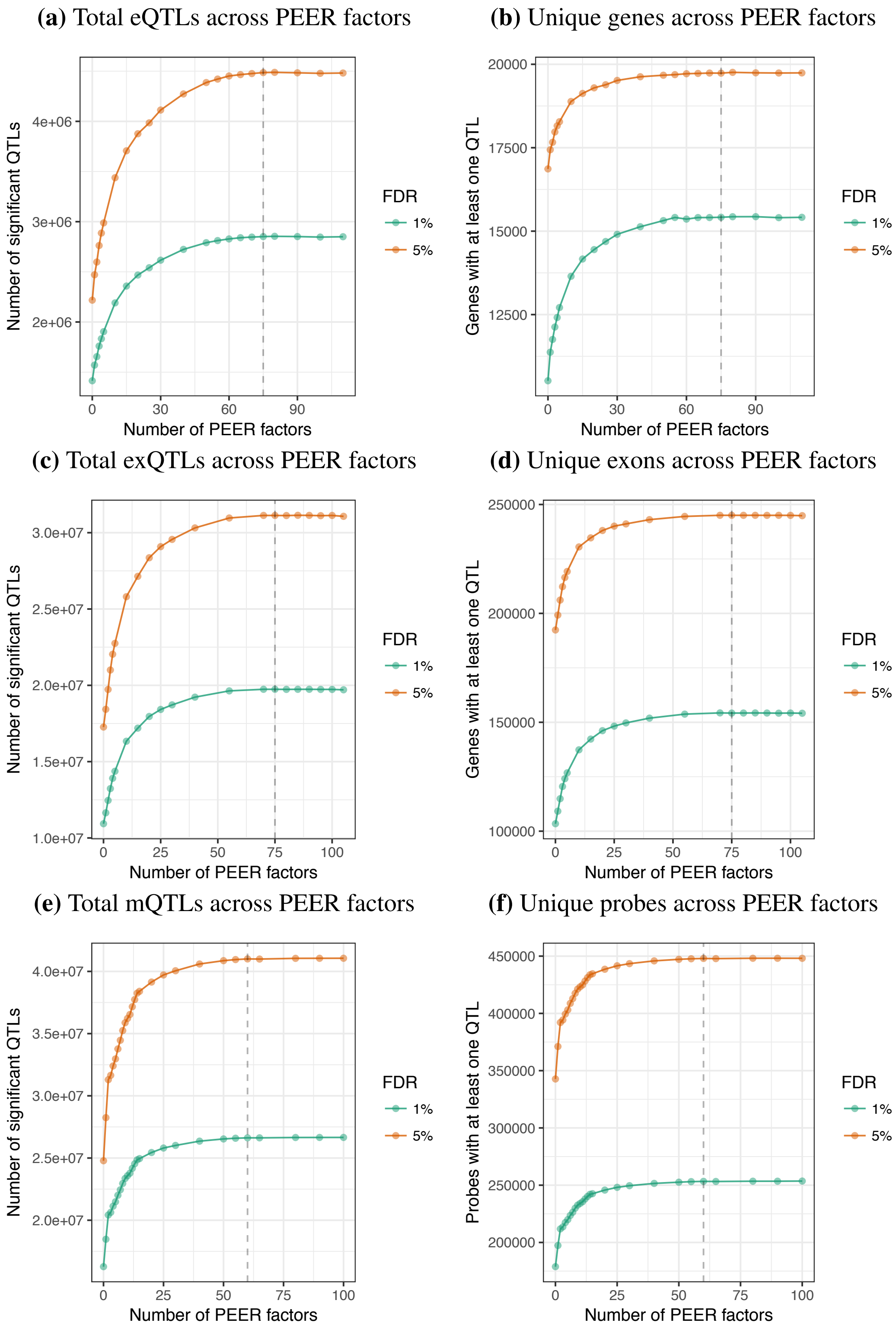


**Figure 4.1** LIMIX QTL pipeline QC. (a) Comparison of $-\log_{10}(\text{p-value})$ from LIMIX to matrix eQTL. (b) Comparison effect sizes (beta) from LIMIX to matrix eQTL.

and 282 for DNAme. Similar to Scott et al. [341], I optimised the number of PEER factors by mapping QTLs, iteratively increasing the number of PEER factors included in the model, before selecting the number of PEER factors that produced the maximal number of significant proximal QTLs. Based on this procedure, I estimated the optimal number of PEER factors to be 75 for eQTLs, 75 for exQTLs, and 60 for mQTLs (Figure 4.2). Using the optimal PEER factors I found 2,851,250 eQTLs (FDR 1%) spanning 15,416 genes, 19,740,690 exQTLs (FDR 1%) spanning 154,229 exon fragments, and 26,622,999 mQTLs spanning 253,343 DNAme sites (FDR 1%).



**(a)** Total eQTLs across PEER factors

**(b)** Unique genes across PEER factors

**(c)** Total exQTLs across PEER factors

**(d)** Unique exons across PEER factors

**(e)** Total mQTLs across PEER factors

**(f)** Unique probes across PEER factors

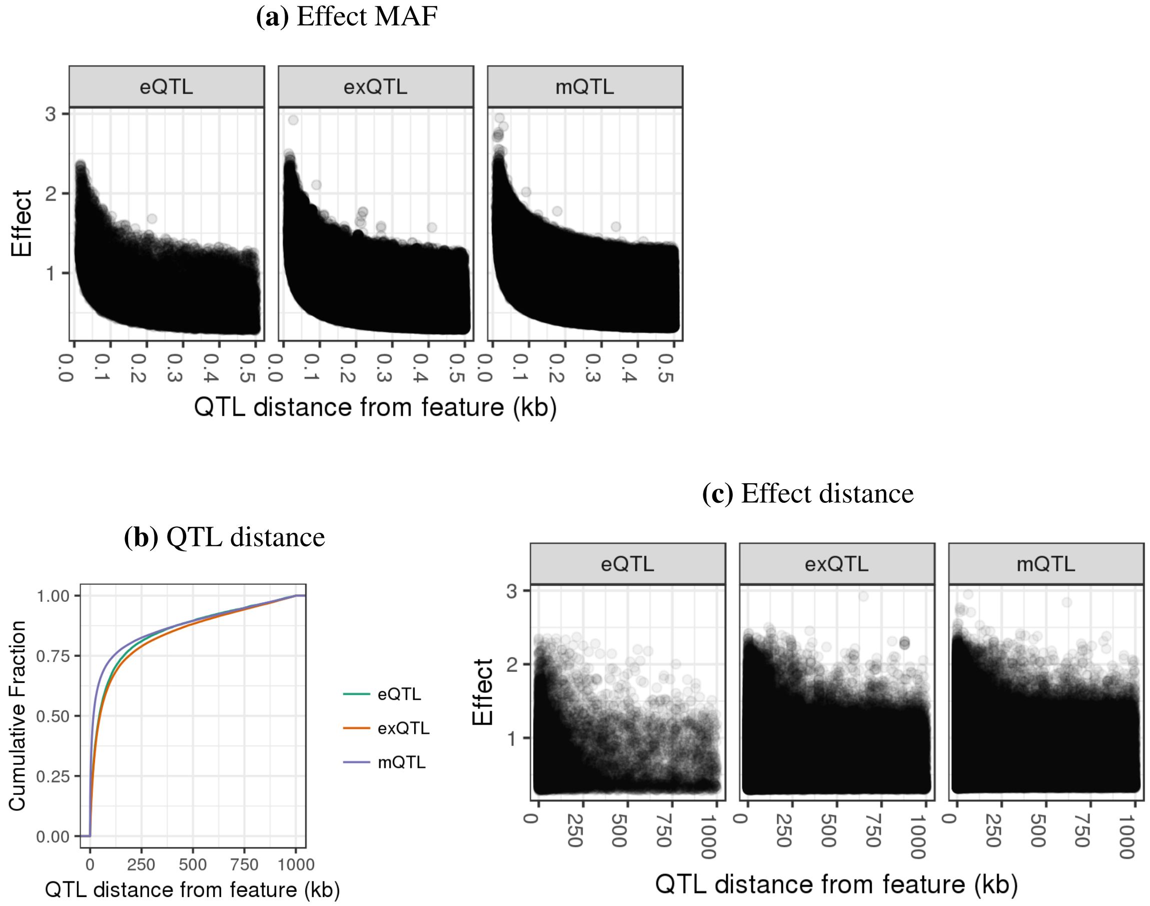
**Figure 4.2** QTL PEER factor optimisation. Plots in the first column show the total number of QTLs discovered. Plots in the second column show the total number of unique features. The dashed line depicts the optimal number of PEER factors. (a-b) eQTL optimised at 75 PEER factors. (c-d) exQTL optimised at 75 PEER factors. (e-f) mQTL optimised at 60 PEER factors.

### 4.3.1 QTL overview

I characterised the general properties of each QTL type (Figure 4.3), taking the best QTL per feature (SNP with minimum p-value; FDR 1%). As a quality control measure, I compared the MAF to effect size, analysing the raw data behind outliers with a larger effect than what was commonly seen at a given MAF. Within the raw data, there were no trends to raise cause for concern in the outliers. Overall, the majority of outliers (7/9) were linked to particularly complex regions of the genome either in the MHC region or in centromere regions.

I compared the distance of the best QTL per feature to the target feature and found that in contrast to eQTLs and exQTLs, mQTLs generally occur closer to the target feature (DNAme site in this case). In addition, I compared the effect size according to distance and found QTLs of larger effects tend to be closer the target feature. This tendency was more pronounced for eQTLs than for exQTLs and mQTLs.

Finally, in order to have a general understanding of overlapping QTL effects, I calculated the overlap of QTLs between QTL types (i.e., the QTL tag SNP is associated with multiple traits), both across all QTLs (FDR 1%) and taking the best QTL per feature, defined as the minimum p-value (Figure 4.4). As expected, I found a high degree of overlap between eQTLs and exQTLs (90% all eQTLs and 58% top eQTLs), and when considering all QTLs, many eQTLs are also mQTLs (92%). Interestingly, I observed more overlap between exQTLs and mQTLs than exQTLs and eQTLs (both panels). Across all exQTLs, 89% are mQTLs while 69% are eQTLs (Figure 4.4a); across top exQTLs, 15% are mQTLs while 10% are eQTLs (Figure 4.4b). This overlap between exQTLs and mQTLs is also reflected in mQTLs where, albeit a small amount of overlap, I found more overlap between mQTLs and exQTLs than mQTLs and eQTLs.



**(a)** Effect MAF

**(b)** QTL distance

**(c)** Effect distance

**Figure 4.3** Comparison of QTL properties. (a) Overview of QTL effects by minor allele frequency. (b) Comparison of QTL distance from feature. (c) Comparison of QTL effects by distance of the tag SNP to the feature.

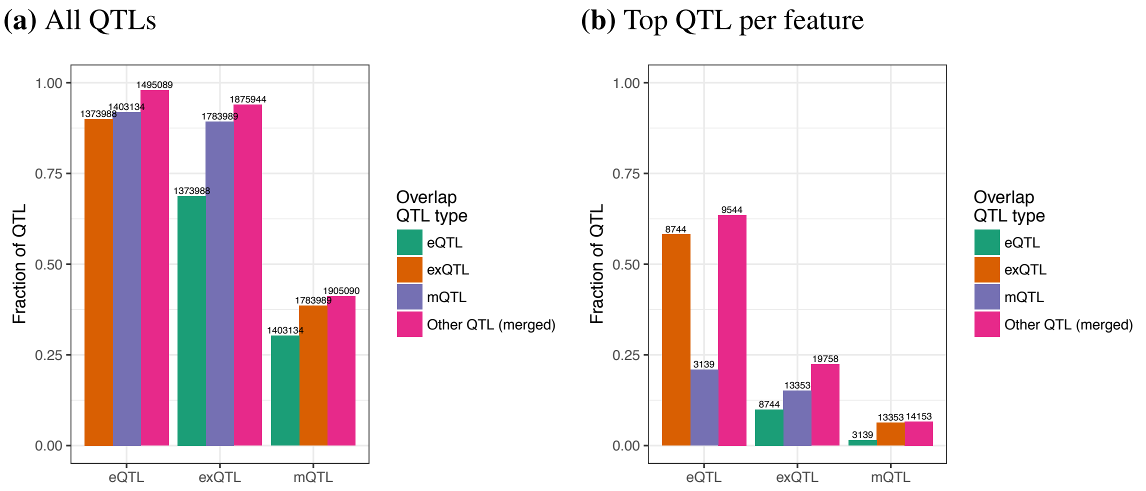


**Figure 4.4** QTL overlap (i.e., SNP associated with multiple traits). "Other QTL (merged)" means a merged set of SNPs from all other molecular traits. As an example, for eQTLs this would be the merged set of exQTLs and mQTLs. Numbers show the total counts. (a) Fraction of SNP overlap across all QTLs (FDR 1%). (b) Fraction of SNP overlap across top QTL per trait.

## 4.3.2 QTL chromatin state enrichments

Narisu Narisu and I calculated enrichments of eQTLs and mQTLs across chromatin states from Varshney et al. [395], described in Section 2.9. In addition, we used enhancer states classified according to length in order to define stretch enhancers, a regulatory element shown to be a signature of tissue-specific active chromatin [287]. For enhancer classifications, active enhancers 1 and 2, weak enhancers, and genic enhancers were merged and classified according to length: typical (< 800 bp), intermediate ($\geq$ 800 and < 3000 bp), and stretch ($\geq$ 3000 bp).

We used GREGOR [337] to calculate the enrichment of QTLs relative to null SNP sets matched for MAF, TSS-distance, and number of LD neighbours. Narisu Narisu ran GREGOR while I led the overall analysis which involved preparing the input files and summarising the results. For both eQTLs and mQTLs, we selected the best QTL per feature, defined as the minimum p-value, and LD pruned ($r^2 < 0.8$) the entire SNP set, keeping the SNP with the minimum p-value per LD block. We used the following GREGOR parameters: $r^2$ threshold = 0.99, LD window size = 1 Mb, and minimum neighbour number = 500. In addition, we grouped QTLs into bins based on (1) the effect size and (2) the specificity index of the feature. For effect size bins, we removed all QTLs with a MAF < 0.2, to avoid including spuriously high effect sizes (Figure 4.3a), and split the data into quintiles, binning the data into 50% overlaps forming 9 total bins. We also performed a similar binning procedure for the specificity index.

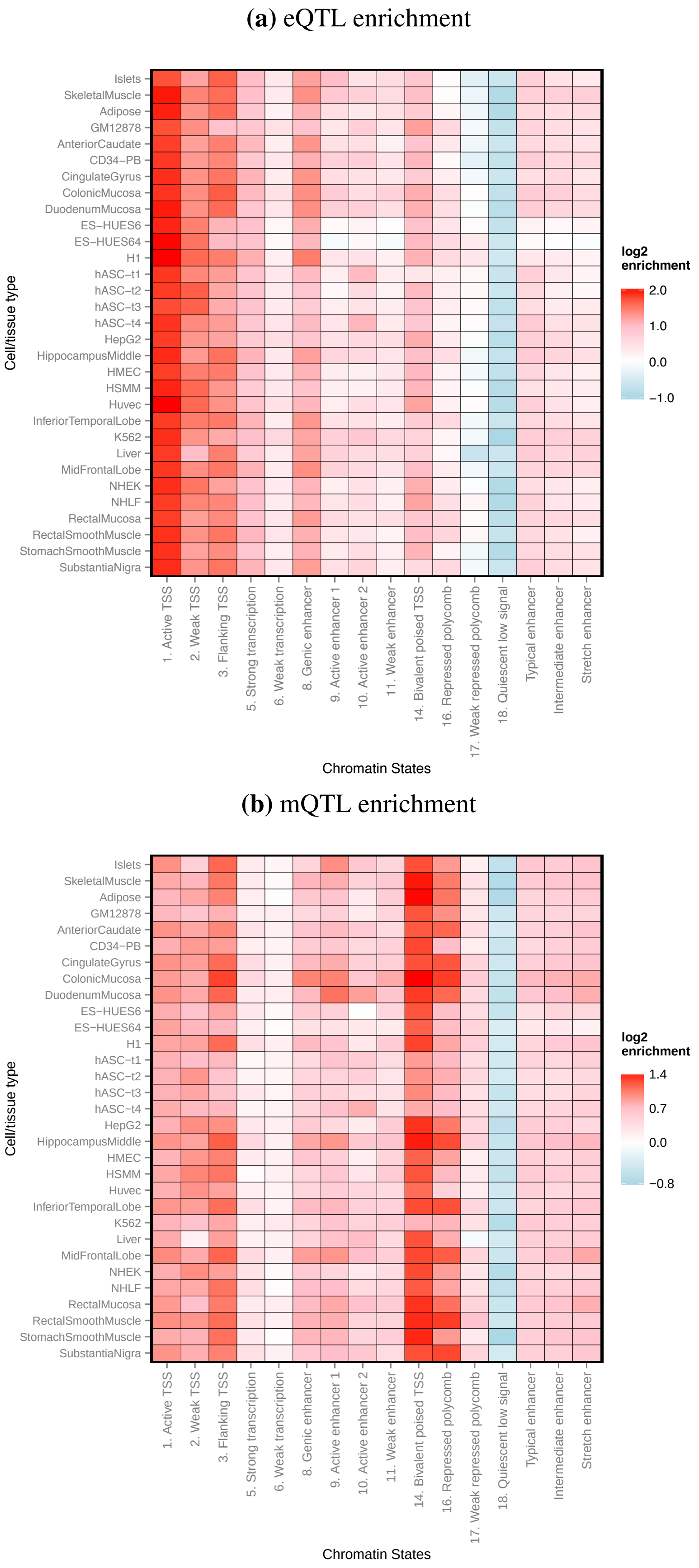
Consistent with previous reports [341, 395], I found eQTLs to be highly enriched in TSS related chromatin states (Figure 4.5), which also supports a model that eQTLs generally perturb protein-DNA interactions of TFs [113]. In addition, as noted in other studies [59, 273], I found mQTLs to be enriched in bivalent/poised TSS states.[1] Both the eQTL and mQTL enrichment patterns were generally similar across cell/tissue types, consistent with findings that suggest many eQTLs [136, 104, 138] and to some extent mQTLs [357, 142] are shared across tissues (note mQTLs are far less characterised due to limited tissue diversity). As stated previously, I found exQTLs showed nearly identical trends in enrichment as eQTLs, with perhaps a slight increased enrichment in genic enhancer states (data not shown).

Next, I focused on skeletal muscle chromatin states and evaluated enrichments of QTLs binned by effect size (Figure 4.6). I found eQTLs of stronger effects were highly enriched in TSS related states. For mQTLs, I found enrichment in TSS related states, most notably bivalent/poised TSS, increased according to effect size. The strongest of these patterns—active TSS for eQTLs and bivalent/poised TSS for mQTLs—were generally consistent across tissues, with slightly more variation for the mQTL enrichments (data not shown).

These trends changed when binning by the skeletal muscle specificity of gene expression or DNAme (using the mESI and MeSS values described in Sections 2.6.3 and 2.7.8 respectively). For eQTLs, I found muscle specific genes are enriched in TSS features, as well as stretch enhancers (Figure 4.7). As noted previously, stretch enhancers mark highly tissue specific chromatin regions [287], and the enrichment of muscle specific eQTLs in such regions further supports this observation. Consistent with our previous report [341], I also found this enrichment is unique to skeletal muscle stretch enhancers compared to other cell and tissue types. I did not observe similar patterns in typical enhancers, which are not necessarily highly tissue specific.

I found mQTLs for DNAme sites specific to skeletal muscle were enriched in skeletal muscle flanking TSS, active enhancer state 1, and bivalent/poised TSS states (Figure 4.8a). Similar to the eQTL analysis, I compared these enrichment patterns across cell/tissue types. Albeit

[1] Note that in Chen et al. [59], the "Repressed Polycomb TSS (H3K27me3, H3K4me3, H3K4me1)" state in Figure 3G corresponds to the "bivalent/poised TSS" state from Varshney et al. [395] which is enriched in the same signals [395, Figure S1C].



**(a)** eQTL enrichment

**(b)** mQTL enrichment

**Figure 4.5** Enrichment of QTLs in chromatin states across cell/tissues types. $\log_2$ fold enrichment of all QTLs in chromatin states across tissues. (a) eQTL enrichment. (b) mQTL enrichment.

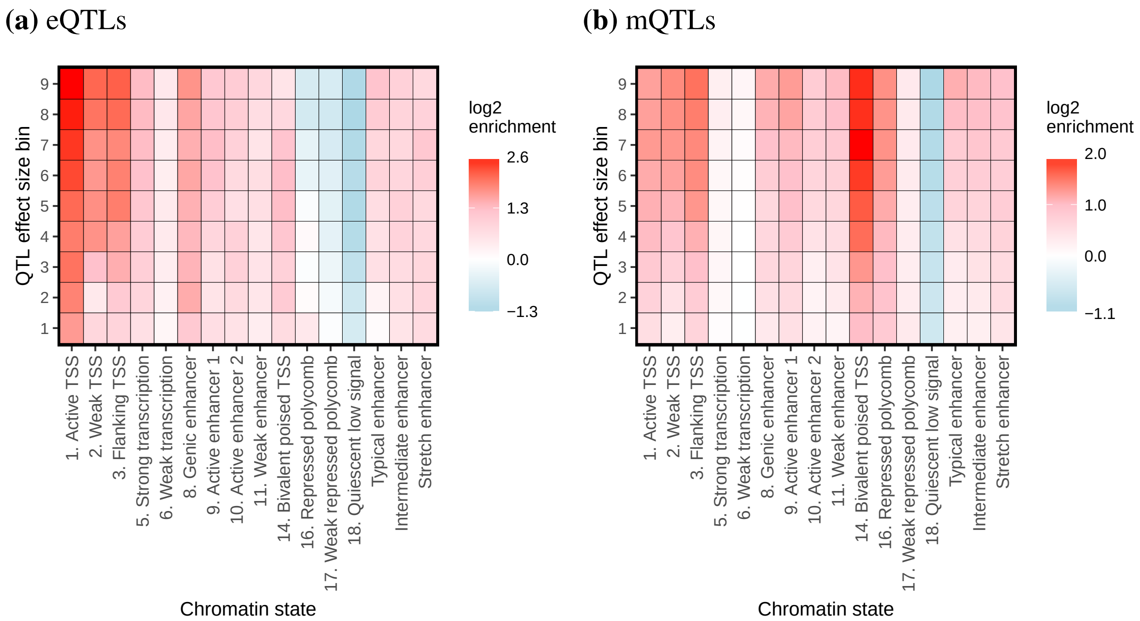
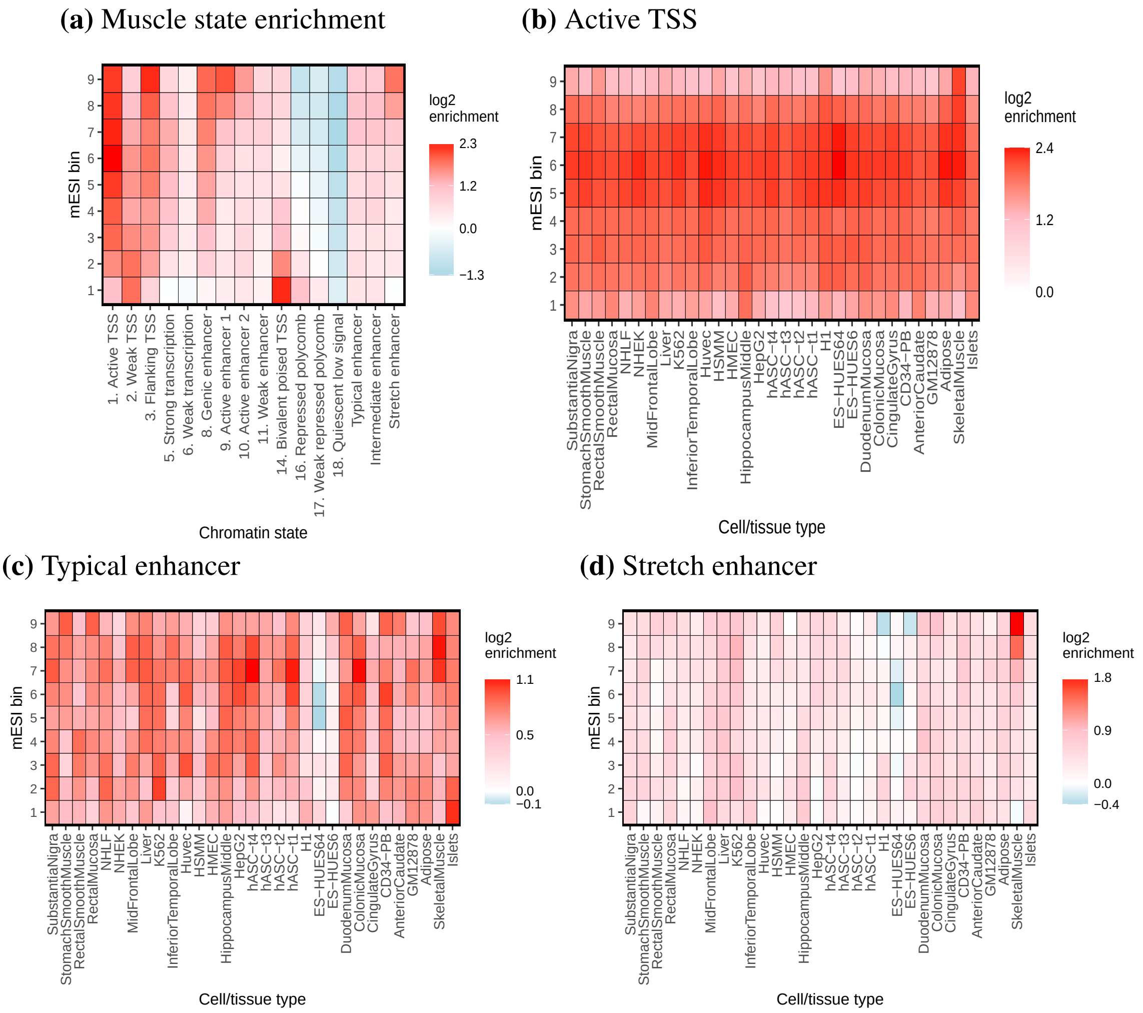


**Figure 4.6** Enrichment of QTLs binned by effect size. (a) Enrichment of eQTLs binned by effect size in muscle chromatin states. (b) Enrichment of mQTLs binned by effect size in muscle chromatin states.

noisy, for flanking TSS I found muscle mQTLs were most strongly enriched in muscle related cell/tissue types (Skeletal Muscle, HSMM, Stomach Smooth Muscle, Rectal Smooth Muscle; Figure 4.8b). I did not observe enrichment in muscle related cell types for active enhancer state 1 (Figure 4.8c) or stretch enhancers (Figure 4.8d). This could potentially be due differences between the cell/tissue types represented in the methylation specificity reference panel (Table 2.6) and those in the chromatin states (Table A.1). The trend may also indicate poor quality methylation specificity scores, as only one sample per tissue type was used (which itself is made of heterogenous tissue); however, as noted in Chapter 2, I do observe correlation between the muscle specificity of methylation and the muscle specificity of gene expression in cases of nearby eQTMs (Figure 2.22).

In addition, I found both general (non-specific) and muscle specific DNAme sites were highly enriched in bivalent/poised TSS states. I focused on the bivalent/poised TSS states and calculated enrichment across tissues. As with the previous analysis, I did not observe muscle specific chromatin state trends (Figure 4.9a). I reasoned this could be due to persistent bivalent/poised states from undifferentiated cells. To test this hypothesis, Narisu Narisu and I subdivided bivalent/poised TSS chromatin states into those that overlapped bivalent/poised TSS states of stem cells (ES-HUES6, ES-HUES64, hASC, and H1) and those that did not. I found that the division of bivalent/poised TSS states in this manner partitioned the low MeSS



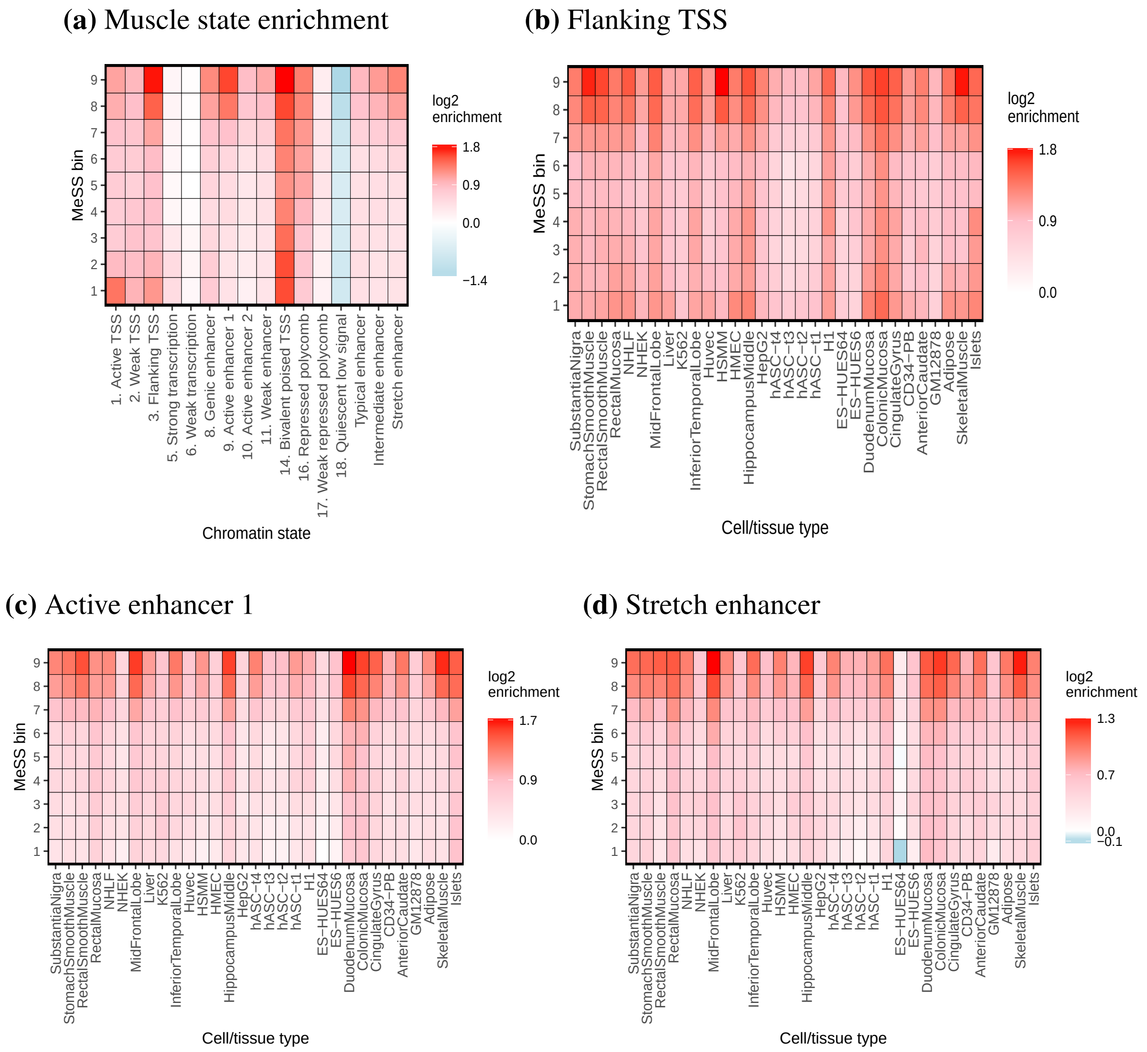
**(a)** Muscle state enrichment

**(b)** Active TSS

**(c)** Typical enhancer

**(d)** Stretch enhancer

**Figure 4.7** Enrichment of eQTLs binned by muscle specificity across cell/tissues types. (a) Enrichment in skeletal muscle states. (b) Enrichment in active TSS across tissues. (c) Enrichment in typical enhancer across tissues. (d) Enrichment in stretch enhancer across tissues.



**(a)** Muscle state enrichment

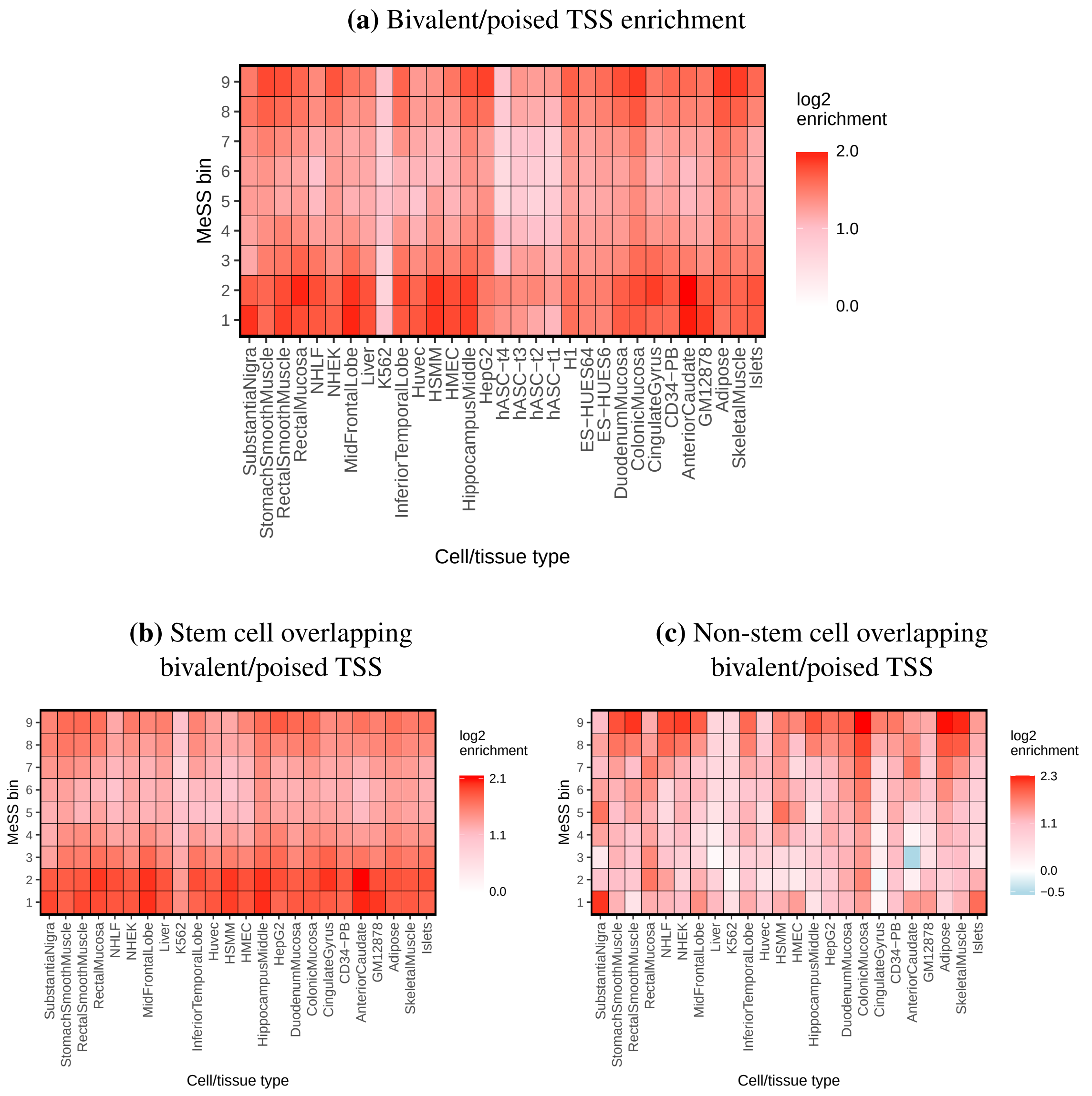
**(b)** Flanking TSS

**(c)** Active enhancer 1

**(d)** Stretch enhancer

**Figure 4.8** Enrichment of mQTLs binned by muscle specificity across cell/tissues types. (a) Enrichment in skeletal muscle states. (b) Enrichment in flanking TSS across tissues. (c) Enrichment in active enhancer 1 across tissues. (d) Enrichment in stretch enhancer across tissues.

bin and high MeSS bin enrichment patterns to some extent; however, it did not elucidate muscle specific chromatin state enrichment patterns (Figures 4.9b, 4.9c).


**(a)** Bivalent/poised TSS enrichment

**(b)** Stem cell overlapping bivalent/poised TSS

**(c)** Non-stem cell overlapping bivalent/poised TSS

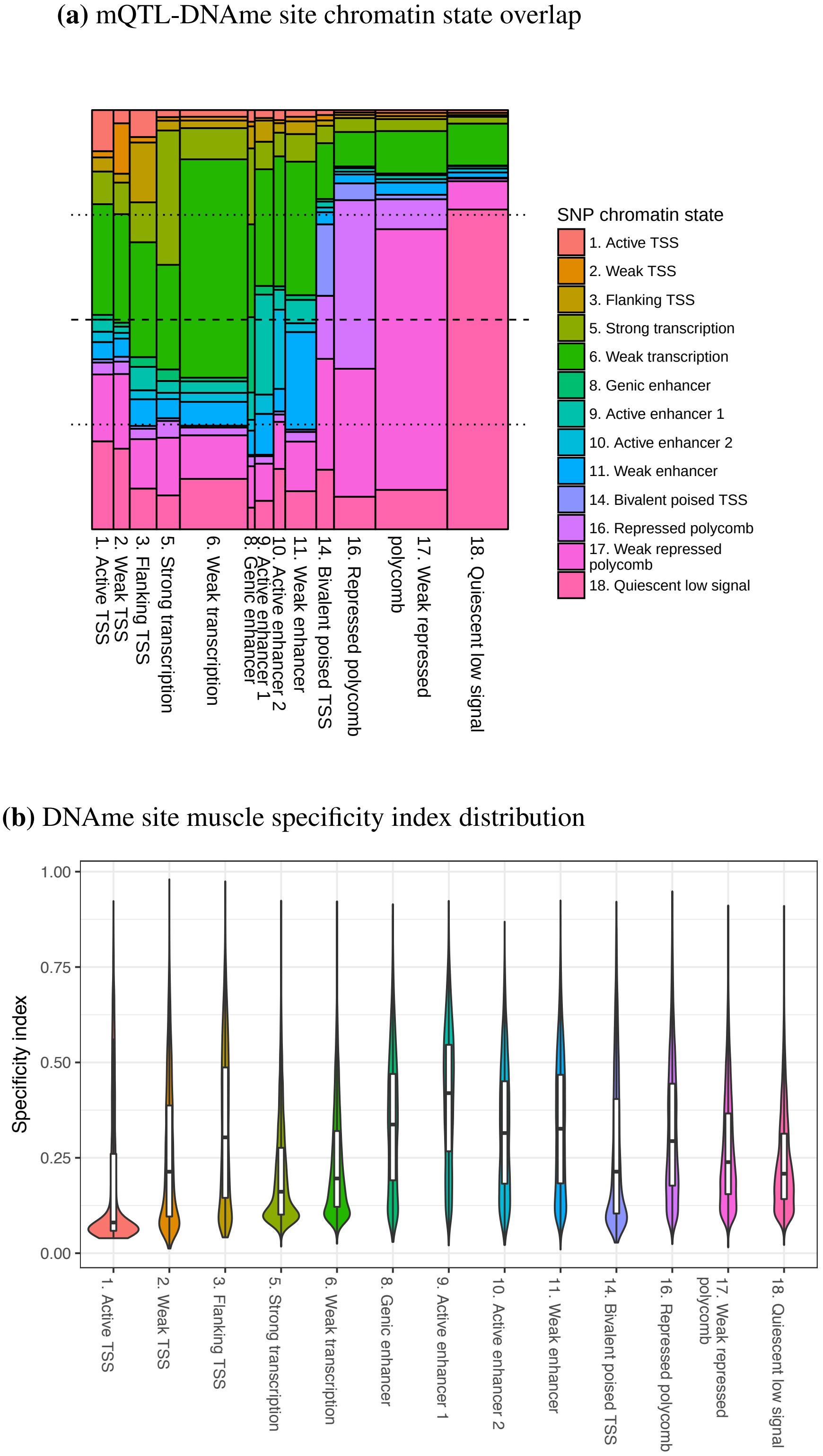
**Figure 4.9** Enrichment of bivalent/poised TSS mQTLs binned by muscle specificity across cell/tissues types. (a) Enrichment in all bivalent/poised TSS. (b) Enrichment in bivalent TSS that overlap stem cell bivalent/poised states. (c) Enrichment in bivalent TSS that do not overlap stem cell bivalent/poised states.

As a final note, the observed mQTL enrichment trends may be partially attributed to properties of the specific genomic regions targeted by the EPIC probes. Rather than an unbiased, genome wide methylation signal, the EPIC array only measures methylation at specific, predetermined regions. In the EPIC array, Illumina expanded the previous 450k array probes, which primarily assayed DNAme sites nearby or within gene bodies and promoters [333, 33], by adding probes targeted to FANTOM5 enhancer regions, ENCODE enhancers, and ENCODE open chromatin regions [261]. This could explain why I observe enrichment of highly specific mQTLs in flanking TSS states of muscle related cell types but not in enhancer states (Figure 4.8)—because DNAme sites at gene related features are generally well represented on the array, while only a fraction of enhancer states are represented. For example, Pidsley et al. [296] report only ~58% of FANTOM5 enhancers are assayed with $\geq$ 1 probe. When considering $\geq$ 2 probes, <10% of FANTOM5 enhancers are assayed.

In addition, probe representation bias may explain the enrichment of MeSS bin 9 mQTLs in active enhancer 1 states, but not in active enhancer 2 states (Figure 4.8a). Compared to active enhancer 2, the active enhancer 1 states are more strongly enriched in H3K27ac and H3K4me1 signals (Varshney et al. [395, Figure S1C]). The observed patterns in enrichment could be because the EPIC probes target enhancers that were stronger enhancer calls, more similar to active enhancer 1 states than active enhancer 2 states. Such a probe bias, coupled with the fact that mQTLs generally occur very close to the target DNAme site (Figure 4.3b), make it likely that probe bias might result in particular QTL chromatin state enrichment trends.

Finally, to better understand the potential effects of probe bias, I evaluated the chromatin state overlap of the mQTL tag SNP, binned by the chromatin state of the DNAme site (Figure 4.10a). I found that the largest fraction of mQTL DNAme sites fall in weak repressed polycomb and weak transcription chromatin states (x axis). As one would expect, given the close distance of mQTLs to the target DNAme site, the mQTL tag SNP most often fell in the same chromatin state as the DNAme site. However, a notable exception to this trend was that a sizeable proportion of mQTL SNPs fell in the weak transcription state across all DNAme site states.

Focusing on mQTL SNPs overlapping bivalent/poised TSS states in Figure 4.10a (purple colour), the DNAme sites tended to fall in either a bivalent/poised TSS state or a repressed polycomb state. Such a split may contribute to the divergent specificity score enrichment in Figure 4.8a. I tested this hypothesis by comparing the distribution of DNAme skeletal muscle specificity scores across chromatin states using only DNAme sites with a significant mQTL (Figure 4.10b). I found that mQTLs where the DNAme site resides in a bivalent/poised TSS state are generally less skeletal muscle specific than those in a repressed polycomb state. These results suggest that some mQTLs perturb DNAme in genomic regions that are specifically repressed in skeletal muscle compared to other cell/tissue types (e.g., *PIEZO1*



**(a)** mQTL-DNAme site chromatin state overlap

**(b)** DNAme site muscle specificity index distribution

**Figure 4.10** mQTL-DNAme site chromatin state overlap. Only DNAme sites of significant (FDR 1%) mQTLs considered. (a) Fraction of mQTLs in skeletal muscle chromatin states sub-divided by probe overlap (x axis) and SNP overlap (y axis). Dashed line at 0.5 and dotted line at 0.25 and 0.75. (b) Distribution of DNAme site specificity of mQTLs across chromatin states.

locus in Section 5.3.2.4). In addition, these trends were unchanged when only selecting DNAme sites whose mQTL SNPs were in bivalent/poised TSS (data not sown).

Collectively these results suggest that indeed, some of the observed enrichment trends of QTLs binned by target DNAme site muscle specificity may be due to the chromatin states of the DNAme site. However, to truly test such a hypothesis, one would need genome wide mQTLs from a less biased assay like WGBS.

### 4.3.3 Dissecting TF effects using QTLs

I also used QTLs to explore the binding effects of TFs on molecular traits, using a pipeline developed by Arushi Varshney (University of Michigan) and Stephen Parker (University of Michigan). The goal of this analysis was to classify skeletal muscle activators or repressors—TFs that tend to increase or decrease gene expression or DNAme when they bind across the genome. To summarise genome wide TF binding effects, Narisu Narisu and I worked together to identify potential instances where a TF binds in skeletal muscle and a QTL perturbs TF binding. Subsequently for each TF, we aggregated the effects of TF binding on a molecular trait (gene expression or DNAme) oriented to the preferred allele across all sites.

For in silico TF binding site (TFBS) predictions, we used data published in Scott et al. [341]. Briefly, we predicted TF binding sites using FIMO [133] with default values. In order to account for TF binding preference among common alleles, we scanned reference and alternate alleles using biallelic SNPs from 1000 Genomes phase 3 (release v5) along with 29 bp of flanking sequence from the GRCh37/hg19 human reference on each side. For in silico scans, we used a library of position weight matrixes (PWMs) from ENCODE [184], JASPAR [241], and Jolma et al. [172]. In order to subset these TF binding predictions to those likely bound in skeletal muscle, we integrated skeletal muscle ATAC-seq data (see Section 2.9) using CENTIPEDE [297] to call ATAC-seq footprints. We considered a predicted TFBS bound if the CENTIPEDE posterior probability was $> 0.99$ and the motif coordinates were contained within an ATAC-seq peak. We called these bound instances “TF footprints”, defined as a FIMO predicted TFBS plus chromatin accessibility data that has a cleavage pattern indicative of TF binding.

Many proximal QTLs are thought to act by perturbing the binding of regulatory TFs [112]. Therefore, one would expect QTLs to be enriched in overlaps of the binding sites of key regulatory TFs. To reduce our search space to key skeletal muscle TFs for further analysis, we calculated QTL enrichments in TFBSs for both eQTLs and mQTLs (FDR 1%), selecting the best SNP per feature (gene or DNAme site), and pruning the SNP list for $r^2 < 0.8$, keeping the SNP with the minimum p-value per LD block. For the enrichment calculations, we used GREGOR as described in the chromatin state enrichment Section 4.3.2. In addition, I removed general, low information content PWMs with a total information content < 10, as SNP effects on these TFBS predictions would be minimal (first quartile = 10.39; Figure 4.11). From the remaining PWMs, I found a strong correlation between the enrichment p-values for eQTLs and mQTLs, suggesting that a common set of key TFs drive a set of mQTL and eQTL signals. After using the Bonferroni correction method to control for the number of tests performed with eQTLs and mQTLs separately, I found 641 and 671 significantly enriched TF footprints (Bonferroni p-value $\leq 0.05$) respectively for eQTLs and mQTLs, which were further analysed for aggregate QTL effects.

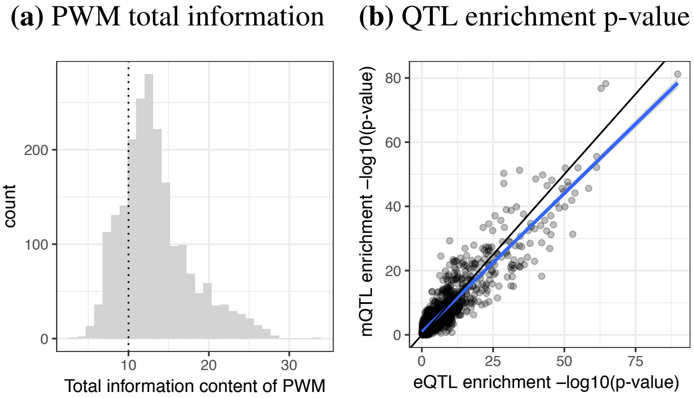

**(a)** PWM total information **(b)** QTL enrichment p-value

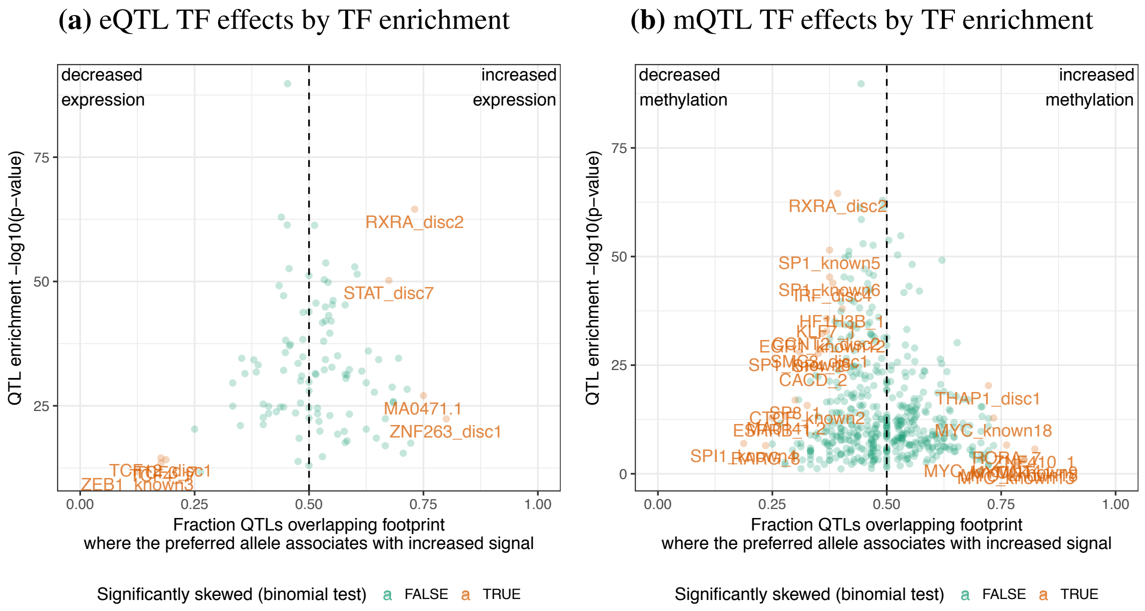
**Figure 4.11** QTL TF filters and comparison. (a) Distribution of total information content per PWM. Dashed line depicts cutoff at 10. (b) Comparison of $-\log_{10}$(p-value) of TF enrichment between eQTLs (x axis) and mQTLs (y axis).

To catalogue potential activators or repressors for each enriched TF footprint, Narisu Narisu and I summarised the effect of TF binding by aggregating the direction of effect across QTL overlaps, orienting the QTL effect to the allele that best matches the PWM. First, we intersected the lead (smallest p-value) QTL SNP and SNPs in LD ($r^2 > 0.99$) with TF footprints. For each SNP-TF footprint, we recorded the binding probability of all bases (A,T,G,C) at the particular SNP position in the PWM. Then, in order to select for cases where a QTL SNP is likely to affect TF binding, I removed SNP-TF footprint pairs where the information content of the SNP position was < 1.0 bits and one of the QTL alleles was not the allele with the greatest binding probability at that specific position. For instance, I removed cases where the QTL tag SNP was A/G and the allele with the highest binding probability at the specific site in the PWM was T, because in such a scenario it is unlikely that the A/G allele will affect TF binding in an allele specific manner. In order to have confident effect measurements, I removed instances where the SNP MAF was < 0.2, which stabilised the range of effect sizes (Figure 4.3a).

At a single TF footprint, there could exist multiple QTLs, which have the potential to skew genome wide TF binding effect estimates. For example, suppose we have 10 QTL-TF pairs, but these actually only span 2 TF footprints, each with 5 separate QTLs. While such an example seems farfetched, it demonstrates that if I were to naively consider all QTL-TF pairs, I may not have a confident genome wide estimates of TF binding effects, for in this case I would be "aggregating" over 2 total TF footprints. To resolve such issues, I performed filters to select one QTL per TF footprint. For each TF footprint, I selected the single SNP with maximum binding probability, across all SNPs that overlapped the PWM. In cases where multiple SNPs within the PWM had the same maximum binding probability, I selected the one with the largest effect size as a QTL. If a single QTL TF footprint pair was not identified through the above filters, I randomly selected one QTL for a TF footprint. This was a rare event, occurring for 1 TF footprint with eQTLs and 3 TF footprints with mQTLs. Additionally, in some cases with highly repetitive motifs, a single QTL overlapped multiple predicted TFBSs for the same TF on the positive and negative strand. For instance, the *MPP7*-eQTL rs1148181 overlaps two TAL1_known5 binding predictions (1) chr10:28616047-28616059 - strand and (2) chr10:28616045-28616057 + strand. In such cases, I randomly selected one instance, so that each QTL was considered once per motif. Such cases were infrequent, amounting to 587 cases out of 9,888 eQTL-TF overlaps and 1,627 cases out of 30,508 for mQTLs.

After identifying one QTL for each TF footprint, I oriented the QTL effect according to the allele that best matched the PWM, and for each motif, I summarised the fraction of QTL TF overlaps where the QTL shows a positive effect (increased gene expression or DNAme). I dropped cases where < 15 overlaps occurred and used the binomial test to calculate TFs that showed significant skewing (p-value < 0.05).

I compared the number of QTL overlaps to the fraction of overlaps where TF binding shows a positive effect (Figure 4.12). I found eQTLs were generally evenly distributed between positive and negative, while mQTLs were shifted towards a negative effect. This shift suggests that TF binding generally co-occurs with a decrease in nearby DNAme (as shown in Figure 4.3b, mQTLs are generally very close to their target DNAme site).


**(a)** eQTL TF effects by TF enrichment

**(b)** mQTL TF effects by TF enrichment

**Figure 4.12** QTL TF effects. For all plots the x axis shows the fraction of cases where the preferred allele of the PWM results in increased signal (either gene expression or methylation). (a) $-\log_{10}$(p-value) of TF mQTL enrichment (y axis) by fraction increased (x axis). (b) $-\log_{10}$(p-value) of TF eQTL enrichment (y axis) by fraction increased (x axis).

In total, I identified 21 TF motifs that show evidence of skewing using either eQTLs or mQTLs (Table 4.1). Many of these TFs are linked to known muscle biology. For instance, I found RXRA acts as a transcriptional activator, increasing gene expression while decreasing methylation, consistent with the UniProt annotation (http://www.uniprot.org/uniprot/P19793). RXRA is part of the tetinoid X receptor (RXR) family, which is known to form heterodimers with many other proteins (e.g., PPAR), play a role in tissue development, and help regulate metabolic processes in developed tissues, including skeletal muscle (reviewed in [380]).

I found EGR1 is linked to slightly increased expression and decreased DNAme, which is consistent with a recent study that reports an activating and important role of EGR1 in differentiation of bovine skeletal muscle satellite cells potentially through MyoG activation [441]. Likewise, I found a motif for STAT proteins, STAT_disc7, is linked with increased expression and decreased DNAme. The JAK/STAT pathway is involved in myogenesis [387, 429, 376], and more recently Jak3/STAT3 has been linked to skeletal muscle glucose uptake through GLUT4 translocation [192]. I also found ZEB1 acts as a repressor, consistent with UniProt (http://www.uniprot.org/uniprot/P37275), as well as previous studies which document repressor activity [420, 285, 331] and the importance of ZEB1 in regulating skeletal muscle differentiation [352].

In addition, I found THAP1 is associated with decreased gene expression and increased DNAme. THAP1 mutations have been linked to dystonia, a neurological disorder characterised by muscle spasms (reviewed in [203]), perhaps by repression of another dystonia related gene, TOR1A, by THAP1 [178]. Such a repressor role is consistent with the repressor function I observe. Additionally, related THAP proteins have been shown to function as repressors. For instance THAP5, important for apoptosis in cardiomyocytes [17], has been reported to act as a repressor [18], as well as THAP7 [232]. Together these support the role of THAP1 as a repressor in skeletal muscle, and perhaps important to skeletal muscle in relationship to dystonia.

Not all of the binding effect annotations matched the predicted effect. For instance, ZNF263 is annotated a repressor in UniProt (http://www.uniprot.org/uniprot/O14978), while in this analysis, it appears to act as an activator, increasing gene expression and decreasing methylation. The fact that I observe concordant activator trends separately in both gene expression and DNAme suggests that, at least in skeletal muscle, ZNF263 acts as an activator. Moreover, when considering the trends across all TF motifs considered (significantly enriched in QTL and $\geq$ 15 total counts after filters), I found a significant trend ($r = -0.38$, p-value = 4.22x10$^{-5}$) of complementary binding effects across eQTLs and mQTLs, whereby a TF that increases expression tends also to decrease methylation or vice versa (Figure 4.13). Together these results suggest this analysis uncovers real, biological activators or repressor trends of TF function.

However, caveats to this analysis should be stressed. First, it cannot be assumed that TFs always function in one manner, as studies have shown TFs can exhibit a “dual role”—acting both as a transcriptional activator or repressor depending on the genomic context [27, 328]. Second, TF binding predictions were based purely on sequence motifs and muscle ATAC-seq patterns (to identify TFs likely bound in muscle). A growing body of evidence suggest many TFs, not just those with methyl-CpG binding domains, recognise DNAme patterns and may bind to alternative motifs in the presence of DNAme (reviewed in [445]). Thus, instances where DNAme directly affects TF binding may be poorly captured. Finally, the DNAme patterns in this analysis are based on aggregate proximal mQTL effects. Generally, such effects are assumed to occur because a proximal variant perturbs TF binding which affects DNAme, making the DNAme signal a consequence of TF binding, not causal. How (or if) this effect on DNAme is then linked to gene expression is not entirely straightforward

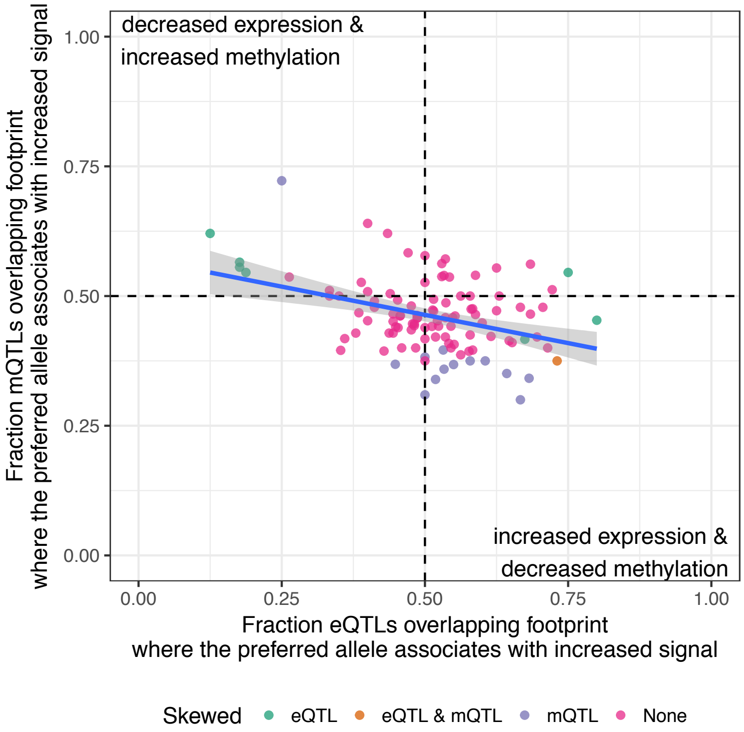


**Figure 4.13** Comparison of eQTL and mQTL TF effects.

and cannot be assumed (see Section 1.4.2.3). The mQTL DNAme site may directly change expression by perturbing the binding of a TF regulating gene expression, as described above. Alternatively, the mQTL may also be an eQTL such that the variant perturbs the binding of a TF that directly regulates gene expression. In such a scenario, DNAme (or lack thereof) may simply function to promote/stabilise the TF-DNA interaction or higher chromatin structure and not be directly causal to or necessary for the gene expression effect. Of course, genome wide there are likely instances of both models, as well as other more complex scenarios. Regardless, these results do not address such questions of expression-DNAme relationships, but describe general effects of predicted TF binding on aggregate expression and DNAme signals.

## 4.4 QTL summary

In summary, I mapped QTLs across three molecular traits: gene expression, exon expression, and DNAme. I found that while many eQTLs are identified as exQTLs, several exQTLs are not identified as eQTLs, perhaps due to transcript specific effects identified when analysing expression at the exon level. I also found that many gene expression QTLs (eQTLs and exQTLs) show some effect on DNAme, although this effect may not be the strongest genetic effect at a DNAme site. I analysed chromatin state enrichment trends, confirmed results from previous studies [341, 395, 59, 273], and showed that enrichment of mQTLs is, to some extent, effected by the genomic location of the DNAme probes. Finally, I used eQTLs and mQTLs to dissect effects of TFs binding predictions in skeletal muscle, classifying activators and repressors. I found that in cases where TF binding was associated with increased DNAme, TF binding was also associated with decreased expression (and vice versa).

| | TF | eQTL overlap (n) | mQTL overlap (n) | eQTL fraction positive | mQTL fraction positive | eQTL skew p-value | mQTL skew p-value | eQTL fold enrich | mQTL fold enrich | eQTL enrichment p-value | mQTL enrichment p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ZNF263_disc1 | 20 | 75 | 0.80 | 0.43 | 1.18e-02 | 2.48e-01 | 2.01 | 0.80 | 4.33e-23 | 5.39e-16 |
| 2 | TCF12_disc1 | 17 | 45 | 0.18 | 0.56 | 1.27e-02 | 5.51e-01 | 2.44 | 1.45 | 3.06e-15 | 1.83e-23 |
| 3 | TCF4_1 | 17 | 46 | 0.18 | 0.57 | 1.27e-02 | 4.61e-01 | 2.23 | 1.41 | 2.94e-14 | 1.19e-24 |
| 4 | TCF3_7 | 16 | 44 | 0.19 | 0.55 | 2.13e-02 | 6.52e-01 | 2.34 | 1.40 | 6.77e-15 | 6.2e-23 |
| 5 | ZEB1_known3 | 16 | 29 | 0.19 | 0.62 | 2.13e-02 | 2.65e-01 | 2.24 | 1.28 | 1.75e-12 | 5.66e-18 |
| 6 | RXRA_disc2 | 26 | 112 | 0.73 | 0.37 | 2.9e-02 | 5.89e-03 | 2.09 | 1.09 | 2.95e-65 | 5.85e-79 |
| 7 | STAT_disc7 | 43 | 120 | 0.67 | 0.42 | 3.15e-02 | 8.24e-02 | 2.02 | 0.88 | 6.04e-51 | 1.32e-40 |
| 8 | MA0471.1 | 20 | 44 | 0.75 | 0.55 | 4.14e-02 | 6.52e-01 | 2.18 | 0.85 | 9.33e-28 | 6.67e-16 |
| 9 | THAP1_disc1 | 16 | 36 | 0.25 | 0.72 | 7.68e-02 | 1.13e-02 | 2.11 | 0.97 | 4.91e-21 | 1.86e-19 |
| 10 | SMC3_disc1 | 22 | 82 | 0.68 | 0.35 | 1.34e-01 | 1.06e-02 | 2.06 | 1.27 | 1.77e-29 | 5.05e-51 |
| 11 | SP4_2 | 28 | 77 | 0.64 | 0.35 | 1.85e-01 | 1.17e-02 | 1.89 | 0.98 | 1.86e-28 | 2.31e-36 |
| 12 | SP8_1 | 18 | 40 | 0.67 | 0.30 | 2.38e-01 | 1.66e-02 | 1.79 | 0.85 | 1.08e-17 | 6.53e-19 |
| 13 | SP1_known5 | 38 | 104 | 0.61 | 0.38 | 2.56e-01 | 1.38e-02 | 2.02 | 0.93 | 3.3e-52 | 3.73e-47 |
| 14 | SP1_known6 | 38 | 96 | 0.58 | 0.38 | 4.18e-01 | 1.84e-02 | 2.08 | 0.92 | 5.37e-46 | 7.63e-38 |
| 15 | EGR1_known12 | 30 | 78 | 0.57 | 0.36 | 5.85e-01 | 1.69e-02 | 2.01 | 0.79 | 5.84e-33 | 1.76e-21 |
| 16 | CCNT2_disc2 | 40 | 106 | 0.55 | 0.37 | 6.36e-01 | 8.42e-03 | 1.96 | 0.87 | 1.77e-33 | 1.74e-28 |
| 17 | KLF7_1 | 29 | 76 | 0.45 | 0.37 | 7.11e-01 | 2.86e-02 | 1.99 | 0.95 | 1.97e-36 | 2.62e-38 |
| 18 | IRF_disc4 | 40 | 102 | 0.50 | 0.39 | 1e+00 | 3.71e-02 | 1.98 | 0.95 | 1.16e-44 | 3.37e-47 |
| 19 | HF1H3B_1 | 47 | 149 | 0.51 | 0.42 | 1e+00 | 4.89e-02 | 1.99 | 0.80 | 1.02e-38 | 3.38e-29 |
| 20 | SP1_known8 | 18 | 42 | 0.50 | 0.31 | 1e+00 | 1.95e-02 | 2.05 | 0.91 | 1.03e-28 | 6.13e-25 |
| 21 | CACD_2 | 27 | 56 | 0.52 | 0.34 | 1e+00 | 2.22e-02 | 2.22 | 0.96 | 2.94e-25 | 1.75e-20 |

**Table 4.1** TF QTL effects.

# Chapter 5

# Effects of GWAS loci on molecular traits

## 5.1 Introduction

In the previous chapter, I discovered molecular trait QTLs and analysed aggregate trends using the single locus most strongly associated with a molecular trait. In order to investigate the effects of GWAS loci on molecular traits, I could identify cases where the previously generated QTLs and the GWAS tag SNP are in high LD. However, if I simply took all SNPs associated with a molecular trait and overlapped them with GWAS tag SNPs, I would not guard against cases where the GWAS SNP accounts for variance that can be attributed to other statistically independent variants in the region. In such instances, after conditioning on the alternative, independent loci, the effect of the GWAS tag SNP could be drastically reduced or removed entirely. Furthermore, using such methods would overlook instances where a GWAS locus has nuanced effects on a molecular trait that are only apparent after conditioning on other loci associated with a molecular trait.

In order to avoid such issues, I performed a conditional QTL analysis, described in this chapter, where I calculated the effect of the T2D and T2D-related GWAS loci conditioned on all other significant QTLs within the region. Based on this analysis, I found numerous QTLs that overlap GWAS loci. I summarise the trends of the most strongly associated loci through a series of vignettes. These vignettes also serve to generate hypotheses for further experimental follow up.

In addition, within these results, I also found many cases where a GWAS tag SNP was associated with both gene expression and DNAme, which were themselves associated with each other (i.e., eQTM). Such overlaps elicit questions in regards to the relationship between genetic effects on gene expression and DNAme. Does a variant affect gene expression and DNAme independently? Does a variant affect gene expression through DNAme? Or, does a variant affect DNAme through expression? To address these questions, the final part of this chapter describes a genome-wide mediation test of the genetic effects on gene expression and DNAme.

## 5.2 Conditional QTL mapping

As described in the introduction to this thesis, the overarching goal of this study is to interrogate the genetic effects of T2D and T2D-related GWAS loci on skeletal muscle molecular traits. To do this, I used an approach from Scott et al. [341] and calculated the effect of the GWAS tag SNP conditioned on all other significant, proximal QTLs. For each GWAS locus I fit two models: (1) a marginal model without any conditional SNPs (identical to Equation 4.1 in Chapter 4) and (2) a conditional model, herein described. Let $y$ be rank-based inverse normalised PEER residuals of feature $j$ across individuals, using the optimised number of PEER factors (75 for eQTLs/exQTLs and 60 for mQTLs). I calculated the conditional GWAS tag SNP effect using the following linear model:

$$y_j = \underbrace{\alpha_j \mathbb{1}}_{\text{intercept}} + \underbrace{\gamma_j Z}_{\text{SNP covariates}} + \underbrace{\beta_j g}_{\text{GWAS SNP eff.}} + \underbrace{\psi_j}_{\text{noise}}, \quad \psi_j \sim \mathcal{N}(0, \sigma_e^2 I) \tag{5.1}$$

Where $\alpha_j$ is the intercept, $Z$ denotes the matrix design of covariates from SNP $i$ to SNP $k$, $g$ the genotype vector of the GWAS tag SNP, $\psi_j$ Gaussian noise, and $\gamma_j$ and $\beta_j$ the effects of SNP covariates and the genotype effect respectively. I controlled for the number of tests per QTL type (gene expression, exon, DNAme) using Storey's FDR [369] and the conditional p-values.

In order to select the SNPs to include in the conditional model, I followed the procedures described in Scott et al. [341]. Briefly, initialising with a GWAS locus represented as the

GWAS tag SNP, I located all possible features (gene expression, exon expression, and DNAme) within a $+/-$ 1 Mb window. After identifying all features, I built the complete conditional model per feature, starting with only the GWAS tag SNP in model and iteratively adding the most strongly associated proximal SNP. I continued this stepwise forward selection process until the minimum p-value across all proximal SNPs was > a threshold. I set the threshold using the p-value from the primary QTL analysis corresponding to a 5% FDR (0.0026 for eQTLs, 0.0012 for exQTLs, and 0.00058 for mQTLs). I chose a 5% FDR instead of 1% as it would potentially allow more SNPs into the model and therefore be more conservative. The entire algorithm is outlined below in pseudocode.

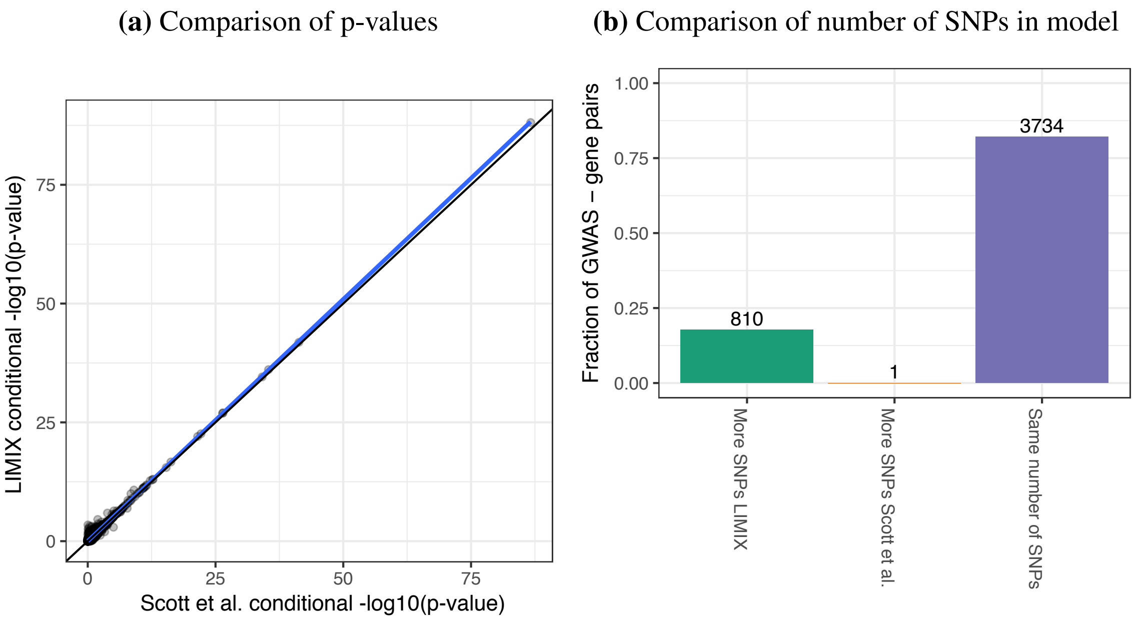
**Algorithm 1** GWAS conditional QTL method

```
model ← $y_j = \alpha_j \mathbb{1} + \beta_j g_{\mathrm{GWAS}} + \psi_j$        ▷ Equation 4.1
$p_{\mathrm{min}}$ ← threshold        ▷ Force first iteration regardless of initial minimum p
while $p_{\mathrm{min}} \leq$ threshold do        ▷ Add additional variants to model
    model ← model $+ \beta_j g_{\mathrm{p\ min}}$        ▷ Update model with minimum variant
    $p_{\mathrm{min}} \leftarrow min(p_{\mathrm{model}})$        ▷ Update minimum p across proximal variants
end while
return model        ▷ Model now contains all conditional variants
```

After constructing the complete model, I calculated the effect of the GWAS tag SNP conditioned on all other SNPs in the model, as described in Equation 5.1.

### 5.2.1 Pipeline validation

Similar to the QTLs from Chapter 4, I first validated the conditional QTL LIMIX pipeline using the Scott et al. [341] data freeze (Section 2.8). I compared the LIMIX conditional QTLs to the Scott et al. [341] conditional QTLs (Figure 5.1), which were called using an ordinary least squares (OLS) model implemented in the Python statsmodels package (http://www.statsmodels.org). Overall, the LIMIX conditional QTLs were nearly identical to those from Scott et al. [341], except for a few cases. For nearly all cases where the two methods varied, the LIMIX model included a few additional SNPs beyond the SNPs included in the OLS model. In these instances, the OLS method encounters an iteration where the minimum p-value across all SNPs is > the cutoff and therefore ceases to add SNPs to the model; however, for that same iteration LIMIX continues, because the minimum p-value is $\leq$ the cutoff. Similar to what was observed with the matrix eQTL/LIMIX comparison from Chapter 4, the LIMIX p-values are slightly smaller, perhaps due to the likelihood-ratio test used by LIMIX. The overall effect is that SNPs with p-values just barely over the threshold in OLS are included when using LIMIX. The one instance where more SNPs were in the OLS model, occurred with rs849134 and *HOXA10*, were LIMIX included 6 SNPs and Scott et al. [341] included 7. In this instance, the LIMIX SNP choices were identical to Scott et al. [341] in all but 2 of the total 6 SNPs (in addition to the 7th SNP included in OLS but not LIMIX). I continued with the LIMIX method as it (1) slightly increased power, and (2) is much faster computationally, which is critical for conditional mQTL mapping as there are many more EPIC methylation sites than genes.


**(a)** Comparison of p-values

**(b)** Comparison of number of SNPs in model

**Figure 5.1** LIMIX conditional QTL pipeline QC. (a) Comparison of conditional $-\log_{10}$(p-value) from LIMIX to Scott et al. [341]. (b) Comparison of the number of SNPs included in the final model.

## 5.3 Conditional QTL results: T2D and T2D-related traits

Using the input data from the primary QTL mapping described in Chapter 4, I mapped conditional eQTLs, exQTLs, and mQTLs for T2D and T2D-related traits from a manually curated T2D database and the EBI-NHGRI GWAS catalogue (v1.0.1 downloaded 24 April 2017) [231]. For T2D-related traits I included entries with glucose, insulin, A1C, HOMA-B, HOMA-IR, metabolic syndrome, and metabolic rate measurements in the "MAPPED_TRAIT" variable of the EBI-NHGRI GWAS catalogue. I used all unique entries, not pruning for LD, and testing all proximal features, regardless of whether or not they had a QTL in the primary analysis (Equation 4.1).

In the final models, I found expression related QTLs had on average more conditional SNPs than mQTLs, with a mean of 3.6, 1.8, and 1.1 conditional SNPs for gene, exon, and DNAme molecular traits respectively (Figure 5.2). The eQTL and mQTL with the most SNPs in the final model was the T2D GWAS tag SNP rs1133146 with 23 SNPs for *NDUFA3* and 15 SNPs for cg04599149 (maximum of 14 SNPs for a *NDUFA3* exon). The exQTL with the most SNPs was 16 SNPs for rs2302063 and an *AES* exon. However, none of these genes, exons, or DNAme sites were significantly associated with the GWAS tag SNP in the final conditional model, where the tag SNP association is evaluated conditioned on the other SNPs (FDR 1%). Of the significant conditional QTLs, the cases with the most conditional SNPs were rs12933472-*CDH13* (eQTL) with 17 SNPs, rs2946504-*FAM66D* and rs12933472-*PLCG2* both with 11 SNPs (exQTL), and rs895636-cg15044760 (near *SIX3*) with 13 SNPs (mQTL). For all 4 of these cases, before conditioning on additional SNPs, none of the associations were significant; the GWAS SNP associations became significant only after conditioning on these additional SNPs.

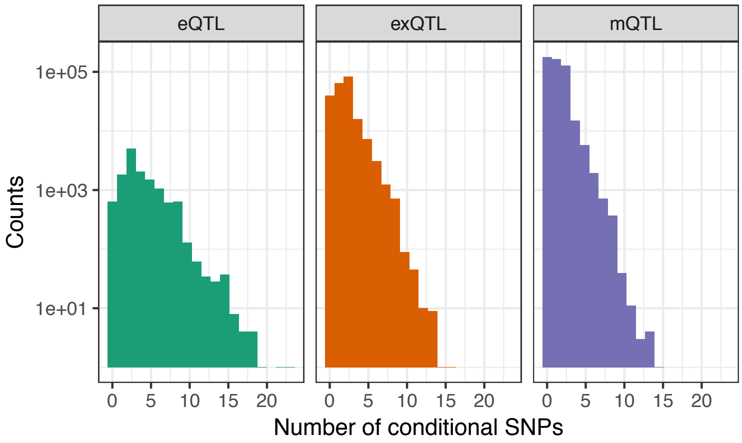


**Figure 5.2** Number of conditional SNPs in model.

Across all QTLs, I found many cases where conditioning on additional SNPs removed the marginal association, as well as cases where including the additional SNPs strengthened the conditional association compared to the marginal association (Figure 5.3). As shown in Figure 5.3, there are many points (GWAS SNP-molecular trait pairs; see figure legend) that aggregate along the x axis with a small $-\log_{10}(\text{p-value}_{\text{conditional}})$ and a large $-\log_{10}(\text{p-value}_{\text{marginal}})$. These points are instances where the GWAS tag SNP was associated with the molecular trait before conditional analysis (small marginal p-value); however, after conditioning on other SNPs in the region, the association of the GWAS locus was removed. Although less frequent, there are also instances where the conditional association was approximately equal to or greater than the marginal association. Such cases suggest the locus tagged by the GWAS SNP significantly affects the target molecular trait. Instances where the association was increased in the conditional model indicate that the additional SNPs helped to isolate the GWAS locus effect by removing variability caused by additional genetic effects.

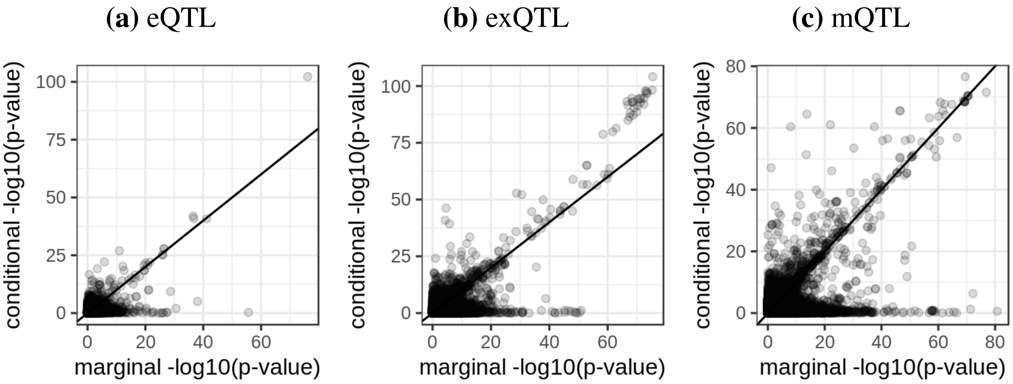


**Figure 5.3** Conditional QTL results. Each point represents pairing of a GWAS tag SNP and a molecular trait feature (gene, exon, or DNAme site). Comparison of $-\log_{10}$(p-value) from the marginal model (x axis) to the conditional model (y axis). (a) Genes (eQTLs). (b) Exons (exQTLs). (c) DNAme sites (mQTLs).

Overall at a FDR 1%, I found 337 significant eQTLs, 2,351 exQTLs, and 2,414 mQTLs (Table 5.1). The top 25 results for each molecular trait based on the conditional p-value are recorded in Tables 5.2, 5.3, 5.4.

| QTL type | GWAS-trait pairs | Significant GWAS-trait pairs |
|---:|---:|---:|
| eQTL | 13629 | 337 |
| exQTL | 215510 | 2351 |
| mQTL | 492105 | 2414 |

**Table 5.1** Summary of conditional GWAS QTLs.

## 5.3.1 Conditional QTL overview

I characterised each of these top results (Tables 5.2, 5.3, 5.4) by analysing the molecular trait relationships across gene, exon, and DNAme signals (i.e., conditional QTL maps and eQTMs/exQTMs). However, before considering these results, two caveats should be highlighted. First, the method I have used does not fine map candidate functional variants, meaning within the results, the actual functional variant(s) could be in high LD with the GWAS tag SNP. Second, the method I have used does not accurately model instances where a GWAS locus harbours multiple, conditionally independent causal variants. Nonetheless, given these caveats, the results I now describe are useful for identifying overlapping GWAS and molecular trait genetic signals, as described in the introduction to this chapter.

Unsurprisingly, I found that nearly all of the top eQTLs also had exQTLs (and vice versa). In most cases, a subset of exons were more strongly associated than the overall gene level association. This could be because the exon level expression reduces additional variability introduced to gene level expression by aggregating over multiple transcripts. Thus, if a QTL affects the abundance of one particular transcript but not another, the genetic effect may be less strong at the gene level (which aggregates over all transcripts) than at the exon level (which would capture expression patterns at exons unique to a specific transcript). The primary exceptions to this trend were lowly expressed genes where only gene level expression passed minimal expression filters (not exons).

In cases where the GWAS locus was both strongly associated with gene expression (gene level or exon level) and DNAme, which were frequent, I generally found the expression feature was also associated with the DNAme signal (i.e., the DNAme site was an eQTM or exQTM). For instance, just within the top conditional mQTLs (Table 5.4) there were several cases (*WFS1*, *FADS1*, *TMEM99*, and *CAMK2B*) where the mQTLs were also top eQTLs or exQTLs (Tables 5.2, 5.3), and the DNAme site was also associated with the implicated gene. As described in Section 5.4, I found such cases were often driven by strong, independent genetic effects on both gene expression and DNAme.

Finally, there were several loci that showed high complexity, where the GWAS locus was strongly associated with multiple genes. For instance, the strongest eQTL and exQTL was an association between the 2-hour glucose variant, rs1019503, and the *ERAP2* gene (gene level cond. p = 6.01x10$^{-103}$).[1] However, this variant was also strongly associated with *CTD-2260A17.2* and *LNPEP* (cond. p = 1.14x10$^{-41}$ and cond. p = 1.65x10$^{-41}$, respectively), making it difficult to discern the regulatory effects linked to 2-hour glucose. Several other loci also showed similar patterns of multiple effects including *RCCD1* (with *PRC1-AS1*), *FADS1* (with *TMEM258* and *FADS2*), *RHOA* (with *NICN1* and many other genes), *CCHCR1* (with *TNXB*), *CYP21A2* (with *HLA-B* and many other genes), *TMEM99* (with *KRT10*), *INTS8* (with *PLEKHF2*), as well as several others. For further characterisation and follow up,

[1] This locus that has been identified across a number of tissues including islets [95, 393, 395] and LCLs [71]. For LCLs, rs2248374 (in high LD with rs1019503; r$^2$ > 0.99 1000GENOMES:phase_3:FIN) has been shown to affect splicing and create a premature stop codon. The isoform created is targeted for nonsense-mediated decay (reviewed in [234]).

I focused on cases where both the conditional eQTLs and exQTLs clearly indicated one primary gene—except for instances DNAme instances where no gene was indicated, as described below.

### 5.3.2 Conditional QTL vignettes and trends

In the following section, I present a series of vignettes or case studies at several GWAS loci and attempt to infer candidate causal mechanisms at each locus. The observations in this section serve to generate hypotheses for further experimental validation, beyond the scope of this thesis.

Focusing in instances where one gene was clearly implied (or no genes in the case of some DNAme signals), I further characterised these loci by integrating many sources of genomic and molecular trait information (CGIs, TF binding, chromatin states, ATAC-seq signal, etc.) as well as detailed literature review. I present these results below, summarising them through a series of vignettes, grouped according to their molecular trait trends. In these vignettes, I also highlight my own hypotheses on how each locus may contribute to disease risk, which proved to be rather unique to each locus. Such results suggest that automating GWAS loci hypothesis generation is not plausible and that understanding how each GWAS locus may contribute to disease risk will require detailed, manual consideration, integrating as many sources of information as possible.

Broadly speaking, scattered throughout these vignettes, one particularly interesting trend emerged: the ability to use DNAme to generate hypotheses in regards to candidate regulatory effects underlying a GWAS locus by honing in on specific genomic regions. For instance, with mQTLs I was able to identify: (1) cases where the underlying molecular mechanism likely involves differential TF binding at the canonical promoter of a gene (which often fell just outside of the core CGI region of the promoter; e.g., *WFS1*, *JAZF1*), (2) cases with candidate alternative promoters or splicing effects (e.g., *CAMK2B*, *FGGY*, *ANK1*), and (3) cases involving distal gene regulation that may indicate chromatin interactions (e.g., *SDHAF4*, *KCNJ11*). Such trends are consistent with a recent study that suggests the regulatory chromatin landscape can be, to some extent, recapitulated using DNAme [106]. Moreover, in a few instances with particularly strong DNAme signals, I found a nearby common variant that may perturb TF binding and is a good candidate for being a causal variant. In one instance, *ANK1*, this candidate variant had already been identified independent of DNAme trends [341] and shown to perturb a TR4 TF binding site in a skeletal muscle promoter that regulates *ANK1* expression.

In addition, I found that few T2D-related GWAS risk loci appear to function exclusively in skeletal muscle. For instance, I found several loci that have an established, T2D-related effect within a non-muscle tissue—often pancreatic islets (e.g., *CAMK2B*/*GCK*, *PROX1*/*PROX1-AS1*, *WFS1*, *JAZF1*, *KCNJ11*, *ACHE*). There are several ways to interpret such results, often varying in a case-by-case manner. It could be that the skeletal muscle effects are totally irrelevant to the GWAS effect (e.g., possibly *ZFAND3*). Alternatively, given that genetic effects on gene expression [136, 104, 138] and potentially DNAme [357, 142] are shared across tissues, the skeletal muscle trends may be informative for narrowing down regulatory effects that are then causal to disease in the right tissue (or environmental) context (e.g., possibly *WFS1*). Finally, many of these cases may describe molecular pleiotropic effects, where the same variant has different effects, depending on the tissue context. It could be that the multiple effects across tissues collectively contribute to disease risk (e.g., possibly *CAMK2B*/*GCK*). Distinguishing between these possibilities is not directly possible within this dataset that focuses primarily on skeletal muscle. However, as the genomics community continues to generate large scale multi-tissue QTL datasets, like GTEx, we will gain a better understanding of the genetic architecture of molecular traits across tissues and be better positioned to interpret molecular effects of GWAS loci.

#### 5.3.2.1 mQTL with no gene related QTL

Within the top mQTLs, I found several cases (e.g., cg03523917, cg20564521, cg08035822, cg13182339) with an extremely strong effect on DNAme, but a very weak or non-existent effect on gene expression at the gene or exon level. For instance with rs4712523 (or any of the other variants in LD with rs4712523 from Table 5.4), I found only one strong DNAme effect, cg03523917 (cond. p = $2.93\text{x}10^{-69}$), and no significant effect on gene or exon expression both in the marginal and conditional analysis (Figure 5.4). In contrast to expression, rs4712523 was also slightly associated, but significant nonetheless, with 3 additional DNAme sites within the region. rs4712523 and the various other tag SNPs in LD (from various GWA studies) fall in *CDKAL1* introns. Likewise, cg03523917 lies in a *CDKAL1* intron that is in a weak transcription chromatin state in skeletal muscle. Across the various ENCODE cell types, there are 6 ChIP-seq peaks immediately before cg03523917, but none overlapping the DNAme site. Several common variants fall in these ChIP-seq peaks;

the closest common variant to cg03523917 is rs9368218 which is in high LD with rs4712523 ($r^2 > 0.99$ 1000GENOMES:phase_3:FIN). Although speculative, these mQTLs without clear gene effects may identify regulatory effects within a region that are poised and then amplified within the right tissue or potentially environmental context, like the caQTLs described in the introduction to this thesis [3]. However, without additional experiments the nature of these effects is difficult to interpret.

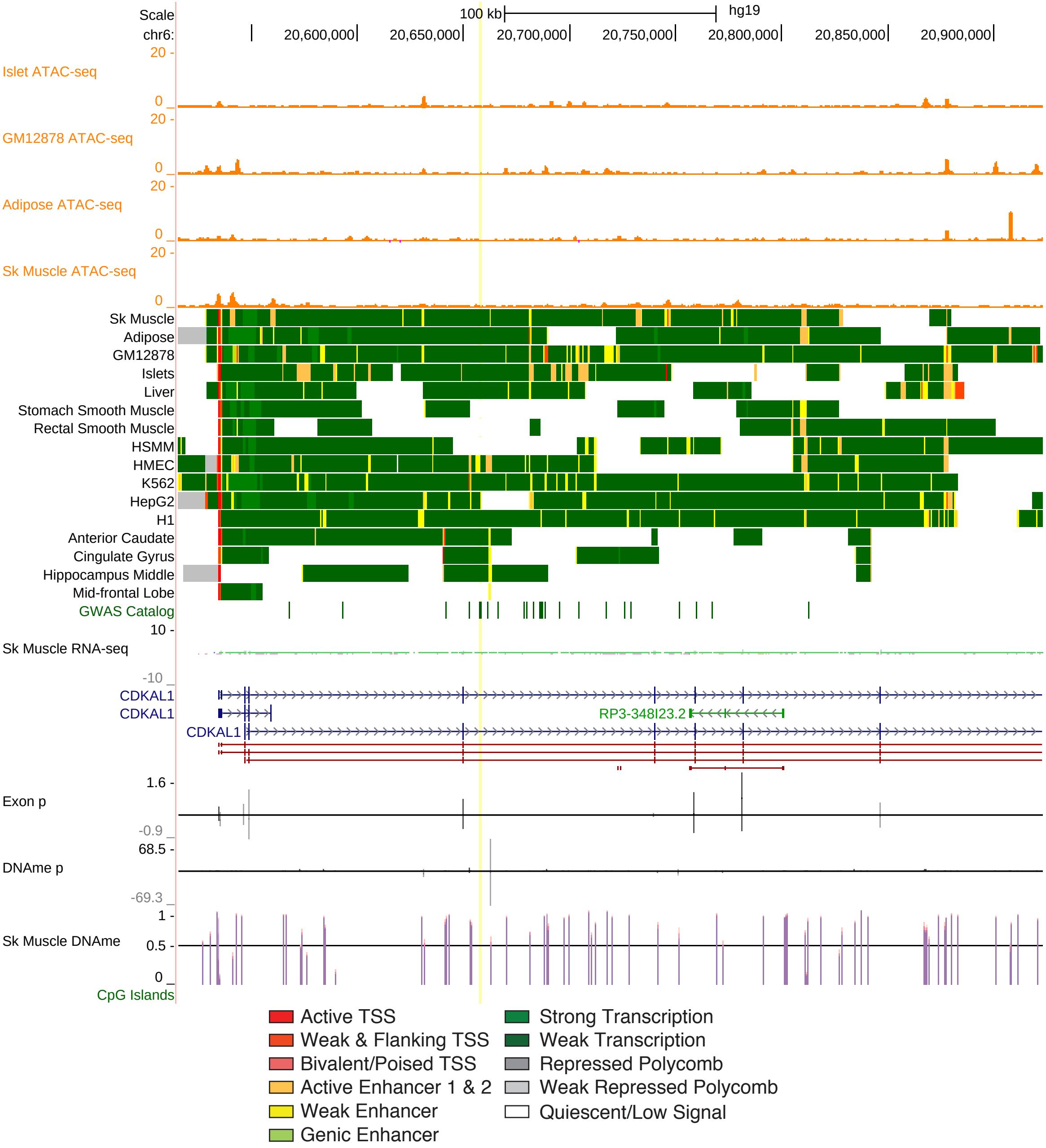


**Figure 5.4** *CDKAL1* locus. Light yellow vertical line shows the position of rs4712523. The top orange tracks show ATAC-seq data. The middle tracks with only cell/tissue type names are chromatin states. These tracks are followed by the average skeletal muscle RNA-seq signal, the gene annotation, the exon marginal/conditional p-values, the DNAme marginal/conditional p-values, the average skeletal muscle DNAme signal, and finally CGI annotation. Both the marginal and conditional p-values are displayed as vertical bars in the "Exon p" and "DNAme p" tracks. The marginal p-values are shown below the horizontal line ($\log_{10}$ scale). The conditional p-values are shown above the horizontal line ($-\log_{10}$ scale). Note that exon fragments were tested. There may be more exon bars than exons in the gene annotation track as each unique exon fragment constitutes a bar. Also, due to minimum expression requirements, not all exons were tested for associations. In cases where an exon was not tested, there will not be a vertical line in the "Exon p" track.

#### 5.3.2.2 exQTL with no eQTL

I also identified three cases within the top QTL results, *CAMK2B, FGGY,* and *PROX1-AS1*, that exhibit a strong conditional exQTL signal, but no eQTL signal. This pattern could be due to specific alternative transcript effects that are lost when analysing aggregate gene level signal.

***CAMK2B* (*GCK* T2D locus)** In the case of *CAMK2B* and rs3757840 (which occurs in *GCK* and is associated with T2D), I found rs3757840 is significantly associated with gene level and exon level expression of *YKT6*, a gene with only a few exons that occurs between *GCK* and the 3' end of *CAMK2B* (minimum exon cond. p = 6.72x10$^{-18}$; Figure 5.5). However, rs3757840 is most strongly associated with several *CAMK2B* exon fragments near the 3' end of *CAMK2B*, the strongest of which is an exon fragment within ENSE00001744696 (cond. p = 2.96x10$^{-26}$; chr7:44259016-44259028).[2] Furthermore, rs3757840 is a strong conditional mQTL for 7 DNAme sites, primarily located at the 3' end of *CAMK2B*. Notably, none of these DNAme sites are associated with gene level or exon level expression of *YKT6*. The only DNAme sites associated with any expression related feature are the two DNAme sites with the strongest mQTL effect, cg06032855 (cond. p = 3.87x10$^{-61}$) and cg21330313 (cond. p = 1.59x10$^{-17}$). Both cg06032855 and cg21330313 are exQTMs for the *CAMK2B*:ENSE00001744696 fragment. The weaker DNAme site, cg21330313, lies immediately before the *CAMK2B*:ENSE00001744696 fragment, while the stronger site, cg06032855, lies in a CGI (chr7:44259670-44259923) within a preceding exon, *CAMK2B*:ENSE00001522684 (whose expression levels are not associated with rs3757840). Both DNAme sites are in a weak transcription chromatin state in a variety of tissues including skeletal muscle. Within *CAMK2B*:ENSE00001522684 and closest to cg06032855 (the strongest DNAme association) are two synonymous variants, rs1127065 and rs1065359, of which rs1065359 is somewhat in LD with rs3757840 ($r^2$ = 0.73 1000GENOMES:phase_3:FIN). Finally, in H1 there is a REST binding site from ChIP-seq data at the start of *CAMK2B*:ENSE00001522684 (chr7:44259848-44259951). REST binding has been linked to increased levels of methylation [364]. Such patterns are consistent with cg06032855 which shows an average methylation of 0.83 across all samples (standard deviation 0.06).

[2] All coordinates are on GRCh37/hg19.

At this locus there are many indications of tissue specificity. As described earlier in this thesis, the Collins laboratory and others recently identified stretch/super enhancers—large enhancers that mark genomic regions associated with tissue specific activity, likely composed of a series of many TF binding events in a restricted region [287, 421, 229]. In our stretch enhancer paper, we noted a large stretch enhancer round *GCK* in islets [287]. Indeed, *GCK* expression is highly specific to islets [395] and absolutely critical for glucose sensing and insulin response in islets (reviewed in [242]). In addition, *GCK* is also expressed in liver, where it regulates glucose storage, highlighting the possibility of a single genetic effect contributing to disease risk based on physiological effects in multiple tissues (reviewed in [242, 392]).

Within this region, I also found several intragenic *CAMK2B* skeletal muscle stretch enhancers, one of which (chr7:44255000-44258800) falls just beyond the *CAMK2B* exon strongly associated with rs3757840 (there is an additional noteworthy skeletal muscle stretch enhancer, chr7:44299400-44323400, that also shows some activity in brain and islets). In addition, *CAMK2B* is expressed at very high levels in skeletal muscle in a tissue specific manner (mESI 0.57). Interestingly, noting the surrounding chromatin states, the most muscle specific *CAMK2B* chromatin patterns fall upstream of the observed splicing effect. While poorly explored in the context of T2D (given such a clear and T2D-relevant effect at *GCK*), *CAMK2B* is linked to the response of skeletal muscle to exercise [323]. Also in skeletal muscle, CAMK2B is reported to interact with IRS1, a key mediator in the insulin signalling cascade [55]. Therefore, *CAMK2B* may constitute an additional, molecular pleiotropic effect contributing to T2D risk at this locus.

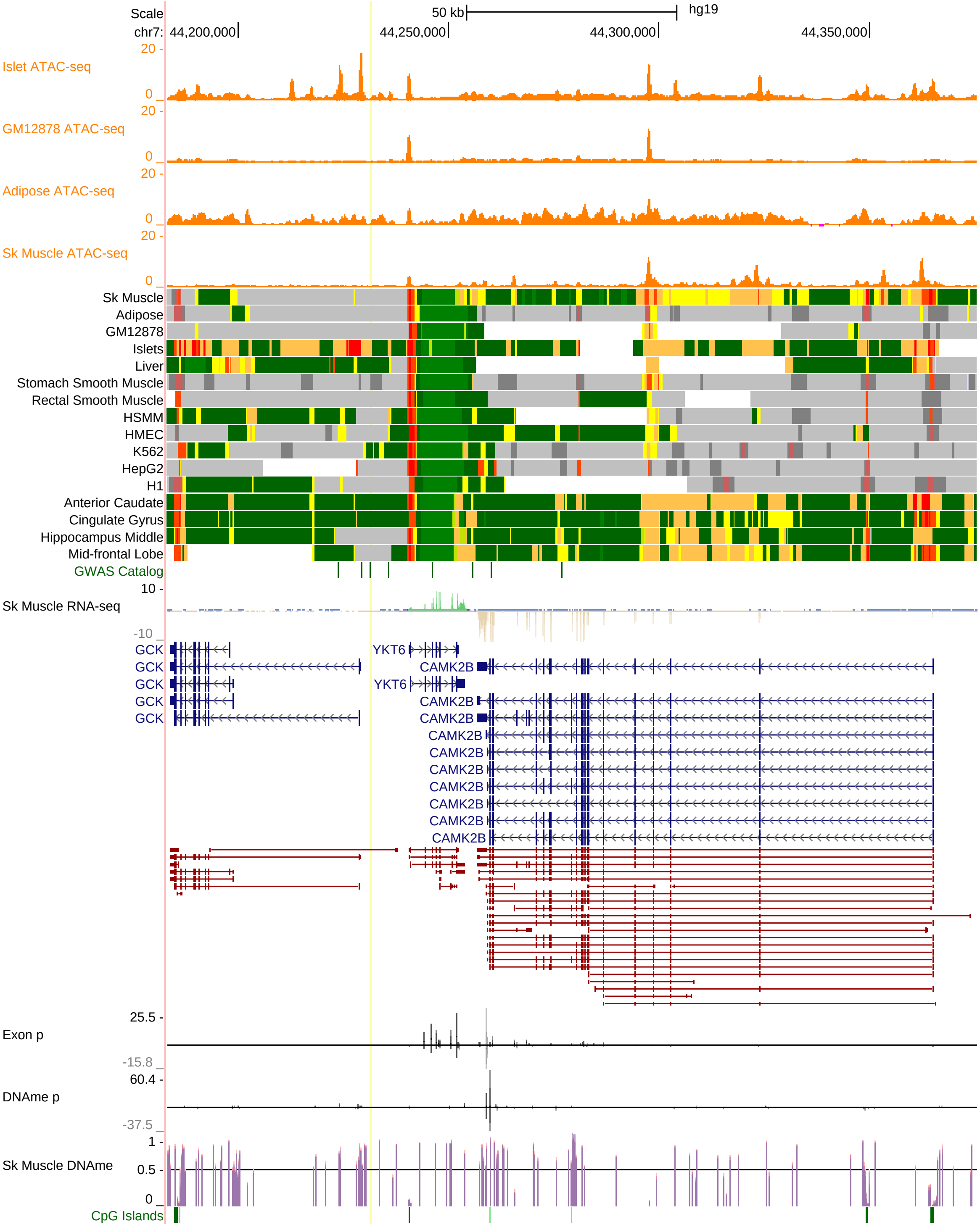


**Figure 5.5** *CAMK2B* (*GCK* T2D locus). Light yellow vertical line shows the position of rs3757840. I found rs3757840 was strongly associated with a *CAMK2B* exon and a nearby DNAme site. There are also several associations between rs3757840 and *YKT6* exons, but these are not as strong as the *CAMK2B* exon association. Detailed figure legend can be found in Figure 5.4.

***FGGY*** Similar to *CAMK2B*, I found an obesity linked SNP, rs835367 (which falls in the *FGGY* promoter in a striking dip between two skeletal muscle ATAC-seq peaks), is strongly associated with three 5' end *FGGY* exon fragments (spanning chr1:59787208-59805741 corresponding to ENSE00003609460 and ENSE00003592558), but not overall *FGGY* expression (Figure 5.6). Layering in the DNAme information revealed two associated intergenic DNAme sites: cg09869950 (cond. p = 2.19x10$^{-23}$) and cg18580450 (cond. p = 5.36x10$^{-10}$). cg09869950 occurs immediately before *FGGY*:ENSE00003609460, the strongest exon association with rs835367 (cond. p = 1.49x10$^{-53}$), and is an exQTM for the *FGGY*:ENSE00003609460 exon fragment. cg09869950 falls in a skeletal muscle intergenic weak enhancer state, surrounded by two weak TSS states, as well as multiple common variants and TF ChIP-seq peaks. The chromatin states surrounding *FGGY* show muscle specific patterns, which is consistent with the high muscle specificity index for this gene (mESI 0.57). *FGGY* is a poorly studied gene, although very preliminary evidence in mice suggests it is linked to neurogenic skeletal muscle atrophy [76]. This is a potential locus where the obesity risk allele may lead to alternative splicing of the *FGGY* gene.

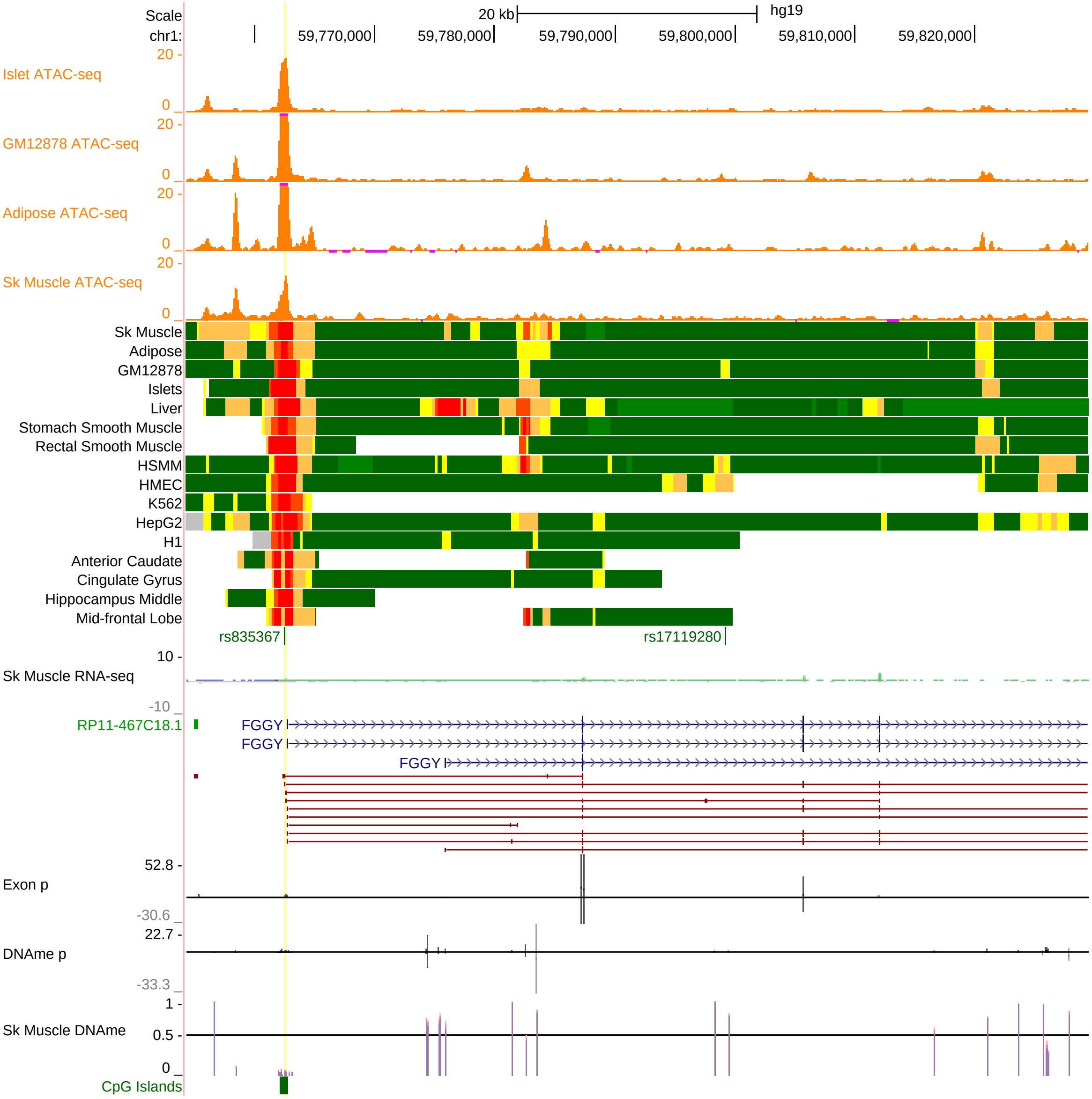


**Figure 5.6** *FGGY* locus. Light yellow vertical line shows the position of rs835367. I found a very strong association with two *FGGY* exon fragments, preceded by a strong DNAme association. Detailed figure legend can be found in Figure 5.4.

***PROX1-AS1*** Finally, at the *PROX1-AS1*/*PROX1* T2D locus, I identified two very strong DNAme signals associated with rs2075423—cg14810798 and cg05052969—that are also eQTMs and exQTMs for *PROX1-AS1* (Figure 5.7). rs2075423 falls in a non-coding exon of *PROX1-AS1* and, similar to *CAMK2B* and *FGGY*, is not associated with gene level *PROX1-AS1* expression after conditional analysis (*PROX1-AS1* cond. p = 0.54; *PROX1* cond. p = 0.03). However, rs2075423 is associated with the furthest 3' end exon of *PROX1-AS1*, ENSE00001785835 (cond. p = 4.43x10$^{-7}$). After conditional analysis, rs2075423 is significantly associated with 7 DNAme probes, localised at the shared *PROX1-AS1* and *PROX1* TSS region.

The two strongest DNAme associations are cg14810798 (cond. p = 2.99x10$^{-71}$) and cg05052969 (cond. p = 1.64x10$^{-68}$). Both cg14810798 and cg05052969 are 17 bp away from each other, at the tail end of a CGI (chr1:214156001-214156851), and fall in a bivalent promoter state common across many cell types, including skeletal muscle. Both probes overlap several ChIP-seq peaks from various ENCODE ChIP-seq experiments, including a CTCF ChIP-seq peak from A549 and HUVEC, where cg05052969 resides within the canonical CTCF motif (chr1:214156841-214156855). Immediately before this motif, 18 bp from cg05052969, is a common variant, rs235924, that is also in high LD with the T2D-tag SNP, rs2075423 ($r^2$ = 0.98 1000GENOMES:phase_3:FIN). rs235924 is the closest common variant to these probe sites and the only common variant that falls in this CTCF peak, potentially suggesting that this variant perturbs CTCF binding or a related TF.

Little is known about the role of *PROX1-AS1*. However, *PROX1* is known to be critical for the development of many tissues including eye, liver, pancreas, and the lymphatic system (reviewed in [89]). In rat, reduced Prox1 expression has also been shown to reduce glucose-stimulated insulin secretion in the INS-1E beta cell line [202]. Interestingly, Prox1 has also recently been shown to be necessary and sufficient for the differentiation of human and rodent muscle myosatellite cells into slow muscle fibres [186]. These resident myosatellite stem cells are critical for healthy muscle function as they enable the repair and regrowth of skeletal muscle in response to injury (myosatellite cells become myoblasts, which then become myocytes). Thus, similar to the *CAMK2B*/*GCK* locus, there is a clear islet-T2D connection at this locus, but also additional pleiotropic effects important to other tissues that could also potentially play a role in T2D pathogenesis.

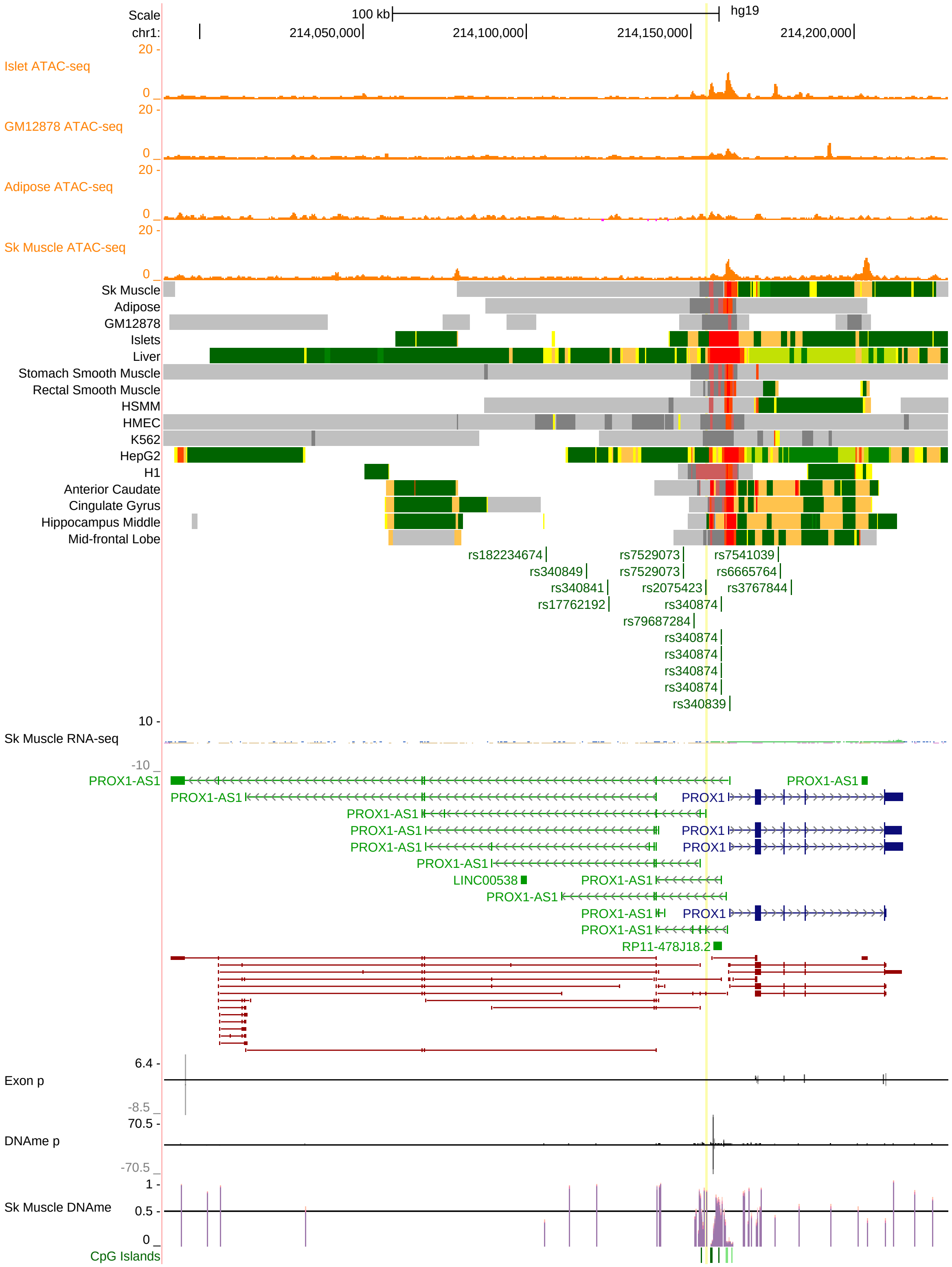


**Figure 5.7** *PROX1-AS1* locus. Light yellow vertical line shows the position of rs2075423. I found one association with the furthest 3' end exon of *PROX1-AS1*. There were also several, very strong DNAme associations clustered at the CGI of a shared *PROX1*/*PROX1-AS1* promoter. Detailed figure legend can be found in Figure 5.4.

#### 5.3.2.3 One clear gene from eQTL and exQTL

Within the top results, there were 9 cases that clearly point to a single gene at the level of whole gene and exon expression. However, several of these cases have rather weak, barely significant, DNAme associations including *SLU7*, *ZNF697*, *GLRX5, ZBTB20.* I did not characterise these cases further, and focused on other 5 associations with stronger DNAme signals.

***WFS1*** In *WFS1*, I found rs4689388 (associated with T2D) has strong effects on expression and DNAme (Figure 5.8). rs4689388 lies immediately before the promoter of *WFS1*. rs4689388 is associated with the total levels of *WFS1* expression (cond. p = 2.02x10$^{-22}$) as well as exons near the 3' end of the transcript (minimum cond. p = 5.34x10$^{-16}$). The GWAS tag SNP is also associated with 12 DNAme probes, the strongest of which cluster around *WFS1* promoter the on either side of a CGI (chr4:6271281-6272182): cg17816406 (cond. p = 2.72x10$^{-72}$), cg17662872 (cond. p = 3.99x10$^{-56}$), cg25554036 (cond. p = 4.04x10$^{-40}$), and cg08703151 (cond. p = 2.11x10$^{-32}$). These probes are also very strong exQTMs for the top exQTL exons near the 3' end. Of the 8 DNAme probes that fall within the CGI, none of them were associated with rs4689388 (minimum cond. p = 0.011). The strongest DNAme association, cg17816406, lies just beyond the canonical CGI, within the first *WFS1* intron. There were no ENCODE TF binding sites overlapping or immediately surrounding cg17816406. The closest common variants are rs11723602 and rs71173429, which is a 1 bp deletion, both in high LD with rs4689388 ($r^2 > 0.93$ 1000GENOMES:phase_3:FIN). In addition, there are several other common indels that lie within the CGI. For instance, near the beginning of the CGI is a frequent 6 bp deletion, rs148797429, that is in LD with the GWAS tag SNP ($r^2 = 0.88$ 1000GENOMES:phase_3:FIN). Finally, in GTEx v6p, rs4689388 is a significant eQTL for *WFS1* in ~8 different tissues, suggesting that the observed effect is not unique to skeletal muscle.

*WFS1* encodes for wolframin, a transmembrane protein that maintains $Ca^{2+}$ homeostasis in the endoplasmic reticulum (ER) [389]. Rare mutations in *WFS1* are of high penetrance, causing Wolfram syndrome (OMIM 222300), a disease characterised by neurological and endocrine dysfunction. Indeed, the first manifestation of Wolfram syndrome is typically diabetes by the age of 6 [389]. In addition, common variants (like rs4689388) have been identified at the *WFS1* locus that are associated with T2D [332, 123].

Both pancreatic beta cells and neuronal cells are particularly sensitive to ER dysfunction. Because of role of wolframin in normal ER function, mutations in *WFS1* have a large impact on these cells, leading to ER stress-associated cell death [389]. Such effects explain the clinical manifestations of Wolfram syndrome, involving neurological and endocrine dysfunction. They also suggest that beta cells are the critical tissue of origin for the observed *WFS1* T2D associations.

Given the general importance of ER function across all cell types, one would expect to find similar transcription and translation programs of key ER genes/proteins in most tissues. While the transcriptional rate or overall protein abundance may be dialed up or down, the overall programs themselves would not change. Indeed, *WFS1* fits such a model as there are very few alternative splice isoforms of *WFS1*. Thus, if a variant altered the transcription rate of a key gene like *WFS1*, one would expect to find similar effects across most cell types (since most eukaryotic cells require a functioning ER), while the actual effect most crucial to disease might be isolated to one cell type. With knowledge of *WFS1* function from Wolfram syndrome, the effects of rs4689388 on transcription, and the fact that the DNAme signals localise around the canonical promoter, it seems likely that these skeletal muscle effects are representative for many tissues, including beta cells. A reasonable hypothesis is that while the rs4689388 risk haplotype has similar molecular effects across all tissues, it gives rise to disease specifically through beta cells, given the crucial role of beta cells in endocrine function. In such a model, mutations that seriously damage *WFS1* function (for instance by changing the protein sequence) rapidly give rise to diabetes through increased beta cell death; however, variants that simply perturb the overall abundance levels increase the fragility of an individual's beta cell population that over time, combined with other genetic and environmental factors, predisposes one to T2D.

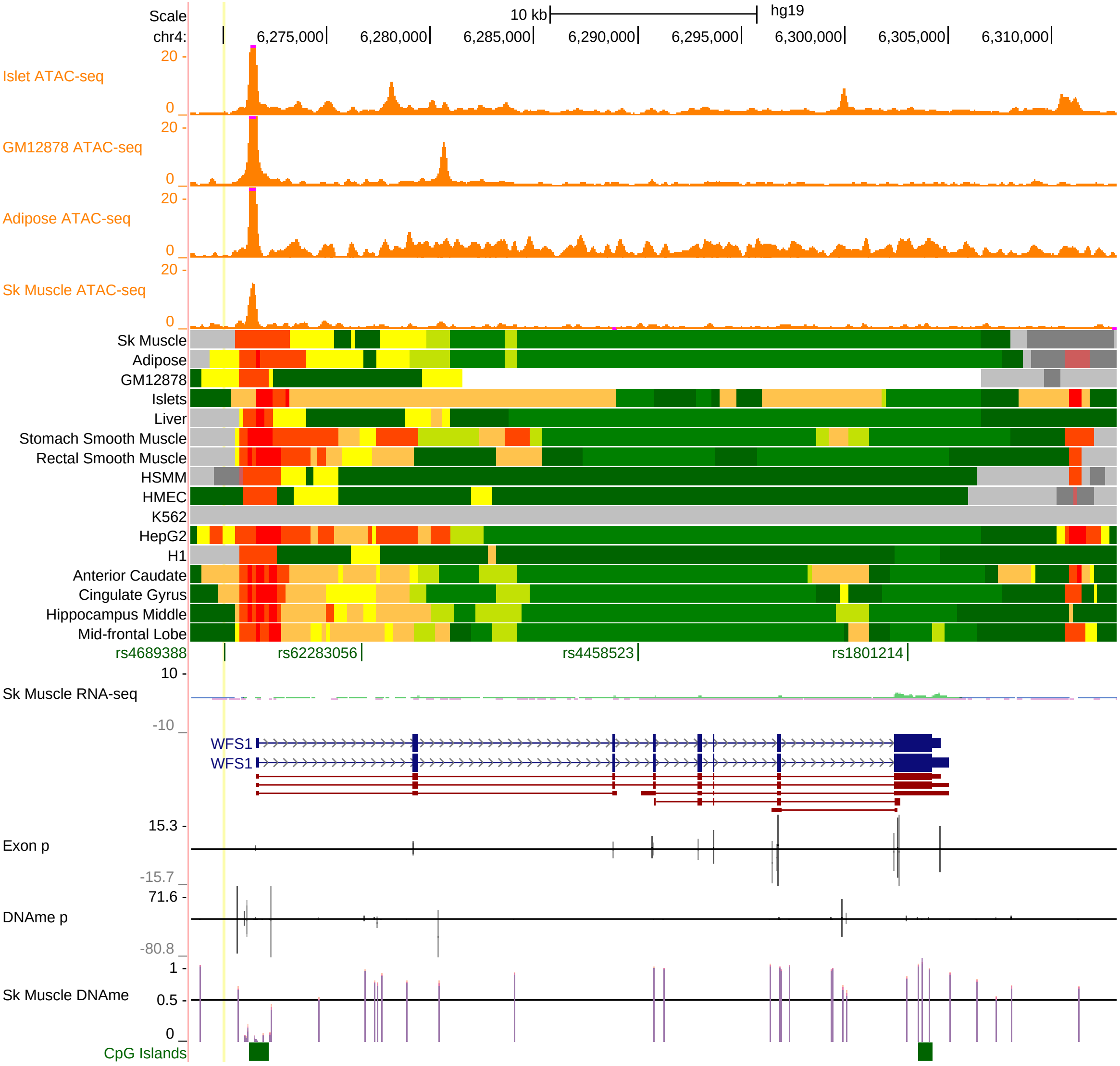


**Figure 5.8** *WFS1* locus. Light yellow vertical line shows the position of rs4689388. In addition to a *WFS1* eQTL effect, I found strong associations with exons near the 3’ end of *WFS1* (the most highly expressed exons). There were also very strong DNAme associations around either side of a CGI in the *WFS1* promoter. Detailed figure legend can be found in Figure 5.4.

***JAZF1*** rs849135 (which occurs in a *JAZF1* 5' intron) is associated with T2D and is significantly associated with gene level and exon level *JAZF1* expression (Figure 5.9). Like *WFS1*, the strongest exon associations occur at the 3' end of the *JAZF1* transcript and the strongest DNAme associations occur around the promoter. Of the 10 significantly associated DNAme sites, all of the strongest sites cluster around the promoter and are exQTMs for the exon most strongly associated with rs849135: cg01883759 (cond. $p = 8.50x10^{-62}$), cg21912938 (cond. $p = 1.41x10^{-40}$), and cg02010481 (cond. $p = 1.96x10^{-24}$). Both of the top associations, cg01883759 and cg21912938, occur 18 bp from each other and lie right before an annotated CGI (chr7:28220015-28220534). Within the ENCODE ChIP-seq data, these two probes overlap a MAZ (HeLa-S3) and EZH2 (GM12878 and HepG2) peak, of which EZH2 has been associated with DNAme through the recruitment of DNMTs [398]. None of the common variants near this cluster of DNAme probes are in high LD with the GWAS tag SNP (rs849135), perhaps indicating more distal regulatory effects.

JAZF1 (aka TIP27) primarily functions as a repressor of TR4 (aka NR2C2) mediated transactivation [270]. TR4 is a TF that targets many genes related to metabolism [277], including the highly muscle specific *ANK1* gene described below (Section 5.3.2.3). In mice, Jazf1 overexpression has been shown to suppress lipogenesis in adipose [256], inhibit gluconeogenesis in liver [169], block myoblast differentiation [436], and enhance overall insulin sensitivity [435]. Using MIN6 cells (a mouse beta cell line), PDX1, a very important pancreatic islet TF, has also been shown to preferentially bind to rs1635852 in the first *JAZF1* intron, also in LD with rs849135 ($r^2 = 0.75$ 1000GENOMES:phase_3:FIN) [105]. In human, higher levels of *JAZF1* islet expression are associated with increased insulin secretion, and T2D patients have lower *JAZF1* islet expression [383], which is consistent with T2D risk effect and the observed effect of rs849135 in islets [105]. Additionally, in GTEx v6p, both rs849135 and rs1635852 are significantly associated with *JAZF1* expression and have similar effects (T2D risk allele reduces expression) in several tissues including skeletal muscle, pancreas, adipose, artery, and colon. Clearly, there is a connection between islet *JAZF1* expression and T2D. However, given the importance of *JAZF1* in various T2D-related tissues, it seems likely that the T2D risk haplotype effects *JAZF1* expression in multiple tissues, possibly including skeletal muscle, all of which contribute to T2D pathogenesis.

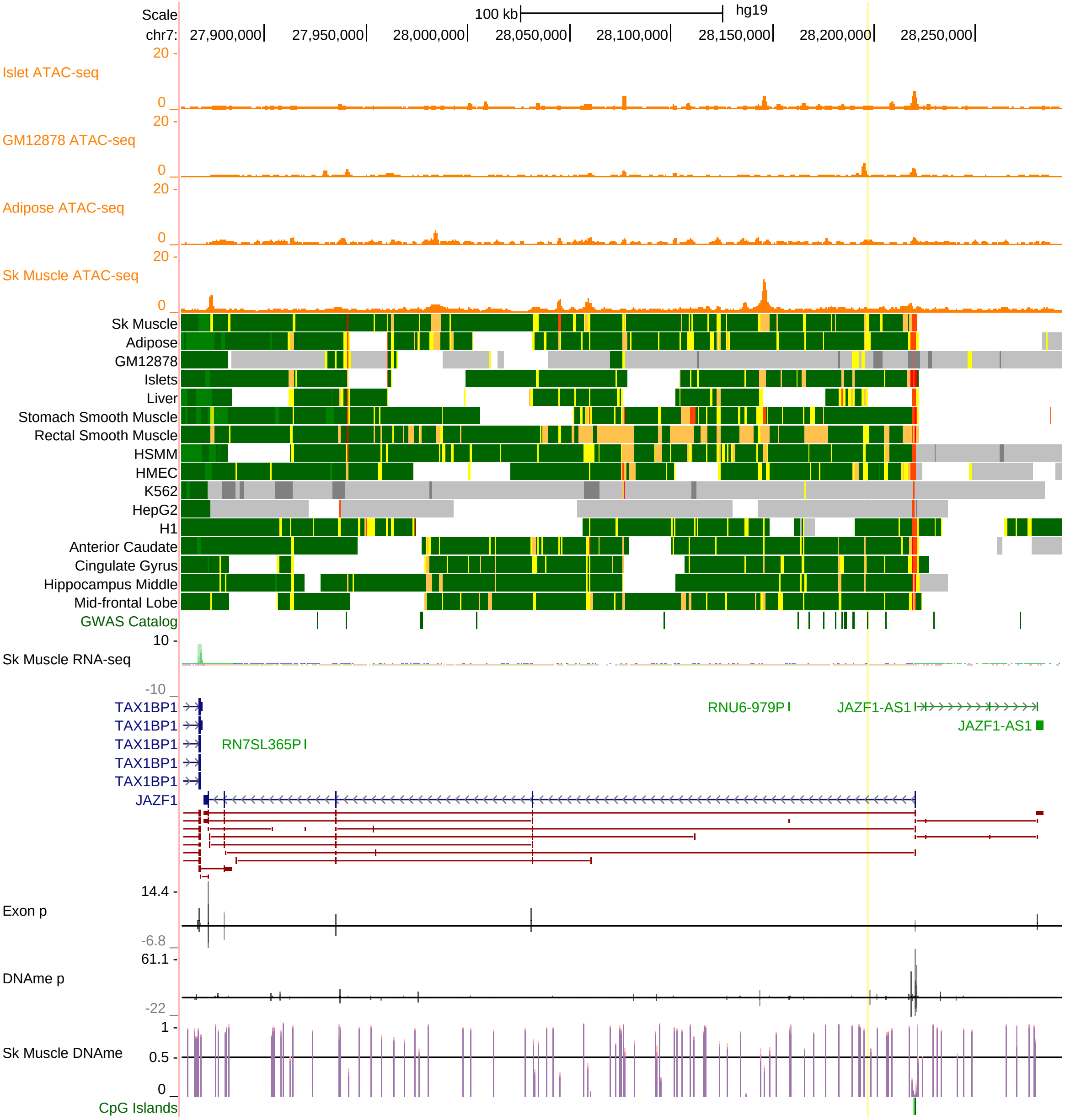


**Figure 5.9** *JAZF1* locus. Light yellow vertical line shows the position of rs849135. In addition to a *JAZF1* eQTL effect, I found strong associations with exons near the 3' end of *JAZF1*. I found several strongly associated DNAme signals around the shared *JAZF1*/*JAZF1-AS1* promoter. The strongest DNAme associations lie right before a CGI. Detailed figure legend can be found in Figure 5.4.

***ZFAND3*** I also identified rs9470794 (associated with T2D) as a *ZFAND3*-eQTL, *ZFAND3*-exQTL, and mQTL. rs9470794 lies in a 3' *ZFAND3* intron and is most strongly associated with the final, 3' *ZFAND3* exon fragments (Figure 5.10). After conditional analysis there were 9 DNAme sites significantly associated with rs9470794. Unlike *WFS1* and *JAZF1*, the strongest associated DNAme probe, cg22213309 (cond. $p = 5.44x10^{-36}$), did not occur in the promoter, but rather in an intragenic region within *ZFAND3*. cg22213309 is the only significant exQTM for the aforementioned *ZFAND3* exon fragments. cg22213309 occurs in a weak transcription chromatin state in most tissues, including skeletal muscle. However, for a few cell types, such as HMEC and K562, this CpG site occurs in either a weak enhancer or flanking TSS state. The CpG site overlaps 4 ChIP-seq peaks from various cell types, including JUND in K562, which has been associated with DNAme [274].

Very little is known about *ZFAND3* in tissues other than islets, where *ZFAND3* expression has recently been associated with insulin secretion [271]. Furthermore, in islets, a common variant in East Asian populations > 10 kb upstream of the *ZFAND3* promoter, rs58692659, has been shown to disrupt NEUROD1 binding, a key islet TF, thereby abolishing an important islet enhancer [288]. This variant is in high LD with the T2D tag SNP, rs9470794, which was discovered to be associated with T2D in East Asians [63]. The genetics of European populations support this being the causal SNP, as rs58692659 is monomorphic in European populations and no SNPs in this region (including rs9470794) are associated with T2D in Europeans (rs9470794 DIAGRAM $p = 0.8$ [63]). Therefore, in this instance the relevance of the rs9470794 association with *ZFAND3* in skeletal muscle is likely of little importance to T2D aetiology.

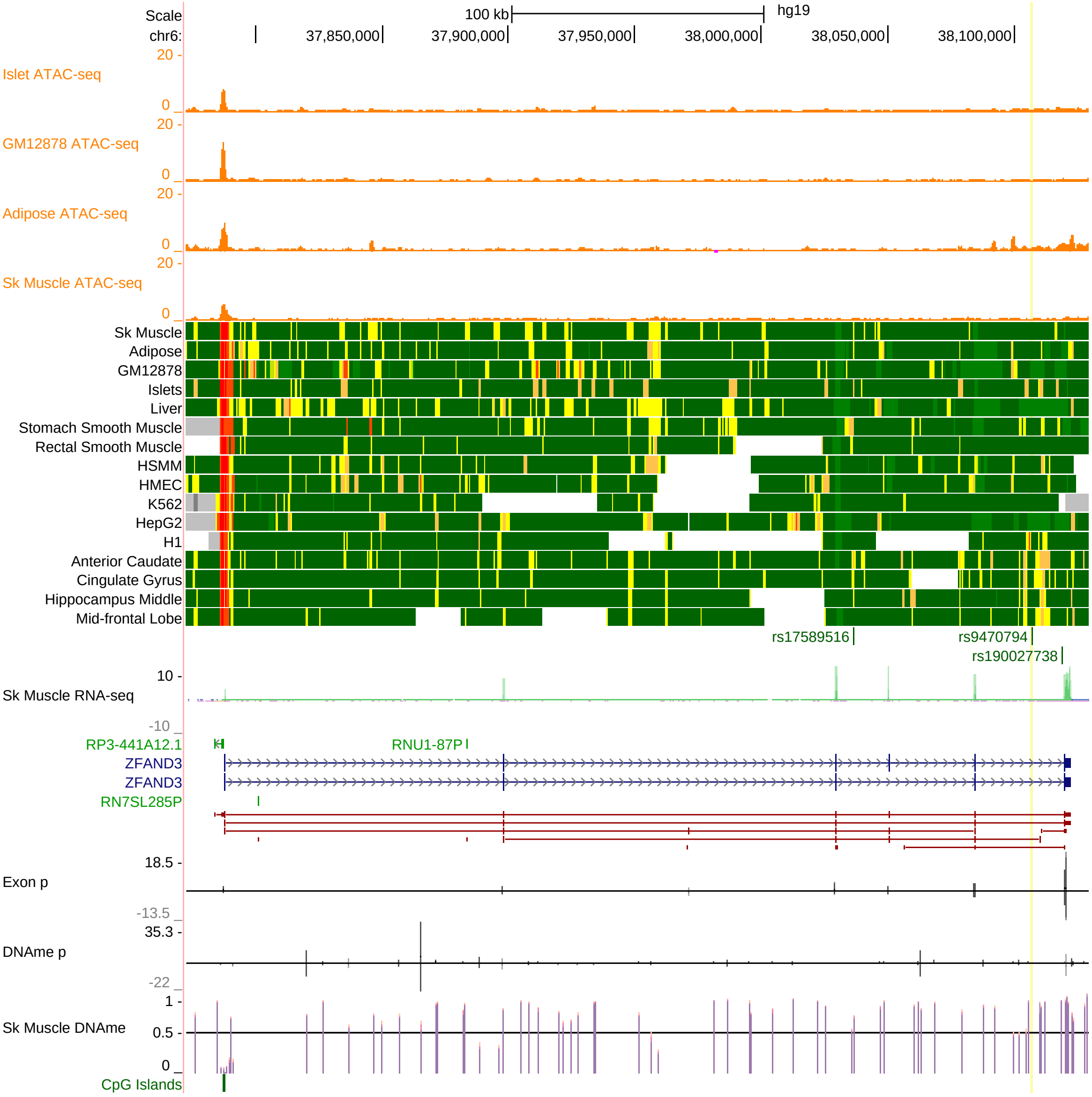


**Figure 5.10** *ZFAND3* locus. Light yellow vertical line shows the position of rs9470794. In addition to a *ZFAND3* eQTL effect, I found associations with 3' *ZFAND3* exon fragments. There was also one strong DNAme association that occurred some distance away in an intergenic *ZFAND3* region near *RNU1-87P*. This DNAme site is also the only exQTM associated with the aforementioned *ZFAND3* exon fragments. Detailed figure legend can be found in Figure 5.4.

***SDHAF4*** I found rs1048886, which tags a T2D locus, is a strong *SDHAF4*-eQTL, *SDHAF4*-exQTL, and mQTL.[3] However, at this locus, I also observed exceptionally unique DNAme patterns (Figure 5.11). *SDHAF4* is a small gene with one main middle exon (ENSE00001013963). rs1048886 falls in the middle of *SDHAF4*:ENSE00001013963 and is strongly associated with the expression of this exon (cond. $p = 2.93\text{x}10^{-44}$). Unfortunately, near the core exon, there are no DNAme probes, and the DNAme patterns around the surrounding 5' and 3' exons are not associated with rs1048886. However, I did find 5 DNAme probes associated with rs1048886, all of which were skewed upstream of *SDHAF4*.

The three probes most strongly associated with rs1048886 are cg21200554, cg20244062, and cg18341098—which are also exQTMs for *SDHAF4*:ENSE00001013963. Two of these probes, cg21200554 and cg18341098, fall just outside of the main CGI of the *FAM135A* promoter. The third DNAme site, cg20244062 (cond. $p = 2.38\text{x}10^{-13}$), falls at the 3' end of a lincRNA transcript, *RP11-462G2.1* (ENSG00000237643), in an area of active chromatin (TSS or enhancer related state) across all cell types. In islets, however, this DNAme site falls just beyond an active TSS chromatin state in the tail of a very islet specific ATAC-seq peak, suggesting islet specific effects. Indeed, using GTEx, I found *RP11-462G2.1* expression is highly specific to pancreas; however, *SDHAF4* is not. This gene is also expressed in islet RNA-seq data [95]. To date, very little is known about *SDHAF4*, other than it is essential for the assembly of complex II of the mitochondrial electron-transport chain [394]. However, these DNAme trends suggest particularly interesting and complex regulatory effects that may operate in a basal state across tissues and be activated specifically in islets or other pancreatic cell types.

[3] *SDHAF4* is also commonly referred to as *C6orf57*.

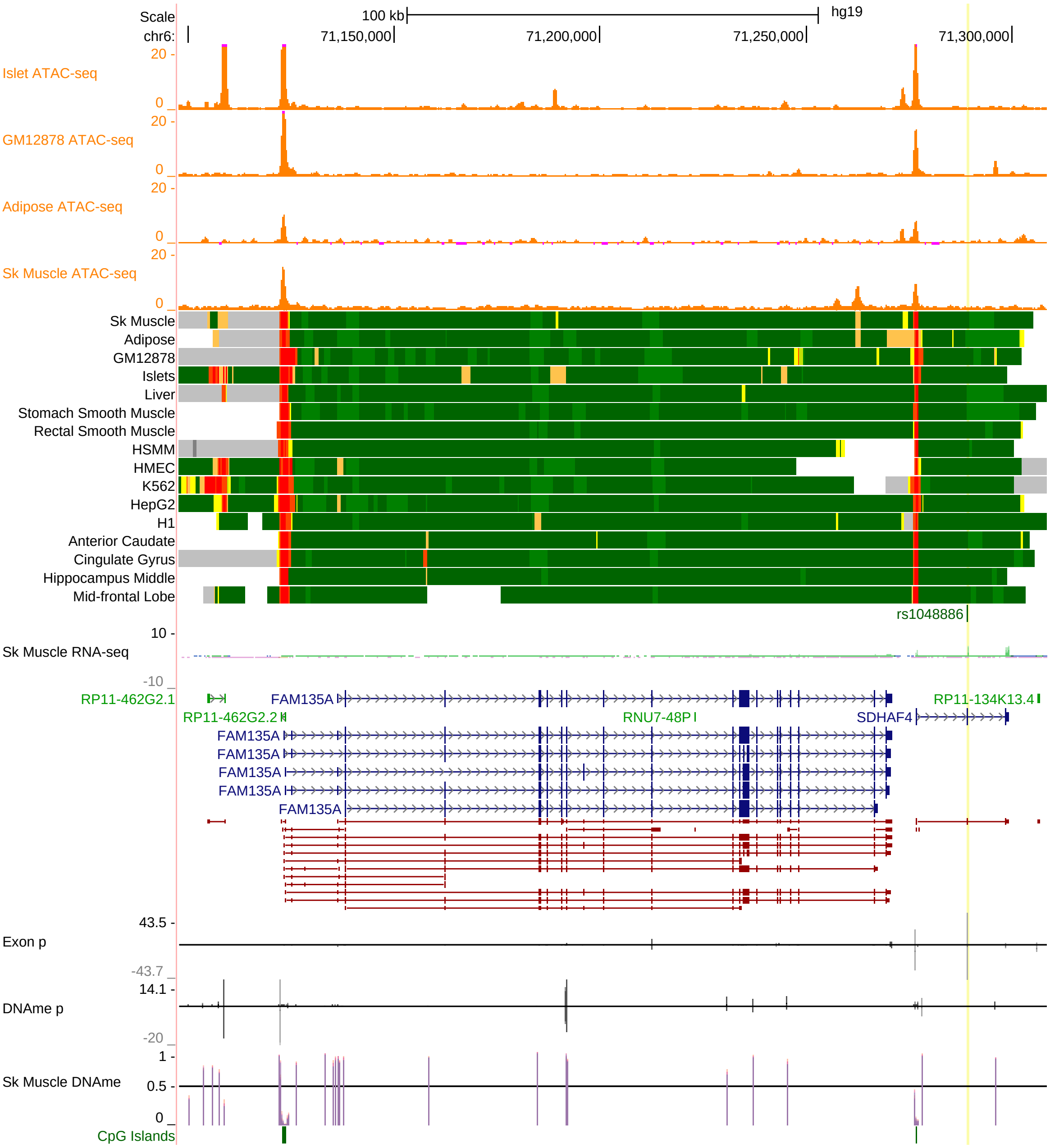


**Figure 5.11** *SDHAF4* locus. Light yellow vertical line shows the position of rs1048886. In addition to a *SDHAF4* eQTL effect, I found a very strong association with the *SDHAF4* exon nearest to rs1048886. However, the strongest DNAme site associated with rs1048886 falls in *RP11-462G2.1*, a lincRNA with a very strong islet ATAC-seq peak. In GTEx, this gene is specifically expressed in pancreas, as well as islets from other datasets. Detailed figure legend can be found in Figure 5.4.

***ANK1*** Consistent with Scott et al. [341], I found the T2D associated rs516946 SNP is a strong conditional GWAS eQTL for the highly muscle specific *ANK1* gene (mESI 0.7; Figure 5.12). With the addition of skeletal muscle DNAme data, I also discovered that rs516946 is a strong conditional mQTL for several highly muscle specific methylation sites (maximum of 0.81 MeSS), many of which are also associated with decreased *ANK1* expression (the top eQTMs are cg12439423, cg01678292, and cg23241016). All of the top 5 DNAme probes (cond. $p < 1.0x10^{-26}$)—cg01678292, cg12439423, cg23241016, cg11479568, and cg17274126—fall within a 230 bp region in a skeletal muscle specific promoter and ATAC-seq peak (chr8:41,522,721-41,522,951). Only one variant common to European populations falls within this region, rs508419.

Interestingly, in Scott et al. [341] we independently identified rs508419 because it is the only SNP in high LD ($r^2 \geq 0.8$) with rs516946 that falls within a skeletal muscle specific promoter. Through in silico TF binding predictions, we computationally predicted this SNP would perturb TR4 binding. We validated this prediction using an electrophoretic mobility shift assay (EMSA) and nuclear extract from human skeletal muscle cells (SkMC), where we observed an allele-specific supershift using the TR4 antibody. The analysis from Scott et al. [341] suggests the rs508419 risk allele disrupts TR4 binding and repression, resulting in increased overall *ANK1* expression and alternative splice isoforms.

*ANK1* is a highly spliced gene, with many alternative transcripts. Long isoforms are highly expressed and specific to the cerebellum, while several alternative short isoforms are highly expressed and specific to skeletal muscle (GTEx v6p). ANK1 links membrane proteins to the spectrin-actin cytoskeleton [252], and rare *ANK1* mutations cause hereditary spherocytosis (OMIM 182900), where erythrocytes (red blood cells) become spherical rather than toroidal, leading to many complications including an increased propensity for hemolysis [293]. ANK1 interacts with obscurin [422, 39], a component of the sarcoplasmic reticulum (SR) [15, 1]. The SR is critical for healthy skeletal muscle function as it regulates the translocation of GLUT4, the primary glucose transport protein in skeletal muscle, to the plasma membrane in response to insulin stimulation (reviewed in [69]). Furthermore, in skeletal muscle, ANK1 also interacts with IRS1 [55], which is critical for the glucose uptake of skeletal muscle cells (SkMCs) in response to insulin stimulation [41]. These observations, combined with the results from Scott et al. [341] clearly point to a skeletal muscle effect linked to T2D through insulin response.

However, GLUT4 translocation is also important to other tissues, like adipose, and indeed adipose and skeletal muscle share some of the proteins involved in GLUT4 trafficking

[268]. Although *ANK1* is not highly expressed in adipose, both rs508419 and rs516946 exert significant effects on *ANK1* in subcutaneous adipose expression (GTEx v6p). Furthermore, an additional locus associated with T2D, rs12549902, that is not in LD with either rs516946 or rs508419 ($r^2 < 0.17$ 1000GENOMES:phase_3:FIN), has been shown to colocalise with an *ANK1*-exQTL in islets [393]. While the islet effects are difficult to interpret, especially since rs12549902 was also associated with *NKX6-3* in gene level analysis [393], the adipose observation suggests that at some GWAS loci, although the primary effect that significantly contributes to disease risk may be isolated to one tissue, effects will be present and detectable in other tissues that share similar molecular biological programs.

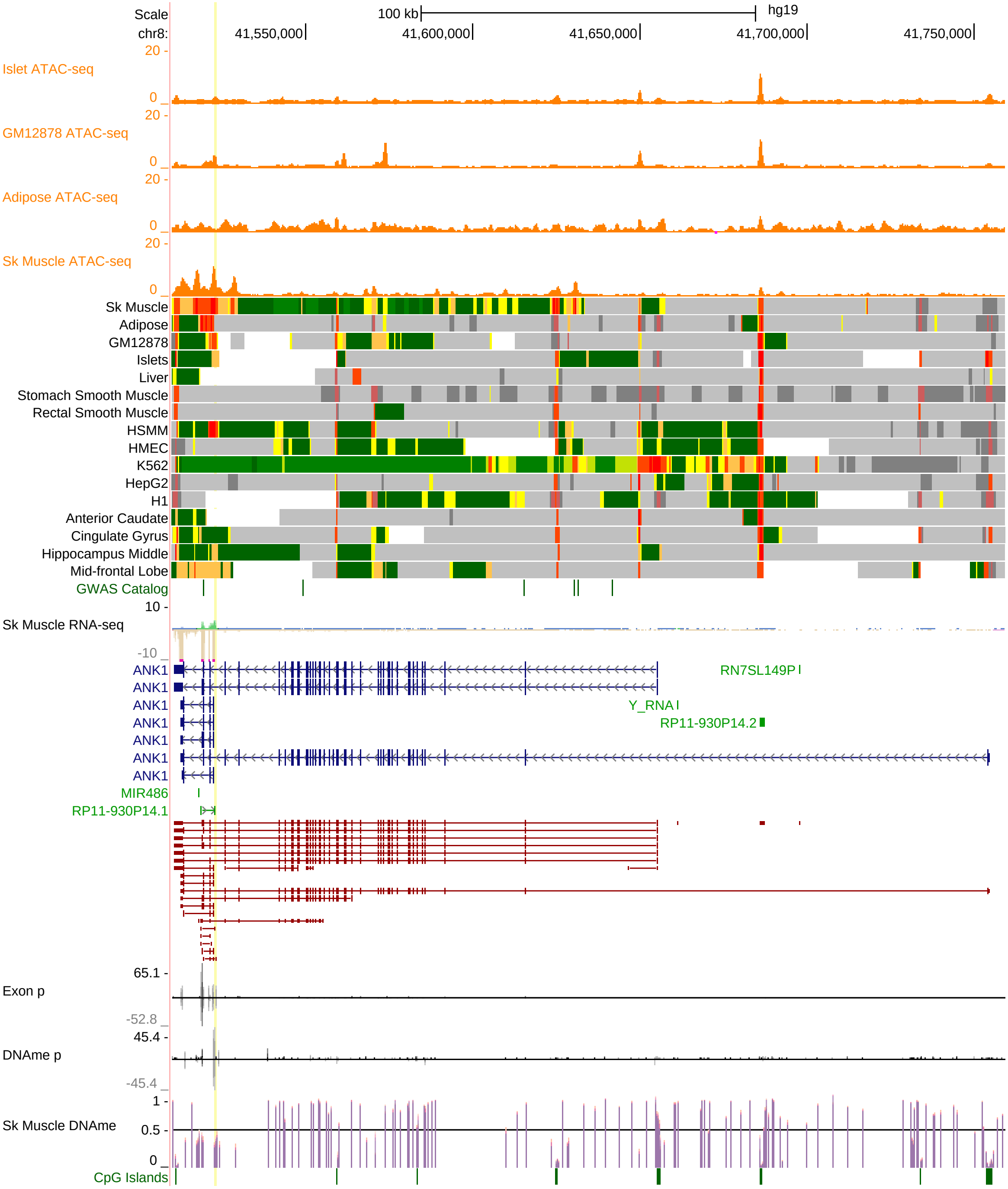


**Figure 5.12** *ANK1* locus. Light yellow vertical line shows the position of rs508419. In Scott et al. [341], we featured this locus and describe how rs508419 (in LD with the T2D SNP rs516946) likely affects a small, muscle specific *ANK1* isoform by altering TR4 binding. With the DNAme data, I found the DNAme sites associated with rs508419 (and rs516946) all fall in the muscle specific promoter near the TR4 binding site. Detailed figure legend can be found in Figure 5.4.

#### 5.3.2.4 Candidate muscle specific effects

Initial results from GTEx [138] and other studies [136, 104] suggest many eQTLs are shared across tissues, and although far less characterised due to limited tissue diversity, studies suggest mQTLs appear to be as well [357, 142]. Therefore, in order to investigate the genetic effects of T2D related loci that are particularly relevant to skeletal muscle, I used the specificity indices described earlier, to identify genes (mESI) and DNAme (MeSS) with muscle specific patterns. For gene expression and DNAme, I used the 90th percentile genome wide (i.e., not only across genes with a QTL) as a cutoff and sorted by the conditional p-value (Table 5.5, 5.6). This cutoff corresponded to a mESI of 0.55 and a MeSS of 0.47.

Across the top eQTL results for muscle specific genes, I observed large regions of active chromatin (TSS or enhancer related chromatin state) very specific to skeletal muscle within the region of the target gene. These regions were often accompanied by specific skeletal muscle ATAC-seq trends. In many cases, however, I did not find strong DNAme effects or did not find a strong effect at both gene and exon level. Such trends often occurred for instances with particularly weak marginal effects, and do not make a compelling case that skeletal muscle is important to the GWAS effect. In addition, trends in the top DNAme loci corresponded to regions that were either specifically activated or repressed in skeletal muscle. Below I highlight the instances of candidate muscle specific effects with clear trends in expression and DNAme.

***KCNJ11*** After *ANK1*, the second strongest conditional association for a highly muscle specific gene was for *KCNJ11* (mESI 0.71) and rs5219 (Figure 5.13). rs5219 is a missense variant in *KCNJ11* where the T allele is associated with increased T2D risk, causing a K23 change in ENST00000339994 and a K29 change in ENST00000528992 (C allele is ancestral, protective against T2D, and corresponds to E23 and E29; see www.type2diabetesgenetics.org). At the gene expression level, before conditional analysis, rs5219 was nominally associated with *KCNJ11* expression (marg. p = 0.023); however, after conditional analysis the association increased drastically (cond. p = $3.09\text{x}10^{-8}$). In addition, rs5219 showed association with *NCR3LG1* (mESI 0.42) and *RP1-239B22.5* (ENSG00000260196; mESI 0.36) before and after conditional analysis. In GTEx v6p, there are significant associations between these genes and rs5219 across various tissues, all of which show decreased expression with the T T2D risk allele.

At the exon level, *KCNJ11* showed similar patterns with little effect before conditional analysis across 4 exon fragments. rs5219 was also an exQTL for *RP1-239B22.5* and *ABCC8* (mESI 0.43). Due to minimum exon expression requirements, only one *NCR3LG1* exon was tested, which showed nominal association (cond. p = 0.0012) yet failed the FDR cutoff of 1% (cond. q = 0.04). The effect of rs5219 on *KCNJ11* expression (gene level) was complex, with the T allele showing a weak positive effect before becoming a much stronger negative effect after conditional analysis (although GTEx did not perform a conditional analysis, a negative effect is consistent with all of the reported GTEx v6p associations). This trend held true for all exons, except the very first (*KCNJ11*:ENSE00002147964), which showed a negative effect before and after conditional analysis.

The exon effects spanned the rather small *KCNJ11* gene. However for *ABCC8*, all of the exon effects were for exons near the 3' end of the *ABCC8* transcript, immediately before *KCNJ11.* Since *ABCC8* and *KCNJ11* lie in sequence on the same DNA strand, it could be that the *ABCC8* exon associations are partially due to Pol II leading up to *KCNJ11* (median TPM 50.77, mean TPM 51.45). The low expression of *ABCC8* (median TPM 0.69, mean TPM 0.87) and the rarity of the *KCNJ11/ABCC8* protein channel in muscle, described subsequently, would support such a hypothesis. Both *RP1-239B22.5* and *NCR3LG1*, lie on the opposite strand; however, given the striking chromatin patterns, there are likely complex muscle chromatin interactions at this locus, which may explain these associations.

In contrast to gene expression, 14 conditionally significant DNAme sites showed consistent trends in the marginal and conditional associations with rs5219. The two strongest associations were for cg11839944 (cond. p = 1.17x10$^{-28}$) and cg09674956 (cond. p = 3.62x10$^{-21}$). cg11839944 occurs in a CGI (chr11:17409453-17409692) within *KCNJ11* at the start of the largest exon (*KCNJ11*:ENSE00001366321), where rs5219 resides. cg11839944 also shows highly muscle specific methylation patterns with a MeSS of 0.62. Between cg11839944 and rs5219, within the CGI, lies cg15432903. Despite being closer to rs5219, cg15432903 shows slightly less association (cond. p = 9.56x10$^{-16}$). Many of the DNAme sites associated with rs5219 fall throughout *KCNJ11*; however, the second strongest site, cg09674956 was some distance away, just outside of the CGI of the *NCR3LG1* promoter (chr11:17373020-17373665), supporting the hypothesis of many chromatin interactions at this locus.

Within this region, the chromatin states showed very specific trends for skeletal muscle. Near the end of *ABCC8*, there is a skeletal muscle specific promoter and ATAC-seq peak and large enhancer. Immediately, after this muscle specific promoter are all of the *ABCC8* exon associations. Shortly thereafter, is *KCNJ11*, which has three active muscle TSS states surrounded by weak and flanking muscle TSS states. Each active TSS corresponds to a muscle ATAC-seq peak that shows distinct trends compared to other cell types. Surrounding these active TSS regions, flanking the ATAC-seq peak, I observed spikes in the DNAme association with cg22710661 after TSS 1 (cond. p = $1.42 \times 10^{-11}$), cg11839944 and cg15432903 after TSS 2 (described above), and cg03864215 before TSS 3 (cond. p = $1.59 \times 10^{-16}$).

The strongest DNAme association (cg11839944) occurred after the middle active TSS state. Both of the outer active TSS states overlap many ChIP-seq peaks from ENCODE cell types, the strongest of which is a strong CTCF peak, common to many cell types. In both cases, the DNAme association occurs at the tails of the CTCF peak. Finally, downstream of the *KCNJ11* gene body, there is a large, highly muscle specific enhancer state that trails well beyond *RP1-239B22.5*.

*KCNJ11* and *ABCC8* respectively encode for Kir6.2 and SUR1—two subunits that form adenosine triphosphate (ATP) sensitive $K^+$ ($K_{ATP}$) plasma membrane channels. There are a variety of $K_{ATP}$ channels, all composed of a pore-forming Kir subunit and an ATP binding cassette (ABC). In addition to *KCNJ11* and *ABCC8*, the two other key $K_{ATP}$ subunit genes are *KCNJ8* (Kir6.1) and *ABCC9* (SUR2A and SUR2B, depending on splicing; reviewed in [100]). In skeletal muscle, the primary $K_{ATP}$ channel is Kir6.2/SUR2A; however, other forms exist, for instance involving SUR1 or SUR1/SUR2A hybrids, that relate to skeletal muscle type (fast or slow twitch; reviewed in [100]).

In relation to diabetes, several associations with T2D and related traits have been identified in and around *KCNJ11* and *ABCC8*, which clearly affect pancreatic beta cell function. Protein coding mutations in *KCNJ11* and/or *ABCC8* often underlie rare monogenic forms of diabetes such as neonatal diabetes (diabetes by ~6 months) and MODY (autosomal dominant diabetes by ~25 years of age; reviewed in [415]). These mutations generally result in overactivity of islet Kir6.2/SUR1 $K_{ATP}$ channels, where the channels are constantly open, leading to reduced beta cell insulin secretion [415].

In beta cells, Kir6.2/SUR1 $K_{ATP}$ channels are responsible for enabling insulin release in response to glucose levels (reviewed in [415]). Briefly, in low levels of plasma glucose and consequently low levels of cytosolic ATP, $K_{ATP}$ channels are more likely to be open allowing basal efflux of $K^+$, keeping the membrane potential at -70mV. In high levels of plasma glucose, glucose enters beta cells via GLUT transporters, resulting in increased cytosolic levels of ATP through glucose metabolism. The increased levels of ATP promotes the closing of $K_{ATP}$ channels, which leads to depolarisation of the beta cell membrane (decrease of negative charge within the cell). Depolarisation opens voltage-gated calcium ion channels, increasing cytosolic calcium levels which stimulates the release of stored insulin.

Given the importance of Kir6.2/SUR1 for islet biology, it is not surprising common variants associated with T2D and related traits have also been identified. As noted above, one of the most studied common variants is the missense variant rs5219 (which causes a E23K change in ENST00000339994), identified in this study as having an effect on *KCNJ11* (and *ABCC8*) skeletal muscle expression. This variant is in high LD with rs757110 (which causes a S1369A change in ENST00000389817),[4] a missense variant in *ABCC8*, making it extremely difficult to prioritise a candidate functional variant based on statistics alone [392, 415]. However, functional follow up revealed that in the K23/A1369 risk haplotype, A1369 (C allele of rs757110) is responsible for reducing the ATP sensitivity of $K_{ATP}$, leading to an increased probability of $K_{ATP}$ openness, suppressing insulin secretion [143, 415]. This does not, however, mean that K23 (rs5219) could not have an independent effect that contributes to T2D risk, possibly dependent on tissue context such as skeletal muscle (although it could very well have a islet effect).

Indeed, in context of the Kir6.2/SUR1 channel (which is less frequent in skeletal muscle), K23 increases $K_{ATP}$ sensitivity to intracellular long chain acyl coenzyme A esters (LC-CoA), which are more abundant in obese and T2D patients. This sensitivity change results in the increased likelihood of K23 $K_{ATP}$ channels being open at various ATP concentrations compared to E23 $K_{ATP}$ when LC-CoA is present [311, 310, 397]. Such effects likely persist across all $K_{ATP}$ variants, as LC-CoA $K_{ATP}$ sensitivity has been reported in various cell types, some of which, like ventricular myocytes, are even more sensitive than beta cells [100]. Moreover, in the context of Kir6.2/SUR2A (common to skeletal muscle) K23 increases $K_{ATP}$ sensitivity to intracellular pH levels, such that at more acidic levels (as would happen during exercise) K23 $K_{ATP}$ are more likely to be open than E23 $K_{ATP}$ [215]. These effects may be specific to skeletal muscle, as pH sensitivity has not been observed in cardiac $K_{ATP}$, which also prominently feature Kir6.2/SUR2A [100].

Although not intensely studied, it seems likely genetic effects on Kir6.2 will be particularly important in the context of skeletal muscle, especially since Kir6.2 is the predominant $K_{ATP}$ subunit (which in contrast to beta cells primarily combines with SUR2A or SUR2A/SUR1

[4] For rs757110, the A allele is ancestral, protective against T2D, and corresponds to S1370 in ENST00000302539 and S1369 in ENST00000389817. The C allele is the T2D risk allele and corresponds to A1370 in ENST00000302539 and A1369 in ENST00000389817. There is also evidence of an extremely rare T allele.

hybrid subunits) [100]. In muscle, $K_{ATP}$ channels are generally closed at rest and become open in response to stress, such as exercise [158]. Alterations to $K_{ATP}$ channels have been shown to accelerate muscle fatigue during exercise (reviewed in [100]).

Most relevant in the context of T2D (although the mechanisms are poorly understood), skeletal muscle $K_{ATP}$ openness has been linked to decreased glucose uptake (reviewed in [100]). For instance, Kir6.2$^{-/-}$ mice show increased skeletal muscle glucose uptake in addition to impaired insulin secretion (as well as impaired stress response cardiac and skeletal muscle), while SUR1$^{-/-}$ mice only show impaired insulin secretion (SUR2$^{-/-}$ leads to coronary vasospasm and premature death; reviewed in [100]). Furthermore, *in vitro*, $K_{ATP}$ channel openers have also been shown to suppress skeletal muscle glucose uptake, stimulated by either high glucose concentration or insulin induction. These effects could be pharmacologically reversed by channel blockers [410]. Therefore, effects like K23 that increase $K_{ATP}$ sensitivity to various external molecular modifiers may have important consequences to glucose homeostasis by perturbing skeletal muscle function.

Collectively, these findings highlight the importance of Kir6.2 to proper skeletal muscle function and suggests that the same T2D risk haplotype perturbs glucose homeostasis by reducing beta cell insulin secretion and decreasing skeletal muscle glucose uptake.

Nonetheless, the effect identified by on rs5219 expression is particularly perplexing. $K_{ATP}$ is clearly important for normal skeletal muscle function. Therefore, alterations at the level of protein sequence or protein abundance could potentially perturb normal muscle function. rs5219 certainly affects protein sequence and function. The results presented here also suggests rs5219 (or a variant in high LD) effects *KCNJ11* expression. One hypothesis is that rs5219 (or an LD variant) affects specific *KCNJ11* isoforms. Indeed, this seems supported by the exQTL analysis, where not every *KCNJ11* exon fragment was associated with rs5219. In addition, a previous study [221] identified numerous novel skeletal muscle transcripts surrounding *KCNJ11*, which could explain why cardiac and muscle tissue differ in pH response (these were not included in our gene annotation file as we used the official GENCODE v19 release [146]).

Given the several conditional QTLs (the other SNPs went into the final conditional model), it seems likely that any novel or alternative isoforms affected by rs5219 are not the primary *KCNJ11* transcript. Indeed, with the extensive study of Kir6.2, it seems unlikely that radically novel protein isoforms exist in abundant quantities, suggesting either rare protein forms, protein isoforms with very slight differences, ncRNA, or a combination of all three contribute to the identified signals. In further support of this hypothesis (that the QTL effect is nuanced and not on the primary *KCNJ11* isoform), the first conditional variant was rs10766394 (rs5219 $r^2$ = 0.52 1000GENOMES:phase_3:FIN), the strongest variant association from the primary QTL analysis in Chapter 4. In the primary QTL analysis, the G allele (A/G) results in increased expression (beta = 0.60, p = 3.73x$10^{-13}$). rs10766394 falls just beyond the 3' end of *KCNJ11*, between the end of *KCNJ11* and the start of *RP1-239B22.5*, nestled between two large ATAC-seq peaks common across tissues. In GTEx, rs10766394 has a significant effect across many tissues, but noticeably not muscle. Since there are potential novel transcripts at this locus, I cannot rule out that the rs10766394 effect may change with different gene models for gene quantification using RNA-seq data.

Finally the DNAme signals also underscores the importance of rs5219 on the surrounding epigenomic landscape. In the conditional analysis, rs5219 was most strongly associated with cg11839944 (marg. p = 8.03x$10^{-36}$, marg. beta = -0.96, cond. p = 1.19x$10^{-28}$, cond. beta = -0.86). In the primary QTL analysis, rs5219 was not the strongest QTL (minimum p-value) for any DNAme sites. However for cg11839944, rs5213, a variant in high LD with rs5219 ($r^2$ = 0.95 1000GENOMES:phase_3:FIN), was the top mQTL (beta -0.97, p = 4.16x$10^{-36}$). rs5213 falls immediately after the main *KCNJ11* exon in a 3' UTR of *KCNJ11*, overlapping the tail of an ATAC-seq peak common across cell types (most strong in muscle). As one would expect, this variant also overlaps the ChIP-seq peaks of many TFs common across cell types, including a HSMM CTCF peak. Thus, the gene expression rs5219 association may be driven by rs5213.

Regardless, detailed follow up will be required to fully understand the skeletal muscle effects at this locus. Careful analysis of *KCNJ11* reads (e.g., isoform deconvolution), the surrounding LD structure, and chromatin information, may reveal further insights and guide functional follow up. Most challenging will be to understand effects in the context of T2D, and establish expression effects that contribute to T2D risk independent of the undoubtedly important rs5219 effect on Kir6.2.

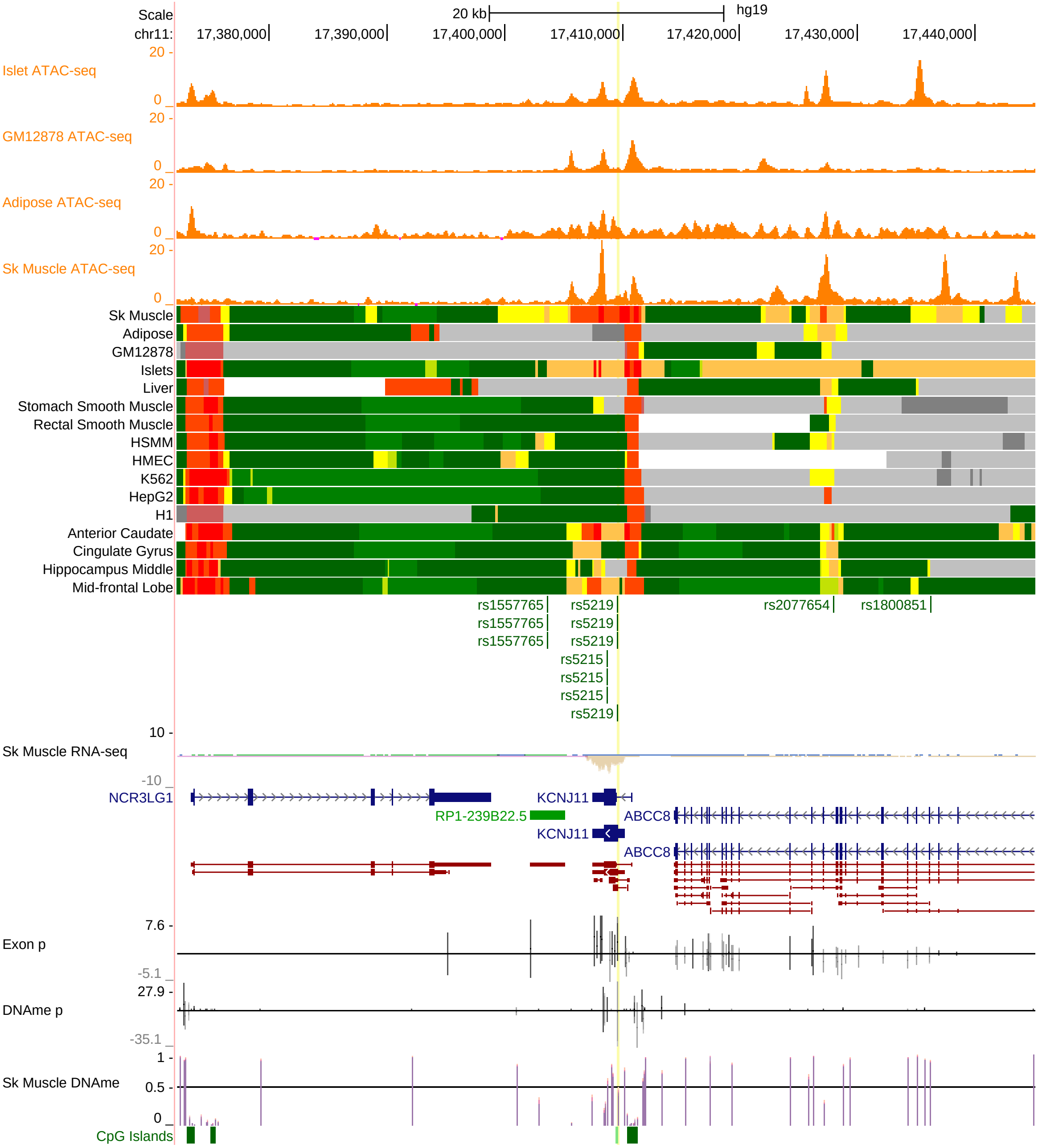


**Figure 5.13** *KCNJ11* locus. Light yellow vertical line shows the position of rs5219. All of the *KCNJ11* expression associations became significant after conditioning on rs10766394, the SNP most strongly associated with *KCNJ11* from the primary QTL analysis in Chapter 4. rs10766394 lies between *KCNJ11* and *RP1-239B22.5*. Unlike the expression trends for rs5219, I found many strong DNAme associations in marginal and conditional analysis. The strongest DNAme association occurred within a *KCNJ11* CGI. There is also a DNAme association in the *NCR3LG1* promoter. Detailed figure legend can be found in Figure 5.4.

***PIEZO1*** Several of top mQTLs with muscle specific DNAme patterns, corresponded to regions with repressed chromatin states specific to skeletal muscle. As expected, these such instances generally had weak or nonexistent effects on gene expression. A representative example of such cases is an HbA1c-associated locus tagged by rs9933309. rs9933309 falls in *PIEZO1* (Figure 5.14). In most cell types, this region is active, especially adipose; however, in skeletal muscle this region is specifically repressed, falling in a weak repressed polycomb chromatin state. Near the 3' end of *PIEZO1*, the skeletal muscle chromatin state changes to a strong transcription state. In that transition region I find one *PIEZO1* exon significantly associated with rs9933309 (cond. $p = 3.65\text{x}10^{-6}$). Otherwise, rs9933309 is associated with several 9 DNAme probes, many of which show muscle specific DNAme trends. Surrounding the strongest DNAme association, cg04602696 (cond. $p = 1.08\text{x}10^{-56}$; MeSS = 0.53), are many TF binding sites and common variants, making it difficult to identify any potentially relevant variants or TFs.

Piezo1 mediates non-selective cation transport involved in mechanosensory signal transduction—a cell's biological response to mechanistic stimulation such as increased pressure (reviewed in [16]). Mutations in *PIEZO1* lead to the autosomal dominant dehydrated hereditary stomatocytosis (OMIM 194380), a disease where faulty ion channels leads to lysis of erythrocytes [5, 11]. In addition, Piezo1 is known to be involved in vascular, genitourinary, and pulmonary biology (reviewed in [16]). Such trends are consistent with the observed patterns of active chromatin states and the high expression of *PIEZO1* in bladder, colon, kidney, lung, oesophagus, spleen, vascular tissue, genital tissue, and adipose (GTEx v6p).

Despite these trends, no gene expression effects were identified for rs9933309 in GTEx v6p. However, *PIEZO1* is a highly spliced gene, so the lack of gene level effects by rs9933309 could be explained by specific transcript effects. Although speculative, given the general importance of *PIEZO1* in diverse tissues, *PIEZO1* may turn out to be a locus where the same molecular effect on *PIEZO1* expression leads to multiple manifestations, depending the tissue, that collectively add to disease risk. Regardless, it seems unlikely that rs9933309 has any disease relevant effect in muscle; however, the stark contrast between the skeletal muscle chromatin states in this region and that of other cell types is particularly striking and may indicate muscle specific higher order chromatin structure.

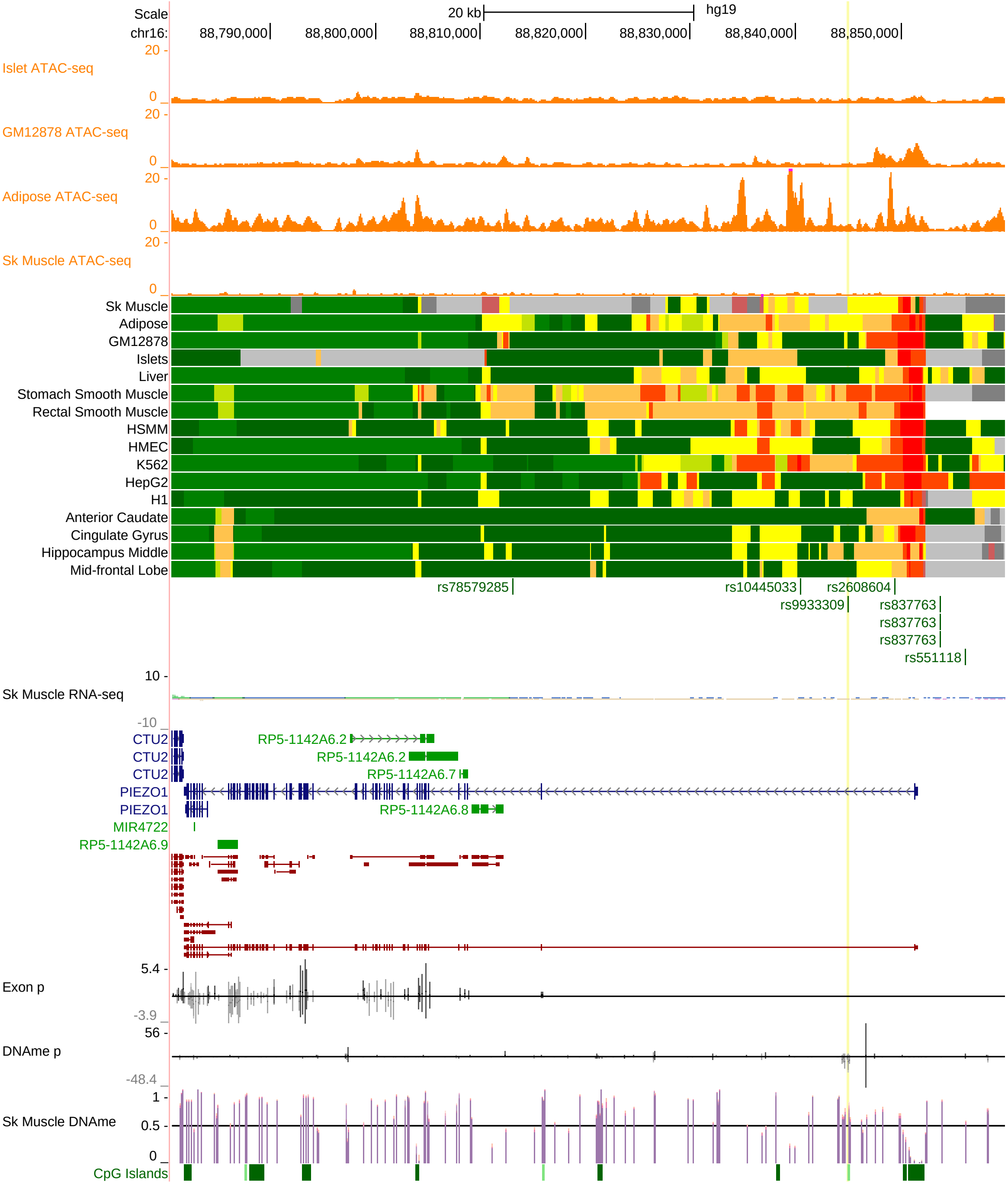


**Figure 5.14** *PIEZO1* locus. Light yellow vertical line shows the position of rs9933309. This *PIEZO1* region is specifically repressed in skeletal muscle. I found one strong DNAme association near the *PIEZO1* TSS. In addition, there was one *PIEZO1* exon near the 3' end (after the muscle chromatin state has transitioned to a strong transcription state) associated with rs9933309. Detailed figure legend can be found in Figure 5.4.

***ACHE*** I found a T2D tag SNP, rs7636, is an eQTL for *ACHE*, a gene with muscle specific expression trends (mESI 0.59). rs7636 is a synonymous *ACHE* coding variant that lies in a muscle specific genic enhancer state and overlaps an intergenic CGI (Figure 5.15). The exon that rs7636 falls in has various lengths depending on the transcript. rs7636 was not associated with expression levels of the exon fragment in which it resides, but was associated with the expression of smaller fragments, or versions of the exon (cond. p = 1.75x10$^{-5}$), indicating transcript specific effects. The rs7636 risk allele, A [353], was associated with decreased overall *ACHE* expression (marg. beta -0.82, cond. beta -0.86). This effect was consistent at the exon level in all but one exon, where risk allele was associated with increased exon expression.

rs7636 was also associated with 6 DNAme probes, the strongest of which, cg15575433 (cond. p = 3.56x10$^{-9}$), occurred at the start of a final 3' exon of two alternate *ACHE* transcripts. Moving towards the TSS from cg15575433, all of the other DNAme sites fall after chr7:100,491,854, the approximate start of several alternate *ACHE* transcripts. These patterns strongly suggest alternate transcript effects. The surrounding chromatin states further support this conclusion. Compared to all other tissues, the TSS related chromatin states for *ACHE* in skeletal muscle (active TSS and flanking TSS) are shifted well beyond chr7:100,491,854, towards the 3' end of *ACHE*. In all other tissues, the TSS related chromatin states stop by ~chr7:100,491,854. There are, however, numerous *ACHE* transcripts (reviewed in [363, 278]), making it difficult to further narrow down effects.

*ACHE* encodes acetylcholinesterase (AChE), which degrades acetylcholine, thereby terminating acetylcholine-mediated neurotransmission which is essential for neuromuscular communication (reviewed in [363]). Since the substrate of AChE, acetylcholine (ACh), was historically the first identified neurotransmitter whose discovery was shortly followed by AChE, immense study has focused on ACh and AChE [363]. Through this research, a variety of AChE inhibitors have been developed. Some AChE inhibitors have been weaponised (e.g., sarin), but others are used medically to treat a variety of diseases, including myasthenia gravis (chronic skeletal muscle weakness), Alzheimer's disease, and glaucoma (reviewed in [66]). In relation to diabetes, pharmacologically induced type 1 diabetic mice show reduced Ache expression in muscle, possibly indicating a mechanism that leads to diabetes related muscle weakness [121]. In addition, accumulation of AChE in mouse islets has been associated with increased beta cell apoptosis and diabetes development [440]. In human, pancreatic alpha cells have been shown to release ACh in response to lowered glucose levels which sensitise beta cells to increases in glucose concentration for insulin release [318]. While undoubtedly important for skeletal muscle biology, such findings suggest the role of AChE in healthy islet function is more directly relevant to T2D. Indeed, phase I clinical trials are underway to test if AChE Inhibitors promote insulin secretion (ClinicalTrials.gov Identifier: NCT03063515).

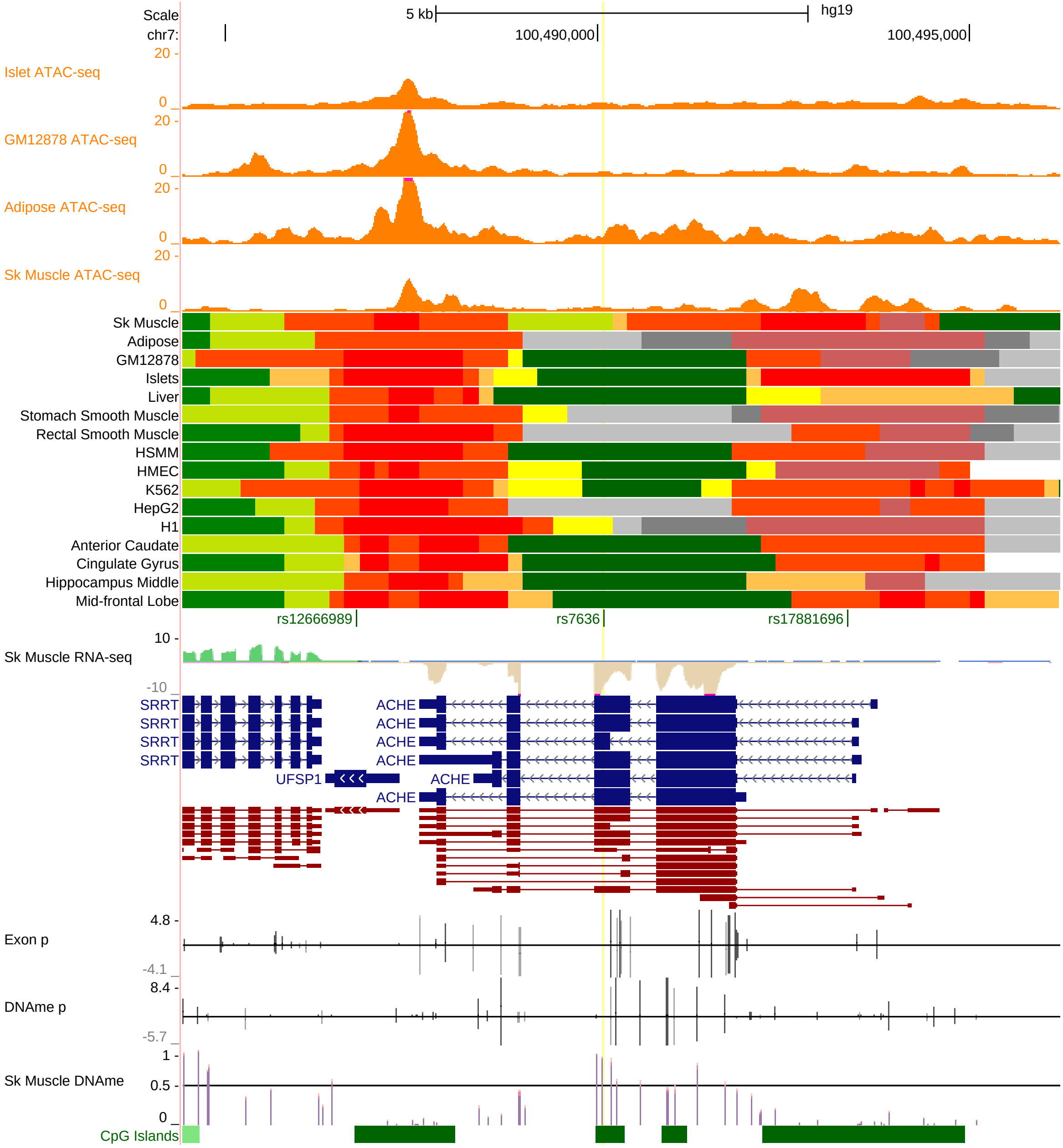


**Figure 5.15** *ACHE* locus. Light yellow vertical line shows the position of rs7636. In addition to a *ACHE* eQTL effect, I found several exon and DNAme associations throughout *ACHE*. The exon and DNAme trends strongly suggest transcript specific effects at *ACHE*. Detailed figure legend can be found in Figure 5.4.

***SPTB*** Finally, I identified three QTLs with strong exon and weak gene level effects for highly muscle specific genes, two of which were mentioned earlier—*CAMK2B* (mESI 0.57) and *FGGY* (mESI 0.57). The third instance was for *SPTB* (mESI 0.75) and rs11158559, where the A allele is associated with increased levels of insulin and leptin [67] (Figure 5.16). Unfortunately, there seems to have been an error in the reporting study as the only discovered alleles for rs11158559 are C/T (there is no A allele as reported by Comuzzie et al. [67]). In the T2D knowledge portal (www.type2diabetesgenetics.org), the minor allele, T, is nominally associated (p = 0.0029) with increased insulin sensitivity index adj age-sex-BMI, so presumably T is the risk allele.

rs11158559 was not associated with gene level *SPTB* expression (marg. p = 0.11, cond. p = 0.90). rs11158559 falls in an *SPTB* intron, immediately after *SPTB*:ENSE00000392079, which is the only exon associated with rs11158559 (cond. p = 2.80x10$^{-16}$). The T allele of rs11158559 is negatively associated with *SPTB*:ENSE00000392079 expression. At the end of the two closest downstream exons, I observe strong DNAme associations within the exon sequence. Unfortunately, there were no DNAme probes near *SPTB*:ENSE00000392079 or within the rs11158559 intron. However, the DNAme probe with the strongest association, cg25083366 (cond. p = 8.20x10$^{-48}$), falls within an exonic CGI (chr14:65239323-65239666) that overlaps a strong CTCF ChIP-seq peak across many cell types. Sadly, there were no variants in LD with rs11158559 that occurred in this binding site.

Spectrin is the primary constituent of the cytoskeleton which is anchored by beta-spectrin to the plasma membrane through ankyrin interactions (reviewed in [233]). There are 5 spectrin genes, each with different tissue specificity patterns. *SPTB* is the only spectrin gene that shows expression patterns specific to skeletal muscle (mESI 0.75). Given the intimate interaction between beta-spectrin and ankyrin, it is not surprising that rare mutations in *SPTB* also give rise to hereditary spherocytosis (OMIM 616649), similar to *ANK1*. However, the aforementioned small, skeletal muscle ANK1 isoform (sAnk1) lacks a spectrin-binding domain and is not thought to have direct interactions with SPTB [189]. Regardless, beta-spectrin is linked to the Golgi complex in skeletal muscle [29], which is important for GLUT4 sequestration and translocation via GLUT4 storage vesicles [31], possibly explaining earlier studies that link beta-spectrin and GLUT4 [70, 388]. These observations suggest a *SPTB* could be linked to skeletal muscle insulin response through similar molecular mechanisms as *ANK1*. If rs11158559 turns out to be a robust insulin association, given the key role of *SPTB* in skeletal muscle insulin response, it seems likely that the disease risk effect at this locus will be in skeletal muscle through *SPTB*.

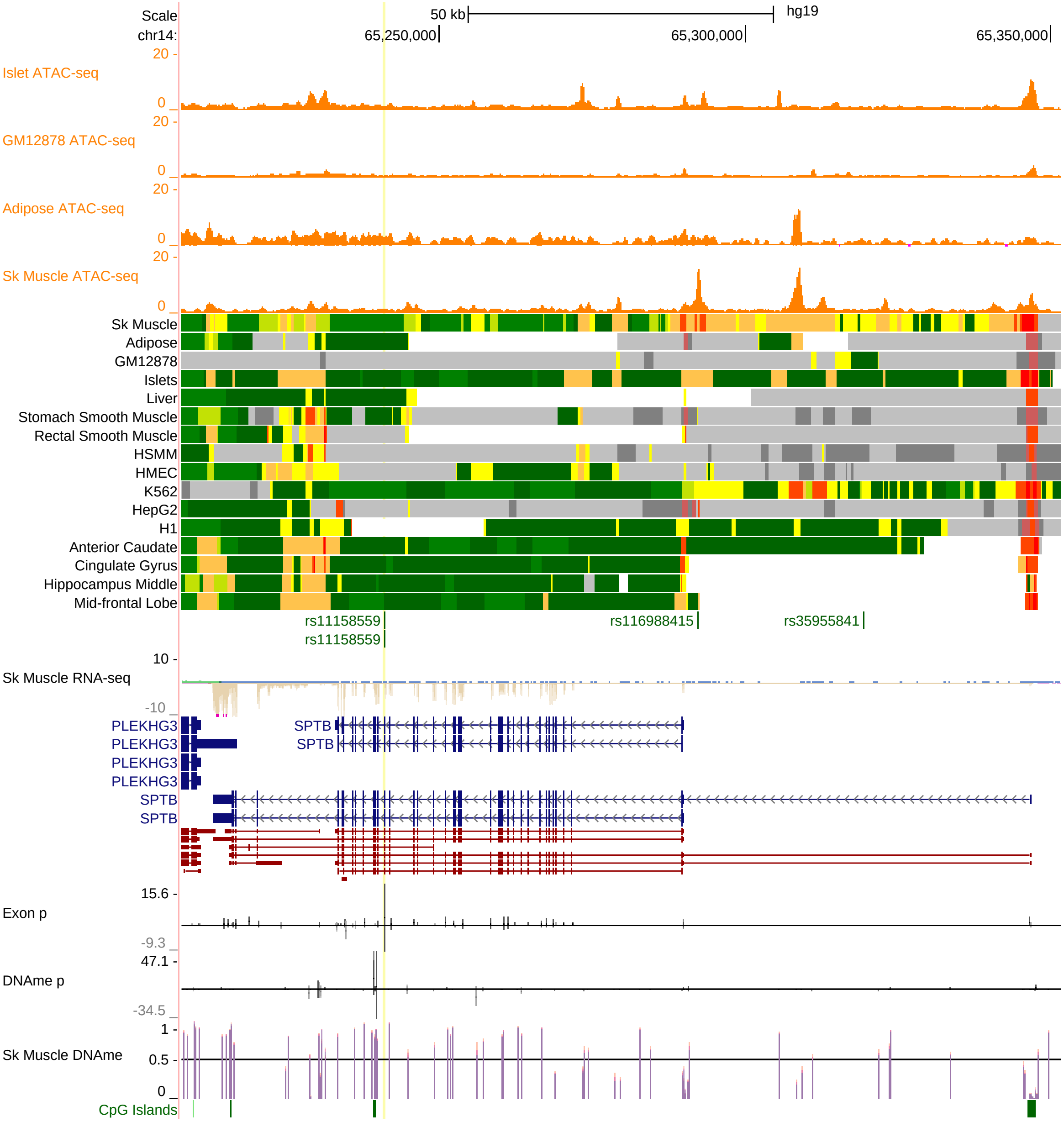


**Figure 5.16** *SPTB* locus. Light yellow vertical line shows the position of rs11158559. rs11158559 was not associated with *SPTB* gene level expression, but was associated with one exon. Near this exon, towards the 3' end of *SPTB* lies the strongest DNAme association, in an exonic CGI. Detailed figure legend can be found in Figure 5.4.

### 5.3.3 Conditional QTL summary

In summary, the results presented catalogue the effects of variants associated with T2D and T2D-related traits in skeletal muscle. Through these vignettes, it is clear that comprehensive QTL analysis across disease related tissues is critical for understanding the pathophysiological effects of GWAS loci.

At the molecular trait level, in some cases, the molecular effect of a GWAS locus will be the same across multiple tissues, perhaps due to common molecular pathways (e.g., *WFS1* and normal ER function). Such effects may manifest themselves as a strong QTL across all tissues, or as a weaker QTL in some tissues (where only basal activation is necessary) that are magnified in certain tissue contexts. In other cases, molecular trait effects may be shared only across a handful of tissues, or be highly specific to a tissue, based on the molecular pathways active in the cell type. In such instances, it seems likely that the same GWAS locus (perhaps even the same causal variant) will affect multiple molecular traits (e.g., genes) depending on the tissue context. In all of these scenarios, narrowing down to functional variants will be important, as many cases may consist of two or more variants in high LD that affect different molecular traits, rather than the same actual SNP causing the effect.

In addition, the various molecular trait effects may physiologically manifest themselves differently based on tissue context. For instance, one could imagine an eQTL that affects the same single gene in a similar molecular manner across tissues, for instance an indel that abolishes the CGI of a promoter. However, the physiological manifestations of that effect could be very different depending on the tissue context. Hypothetically in one tissue, the eQTL effect results in decreased insulin secretion and the other tissue it results in decreased glucose uptake. In both instances, the molecular effect is the same, but the physiological effect is different—suggesting that as a field, distinguishing between molecular pleiotropy and physiological pleiotropy may be important. This underscores the critical importance of function followup across a variety of models (*in vitro* and *in vivo*), in order to understand physiological effects and deduce which (if any) of these effects underlie disease risk.

## 5.4 Molecular trait mediation analysis

### 5.4.1 Molecular trait mediation motivation

As noted throughout the previous section, there are many instances where I found a strong genetic effect on gene expression and DNAme accompanied with a strong association between the two molecular traits. A representative example of such instances is *ANK1*.

Recall *ANK1* is strongly associated with rs508419 which perturbs TR4 binding, leading to increased expression with the T2D risk allele, and possibly altering the abundance of splice isoforms. In addition, rs508419 is also strongly associated with several DNAme sites that exhibit muscle specific patterns. The strongest association was for cg01678292 (marg. p and cond. p = $3.87x10^{-46}$ as there were not other, significant QTLs after conditioning on rs508419), which was also a strong eQTM for *ANK1* expression (as well as several of the other DNAme sites).

In order to better understand the relationship between gene expression and DNAme at this region, I compared the association between *ANK1* and the top mQTL probe (cg01678292) coloured by rs508419 (the TR4 altering SNP), as well as rs508419 effects on *ANK1* expression with and without conditioning on cg01678292 methylation (and vice versa). I found that rs508419 exhibits independent effects on *ANK1* and cg01678292, as conditioning on DNAme did not change the overall effect of rs508419 on *ANK1* (and vice versa; Figure 5.17). Furthermore, these trends were consistent when considering the GWAS tag SNP rs516946, other methylation sites that are also mQTL and eQTM, as well as *ANK1* exons.

### 5.4.2 Molecular trait mediation results

Motivated by the *ANK1* example, I systematically tested the genetic effects of QTLs across all significant eQTMs both at a genome wide scale and focused specifically on T2D and T2D-related trait GWAS QTLs. For the genome wide analysis, I generated SNP, methylation, expression trios following a method similar to Gutierrez-Arcelus et al. [141]. Briefly, to generate trios I selected all eQTMs with an FDR of 1% (n = 38,115), and filtered these eQTMs for those with a significant eQTL and/or mQTL (FDR 1%), selecting 37,438 eQTMs. As QTLs I used the most significant QTL per feature, so that if an eQTM has a significant

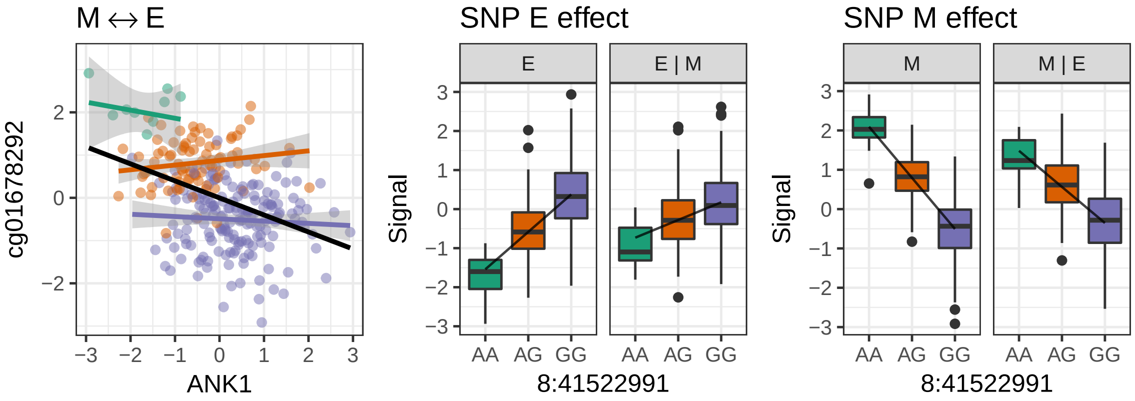


**Figure 5.17** *ANK1* expression and DNAme. (left) The association between *ANK1* and cg01678292 coloured by rs508419 (8:41522991). (middle) The association between rs508419 and *ANK1* expresion with and without conditioning on cg01678292 methylation. (right) The association between rs508419 and cg01678292 DNAme with and without conditioning on *ANK1* expression. All signal shown in PEER residual space.

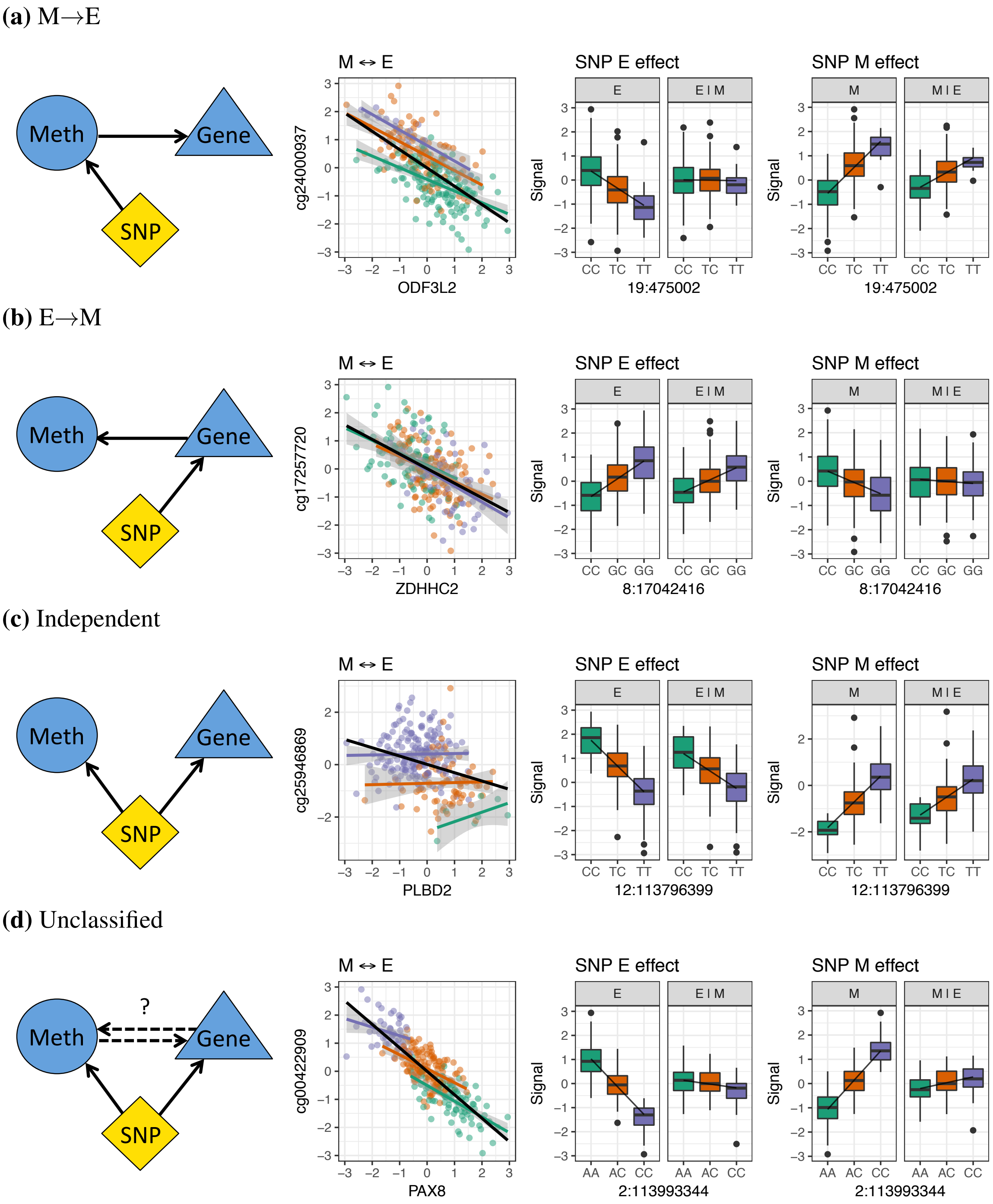
eQTL and mQTL with two different top SNPs, that eQTM would occur twice, once for each SNP. Such cases occurred 31,121 times, generating 68,559 total trios in the final analysis. For the GWAS QTL analysis, I started from overlapping QTLs (i.e., the GWAS SNP) and selected cases where (1) the GWAS SNP had a significant marginal and conditional p-value for expression and/or DNAme (using the p-value thresholds corresponding to an FDR of 1%) and (2) the expression and DNAme signals were linked by a significant eQTM. In the final genome wide trios and GWAS trios, I classified trios based on if the SNP was a significant eQTL, mQTL, or both. For "both QTL significant" trios I did not require the SNP to be the top eQTL and mQTL, only have a significant effect in both (FDR 1%). Note that for defining trios, I used the top QTL SNP, but not for trio classification. Finally, I ignored exQTMs from this analysis simply due to the computational burden of testing all exQTMs using the permutation test, as described below.

For each trio, I used the causal inference test (CIT) R package v2.1 [255] to infer the direction of effect between gene expression and DNAme following the method outlined in Millstein et al. [254]. Briefly, I ran each trio twice, flipping the intermediate and outcome variable between gene expression and methylation, testing for two possible models: (1) expression mediated by DNAme (M→E) where the outcome is expression and the intermediate is DNAme and (2) DNAme mediated by expression (E→M) where the outcome is DNAme and the intermediate is expression. Using 100 permutations to calculate the FDR with the fdr.cit function, I classified each eQTM as M→E, E→M, independent, or unclassified, as performed by Millstein et al. [254], except using q-values instead of p-values. With a q-value threshold of 0.05, I classified cases where (a) M→E q-value < threshold and E→M q-value > threshold as "M→E", (b) M→E q-value > threshold and E→M q-value < threshold as "E→M", (c) M→E q-value > threshold and E→M q-value > threshold as "independent", and finally (d) M→E q-value < threshold and E→M q-value < threshold as "no prediction" or "bi-directional". An example of each scenario from the FUSION data is shown in Figure 5.18.

In both the genome wide and GWAS analysis, I found that the majority of eQTMs show independent effects (Figure 5.19). For the genome wide results, the fraction of independent genetic effects increased according to the eQTM distance of the methylation site to gene TSS (Figure 5.20), consistent with a model of large haplotype effects independently influencing multiple molecular traits. In addition, across all mediation predictions, I found that the eQTMs tended to have a negative effect. Focusing on the most confident mediation predictions where both QTL effects are significant (FDR 1%), I found the fraction of negative eQTMs sharply rose for M→E cases to > 75%, which is consistent with the traditional understanding that in cases where DNAme is actively controlling gene expression, it decreases expression. Interestingly, when considering the best trios, where both QTL effects are significant, I found the M→E cases also had a larger fraction of antisense genes compared to other types of mediation. While intriguing, it is difficult to establish if this is true signal or by random chance, as this fraction is based on 18/106 M→E cases.

For the GWAS QTLs, I discovered one case that was not independent or unclassified, where the effect of rs1019503 on cg27515589 DNAme was predicted to be mediated by the expression of the antisense gene, *CTD-2260A17.2*. However, closer examination of the rs1019503 locus revealed high complexity as rs1019503 is also a strong eQTL for *ERAP2* (cond. p = $6.01 \times 10^{-103}$), *LNPEP* (cond. p = $1.65 \times 10^{-41}$), and *CTD-2260A17.2* (cond. p = $1.14 \times 10^{-41}$), making it difficult to interpret this result.

Overall, the results I found are consistent with the results from Gutierrez-Arcelus et al. [141], who used similar mediation methods to compare exQTL and mQTL effects in fibroblasts, lymphoblastoid cells, and T-cells. Across all tissues the authors found ~50% of cases were independent. More recently, Ng et al. [273] performed a similar analysis in the cortex, dissecting the relationship between gene expression and the epigenome using the intersection between eQTLs, mQTLs, and histone QTLs (hQTLs). The authors found ~85% of cases tested showed independent trends. While the exact fraction of independent relationships differs between studies, presumably attributable to different epigenome assays (both studies use 450k DNAme arrays, DNAme vs histone marks, etc.), eQTM parameters (eQTMs or exQTMs, window size, etc.), trio inclusion criteria, and methods (e.g., CIT with permutations



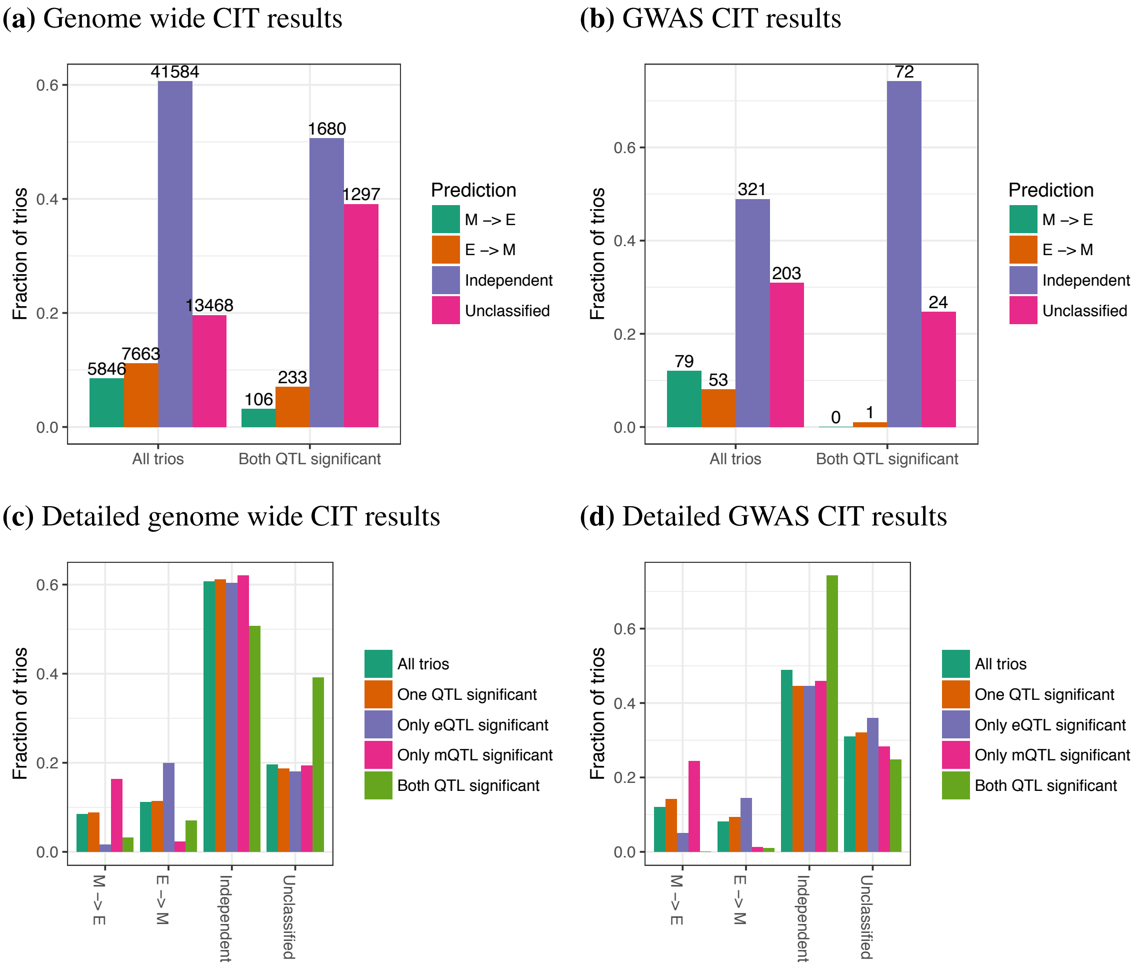
**(a)** M→E

**(b)** E→M

**(c)** Independent

**(d)** Unclassified

**Figure 5.18** Examples of molecular mediation. (a) DNAme → Expression scenario where SNP affects DNAme which changes expression. (b) Expression → DNAme scenario where SNP affects expression which changes DNAme. (c) Independent scenario where the SNP effect on DNAme and expression is statistically independent. (d) No prediction or bi-directional scenario where a DNAme prediction cannot be made.



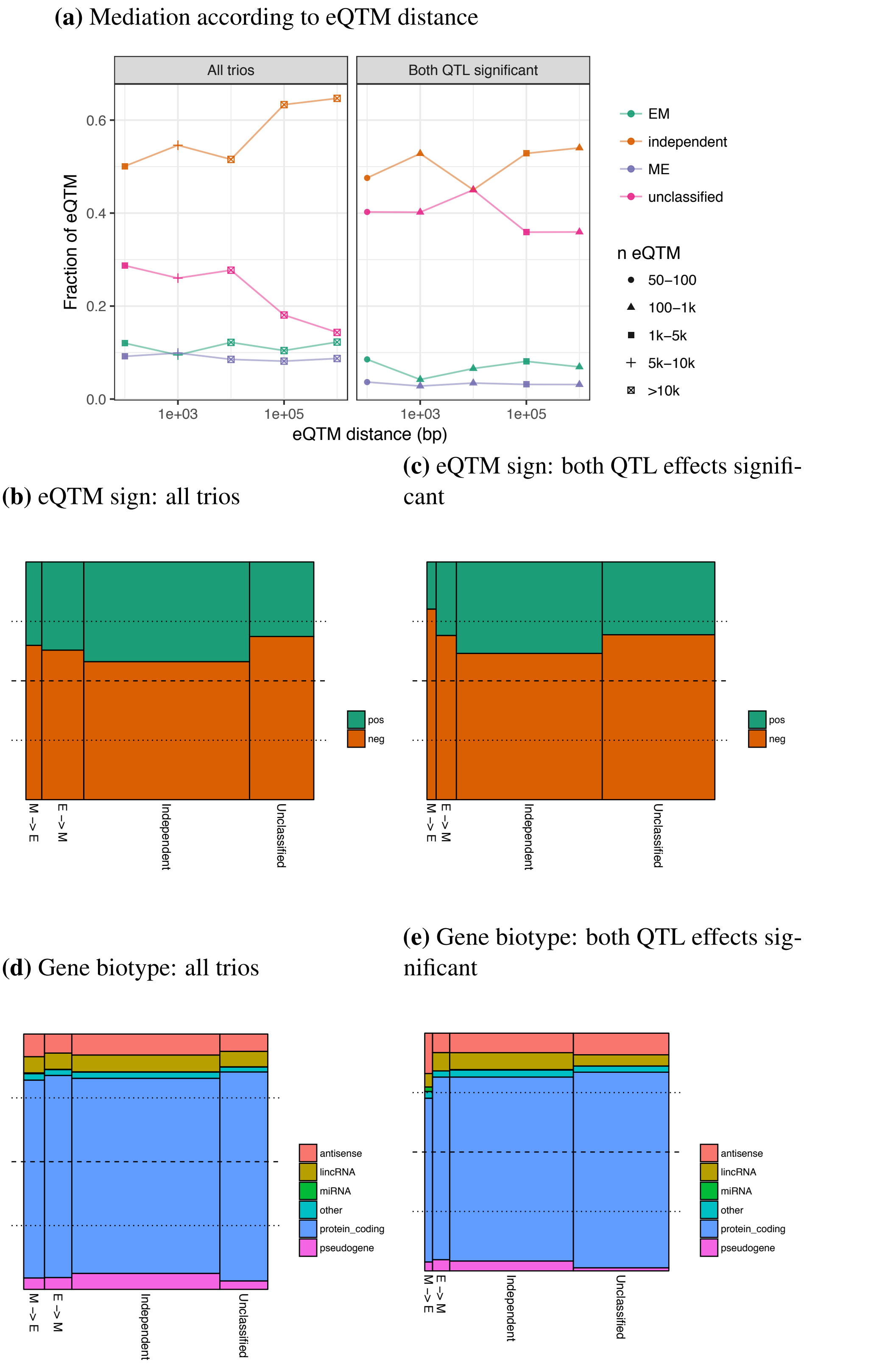
**(a)** Genome wide CIT results

**(b)** GWAS CIT results

**(c)** Detailed genome wide CIT results

**(d)** Detailed GWAS CIT results

**Figure 5.19** Molecular mediation summary. (a) Mediation predictions across all trios, genome wide. (b) Mediation predictions across trios for T2D and glycemic traits GWAS SNPs. (c) Detailed breakdown of genome wide mediation predictions. (d) Detailed breakdown of trios for T2D and glycemic traits GWAS SNPs.



(**a**) Mediation according to eQTM distance

(**b**) eQTM sign: all trios

(**c**) eQTM sign: both QTL effects significant

(**d**) Gene biotype: all trios

(**e**) Gene biotype: both QTL effects significant

**Figure 5.20** Characterisation of genome wide molecular mediation results (not GWAS focused). (a) Mediation fractions binned by eQTM distance. Bins range from 100 bp, 1 kb, 10 kb, 100 kb, to 1 Mb, and are non-overlapping. (b) Sign of eQTM effects split by mediation predictions. (c) Sign of eQTM effects split by mediation predictions across trios where both QTL effects are significant. (d) Gene biotype of eQTM split by mediation predictions across all trios. (e) Gene biotype of eQTM split by mediation predictions across trios where both QTL effects are significant.

or not), there is a clear overall trend of statistically independent genetic effects on gene expression and DNAme.

### 5.4.3 Molecular trait mediation summary

To date, the exploration of mediation between molecular traits is in its infancy, and it is difficult to draw further conclusions about the general relationship between genetic variation, gene expression, and methylation. But it is clear that simplistic views about methylation as a universal primary driver of gene expression (the M→E model), or universal secondary consequence of gene expression levels (the E→M model) cannot be defended in the face of this muscle data set or other similar analyses from other tissues [141, 273].

However, it should be noted that there are significant limitations in the analysis presented. First, the detection of DNAme was performed using arrays. As previously noted (Section 4.3.2) the EPIC array is not an unbiased survey of DNAme and is targeted to specific genomic regions. Compared to the 450k arrays, the EPIC array measures DNAme of more intergenic regions, which may explain the differences between the observed fractions and those from other studies [141, 273], both of which used 450k arrays (in addition to several other different parameters).

In addition, these results may be strongly influenced by the CIT assumption of minimal measurement error, meaning the difference between what is measured and truth. Measurement error could be technical in nature, due to difficulties in measuring a molecular trait (e.g., background noise or batch effects). Alternatively, measurement error may be biological in nature, where the molecular trait is measured perfectly, but the signal of the biological driver is diluted by other biological factors (e.g., tissue heterogeneity). Such departures from the true, underlying signal could lead the CIT to infer the wrong causal direction [153]. To date, the CIT has been the primary method used to investigate mediation between molecular traits [141, 273]. However, given these limitations, in the future it will be important to further develop and apply alternative methods with different assumptions (e.g., Mendelian randomisation methods [153]) to answer questions of molecular mediation and identify robust mediation signals that are consistent across methods and assumptions.

Finally, in addition to leveraging alternative approaches, the analysis I performed could be improved by (1) discovering and estimating effect sizes from different datasets and (2) using multiple SNPs or instrument variables. For genetic instrument analyses, using the same dataset for discovery and effect estimates can lead to biases [50, 49], although for very strong genetic effects (like those typically found for molecular traits) these biases may be minimal [49]. Practically, I had to use the same dataset for discovery and effect estimates as an external skeletal muscle mQTL dataset was not available. In addition, such analyses are most robust when multiple, independent predictor SNPs are used, as they allow one to identify potential aberrant SNP effects and ensure consistency across multiple SNP predictors [152]. To date, these methods have mainly been developed in the context of higher level phenotypes (e.g., metabolites or disease outcomes) where GWA studies have identified multiple, independent loci, scattered throughout the genome. In the context of molecular traits, the extent to which multiple variants meaningfully contribute to molecular traits and the reproducibility of such signals is poorly understood. These questions highlight areas for further research using the FUSION tissue biopsy dataset, as well as other increasingly large datasets like GTEx.

| | Gene | GWAS SNP | Minor allele | Marg. p-value | Cond. p-value | Cond. q-value | Cond. beta | Numb. eQTL | Numb. eQTM | Gene mESI |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | *ERAP2* | rs1019503 | G | 7.64e-77 | 6.01e-103 | 6.35e-99 | -1.2 | 3 | 12 | 0.42 |
| 2 | *RCCD1* | rs79548680 | C | 3.02e-37 | 1.11e-42 | 5.89e-39 | 1.2 | 7 | 6 | 0.42 |
| 3 | *CTD-2260A17.2* | rs1019503 | G | 1.91e-37 | 1.14e-41 | 4.02e-38 | 0.91 | 5 | 7 | 0.44 |
| 4 | *LNPEP* | rs1019503 | G | 6.83e-42 | 1.65e-41 | 4.36e-38 | 0.94 | 3 | 8 | 0.49 |
| 5 | *FADS1* | rs174550 | C | 3.74e-27 | 1.43e-28 | 2.51e-25 | -0.91 | 2 | 2 | 0.45 |
| 6 | *FADS1* | rs174546 | T | 3.74e-27 | 1.43e-28 | 2.51e-25 | -0.91 | 2 | 2 | 0.45 |
| 7 | *SLU7* | rs1895320 | G | 1.02e-11 | 1.11e-27 | 1.68e-24 | 1.1 | 3 | 0 | 0.5 |
| 8 | *ANK1* | rs515071 | A | 1.1e-26 | 6.61e-26 | 7.76e-23 | -0.99 | 2 | 6 | 0.7 |
| 9 | *ANK1* | rs516946 | T | 1.1e-26 | 6.61e-26 | 7.76e-23 | -0.99 | 2 | 6 | 0.7 |
| 10 | *RHOA* | rs11715915 | T | 1.47e-06 | 7.34e-23 | 7.75e-20 | -1 | 3 | 17 | 0.47 |
| 11 | *CCHCR1* | rs3132524 | T | 2.76e-22 | 8.47e-23 | 8.14e-20 | 0.86 | 4 | 28 | 0.49 |
| 12 | *WFS1* | rs4689388 | G | 8.68e-21 | 2.02e-22 | 1.78e-19 | -0.64 | 6 | 8 | 0.4 |
| 13 | *GPN1* | rs3749147 | A | 1.66e-21 | 1.66e-21 | 1.35e-18 | 0.82 | 1 | 11 | 0.49 |
| 14 | *PRC1-AS1* | rs79548680 | C | 3.68e-20 | 2.29e-21 | 1.73e-18 | -0.92 | 4 | 5 | 0.44 |
| 15 | *CYP21A2* | rs3099844 | A | 6.51e-13 | 5.95e-21 | 4.19e-18 | 1.1 | 9 | 15 | 0.35 |
| 16 | *ZNF697* | rs478093 | A | 4.17e-04 | 6.91e-20 | 4.56e-17 | 0.72 | 6 | 4 | 0.46 |
| 17 | *GPN1* | rs13022873 | C | 2.39e-20 | 3.78e-19 | 2.22e-16 | 0.72 | 4 | 11 | 0.49 |
| 18 | *GPN1* | rs1919128 | G | 2.39e-20 | 3.78e-19 | 2.22e-16 | 0.72 | 4 | 11 | 0.49 |
| 19 | *TMEM99* | rs7219451 | C | 2.36e-13 | 4.03e-19 | 2.24e-16 | 0.78 | 2 | 2 | 0.47 |
| 20 | *INTS8* | rs17359493 | G | 2.79e-01 | 1.99e-17 | 1.05e-14 | -2.3 | 10 | 0 | 0.47 |
| 21 | *NICN1* | rs11715915 | T | 3.16e-17 | 3.02e-17 | 1.52e-14 | 0.9 | 2 | 9 | 0.44 |
| 22 | *WFS1* | rs4458523 | T | 1.06e-18 | 1.49e-16 | 7.14e-14 | -0.59 | 4 | 8 | 0.4 |
| 23 | *ZBTB20* | rs73230612 | T | 7.63e-16 | 7.63e-16 | 3.51e-13 | 0.87 | 1 | 0 | 0.52 |
| 24 | *WFS1* | rs1801214 | C | 3.63e-17 | 9.35e-16 | 4.12e-13 | -0.57 | 4 | 8 | 0.4 |
| 25 | *ZFAND3* | rs9470794 | C | 5.86e-09 | 1.02e-15 | 4.32e-13 | 1.1 | 8 | 1 | 0.52 |

**Table 5.2** Conditional GWAS eQTLs. Number eQTM reports the number of DNAme sites associated with a gene (FDR 1%). Effect estimates are oriented to the minor allele.

| | Gene | Exon fragment | GWAS SNP | Minor allele | Marg. p-value | Cond. p-value | Cond. q-value | Cond. beta | Numb. exQTL | Numb. exQTM | Gene mESI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | *ERAP2* | 006 | rs1019503 | G | 3.83e-76 | 7.91e-105 | 1.41e-99 | -1.2 | 6 | 10 | 0.42 |
| 2 | *ANK1* | 010 | rs516946 | T | 1.43e-53 | 8.67e-66 | 5.34e-62 | 1.4 | 5 | 8 | 0.7 |
| 3 | *ANK1* | 010 | rs515071 | A | 1.43e-53 | 8.67e-66 | 5.34e-62 | 1.4 | 5 | 8 | 0.7 |
| 4 | *WARS* | 062 | rs3783347 | T | 2e-61 | 1.98e-64 | 1.18e-60 | 1.5 | 2 | 8 | 0.47 |
| 5 | *ZBTB20* | 051 | rs73230612 | T | 2.99e-52 | 2.22e-57 | 1.13e-53 | 1.5 | 4 | 1 | 0.52 |
| 6 | *HERPUD1* | 008 | rs2217332 | A | 1e-53 | 2.35e-57 | 1.17e-53 | 1.5 | 3 | 0 | 0.48 |
| 7 | *FGGY* | 015 | rs835367 | G | 1.74e-29 | 1.45e-53 | 6.84e-50 | 0.99 | 4 | 2 | 0.57 |
| 8 | *CARD9,DNLZ* | 003 | rs28642252 | A | 2.44e-05 | 5.83e-47 | 2.42e-43 | -2.1 | 6 | 6 | 0.37,0.48 |
| 9 | *LNPEP,AC008865.1* | 031 | rs1019503 | G | 3.68e-41 | 1.99e-46 | 8.06e-43 | 0.97 | 4 | 9 | 0.49,0.47 |
| 10 | *CCHCR1* | 030 | rs3132524 | T | 1.1e-34 | 2.04e-45 | 7.6e-42 | 1.2 | 3 | 27 | 0.49 |
| 11 | *SDHAF4* (*C6orf57*) | 003 | rs1048886 | G | 2.07e-44 | 2.93e-44 | 1.07e-40 | 1.1 | 2 | 3 | 0.49 |
| 12 | *CTD-2260A17.2* | 010 | rs1019503 | G | 5.5e-41 | 9.28e-43 | 3.31e-39 | 0.92 | 3 | 8 | 0.44 |
| 13 | *GLRX5* | 003 | rs7120 | G | 1.59e-06 | 5.17e-40 | 1.65e-36 | -1.5 | 6 | 0 | 0.52 |
| 14 | *RCCD1* | 034 | rs79548680 | C | 2.9e-29 | 1.33e-36 | 3.95e-33 | 1.2 | 4 | 2 | 0.42 |
| 15 | *AMT* | 002 | rs11715915 | T | 1.64e-13 | 3.1e-34 | 8.64e-31 | 1.2 | 2 | 17 | 0.44 |
| 16 | *HLA-C* | 015 | rs3099844 | A | 8e-07 | 4.17e-30 | 1.09e-26 | 1.2 | 5 | 28 | 0.43 |
| 17 | *RHOA* | 001 | rs11715915 | T | 3.36e-07 | 1.95e-26 | 4.65e-23 | -1.1 | 2 | 15 | 0.47 |
| 18 | *CAMK2B* | 011 | rs3757840 | T | 1.51e-10 | 2.96e-26 | 6.96e-23 | 0.84 | 4 | 2 | 0.57 |
| 19 | *HLA-B* | 015 | rs2244020 | G | 1.67e-15 | 3.16e-26 | 7.32e-23 | -0.72 | 9 | 44 | 0.42 |
| 20 | *FADS1,MIR1908* | 001 | rs174546 | T | 3.92e-24 | 3.92e-24 | 8.33e-21 | -0.78 | 1 | 3 | 0.45,0.4 |
| 21 | *FADS1,MIR1908* | 001 | rs174550 | C | 3.92e-24 | 3.92e-24 | 8.33e-21 | -0.78 | 1 | 3 | 0.45,0.4 |
| 22 | *FADS1,MIR1908* | 001 | rs1535 | G | 6.83e-24 | 6.83e-24 | 1.42e-20 | -0.77 | 1 | 3 | 0.45,0.4 |
| 23 | *FADS1,MIR1908* | 001 | rs102275 | C | 1.54e-23 | 1.54e-23 | 3.17e-20 | -0.77 | 1 | 3 | 0.45,0.4 |
| 24 | *SETD9* | 004 | rs10461617 | A | 1.61e-21 | 7e-23 | 1.37e-19 | -0.93 | 2 | 4 | 0.49 |
| 25 | *NDUFAF6* | 015 | rs7845219 | T | 1.32e-18 | 4.26e-22 | 7.85e-19 | -0.71 | 4 | 9 | 0.51 |

**Table 5.3** Conditional GWAS exQTLs. In cases where the exon from two genes overlap, a new fragment of the overlapping region is created. Such cases are reported as lists of the overlapping gene names followed by the exon fragment number. To avoid redundancy, only the GWAS tag SNP-gene pair with the minimum conditional p-value are reported. Number exQTM corresponds to the number of DNAme sites associated with an exon (FDR 1%). Effect estimates are oriented to the minor allele. The mESI values are calculated from gene level expression and in the case of an overlapping fragment from multiple genes, both mESI values are listed, corresponding to the genes in the first column.

| | Probe | eQTM | exQTM | GWAS SNP | Minor allele | Marg. p-value | Cond. p-value | Cond. q-value | Cond. beta | Numb. mQTL | Numb. eQTM | Numb. exQTM | MeSS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | cg20217307 | >3 | >3 | rs3132524 | T | 3.19e-70 | 2.46e-77 | 1.05e-71 | -1.4 | 5 | 6 | 68 | 0.16 |
| 2 | cg17816406 | *WFS1* | *WFS1* | rs4689388 | G | 1.25e-77 | 2.72e-72 | 5.84e-67 | -1 | 2 | 1 | 5 | 0.11 |
| 3 | cg14810798 | *PROX1-AS1* | *PROX1-AS1* | rs2075423 | T | 2.99e-71 | 2.99e-71 | 3.23e-66 | 1.2 | 1 | 1 | 1 | 0.12 |
| 4 | cg19610905 | *FADS1* | >3 | rs174550 | C | 4.16e-71 | 4.16e-71 | 3.23e-66 | -1.2 | 1 | 1 | 10 | 0.063 |
| 5 | cg19610905 | *FADS1* | >3 | rs174546 | T | 4.16e-71 | 4.16e-71 | 3.23e-66 | -1.2 | 1 | 1 | 10 | 0.063 |
| 6 | cg19610905 | *FADS1* | >3 | rs1535 | G | 4.53e-71 | 4.53e-71 | 3.23e-66 | -1.2 | 1 | 1 | 10 | 0.063 |
| 7 | cg19610905 | *FADS1* | >3 | rs102275 | C | 1.71e-69 | 1.71e-69 | 1.05e-64 | -1.2 | 1 | 1 | 10 | 0.063 |
| 8 | cg03523917 | | | rs4712523 | G | 4.99e-70 | 2.93e-69 | 1.1e-64 | 1.1 | 3 | 0 | 0 | NA |
| 9 | cg03523917 | | | rs10946398 | C | 4.99e-70 | 2.93e-69 | 1.1e-64 | 1.1 | 3 | 0 | 0 | NA |
| 10 | cg03523917 | | | rs7754840 | C | 4.99e-70 | 2.93e-69 | 1.1e-64 | 1.1 | 3 | 0 | 0 | NA |
| 11 | cg03523917 | | | rs7772603 | C | 4.99e-70 | 2.93e-69 | 1.1e-64 | 1.1 | 3 | 0 | 0 | NA |
| 12 | cg03523917 | | | rs4712524 | G | 5.6e-70 | 3.09e-69 | 1.1e-64 | 1.1 | 3 | 0 | 0 | NA |
| 13 | cg23819653 | >3 | >3 | rs3132524 | T | 1.91e-61 | 5.17e-69 | 1.71e-64 | -1.4 | 4 | 4 | 120 | 0.16 |
| 14 | cg05052969 | *PROX1-AS1* | | rs2075423 | T | 6.04e-63 | 1.64e-68 | 5.01e-64 | 1.2 | 3 | 1 | 0 | 0.13 |
| 15 | cg02569219 | | | rs4790333 | C | 1.5e-65 | 2.22e-68 | 6.34e-64 | -1.2 | 2 | 0 | 0 | 0.089 |
| 16 | cg16049864 | *NDUFAF6* | *NDUFAF6,INTS8* | rs896854 | T | 2.05e-62 | 2.8e-66 | 6.92e-62 | 1.1 | 2 | 1 | 5 | 0.086 |
| 17 | cg08035822 | *SPPL3* | *SPPL3* | rs7957197 | A | 2.44e-47 | 2.78e-66 | 6.92e-62 | 1.3 | 6 | 1 | 1 | 0.32 |
| 18 | cg08035822 | *SPPL3* | *SPPL3* | rs12427353 | C | 4.02e-47 | 2.91e-66 | 6.92e-62 | 1.3 | 6 | 1 | 1 | 0.32 |
| 19 | cg17985300 | *TMEM99* | *CASC3,TMEM99* | rs7219451 | C | 1.93e-14 | 3.35e-65 | 7.54e-61 | -1.4 | 2 | 1 | 4 | 0.29 |
| 20 | cg20564521 | *TMEM50B* | *TMEM50B* | rs2268241 | A | 3.09e-63 | 6.53e-65 | 1.4e-60 | 1.5 | 3 | 1 | 7 | 0.4 |
| 21 | cg16049864 | *NDUFAF6* | *NDUFAF6,INTS8* | rs7845219 | T | 7.49e-61 | 1.87e-64 | 3.82e-60 | 1.1 | 2 | 1 | 5 | 0.086 |
| 22 | cg13393036 | *NDUFAF6* | *NDUFAF6,INTS8* | rs896854 | T | 1.48e-57 | 5.49e-64 | 1.07e-59 | 1.1 | 4 | 1 | 6 | 0.35 |
| 23 | cg01883759 | *JAZF1* | *JAZF1* | rs849135 | A | 1.03e-22 | 8.51e-62 | 1.58e-57 | -1.2 | 2 | 1 | 4 | 0.078 |
| 24 | cg06032855 | | *CAMK2B* | rs3757840 | T | 3.15e-38 | 3.88e-61 | 6.92e-57 | -1.1 | 6 | 0 | 6 | 0.13 |
| 25 | cg13182339 | *PSMD6-AS2* | | rs831571 | T | 9.47e-09 | 4.04e-61 | 6.93e-57 | 1.3 | 4 | 1 | 0 | 0.17 |

**Table 5.4** Conditional GWAS mQTLs. Number eQTM reports the number of genes associated with a DNAme site (FDR 1%). Effect estimates are oriented to the minor allele. NA MeSS denotes insufficient methylation signal in the reference panel to estimate specificity.

| | Gene | GWAS SNP | Minor allele | Marg. p-value | Cond. p-value | Cond. q-value | Cond. beta | Numb. eQTL | Numb. eQTM | Gene mESI |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | *ANK1* | rs515071 | A | 1.1e-26 | 6.61e-26 | 7.76e-23 | -0.99 | 2 | 6 | 0.7 |
| 2 | *ANK1* | rs516946 | T | 1.1e-26 | 6.61e-26 | 7.76e-23 | -0.99 | 2 | 6 | 0.7 |
| 3 | *ASB16* | rs12602486 | G | 8.78e-02 | 2.21e-08 | 3.44e-06 | 0.95 | 5 | 7 | 0.65 |
| 4 | *KCNJ11* | rs5219 | T | 2.29e-02 | 3.09e-08 | 4.6e-06 | 0.56 | 7 | 1 | 0.71 |
| 5 | *KCNJ11* | rs5215 | C | 2.29e-02 | 3.09e-08 | 4.6e-06 | 0.56 | 7 | 1 | 0.71 |
| 6 | *IDE* | rs7087591 | A | 2.06e-01 | 1.35e-07 | 1.68e-05 | 0.7 | 5 | 0 | 0.55 |
| 7 | *CEP85* | rs6598955 | T | 2.09e-01 | 6.08e-07 | 6.3e-05 | -0.76 | 3 | 0 | 0.59 |
| 8 | *RP3-414A15.10* | rs11159086 | C | 4.48e-02 | 9.24e-07 | 9.13e-05 | -0.46 | 9 | 0 | 0.55 |
| 9 | *IDE* | rs5015480 | T | 7.88e-02 | 2.43e-05 | 1.42e-03 | -0.58 | 5 | 0 | 0.55 |
| 10 | *IDE* | rs1111875 | T | 7.88e-02 | 2.43e-05 | 1.42e-03 | -0.58 | 5 | 0 | 0.55 |
| 11 | *RP11-800A3.4* | rs10898909 | A | 3.81e-03 | 5.68e-05 | 2.82e-03 | 0.35 | 3 | 0 | 0.64 |
| 12 | *ACHE* | rs7636 | A | 3.58e-04 | 7.65e-05 | 3.5e-03 | -0.86 | 3 | 8 | 0.59 |
| 13 | *LPL* | rs7841189 | T | 4.92e-03 | 1.26e-04 | 5.07e-03 | 0.5 | 5 | 2 | 0.57 |
| 14 | *ACADS* | rs7305618 | T | 2.29e-01 | 1.5e-04 | 5.66e-03 | -0.34 | 8 | 1 | 0.58 |
| 15 | *H19* | rs7107784 | G | 4.79e-01 | 1.83e-04 | 6.61e-03 | 1.7 | 8 | 0 | 0.67 |
| 16 | *MYL1* | rs715 | C | 1.15e-01 | 2.12e-04 | 7.41e-03 | -0.25 | 6 | 3 | 0.97 |
| 17 | *LCN8* | rs3829109 | A | 8.16e-02 | 2.39e-04 | 8.19e-03 | 0.16 | 7 | 19 | 0.59 |
| 18 | *THBS4* | rs17823642 | T | 2.57e-01 | 2.41e-04 | 8.25e-03 | 0.79 | 7 | 0 | 0.59 |
| 19 | *TNK2-AS1* | rs843532 | C | 1.5e-01 | 2.54e-04 | 8.62e-03 | 0.4 | 5 | 0 | 0.58 |
| 20 | *LINC00346* | rs12853515 | G | 2.23e-02 | 2.74e-04 | 9.05e-03 | 0.42 | 6 | 0 | 0.74 |
| 21 | *GS1-388B5.1* | rs1107366 | G | 5.17e-01 | 2.85e-04 | 9.3e-03 | 0.19 | 9 | 4 | 0.75 |

**Table 5.5** Conditional GWAS eQTLs for muscle specific genes. Effect estimates are oriented to the minor allele.

| | Probe | eQTM | exQTM | GWAS SNP | Minor allele | Marg. p-value | Cond. p-value | Cond. q-value | Cond. beta | Numb. mQTL | Numb. eQTM | Numb. exQTM | MeSS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | cg04602696 | *PIEZO1* | | rs9933309 | T | 4.31e-49 | 1.08e-56 | 1.55e-52 | 1.1 | 5 | 1 | 0 | 0.53 |
| 2 | cg10251070 | | | rs496300 | C | 3.69e-61 | 1.36e-56 | 1.83e-52 | 1 | 3 | 0 | 0 | 0.6 |
| 3 | cg01678292 | *ANK1* | *ANK1* | rs515071 | A | 3.87e-46 | 3.87e-46 | 3.25e-42 | 1.3 | 1 | 1 | 16 | 0.69 |
| 4 | cg01678292 | *ANK1* | *ANK1* | rs516946 | T | 3.87e-46 | 3.87e-46 | 3.25e-42 | 1.3 | 1 | 1 | 16 | 0.69 |
| 5 | cg12439423 | *ANK1* | *ANK1* | rs516946 | T | 1.03e-41 | 1.03e-41 | 7.03e-38 | 1.2 | 1 | 1 | 17 | 0.73 |
| 6 | cg12439423 | *ANK1* | *ANK1* | rs515071 | A | 1.03e-41 | 1.03e-41 | 7.03e-38 | 1.2 | 1 | 1 | 17 | 0.73 |
| 7 | cg25308173 | | | rs1107366 | G | 8.86e-30 | 8.82e-40 | 5.25e-36 | -0.97 | 3 | 0 | 0 | 0.49 |
| 8 | cg24614755 | *RP11-307C19.1*,★ | *RP11-307C19.1* | rs7119 | T | 6.77e-19 | 3.32e-39 | 1.92e-35 | -1 | 3 | 2 | 1 | 0.58 |
| 9 | cg21926782 | >3 | >3 | rs11715915 | T | 1.09e-11 | 7.43e-39 | 4.19e-35 | -1.3 | 3 | 7 | 40 | 0.69 |
| 10 | cg25729166 | *PEAK1*,♣,♠ | *PEAK1*,♣ | rs7119 | T | 6.57e-21 | 1.6e-36 | 8.17e-33 | -0.97 | 2 | 3 | 2 | 0.6 |
| 11 | cg23241016 | *ANK1* | *ANK1* | rs516946 | T | 1.93e-36 | 1.93e-36 | 9.64e-33 | 1.2 | 1 | 1 | 16 | 0.7 |
| 12 | cg23241016 | *ANK1* | *ANK1* | rs515071 | A | 1.93e-36 | 1.93e-36 | 9.64e-33 | 1.2 | 1 | 1 | 16 | 0.7 |
| 13 | cg03351003 | | | rs7202877 | G | 3.35e-34 | 3.35e-34 | 1.56e-30 | -1.3 | 1 | 0 | 0 | 0.55 |
| 14 | cg00498772 | | | rs8048589 | C | 2.34e-40 | 2.33e-32 | 9.71e-29 | 1.3 | 3 | 0 | 0 | 0.58 |
| 15 | cg03940650 | | | rs4846922 | T | 1.12e-10 | 2.59e-30 | 9.55e-27 | 1.1 | 2 | 0 | 0 | 0.76 |
| 16 | cg11250194 | | | rs1535 | G | 3.61e-02 | 1.92e-29 | 6.79e-26 | -1.2 | 2 | 0 | 0 | 0.49 |
| 17 | cg11250194 | | | rs174546 | T | 3.38e-02 | 2.55e-29 | 8.89e-26 | -1.2 | 2 | 0 | 0 | 0.49 |
| 18 | cg11250194 | | | rs174550 | C | 3.38e-02 | 2.55e-29 | 8.89e-26 | -1.2 | 2 | 0 | 0 | 0.49 |
| 19 | cg11839944 | | | rs5215 | C | 8.03e-36 | 1.19e-28 | 3.97e-25 | -0.86 | 2 | 0 | 0 | 0.62 |
| 20 | cg11839944 | | | rs5219 | T | 8.03e-36 | 1.19e-28 | 3.97e-25 | -0.86 | 2 | 0 | 0 | 0.62 |
| 21 | cg11479568 | *ANK1* | *ANK1* | rs516946 | T | 1.5e-27 | 1.5e-27 | 4.5e-24 | 1 | 1 | 1 | 15 | 0.68 |
| 22 | cg11479568 | *ANK1* | *ANK1* | rs515071 | A | 1.5e-27 | 1.5e-27 | 4.5e-24 | 1 | 1 | 1 | 15 | 0.68 |
| 23 | cg17274126 | *ANK1* | *ANK1* | rs516946 | T | 6.81e-27 | 6.81e-27 | 1.95e-23 | 1 | 1 | 1 | 14 | 0.68 |
| 24 | cg17274126 | *ANK1* | *ANK1* | rs515071 | A | 6.81e-27 | 6.81e-27 | 1.95e-23 | 1 | 1 | 1 | 14 | 0.68 |
| 25 | cg09029193 | | | rs73069940 | G | 3.99e-25 | 8.52e-27 | 2.42e-23 | 1.5 | 2 | 0 | 0 | 0.68 |

**Table 5.6** Conditional GWAS mQTLs for muscle specific DNAme sites. Number eQTM reports the number of genes associated with a DNAme site (FDR 1%). Effect estimates are oriented to the minor allele. NA MeSS denotes insufficient methylation signal in the reference panel to estimate specificity. ★ *RP11-307C19.2*; ♣ *RP11-307C19.1*; ♠ *RP11-307C19.2*.

# Chapter 6

# Interactions between genetic variation and cellular environment in gene expression

## 6.1 Introduction

In this chapter, I investigate gene-environment interactions using the skeletal muscle gene expression data freeze from Scott et al. [341], as described earlier (Section 2.8). This analysis capitalizes on the unique aspect of the FUSION tissue biopsy dataset compared to other QTL datasets like GTEx, namely the availability of rich phenotype information on all participants. I conducted this analysis while the DNAme data was being generated. At the time of writing, this study has been stored on *bioRxiv* [384] and is currently under review.

A substantial fraction of variability in gene expression is controlled by changes in transcription rates, mainly mediated by transcription factor (TF) proteins binding to specific DNA sequence motifs that define regulatory elements [208, 90]. The abundance of such proteins and their regulatory co-factors may in turn be controlled by intrinsic mechanisms inherent to a cell, such as an individual's genetic makeup or regulatory programs specific to a cell type, as well as cellular responses to environmental cues. A regulatory element, defined by the DNA region recognised by a DNA-binding TF and other required transcriptional machinery, may be either intrinsic or environment-dependent. In intrinsic elements, the TF and binding machinery is controlled by cell-intrinsic mechanisms that operate within a closed system and are unresponsive to environment. By contrast, in environment-dependent elements the TF and binding machinery is responsive to an environmental stimulus. Both regulatory element types are susceptible to perturbation by genetic variation because the region recognised by the TF is encoded in the DNA sequence.

In this thesis, I have documented the effects of genetic perturbations of regulatory elements on gene expression—expression quantitative trait loci (eQTLs). Variation in intrinsic regulatory programs is expected to give rise to such "standard eQTLs", identified by modelling genetic effects on gene expression (Equation 4.1 in Chapter 4). However, it is also likely that variation in environment-dependent elements will be detected in standard eQTL studies. For an environment-dependent regulatory variant to pass undetected in a standard eQTL study, the variant must change the relationship between gene expression and environment without altering the mean gene expression levels for each genotype, an unlikely event. Therefore we would expect a subset of eQTLs detected by modelling only genetic effects to also have effects mediated by an environmental context. If one were to model the combined environmental and genetic effects on gene expression, such variants would exhibit interaction effects between genotype and environment (GxE) and could be described as environmental response expression quantitative trait loci (reQTLs), a specific type of eQTL whose effect changes in response to an environmental context. To date, the overlap between standard eQTLs and reQTLs in human is largely unknown, as few studies have co-measured environmental and genetic effects at scale, and the technology for mapping such reQTLs is in its infancy.

In human populations, several GxE signals have been reported across diseases for various quantitative traits (reviewed in [157]), but only a handful of transcriptional reQTLs have been mapped, treating gene expression as a molecular quantitative trait [356, 319, 135, 21, 238, 159, 235, 306, 430, 204, 96, 47, 267, 444, 188]. Indeed GxE effects have primarily been studied in model organisms where the environment and genotype can be controlled [103, 114, 200, 330, 325, 358, 217]. The challenge of mapping reQTLs using transcriptomic data outside of controlled laboratory settings lies in the confounding effects of environmental, biological, and technical factors on gene expression data, and the difficulty in isolating and/or accounting for such effects while preserving effects of the environment of interest.

However, such limitations may be mitigated if a study quantifies gene expression using RNA-seq technology because RNA-seq enables the measurement of allele specific expression (ASE), an alternative readout less prone to the confounders of gene level measurements [56, 188]. By quantifying differences in expression between haplotypes in samples heterozygous for a transcribed allele (tSNP), ASE provides an internally controlled measurement where biological and technical exposures on the cells are essentially identical for both haplotypes. This makes ASE ideal for reQTL mapping since it minimizes batch effects while preserving *cis* mediated environmental effects. Indeed, ASE has been utilised in several studies to identify genome wide GxE effects [135, 47, 267, 188], including Knowles et al. [188], who recently developed the EAGLE method (Environment-ASE through Generalised LinEar modelling), a hierarchical Bayesian model, which I apply in this study.

An additional challenge for GxE studies is validating results, which at one level can be performed within an RNA-seq study by integrating ASE with standard gene expression data between individuals (abbreviated to gene-level expression) so that the two data types serve as orthogonal forms of signal to validate reQTLs. In cases of true *cis* regulation of gene expression, when a TF preferentially binds to one allele, we would expect to observe increased ASE in participants heterozygous for the regulatory SNP. As an example, Figure 6.1 shows the different types of potential regulatory elements and the impact of different polymorphisms in schematic form. At the gene expression level, we would expect a reQTL to have different effects across environmental contexts in a genotype specific manner. In the ASE data, we would expect correlation between ASE and the environment only in individuals heterozygous for both the reQTL-SNP and tSNP. As opposed to standard eQTLs, which can be summarised by box-plots stratified by genotype, reQTLs are best described with a 6-panel regression plot, and examples of expected behaviour from real data are shown in Figure 6.2 to help orient the reader.

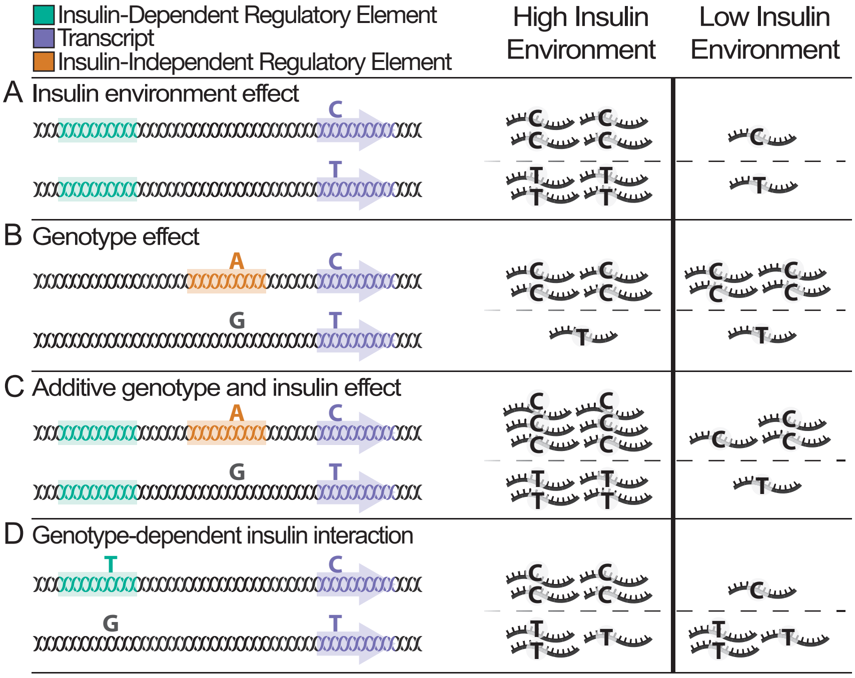
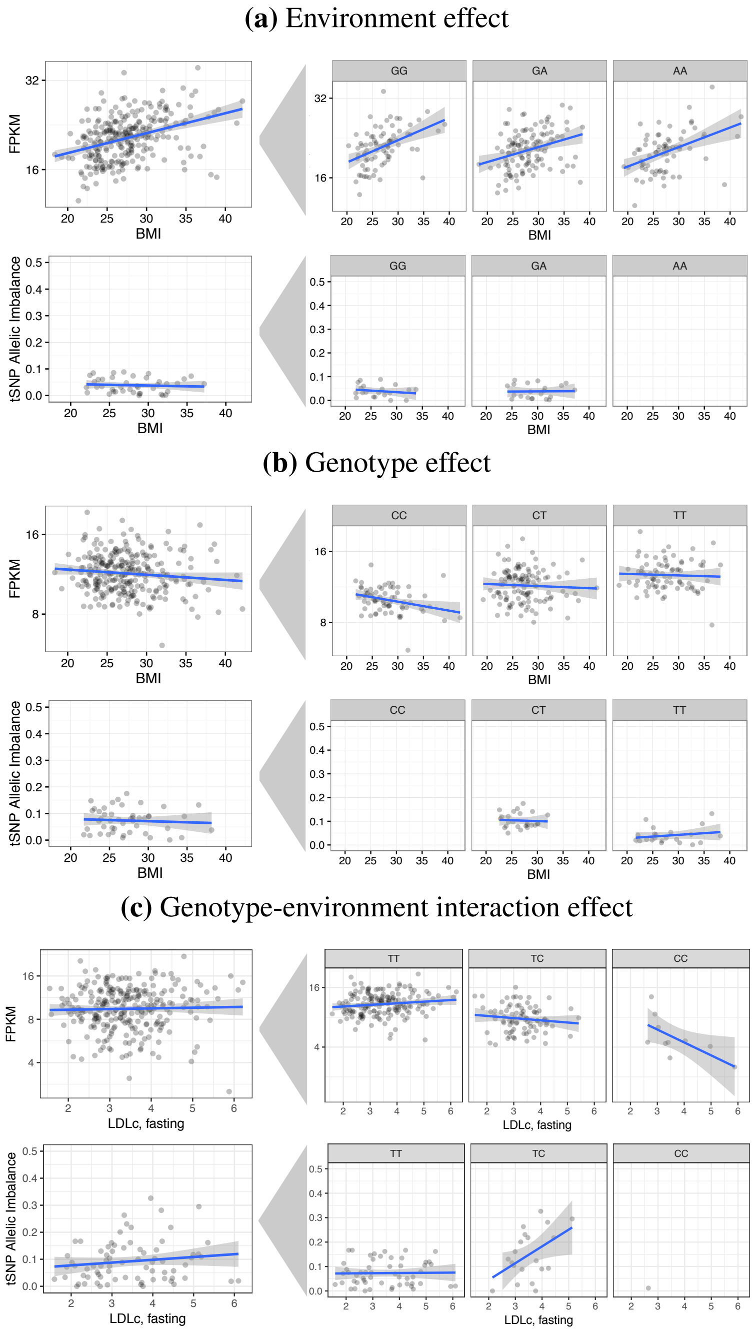

Insulin-Dependent Regulatory Element
Transcript
Insulin-Independent Regulatory Element
High Insulin Environment
Low Insulin Environment
A Insulin environment effect
B Genotype effect
C Additive genotype and insulin effect
D Genotype-dependent insulin interaction
C
T
A
G
C
T

**Figure 6.1** Idealised genetic and environmental effects on gene expression. Blood insulin levels represent a cellular environment for tissues such as skeletal muscle. The left panel depicts a single genome with colour coded genomic elements and various heterozygous sites. The right panel shows the relative transcript abundance for the corresponding locus on the left panel. Some genomic elements contain genetic variants. When the variant is the same color as the element, the element is active. In some cases the variant is black, indicating that the variant renders the regulatory element nonfunctional and only basal transcription occurs. The purple element represents a gene with a transcribed SNP (tSNP), shown in the transcripts. Allele specific expression is calculated across both chromosomes and compared to the high and low environment. (a) When regulated by an insulin-responsive element (green), gene expression changes according to insulin concentrations in the extracellular environment. (b) When regulated by an insulin-independent element (orange) containing genetic variation, gene expression changes according to the presence of a genetic variant (eQTL), but not to insulin levels. The tSNP shows allelic bias due to the eQTL effect, but is not associated with the insulin environment. (c) When regulated by both an insulin-responsive element and an insulin-independent element containing genetic variation, the effects of the insulin environment and the genetic variation on gene expression may be additive, although more complex relationships are possible. The tSNP shows some imbalance due to the eQTL effect and is associated with insulin levels. Such cases may be identified as weak reQTLs. (d) When regulated by an insulin-responsive element containing genetic variation, there may exist an interaction effect between the genetic variant and insulin levels such that changes in gene expression across insulin environments depend on the genetic variant. The tSNP shows allelic imbalance associated with insulin levels due to the reQTL effect. One of several possible interaction effects depicted.



**(a)** Environment effect

**(b)** Genotype effect

**(c)** Genotype-environment interaction effect

**Figure 6.2** Illustrative examples of genetic and environmental effects from the FUSION dataset. (a) Example of a pure environment effect in *SZRD1*—rs12568938 regulatory SNP (rSNP) and rs7529767 transcribed SNP (tSNP). *SZRD1* expression is associated with BMI, and the rSNP does not affect gene expression. The relationship between *SZRD1* and BMI does not change across the rSNP alleles, and BMI is not associated with allelic imbalance. (b) Example of a pure genetic effect in *RBM6*—rs9881008 regulatory locus and rs2023953 tSNP. BMI is not associated with RBM6 expression or allelic imbalance. The rSNP alleles are associated with *RBM6* expression and allelic imbalance is increased in samples heterozygous for the rSNP. (c) Example of a GxE effect in *FHOD3*—rs17746240 regulatory locus and rs72895597 tSNP. The relationship between LDLc and *FHOD3* expression changes according to the rSNP allele as well as the overall expression abundance levels. LDLc is only associated with allelic imbalance in heterozygous individuals, where preferential TF binding could occur.

In this study, I explore the opportunities and challenges for reQTL mapping and replication using gene-level expression and ASE data. This study capitalises on the rich clinical phenotypes in the FUSION tissue biopsy study [341], spanning blood metabolites, anthropometric measurements, and medication (Table 6.2). Collectively, I treat all clinical phenotypes as "environmental traits" since I model skeletal muscle gene expression and therefore the response of a population of cells to the surrounding cellular environment—adjacent cells, extracellular matrix, blood plasma, and interstitial fluid—approximated by each phenotype.

As one clear limitation is sample size, I reduced the multiple testing burden by only testing eQTLs for GxE signals, based on the assumption outlined above that at least some of the strongest reQTLs will also show effects on mean gene expression when stratified by genotype and be detected also as eQTLs. With a well-calibrated statistical test, I identified 12 GxE signals that span 10 candidate reQTLs at a trait-specific FDR of 10%. Replication of such findings is challenging because of the lack of human studies on equivalent tissues with equivalent environmental measurements; however, two of the three testable traits shared with the larger GTEx study showed non-random aggregate replication, although the need to restrict to heterozygous individuals limits the extent of this replication. This study highlights the utility of ASE based GxE analysis in observational studies, and emphasises the need for large RNA-seq cohorts with standardised clinical phenotypes to enable study comparison and replication.

## 6.2 reQTL mapping

### 6.2.1 ASE processing

Brooke Wolford (NIH / University of Michigan) generated the ASE data for autosomal, protein coding genes (GENCODE v19 [146]) as described previously [341], and I subsequently performed part of the ASE filtering. Briefly, we quantified strand-specific read coverage of SNPs using SAMtools mpileup v0.1.18 [214], requiring a minimum mapping quality of 255, minimum base quality of 20, and that reads mapped in a proper pair. We also removed reads that failed vendor quality checks or that were not the primary alignment. We excluded SNPs in ENCODE blacklist regions [90] and any SNP within 101 bp of an indel greater than 4 bp or overlapping an indel of any length. We followed procedures from Lappalainen et al. [201] to remove tSNPs that exhibited mapping bias based on 101 bp simulated reads, dropping

SNPs with a total simulated coverage of < 193 or > 202, and removing SNPs with simulated $\text{count}_{\text{allele}}/\text{count}_{\text{total}}$ deviating from 0.5 by $\geq 5\%$. We removed tSNPs per sample with < 30 total reads. We subsequently required that tSNPs were heterozygous in $\geq 20$ samples. From the remaining 25,913 autosomal tSNPs, I discarded 1,254 tSNPs where one or more samples exhibited near mono-allelic expression, defined as $|0.5-(\text{count}_{\text{alternate SNP}}/\text{count}_{\text{total}})| > 0.4$. Altogether, I considered 24,659 tSNPs to map candidate reQTLs.

### 6.2.2 reQTL pipeline

Similar to Knowles et al. [188], I mapped reQTLs by separately modelling gene level expression and ASE, and subsequently combined p-values using Fisher's combined test. As candidate reQTLs to test for GxE effects in the ASE and gene expression data across all clinical traits, I considered the most significant skeletal muscle eQTL (FDR 5%) per gene for 14,080 autosomal, protein coding genes with at least one significant eQTL from Scott et al. [341]. I tested for interactions of these SNP-gene pairs with 17 clinical phenotypes (Table 6.2; Section 2.3), modelling the impact of genotype effects on gene level expression and ASE levels. I inverse normalised all continuous traits. Blood pressure measurements were missing from 2 participants, whose samples were dropped when analysing blood pressure traits. Prior to fitting models, I regressed all continuous traits on age, age$^2$, and sex, except for age where I regressed only on sex.

For ASE data, I used EAGLE [188]. For sample $i$ and tSNP $s$, I mapped GxE signals by fitting the model:

$$\begin{aligned}
\min(y_{is}, n_{is}-y_{is}) &\sim \text{Binomial}[n_{is}, \sigma(\gamma_s^{(e)} e_{is} + \gamma_s^{(h)} h_{is} + \mu_s + \varepsilon_{is})] && (H_0) \\
\min(y_{is}, n_{is}-y_{is}) &\sim \text{Binomial}[n_{is}, \sigma(\underbrace{\gamma_s^{(e)} e_{is}}_{\text{env. eff.}} + \underbrace{\gamma_s^{(h)} h_{is}}_{\text{genetic eff.}} + \underbrace{\beta_s^{(eh)} e_{is} h_{is}}_{\text{GxE eff.}} + \underbrace{\mu_s}_{\text{intercept}} + \underbrace{\varepsilon_{is}}_{\text{noise}})] && (H_1)
\end{aligned} \tag{6.1}$$

Here $n_{is}$ and $y_{is}$ denote the total and alternative read count for individual $i$ at tSNP $s$, $e_{is}$ the environment, $h_{is}$ the indicator that the eQTL is heterozygous, $\mu_s$ an intercept term to take into account unexplained allelic imbalance unrelated to the environment, $\sigma(x) = 1/(1+e^{-x})$ the logistic function, $\varepsilon_{is}|v \sim N(0, v_s)$ a per individual per tSNP random effect modelling overdispersion. The variance ($v_s$) is given an inverse gamma prior *IG(a, b)*. I learned the hyperparameters $a$, $b$ for this distribution across all tSNPs after filters, estimating them to be 1.80, 0.0024 respectively. In addition, $\gamma_s^{(e)}$, $\gamma_s^{(h)}$ and $\beta_s^{(eh)}$ denote the effect sizes of the environment, eQTL heterozygosity status, and SNP $*$ environment interaction, respectively. I tested the null hypothesis $\beta_s^{(eh)} = 0$ using a likelihood ratio test. As covariates, I included the first two PCs calculated across all genotypes, consistent with Scott et al. [341]. In this analyses, I required $\geq 15$ homozygous and $\geq 15$ heterozygous samples for the eQTL tag SNP and, in the case of dichotomous variables, no group was formed with < 5 samples. With these filters, I could only test for reQTL effects in a subset of genes that differed according to clinical trait in the case of discrete variables where the total sample size was not constant due to missing data (Figure 6.3).

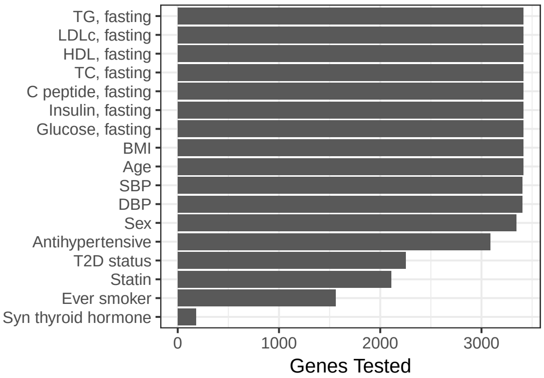


**Figure 6.3** Total number of tested genes across traits. Total number of genes in FUSION considered for each clinical trait.

I also mapped GxE interaction effects for each candidate reQTL in total gene expression data using a linear model for expression levels, testing interactions for each gene-environment pair. Let $y_j$ be a vector of inverse normalised FPKMs for gene $j$ across individuals. Consider the following linear genetic model of gene expression:

$$y_j = \underbrace{\alpha_j^{(i)}\mathbb{1}}_{\text{intercept}} + \underbrace{\alpha_j^{(Z)}Z}_{\text{covariates}} + \underbrace{\gamma_j^{(e)}e}_{\text{env. eff.}} + \underbrace{\gamma_j^{(g)}g}_{\text{genetic eff.}} + \underbrace{\beta_j g \odot e}_{\text{GxE eff.}} + \underbrace{\psi_j}_{\text{noise}}, \quad \psi_j \sim \mathcal{N}(0, \sigma_e^2 I) \tag{6.2}$$

Here $\alpha_j^{(i)}$ is the intercept, $Z$ denotes the matrix design of fixed effect confounding covariates, $e$ and $g$ the environment and genotype vector, $g \odot e$ their element-wise product,$\psi_j$ Gaussian noise, and $\alpha_j^{(Z)}$, $\gamma_j^{(e)}$, $\gamma_j^{(g)}$ and $\beta_j$ the effects of covariates, environment, genotype, and the genotype $*$ environment interaction respectively.

To capture hidden variation in gene expression data, I used PEER v1.0 [366, 367] as described previously [341] to learn latent factors. For covariates in the GxE interaction model, I included sequencing batch, the first two genotype PCs, and the first two PEER factors, as a recent report suggests two PEER factors capture the majority of technical variation, preserving biological effects [216]. I additionally included age and sex as covariates when either trait was not considered as an environmental trait. The GxE model was implemented in LIMIX v0.7.6 [222]. I combined the ASE p-values and gene expression p-values using Fisher's combined test. I controlled for FDR per environment using the Benjamini-Hochberg procedure [32].

## 6.3 reQTL results

After fitting the models, the resulting p-value distributions were well calibrated (Figure 6.4), with the vast majority of tested SNPs consistent with the null distribution. Using a 10% FDR per trait, I identified 10 candidate reQTLs across 6 traits (12 unique gene-environment trait pairs; Figure 6.5; Table 6.1; Table 6.3). Of the clinical variables considered, sex is unique in that GxE sex signals could be due to environmental (for example, circulating sex hormones) or intrinsic, within cell, effects due to differences in gene expression from the sex chromosomes. In addition, I note that I did not find strong correlation between GxE signals of ASE and gene-level models (Table 6.1; Table 6.3), which may indicate power limitations due to sample size.

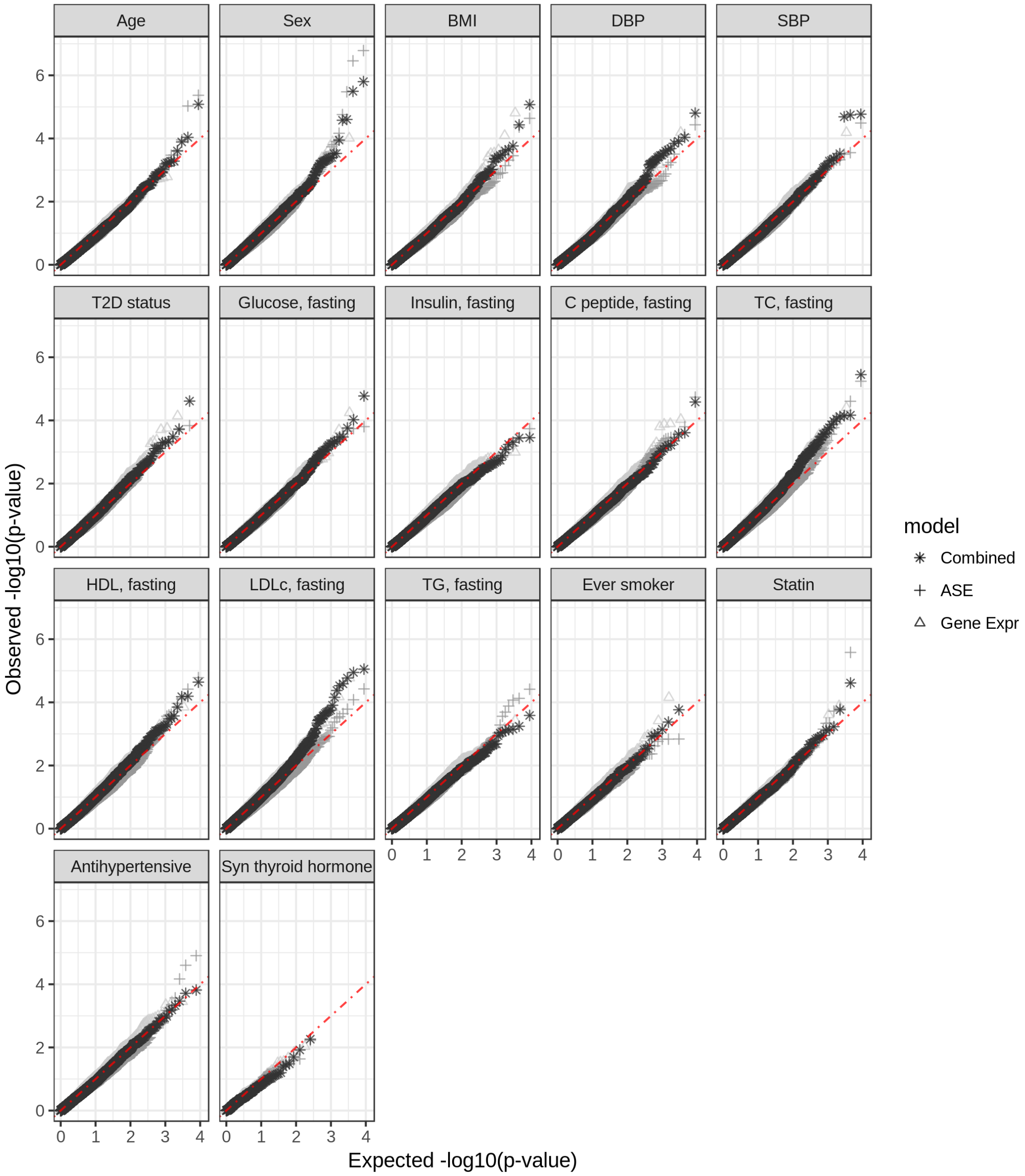


**Figure 6.4** GxE qq-plots across traits. QQ-plots of GxE signal discovery across clinical traits.

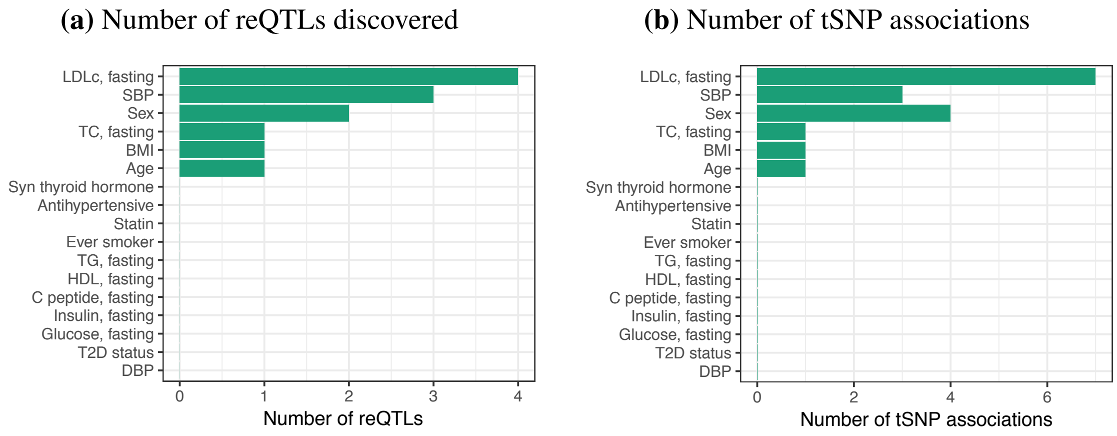


**Figure 6.5** GxE signals. (a) Number of reQTLs per clinical variable (10% FDR). (b) Number of tSNP-environment associations per clinical variable (10% FDR).

| | Clinical trait | Gene | Chr | tSNP position | reQTL alleles (ref/alt) | reQTL position | ASE p-value | Gene p-value | Combined p-value | Combined q-value |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Age | *PCNT* | 21 | 47786817 | G/T | 47823229 | 4.29e-6 | 1.25e-1 | 8.28e-6 | 0.07 |
| 2 | Sex | *BSG* | 19 | 582775 | T/C | 572878 | 1.75e-5 | 1.00e-1 | 2.50e-5 | 0.06 |
| 3 | Sex | *NRAP* | 10 | 115412793 | C/T | 115385650 | 1.65e-7 | 5.61e-1 | 1.59e-6 | 0.01 |
| 4 | BMI | *DAGLB* | 7 | 6449272 | C/T | 6476915 | 3.54e-2 | 1.55e-5 | 8.48e-6 | 0.08 |
| 5 | SBP | *ELP2* | 18 | 33750046 | T/G | 33743660 | 3.24e-5 | 3.58e-2 | 1.70e-5 | 0.06 |
| 6 | SBP | *FHOD3* | 18 | 34324091 | T/C | 33970347 | 2.82e-4 | 5.07e-3 | 2.06e-5 | 0.06 |
| 7 | SBP | *IGF2R* | 6 | 160453978 | T/C | 160379096 | 1.34e-3 | 9.18e-4 | 1.80e-5 | 0.06 |
| 8 | TC, fasting | *AGMAT* | 1 | 15909850 | T/C | 15918676 | 2.52e-3 | 8.60e-5 | 3.54e-6 | 0.03 |
| 9 | LDLc, fasting | *AGMAT* | 1 | 15909850 | T/C | 15918676 | 1.20e-3 | 4.82e-4 | 8.88e-6 | 0.05 |
| 10 | LDLc, fasting | *DEPTOR* | 8 | 121061879 | G/T | 120930135 | 4.43e-2 | 1.69e-5 | 1.13e-5 | 0.05 |
| 11 | LDLc, fasting | *FHOD3* | 18 | 34232657 | T/C | 33970347 | 6.78e-3 | 4.54e-4 | 4.21e-5 | 0.06 |
| 12 | LDLc, fasting | *TMEM261* | 9 | 7799653 | A/G | 7830189 | 8.31e-5 | 1.39e-2 | 1.69e-5 | 0.05 |

**Table 6.1** Summary of most significant tSNP for each reQTL-gene pair. Coordinates based on GRCh37/hg19. The three p-value columns record the ASE, whole gene expression level, and combined p-value respectively. The combined p-values are used for q-value calculation. Results with all reQTL-tSNP pairs are recorded in Table 6.3

### 6.3.1 GTEx replication

I conducted a replication study using data from the GTEx phs000424.v6.p1 dbGaP release. I used the preprocessed, imputed genotypes and the precomputed skeletal muscle gene expression and ASE across imputed genotypes. The GTEx samples were collected post-mortem and do not have available many of the traits assayed in the FUSION samples. Of the clinical variables measured in the FUSION dataset, four were also recorded in the GTEx dataset—age, sex, BMI, and T2D status—from which I excluded age as the distribution was significantly different between FUSION and GTEx (Table 6.2).

Notably, besides the differences in collected phenotype information and age distribution, the GTEx data differ from the FUSION data in four other relevant ways: (1) FUSION is drawn from a more genetically homogenous population (Finland); (2) FUSION is sequenced to mean depth of 91.3M reads per sample compared to 82.1M reads per sample in GTEx; (3) FUSION uses a 100 bp strand specific, paired-end read protocol for RNA-seq and GTEx uses 76 bp non-strand specific, paired-end RNA-seq; and (4) the computational analysis pipelines are different for read mapping, expression abundance quantification, and ASE calculations [138].

Within the GTEx dataset, I tested for GxE effects with the FUSION eQTL SNPs, using the ASE interaction and gene expression interaction models described in Section 6.2.2. Because my goal was replication of the FUSION genotype-environment interactions I did not require the FUSION eQTL to be significant in the GTEx dataset. For the GTEx ASE interaction model, I included the first three genotype PCs as covariates, as was used previously by the GTEx consortium [138], and for the gene expression interaction model, I included age, sex, expression batch, the first three genotype PCs, and the first two PEER factors from the GTEx data release as covariates. I tested reQTL-tSNP pairs in GTEx with sufficient double heterozygotes to pass the filters applied in the FUSION dataset (Section 6.2.2). For genes with multiple tSNPs, I selected the minimum reQTL p-value per gene for the GTEx and FUSION datasets separately.

Treating the FUSION data as a discovery dataset, I calculated the replication rate across varying p-value threshold cutoffs, as done by Knowles et al. [188]. Briefly, I selected $n$ FUSION hits at a given p-value cutoff from $N$ total shared reQTLs without replacement, stopping when $n < 10$. At each cutoff, I calculated $k$, the number of FUSION hits that replicate in GTEx (GTEx p-value < 0.01), out of the total number of nominally significant

GTEx hits, *K*. Using the mean, *K/N*, and the hypergeometric distribution, I estimated two standard deviations from the null distribution.

Despite significant differences in cohort populations, laboratory techniques, and analysis pipelines, I found a trend in the replication rate of BMI and sex that increases with the significance of the reQTL in the FUSION discovery dataset (Figure 6.6). This trend was not observed in T2D, perhaps due to different criteria for inclusion of individuals with T2D. The FUSION tissue study only included individuals with newly diagnosed T2D, not yet treated with antihyperglycemic medications (see Section 2.3 and Scott et al. [341]). By contrast, GTEx individuals may have had longstanding and heavily treated T2D [182, 137].

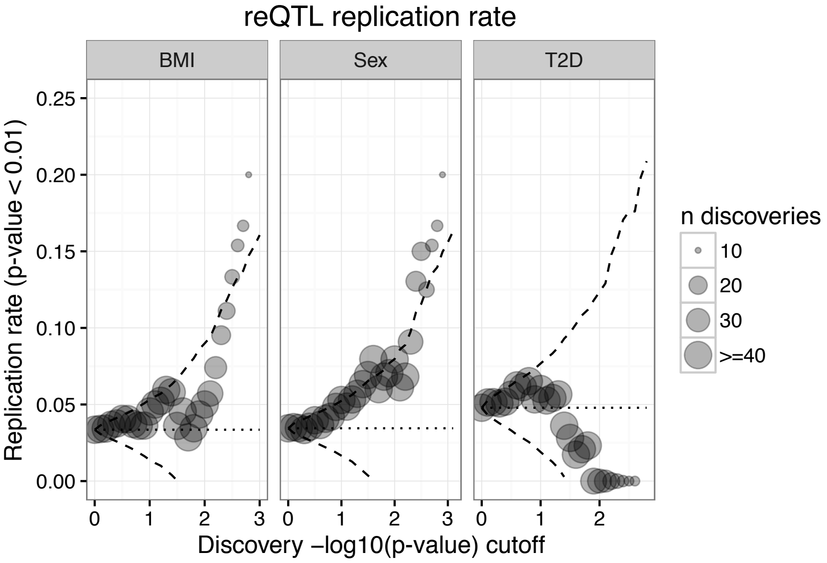
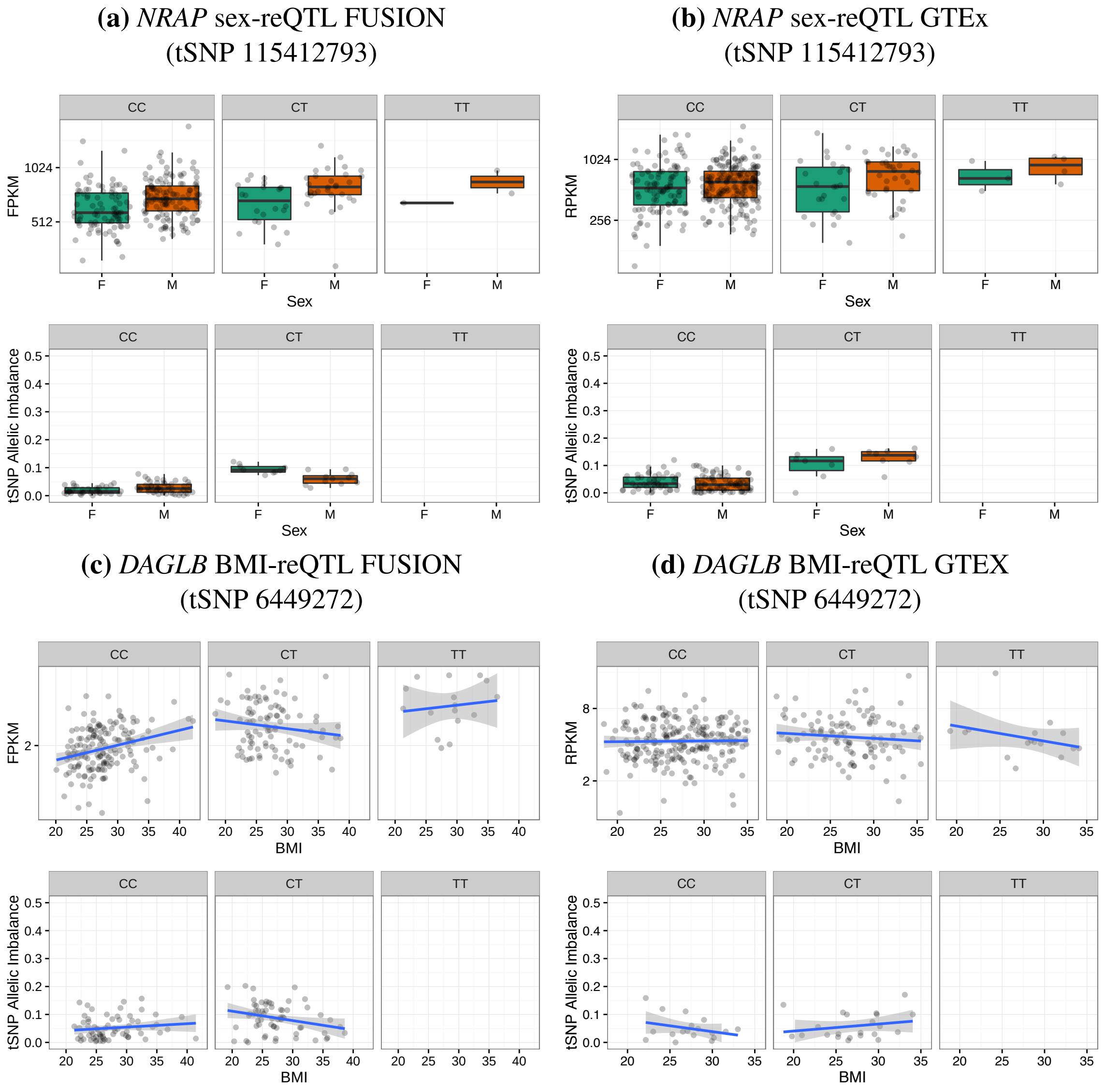


**Figure 6.6** GTEx Replication. Replication rate (y axis) as a function of FUSION reQTL p-value cutoff (x axis). Dashed line represents two standard deviations from the null distribution, calculated using the hypergeometric distribution.

Although this bulk replication is reassuring, closer inspection of the BMI and sex trends revealed that two pairs of genes are driving the observed trend in both BMI and sex, highlighting the need of large sample sizes for such GxE analyses. To this point, only two significant reQTL-tSNP pairs from FUSION met the tSNP filtering criteria in GTEx, neither of which showed similar GxE effects, potentially indicating false positives (Figure 6.7).

Finally, because I selected the minimum reQTL-tSNP pair per gene, it is possible that genes with more tSNPs will be more likely to show significant results. Therefore, I calculated the average tSNPs for the replicated and not replicated reQTL sets to explore if sampling from a larger number of transcribed SNPs was responsible for the observed trends (Figure 6.8).



**(a)** *NRAP* sex-reQTL FUSION (tSNP 115412793)

**(b)** *NRAP* sex-reQTL GTEx (tSNP 115412793)

**(c)** *DAGLB* BMI-reQTL FUSION (tSNP 6449272)

**(d)** *DAGLB* BMI-reQTL GTEX (tSNP 6449272)

**Figure 6.7** Comparison of candidate FUSION reQTLs to GTEx. (a) *NRAP* sex-reQTL in FUSION (b) *NRAP* sex-reQTL in GTEx (c) *DAGLB* BMI-reQTL in FUSION (d) *DAGLB* BMI-reQTL in GTEx.

I found finding that sampling of tSNPs was not responsible for trends at the lower p-value thresholds.

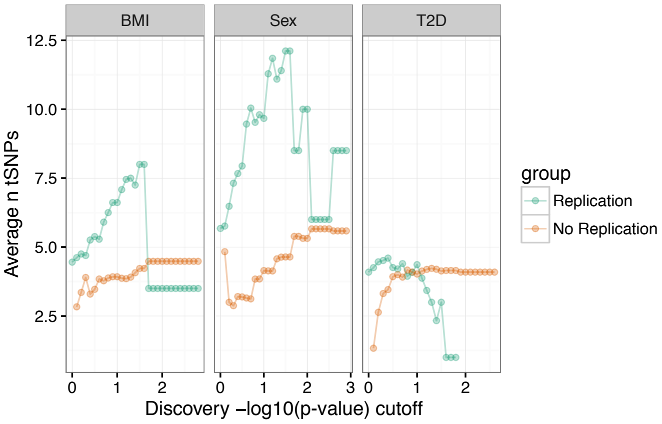
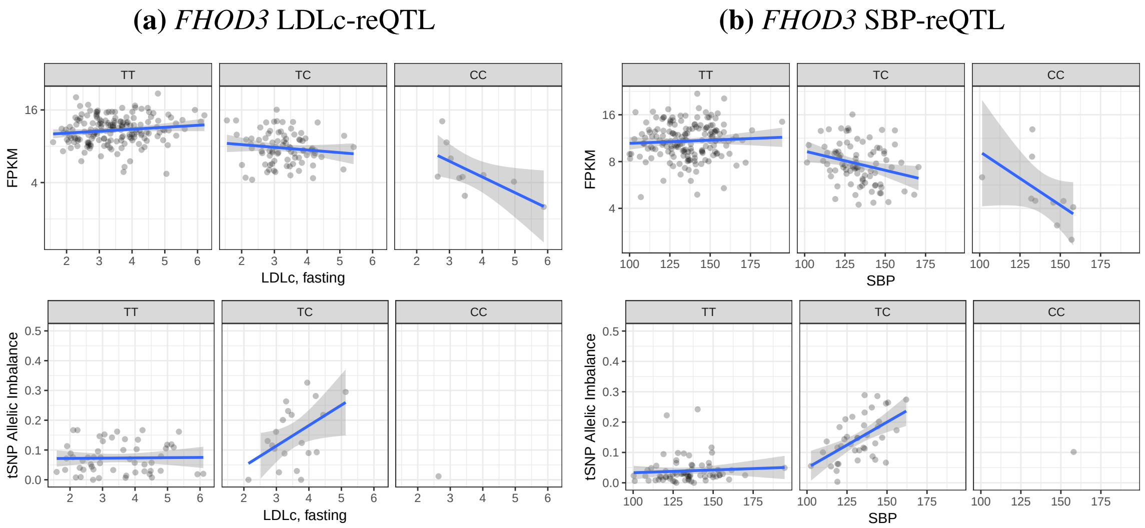


**Figure 6.8** FUSION-GTEx replication number of tSNPs. Average number of tSNPs in the genes with signals that replicated (Replication group) and signals that did not replicate (No Replication).

### 6.3.2 Specific reQTL example: *FHOD3*

Despite the small number of reported hits and replication challenges, I found some putative reQTLs with clear, consistent GxE effects in both gene expression and ASE data. The most clear, consistent example is *FHOD3*, formin homology 2 domain containing 3. *FHOD3* is essential for myofibril formation and repair, forming a doughnut shaped dimer, capable of moving along and extending actin filaments (reviewed in [290, 130, 53]). *FHOD3* is critical for heart development and function in mouse [179, 322] and fly [424], and exhibits tissue specific splicing patterns [180, 164] shown to enable myofibril targeting in striated muscle [164, 165].

I observed a GxE effect for *FHOD3* with both low-density lipoprotein cholesterol (LDLc) levels and systolic blood pressure (SBP; Figure 6.9). The LDLc association was discovered separately in the ASE of two tSNPs, spanning different exons, while the SBP association was discovered with an additional tSNP, falling in an exon separate from the LDLc tSNPs (Table 6.3). In addition, although not significant in the FUSION dataset, a GxE effect with BMI and *FHOD3* was one of the main drivers of the observed GTEx BMI replication trend (2.47x$10^{-4}$ FUSION and 8.40x$10^{-4}$ GTEx—minimum combined p-value across tSNPs). Evaluation of the raw data showed modest replication of the *FHOD3*-BMI signal between the FUSION and GTEx datasets (Figure 6.10).


**(a)** *FHOD3* LDLc-reQTL

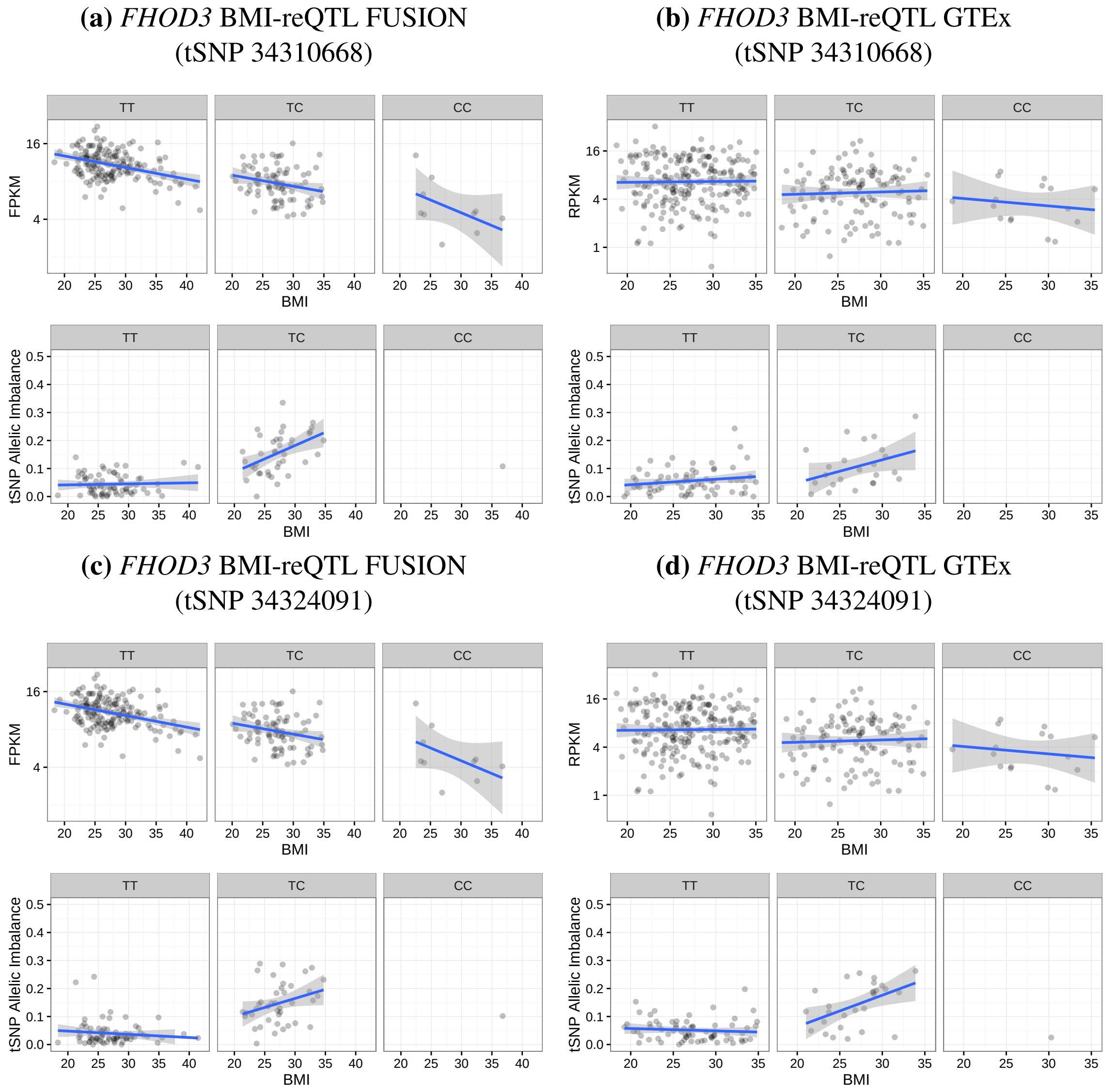
**(b)** *FHOD3* SBP-reQTL

**Figure 6.9** *FHOD3* reQTL. The data for each of the three possible reQTL genotypes are presented in separate plots (columns). The top row plots show the relationship between gene expression (y axis) and the clinical variable (x axis). The bottom row plots show the relationship between the allelic imbalance of the tSNP and the clinical variable (x axis). Note the bottom row has fewer samples because it is limited to samples heterozygous for the tSNP. (a) LDLc GxE effect with rs72895597 (18:34232657) as the tSNP (b) SBP GxE effect with rs2303510 (18:34324091) as the tSNP.

As described earlier (Section 2.6.3), I calculated mESI values across 49 diverse tissues from GTEx. I binned these scores into deciles such that genes in the 1st decile are uniformly, lowly expressed and genes in the 10th decile are highly, specifically expressed in skeletal muscle. I found *FHOD3* expression to be highly specific to skeletal muscle, falling in the 10th decile (mESI 0.56). Note this is slightly different from Taylor et al. [384] which used 16 tissues from the Illumina body map reference panel, finding the mESI decile to be the 9th decile. I believe GTEx to be a better estimate as the expression is aggregated over multiple tissue samples and the GTEx dataset contains more tissues.

In order to understand the genomic context of the region, I integrated the skeletal muscle chromatin states [395], described in Section 2.9. The reQTL tag SNP (rs17746240) and rs2037043, an additional SNP in high LD ($r^2 > 0.99$ 1000GENOMES:phase_3:FIN), overlap a skeletal muscle stretch enhancer (Figure 6.11), a regulatory element shown to be a signature of tissue-specific active chromatin [287]. In addition, these variants fall in two distinct ATAC-seq peaks unique to skeletal muscle, an indicator of open chromatin.

Both SNPs affect predicted TF binding sites (Section 4.3.3), as measured by the delta score, $-\log_{10}(\text{p-value}_{\text{alternate allele}}) - -\log_{10}(\text{p-value}_{\text{reference allele}})$. rs17746240 disrupts motifs for the GATA protein family, TBX5, and EP300. Within the FUSION skeletal muscle data, I



**(a)** *FHOD3* BMI-reQTL FUSION (tSNP 34310668)

**(b)** *FHOD3* BMI-reQTL GTEx (tSNP 34310668)

**(c)** *FHOD3* BMI-reQTL FUSION (tSNP 34324091)

**(d)** *FHOD3* BMI-reQTL GTEx (tSNP 34324091)

**Figure 6.10** Comparison of *FHOD3* BMI-reQTL in FUSION and GTEx. (a) *FHOD3* BMI-reQTL in FUSION with rs3744903 (18:34310668) as the tSNP (b) FHOD3 BMI-reQTL in GTEx with rs3744903 (18:34310668) as the tSNP (c) *FHOD3* BMI-reQTL in FUSION with rs2303510 (18:34324091) as the tSNP (d) *FHOD3* BMI-reQTL in GTEx with rs2303510 (18:34324091) as the tSNP.

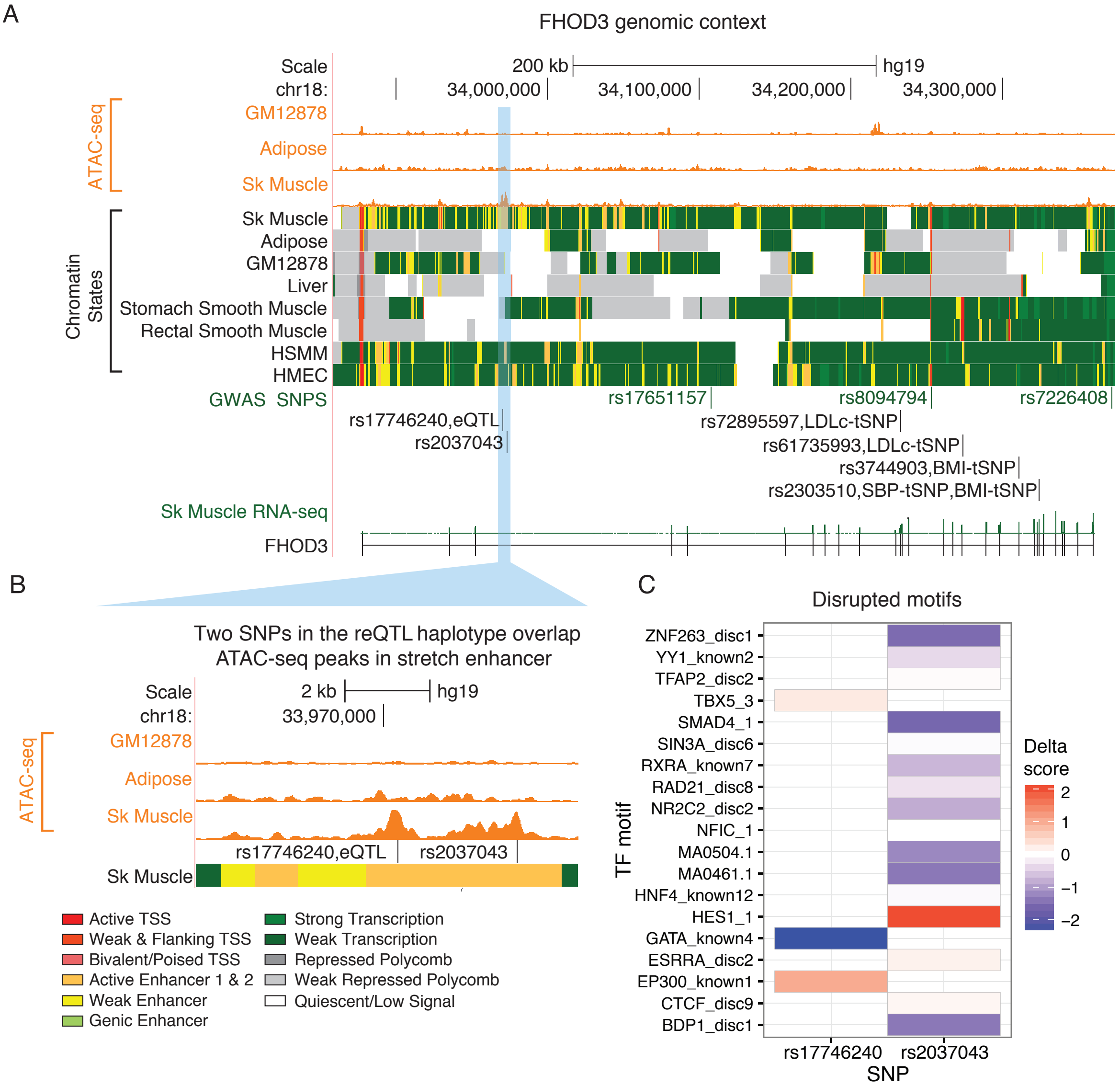


**Figure 6.11** *FHOD3* locus. (a) Top wiggle tracks show ATAC-seq signal in multiple cell types, followed by ChromHMM chromatin state tracks. Beneath are *FHOD3* GWAS loci and the SNPs from this study (reQTL and tSNP). The bottom track shows the FUSION *FHOD3* RNA-seq signal. (b) ATAC-seq signal highlights potential regulatory regions with the skeletal muscle stretch enhancer. (c) Effects of SNPs overlapping ATAC-seq peaks in the reQTL haplotype on in silico predicted TF binding.

find GATA2, GATAD1, GATAD2A, GATAD2B, and EP300 to be expressed (median FPKM > 1). The other variant, rs2037043, disrupts many motifs of which ZNF263, YY1AP1, YY1, SMAD4, SIN3A, RXRA, RAD21, NR2C2AP, NR2C2, NFIC, HES1, ESRRA, CTCF, and BDP1 are expressed in skeletal muscle (median FPKM > 1), making it difficult to identify a specific TF.

## 6.4 reQTL summary

Understanding the genetic regulators of molecular responses to environment, both at the cellular and organismal level, is essential for a complete understanding of the relationship between genotype and phenotype. Environmental influences are a critical part of human disease aetiology, but are far harder to study than intrinsic genetic factors. RNA-seq technology provides an information-dense molecular readout that includes ASE, an internally controlled experiment that minimizes technical artefacts by comparing read counts *within* samples instead of *between* samples [56, 188]. Because ASE reduces confounding effects present in gene-level data that are difficult to distinguish from environmental effects, ASE is an ideal molecular readout for probing GxE effects. This study, which is amongst the first to leverage ASE in humans to map trait specific GxE effects [135, 267, 188], demonstrates both the potential and the limitations for using ASE to unravel complex gene-environment regulatory structures. Using a well-calibrated model, I found a handful of reQTLs that show some level of bulk replication. Despite the low level of discovery in this study, which I believe is primarily limited by sample size, this success suggests that at least some eQTLs are likely to be in fact reQTLs.

This study also highlights several challenges associated with using ASE signal for mapping regulatory loci. Such analyses require sufficient sampling of double heterozygotes of the reQTL and tSNP, and therefore large sample sizes are required for a well-powered study. Another limitation of ASE is that it can only be used to identify *cis* effects. Previous studies indicate that many reQTLs operate distally, in *trans*, on highly regulated genes with more opportunities in the regulatory chain for genetic perturbation [217, 358, 356, 319]. Because this method requires ASE, I could only assay local, *cis* effects, and therefore may miss many large *trans* effects.

In the future as a community, we will need larger studies of specific human tissues with co-measured genetic, molecular, and clinical information. The possibility of mapping reQTLs underscores the importance of detailed characterisation of study participants, especially when integrating molecular and genetic data with detailed clinical information. This becomes particularly relevant for replication studies, and argues for the standardisation of a core set of phenotypes and environmental exposures between large cohorts. In addition, further development of statistical models to accommodate technology developments will be needed—for instance the integration of perfectly phased tSNP allele counts within a gene, made possible by long reads.

| Trait | FUSION | GTEx |
|---:|:--|:--|
| N | 267 | 360 |
| Sex = M (%) | 157 (58.8) | 228 (63.3) |
| Age (mean (sd)) | 60.15 (6.99) | 51.81 (12.8) |
| BMI (kg/m$^2$; mean (sd)) | 27.55 (4.17) | 27.22 (4.1) |
| Fasting High-density Lipoprotein (mmol/l; mean (sd)) | 1.45 (0.36) | |
| Fasting Low-density Lipoprotein (mmol/l; mean (sd)) | 3.41 (0.89) | |
| Fasting Triglycerides (mmol/l; mean (sd)) | 1.40 (0.81) | |
| Fasting Total Cholesterol (mmol/l; mean (sd)) | 5.49 (1.04) | |
| Systolic Blood Pressure (mmHg; mean (sd)) | 135.04 (16.05) | |
| Diastolic Blood Pressure (mmHg; mean (sd)) | 83.20 (9.45) | |
| Fasting Serum Insulin (mU/l; mean (sd)) | 8.65 (5.27) | |
| Fasting Serum C-peptide (pmol/l; mean (sd)) | 710.05 (284.40) | |
| Fasting Plasma Glucose (mmol/l; mean (sd)) | 6.26 (0.78) | |
| Ever Smoker = Y (%) | 37 (13.9) | |
| Antihypertensive = Y (%) | 83 (31.2) | |
| Statin = Y (%) | 49 (18.4) | |
| Synthetic Thyroid Hormone = Y (%) | 19 (7.1) | |
| Oral Glucose Tolerance Test Status (%) | | |
| Normal Glucose Tolerance (NGT) | 97 (36.3) | 230 (77.97) |
| Impaired Fasting Glucose (IFG) | 35 (13.1) | |
| Impaired Glucose Tolerance (IGT) | 69 (25.8) | |
| Type 2 Diabetes (T2D) | 66 (24.7) | 64 (21.69) |

**Table 6.2** GxE clinical traits. Phenotype information used as traits from the FUSION tissue biopsy study participants and GTEx skeletal muscle participants. For T2D status in GTEx, only T2D status available, non-T2D participants presumed to be NGT. In some cases, the GTEx T2D status was missing (NA), therefore T2D fraction calculated over non-missing data. Skeletal muscle data from Scott et al. [341], an earlier freeze of the dataset analysed in this thesis.

|  | Clinical trait | Gene | Chr | tSNP position | reQTL alleles (ref/alt) | reQTL position | ASE p-value | Gene p-value | Combined p-value | Combined q-value |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Age | *PCNT* | 21 | 47786817 | G/T | 47823229 | 4.29e-6 | 1.25e-1 | 8.28e-6 | 0.07 |
| 2 | Sex | *BSG* | 19 | 582775 | T/C | 572878 | 1.75e-5 | 1.00e-1 | 2.50e-5 | 0.06 |
| 3 | Sex | *NRAP* | 10 | 115409840 | C/T | 115385650 | 3.33e-6 | 5.61e-1 | 2.65e-5 | 0.06 |
| 4 | Sex | *NRAP* | 10 | 115411598 | C/T | 115385650 | 3.48e-7 | 5.61e-1 | 3.21e-6 | 0.01 |
| 5 | Sex | *NRAP* | 10 | 115412793 | C/T | 115385650 | 1.65e-7 | 5.61e-1 | 1.59e-6 | 0.01 |
| 6 | BMI | *DAGLB* | 7 | 6449272 | C/T | 6476915 | 3.54e-2 | 1.55e-5 | 8.48e-6 | 0.08 |
| 7 | SBP | *ELP2* | 18 | 33750046 | T/G | 33743660 | 3.24e-5 | 3.58e-2 | 1.70e-5 | 0.06 |
| 8 | SBP | *FHOD3* | 18 | 34324091 | T/C | 33970347 | 2.82e-4 | 5.07e-3 | 2.06e-5 | 0.06 |
| 9 | SBP | *IGF2R* | 6 | 160453978 | T/C | 160379096 | 1.34e-3 | 9.18e-4 | 1.80e-5 | 0.06 |
| 10 | TC, fasting | *AGMAT* | 1 | 15909850 | T/C | 15918676 | 2.52e-3 | 8.60e-5 | 3.54e-6 | 0.03 |
| 11 | LDL, fasting | *AGMAT* | 1 | 15909850 | T/C | 15918676 | 1.20e-3 | 4.82e-4 | 8.88e-6 | 0.05 |
| 12 | LDL, fasting | *DEPTOR* | 8 | 121061879 | G/T | 120930135 | 4.43e-2 | 1.69e-5 | 1.13e-5 | 0.05 |
| 13 | LDL, fasting | *DEPTOR* | 8 | 121062077 | G/T | 120930135 | 1.25e-1 | 1.69e-5 | 2.97e-5 | 0.05 |
| 14 | LDL, fasting | *DEPTOR* | 8 | 121062555 | G/T | 120930135 | 1.08e-1 | 1.69e-5 | 2.60e-5 | 0.05 |
| 15 | LDL, fasting | *FHOD3* | 18 | 34232657 | T/C | 33970347 | 6.78e-3 | 4.54e-4 | 4.21e-5 | 0.06 |
| 16 | LDL, fasting | *FHOD3* | 18 | 34273279 | T/C | 33970347 | 1.16e-2 | 4.54e-4 | 6.94e-5 | 0.09 |
| 17 | LDL, fasting | *TMEM261* | 9 | 7799653 | A/G | 7830189 | 8.31e-5 | 1.39e-2 | 1.69e-5 | 0.05 |

**Table 6.3** All reQTL-tSNP pairs FDR 10%.

# Chapter 7

# Conclusion

## 7.1 Concluding summary

In this thesis, I analysed the relationship between gene expression, DNAme, and genetic variation in skeletal muscle. The key tasks in my research involved data curation and QC, latent factor analysis, mapping molecular trait associations, and characterising these associations.

Chapter 2 details the extensive QC measures taken in this study which I led and conducted with help from collaborators across several laboratories. In the molecular trait data, we identified and removed outliers across many technical and biological measurements. In the RNA-seq QC measures, one interesting finding was that samples with higher levels of estimated tissue heterogeneity were outliers in transcriptional diversity measures (Figure 2.3). This result suggests stringent transcriptional diversity filters are an important step in analysing RNA-seq data from tissue samples that are potentially heterogeneous, especially when transcriptional tissue deconvolution analysis is not feasible (for instance, if there is no appropriate reference transcriptome dataset). In addition, for the DNAme QC measures, I demonstrated the importance of many layered QC steps, as samples may appear to be of high quality in one measurement but then fail in subsequent measurements. In particular, I found analysing the global DNAme profile of samples by comparing median raw signal intensities (Figure 2.6), the overall DNAme distribution (Figure 2.15), and the DNAme profiles across tissues through PCA (Figure 2.18) to be important in flagging low quality samples. Collectively, these extensive steps required for DNAme QC highlight substantial challenges that would need to be solved if such data were ever to be useful routinely in the clinic, for instance as a biomarker.

In Chapter 3, I analysed the relationship between gene expression and DNAme by mapping eQTMs and exQTMs. I found evidence of latent correlation between these molecular traits (Figure 3.2) and showed that, despite all samples having > 90% estimated skeletal muscle fraction, tissue heterogeneity likely constitutes a component of this correlation (Figures 3.1, 3.5). Such patterns could, for instance, be due to cell/tissue types poorly represented in the reference panels used for the tissue deconvolution analysis. This observation highlights the importance of the Human Cell Atlas project [308], which aims to build reference maps at the level of single cells as opposed to bulk tissue. The observed latent correlation is also particularly important as studies move toward building a comprehensive molecular profile of samples across multiple molecular traits. To date, many studies have focused on genetic associations with molecular traits, in which case the genotypes would not be affected by tissue heterogeneity or environment. However, in instances where the correlation between molecular profiles is analysed, both of which may be affected by hidden variables, a thorough latent factor analysis will be particularly important. In addition, I reproduced the well known context specific effects of eQTMs, where DNAme in areas of active chromatin tends to repress gene expression (Figure 3.8).

Building on a previous eQTL study using an earlier freeze of this dataset [341], I mapped genetic effects on gene expression and DNAme in Chapter 4. I showed that compared to eQTLs and exQTLs, mQTLs tend to occur slightly closer to the target DNAme site (Figure 4.3b). In addition, by comparing QTL overlaps, I demonstrated that nearly all loci that affect gene expression at some level also affect DNAme, although the effect may not be the strongest QTL for the target DNAme site (Figure 4.4). With a collaborator, I also characterised enrichment trends of QTLs in chromatin states, reproducing previous findings for eQTLs [341, 395] and mQTLs [59, 273]. However, this analysis also suggests that mQTL enrichment trends may, to some extent, be influenced by the specific genomic context of the probe targets (Figure 4.10a). Finally, a collaborator and I also used QTLs overlapping predicted TF binding sites to identify TFs that likely serve as activators or repressors in skeletal muscle (Table 4.1).

In Chapter 5, I identified effects of GWAS loci for T2D and T2D-related traits on gene expression and DNAme in skeletal muscle, finding numerous significant associations (Table 5.1). I prioritised and characterised the strongest associations as well as the associations with molecular traits that exhibit skeletal muscle specific trends. Using gene expression, I reproduced findings from Scott et al. [341], which analysed gene expression data from an earlier version of this dataset. In addition, I found that including DNAme enhanced these previous findings by guiding the identification of candidate regulatory effects at a GWAS locus. For instance, using DNAme, I inferred instances where the regulatory effects underlying a disease-associated locus likely involve TF binding at a canonical promoter, TF binding at an alternative promoter or splice site, and even instances of distal regulation. In some cases, I found extremely strong DNAme effects with a nearby common variant in high LD with the GWAS tag SNP. Such instances may indicate the identification of a functional variant underlying disease risk. As a proof of principle, I found the DNAme patterns around the highly muscle specific *ANK1* gene suggest that rs508419 may be a causal variant driving the T2D GWAS signal (out of 15 variants in LD, $r^2 \geq 0.8$). In Scott et al. [341], this variant was identified independent of DNAme and shown to affect TR4 binding in a skeletal muscle promoter, leading to changes in *ANK1* expression. This example suggests that DNAme can indeed be used as a proxy for the identification of important regulatory events and to distinguish candidate underlying, causal variants from other variants in high LD.

In addition, the inclusion of DNAme enabled the identification of instances where a variant has a strong effect on gene expression and DNAme, which are themselves associated (i.e., are eQTM). Using *ANK1* as a representative example, I showed that the observed eQTM associations appear to be driven primarily by independent, strong genetic effects on both gene expression and DNAme (Figure 5.17). To test such trends genome wide, I performed a mediation analysis across all eQTMs with a QTL and found that for the majority of cases, the association between expression and DNAme is driven by independent genetic effects (Figure 5.19), consistent with the results of two recent studies [141, 273]. These results represent a compelling repudiation of other models that assume DNAme generally drives changes in gene expression level, or vice versa, and highlight the complexity of relationship between expression and DNAme.

Finally, in Chapter 6, I mapped environmental response QTLs, utilising the many phenotypes of the FUSION participants. Although underpowered, these results showed some level of replication (Figure 6.6), including a compelling example with the highly muscle specific *FHOD3* gene (Figures 6.10, 6.11).

## 7.2 Future directions

Collectively, these results summarise the relationships between gene expression, DNAme, and genetic variation in skeletal muscle. They also suggest avenues for further exploration. As I have described, I found several instances where DNAme provided insight into the regulatory architecture around a GWAS locus and pinpointed candidate variants for functional follow up. Such observations suggest that DNAme may be a useful proxy to assay molecular events that shape the regulatory landscape of a region (e.g., TF binding), and that measuring DNAme signals genome wide through sequencing (WGBS) would be a worthwhile investment since arrays are limited by a set of predefined probes. These predefined probes restrict the identification of candidate functional variants (since one does not know if there is a stronger association at a DNAme site not assayed by the array) as well as distal regulatory effects (since the EPIC array poorly captures important intergenic regulatory elements like enhancers). In addition, WGBS would allow for comprehensive analysis of non-CpG DNAme and validation of predicted *cis* regulatory effects through allele specific methylation.

In the broader context of the genetics of T2D and T2D-related traits, the fact that I found skeletal muscle QTLs for many GWAS loci with clear T2D-linked effects in alternative tissues (e.g., islets), highlights the paramount importance of multi-omic datasets across many tissues, integrated with genetic information. When faced with multiple effects of the same GWAS locus across different tissues (possibly differing in magnitude of effect or in a tissue specific manner), identifying and prioritising which effects may be causal for disease poses a significant challenge. Literature review, comparative genomics, multi-phenotype genetics (e.g., phenome-wide association studies), multi-tissue functional genomics (e.g. open chromatin profiles, chromatin state profiles, chromatin interactions), and multi-tissue molecular trait genetics (e.g., eQTL, mQTL, caQTL studies) can help refine candidate causal variants, identify tissue(s) of action, as well as generate hypotheses that describe underlying molecular and physiological mechanisms.

For this reason, ongoing efforts to collect and build molecular trait QTL databases of key T2D tissues, such as islet, muscle, liver, adipose, and brain, are extremely important. Focusing on gene expression, GTEx will provide an excellent resource for many of these tissues, and in cases of tissue overlap, other datasets can be used in conjunction with GTEx to perform replication analyses, as demonstrated by Scott et al. [341]. However, for certain difficult to obtain tissues, like pancreatic islet beta cells, resources like GTEx will need to be supplemented with external datasets. In addition, expanding the current gene expression focused datasets, with additional molecular traits, like DNAme, may prove to be very useful for identifying regulatory effects that underlie a GWAS locus, as described above.

Given finite sample quantities, generating QTL maps across many molecular traits is generally infeasible. However, as recently demonstrated by the Human Induced Pluripotent Stem Cell Initiative (HipSci) [185], iPS technologies can be used to generate pluripotent stem cell reference panels, which provide an unlimited source of cellular material for molecular assays across cell types (after differentiation). Although expensive, investments in such resources (as FUSION has already made with a pilot project of 50 samples from this study) may yield high dividends and enable studies to understand the effects of GWAS loci, as well as genetic variation in general, on multiple molecular traits across various tissues in diverse environmental contexts (e.g., glucose stimulation for islets). Moreover, such approaches could be made even more powerful by using new single-cell molecular trait assays, thereby accounting for heterogeneity of cell states in cultures.

Hypotheses generated from such high throughput analyses cannot fully fill the knowledge gap from "genotype to phenotype" and must be followed up by low throughput functional studies. Such studies can take on many forms and, out of necessity, will vary case-by-case. For instance, candidate allele specific protein-DNA interactions can be validated *in vitro* through electrophoretic mobility shift assays [341]. More extensive studies of candidate phenotypic effects can be performed *in vivo* using transgenic model organisms, such as zebrafish and mice [344, 91]. Genome editing technologies (e.g., CRISPR/Cas9) can expedite model organism engineering as well as enable the testing of specific mutations in model cell lines, differentiated iPS cells, or even iPS cells during differentiation [355, 347, 154]. This is a particularly powerful approach as detailed multi-omic experiments and measurements of cell physiology can be performed in tandem with controlled environmental perturbations. Given the extensive labour and time required to perform such studies, it is important to leverage high throughput methods (as described earlier) to guide specific, targeted experiments. Nonetheless, such work is critical in order to fully validate hypotheses, establish a complete biological picture, and ultimately drive translational research.

T2D is a major challenge for global healthcare, being the 6th leading cause of death worldwide, with hundreds of billions of dollars in associated health, social, and economic costs [161, 426]. Generating a complete catalogue of the genetic risk factors contributing to T2D constitutes a crucial component of efforts to develop efficacious T2D treatments, as genetic associations offer clues to biological mechanisms underlying disease (since there is no reverse causation, common to epidemiological studies). This year, 2017, marks the

10 year anniversary of the first T2D GWAS publications [340, 82, 419, 438]. The progress made in unravelling the genetic architecture of T2D over the past decade is extraordinary. With the ever decreasing costs of sequencing [131], identifying the full repertoire of T2D genetic risk factors stands on the horizon. Yet, functional interpretation of these many genetic effects still poses a significant challenge. The analysis presented in this thesis contributes to ongoing efforts to understand the impact of disease-associated variants on molecular traits. Undoubtedly, through the collective and collaborative efforts of the global scientific community, these challenges will be met—connecting genotype to phenotype, suggesting targetable pathways for therapeutic development, and improving lives of individuals and families around the world.

# References

[1] M. A. Ackermann, A. P. Ziman, J. Strong, Y. Zhang, A. K. Hartford, C. W. Ward, W. R. Randall, A. Kontrogianni-Konstantopoulos, and R. J. Bloch. Integrity of the network sarcoplasmic reticulum in skeletal muscle requires small ankyrin 1. *J Cell Sci*, 124(21):3619–3630, November 2011.

[2] D. Adams, L. Altucci, S. E. Antonarakis, J. Ballesteros, S. Beck, A. Bird, C. Bock, B. Boehm, E. Campo, A. Caricasole, F. Dahl, E. T. Dermitzakis, T. Enver, M. Esteller, X. Estivill, A. Ferguson-Smith, J. Fitzgibbon, P. Flicek, C. Giehl, T. Graf, F. Grosveld, R. Guigo, I. Gut, K. Helin, J. Jarvius, R. Küppers, H. Lehrach, T. Lengauer, A. Lernmark, D. Leslie, M. Loeffler, E. Macintyre, A. Mai, J. H. A. Martens, S. Minucci, W. H. Ouwehand, P. G. Pelicci, H. Pendeville, B. Porse, V. Rakyan, W. Reik, M. Schrappe, D. Schübeler, M. Seifert, R. Siebert, D. Simmons, N. Soranzo, S. Spicuglia, M. Stratton, H. G. Stunnenberg, A. Tanay, D. Torrents, A. Valencia, E. Vellenga, M. Vingron, J. Walter, and S. Willcocks. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol*, 30(3):224–226, March 2012.

[3] K. Alasoo, J. Rodrigues, S. Mukhopadhyay, A. J. Knights, A. L. Mann, K. Kundu, HipSci Consortium, C. Hale, G. Dougan, and D. J. Gaffney. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat Genet*, January 2018.

[4] F. W. Albert and L. Kruglyak. The role of regulatory variation in complex traits and disease. *Nat Rev Genet*, 16(4):197–212, April 2015.

[5] J. Albuisson, S. E. Murthy, M. Bandell, B. Coste, H. Louis-Dit-Picard, J. Mathur, M. Fénéant-Thibault, G. Tertian, J. de Jaureguiberry, P. Syfuss, S. Cahalan, L. Garçon, F. Toutain, P. Simon Rohrlich, J. Delaunay, V. Picard, X. Jeunemaitre, and A. Patapoutian. Dehydrated hereditary stomatocytosis linked to gain-of-function mutations in mechanically activated PIEZO1 ion channels. *Nat Commun*, 4:1884, 2013.

[6] F. Allum, X. Shao, F. Guénard, M. Simon, S. Busche, M. Caron, J. Lambourne, J. Lessard, K. Tandre, A. K. Hedman, T. Kwan, B. Ge, Multiple Tissue Human Expression Resource Consortium, L. Rönnblom, M. I. McCarthy, P. Deloukas, T. Richmond, D. Burgess, T. D. Spector, A. Tchernof, S. Marceau, M. Lathrop, M. Vohl, T. Pastinen, and E. Grundberg. Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nat Commun*, 6:7211, May 2015.

[7] P. Almgren, M. Lehtovirta, B. Isomaa, L. Sarelin, M. R. Taskinen, V. Lyssenko, T. Tuomi, L. Groop, and Botnia Study Group. Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia study. *Diabetologia*, 54(11): 2811–2819, November 2011.

[8] J. C. Alwine, D. J. Kemp, and G. R. Stark. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A*, 74(12):5350–5354, December 1977.

[9] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Res*, 22(10):2008–2017, October 2012.

[10] S. Anders, P. T. Pyl, and W. Huber. HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, January 2015.

[11] I. Andolfo, S. L. Alper, L. De Franceschi, C. Auriemma, R. Russo, L. De Falco, F. Vallefuoco, M. R. Esposito, D. H. Vandorpe, B. E. Shmukler, R. Narayan, D. Montanaro, M. D'Armiento, A. Vetro, I. Limongelli, O. Zuffardi, B. E. Glader, S. L. Schrier, C. Brugnara, G. W. Stewart, J. Delaunay, and A. Iolascon. Multiple clinical forms of dehydrated hereditary stomatocytosis arise from mutations in PIEZO1. *Blood*, 121 (19):3925–35, S1, May 2013.

[12] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics*, 30(10): 1363–1369, May 2014.

[13] E. A. Ashley. Towards precision medicine. *Nat Rev Genet*, 17(9):507–522, August 2016.

[14] Q. Ayub, L. Moutsianas, Y. Chen, K. Panoutsopoulou, V. Colonna, L. Pagani, I. Prokopenko, G. R. S. Ritchie, C. Tyler-Smith, M. I. McCarthy, E. Zeggini, and Y. Xue. Revisiting the thrifty gene hypothesis via 65 loci associated with susceptibility to type 2 diabetes. *Am J Hum Genet*, 94(2):176–185, February 2014.

[15] P. Bagnato, V. Barone, E. Giacomello, D. Rossi, and V. Sorrentino. Binding of an ankyrin-1 isoform to obscurin suggests a molecular link between the sarcoplasmic reticulum and myofibrils in striated muscles. *J Cell Biol*, 160(2):245–253, January 2003.

[16] S. N. Bagriantsev, E. O. Gracheva, and P. G. Gallagher. Piezo proteins: regulators of mechanosensation and other cellular processes. *J Biol Chem*, 289(46):31673–31681, November 2014.

[17] M. P. Balakrishnan, L. Cilenti, Z. Mashak, P. Popat, E. S. Alnemri, and A. S. Zervos. THAP5 is a human cardiac-specific inhibitor of cell cycle that is cleaved by the proapoptotic Omi/HtrA2 protease during cell death. *Am J Physiol Heart Circ Physiol*, 297(2):H643–53, August 2009.

[18] M. P. Balakrishnan, L. Cilenti, C. Ambivero, Y. Goto, M. Takata, J. Turkson, X. S. Li, and A. S. Zervos. THAP5 is a DNA-binding transcriptional repressor that is regulated in melanoma cells during DNA damage-induced cell death. *Biochem Biophys Res Commun*, 404(1):195–200, January 2011.

[19] N. E. Banovich, X. Lan, G. McVicker, B. van de Geijn, J. F. Degner, J. D. Blischak, J. Roux, J. K. Pritchard, and Y. Gilad. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*, 10(9):e1004663, September 2014.

[20] D. P. Barlow and M. S. Bartolomei. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol*, 6(2), February 2014.

[21] L. B. Barreiro, L. Tailleux, A. A. Pai, B. Gicquel, J. C. Marioni, and Y. Gilad. Deciphering the genetic architecture of variation in the immune response to mycobacterium tuberculosis infection. *Proc Natl Acad Sci U S A*, 109(4):1204–1209, January 2012.

[22] R. Barrès, M. E. Osler, J. Yan, A. Rune, T. Fritz, K. Caidahl, A. Krook, and J. R. Zierath. Non-CpG methylation of the PGC-1alpha promoter through DNMT3B controls mitochondrial density. *Cell Metab*, 10(3):189–198, September 2009.

[23] R. Barrès, J. Yan, B. Egan, J. T. Treebak, M. Rasmussen, T. Fritz, K. Caidahl, A. Krook, D. J. O'Gorman, and J. R. Zierath. Acute exercise remodels promoter methylation in human skeletal muscle. *Cell Metab*, 15(3):405–411, March 2012.

[24] R. Barrès, H. Kirchner, M. Rasmussen, J. Yan, F. R. Kantor, A. Krook, E. Näslund, and J. R. Zierath. Weight loss after gastric bypass surgery in human obesity remodels promoter methylation. *Cell Rep*, 3(4):1020–1027, April 2013.

[25] B. Bateson. *William Bateson, Naturalist: His Essays and Addresses Together with a Short Account of His Life*. Cambridge University Press, Cambridge, 1928.

[26] W. Bateson. *Mendel's principles of heredity: A defence*. Cambridge University Press, Cambridge, 1902.

[27] D. C. Bauer, F. A. Buske, and T. L. Bailey. Dual-functioning transcription factors in the developmental gene network of Drosophila melanogaster. *BMC Bioinformatics*, 11:366, July 2010.

[28] G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in neurospora. *Proc Natl Acad Sci U S A*, 27(11):499–506, November 1941.

[29] K. A. Beck. Spectrins and the golgi. *Biochim Biophys Acta*, 1744(3):374–382, July 2005.

[30] M. Becker-André and K. Hahlbrock. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). *Nucleic Acids Res*, 17(22):9437–9446, November 1989.

[31] J. P. Belman, E. N. Habtemichael, and J. S. Bogan. A proteolytic pathway that controls glucose uptake in fat and muscle. *Rev Endocr Metab Disord*, 15(1):55–66, March 2014.

[32] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[33] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J. Fan, and R. Shen. High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295, October 2011.

[34] A. P. Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*, 8(7):1499–1504, April 1980.

[35] C. Bock. Analysing and interpreting DNA methylation data. *Nat Rev Genet*, 13(10): 705–719, October 2012.

[36] M. T. Bolisetty, G. Rajadinakaran, and B. R. Graveley. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biol*, 16(1):204, September 2015.

[37] M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, and S. Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 336(6083):934–937, May 2012.

[38] M. J. Booth, T. W. B. Ost, D. Beraldi, N. M. Bell, M. R. Branco, W. Reik, and S. Balasubramanian. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc*, 8(10):1841–1851, October 2013.

[39] M. A. Borzok, D. H. Catino, J. D. Nicholson, A. Kontrogianni-Konstantopoulos, and R. J. Bloch. Mapping the binding site on small ankyrin 1 for obscurin. *J Biol Chem*, 282(44):32384–32396, November 2007.

[40] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, 32(3):314–331, May 1980.

[41] K. Bouzakri, A. Zachrisson, L. Al-Khalili, B. B. Zhang, H. A. Koistinen, A. Krook, and J. R. Zierath. siRNA-based gene silencing reveals specialized roles of IRS-1/Akt2 and IRS-2/Akt1 in glucose and lipid metabolism in human skeletal muscle. *Cell Metab*, 4(1):89–96, July 2006.

[42] T. Boveri. *Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns*. G. Fischer, Jena, 1904.

[43] E. A. Boyle, Y. I. Li, and J. K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, June 2017.

[44] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, April 2002.

[45] S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81(5):1084–1097, November 2007.

[46] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*, 10(12):1213–1218, December 2013.

[47] A. Buil, A. A. Brown, T. Lappalainen, A. Viñuela, M. N. Davies, H.-F. Zheng, J. B. Richards, D. Glass, K. S. Small, R. Durbin, T. D. Spector, and E. T. Dermitzakis. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet*, 47(1):88–91, January 2015.

[48] M. Bulmer. Galton's law of ancestral heredity. *Heredity*, 81(5):579–585, November 1998.

[49] S. Burgess and S. G. Thompson. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Stat Med*, 30(11):1312–1323, May 2011.

[50] S. Burgess, S. G. Thompson, and CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol*, 40(3): 755–764, June 2011.

[51] W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, December 2012.

[52] A. Byrne, A. E. Beaudin, H. E. Olsen, M. Jain, C. Cole, T. Palmer, R. M. DuBois, E. C. Forsberg, M. Akeson, and C. Vollmers. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat Commun*, 8:16027, July 2017.

[53] K. G. Campellone and M. D. Welch. A nucleator arms race: cellular control of actin assembly. *Nat Rev Mol Cell Biol*, 11(4):237–251, April 2010.

[54] E. Cannavò, N. Koelling, D. Harnett, D. Garfield, F. P. Casale, L. Ciglar, H. E. Gustafson, R. R. Viales, R. Marco-Ferreres, J. F. Degner, B. Zhao, O. Stegle, E. Birney, and E. E. M. Furlong. Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature*, 541(7637):402–406, January 2017.

[55] M. Caruso, D. Ma, Z. Msallaty, M. Lewis, B. Seyoum, W. Al-janabi, M. Diamond, A. B. Abou-Samra, K. Hojlund, R. Tagett, S. Draghici, X. Zhang, J. F. Horowitz, and Z. Yi. Increased interaction with insulin receptor substrate 1, a novel abnormality in insulin resistance and type 2 diabetes. *Diabetes*, 63(6):1933–1947, June 2014.

[56] S. E. Castel, A. Levy-Moonshine, P. Mohammadi, E. Banks, and T. Lappalainen. Tools and best practices for data processing in allelic expression analysis. *Genome Biol*, 16(1):195, September 2015.

[57] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):7, February 2015.

[58] J. Charlet, C. E. Duymich, F. D. Lay, K. Mundbjerg, K. Dalsgaard Sorensen, G. Liang, and P. A. Jones. Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. *Mol Cell*, 62(3):422–431, May 2016.

[59] L. Chen, B. Ge, F. P. Casale, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yan, K. Kundu, S. Ecker, A. Datta, D. Richardson, F. Burden, D. Mead, A. L. Mann, J. M. Fernandez, S. Rowlston, S. P. Wilder, S. Farrow, X. Shao, J. J. Lambourne, A. Redensek, C. A. Albers, V. Amstislavskiy, S. Ashford, K. Berentsen, L. Bomba, G. Bourque, D. Bujold, S. Busche, M. Caron, S. Chen, W. Cheung, O. Delaneau, E. T. Dermitzakis, H. Elding, I. Colgiu, F. O. Bagger, P. Flicek, E. Habibi, V. Iotchkova, E. Janssen-Megens, B. Kim, H. Lehrach, E. Lowy, A. Mandoli, F. Matarese, M. T. Maurano, J. A. Morris, V. Pancaldi, F. Pourfarzad, K. Rehnstrom, A. Rendon, T. Risch, N. Sharifi, M. Simon, M. Sultan, A. Valencia, K. Walter, S. Wang, M. Frontini, S. E. Antonarakis, L. Clarke, M. Yaspo, S. Beck, R. Guigo, D. Rico, J. H. A. Martens, W. H. Ouwehand, T. W. Kuijpers, D. S. Paul, H. G. Stunnenberg, O. Stegle, K. Downes, T. Pastinen, and N. Soranzo. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, 167(5):1398–1414.e24, November 2016.

[60] Y. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 8(2):203–209, February 2013.

[61] V. G. Cheung, L. K. Conlin, T. M. Weber, M. Arcaro, K. Jen, M. Morley, and R. S. Spielman. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*, 33(3):422–425, March 2003.

[62] V. G. Cheung, R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, and J. T. Burdick. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437(7063):1365–1369, October 2005.

[63] Y. S. Cho, C. Chen, C. Hu, J. Long, R. T. H. Ong, X. Sim, F. Takeuchi, Y. Wu, M. J. Go, T. Yamauchi, Y. Chang, S. H. Kwak, R. C. W. Ma, K. Yamamoto, L. S. Adair, T. Aung, Q. Cai, L. Chang, Y. Chen, Y. Gao, F. B. Hu, H. Kim, S. Kim, Y. J. Kim, J. J. Lee, N. R. Lee, Y. Li, J. J. Liu, W. Lu, J. Nakamura, E. Nakashima, D. P. Ng, W. T. Tay, F. Tsai, T. Y. Wong, M. Yokota, W. Zheng, R. Zhang, C. Wang, W. Y. So, K. Ohnaka, H. Ikegami, K. Hara, Y. M. Cho, N. H. Cho, T. Chang, Y. Bao, A. K. Hedman, A. P. Morris, M. I. McCarthy, DIAGRAM Consortium, MuTHER Consortium, R. Takayanagi, K. S. Park, W. Jia, L. Chuang, J. C. N. Chan, S. Maeda, T. Kadowaki, J. Lee, J. Wu, Y. Y. Teo, E. S. Tai, X. O. Shu, K. L. Mohlke, N. Kato, B. Han, and M. Seielstad. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet*, 44(1):67–72, December 2011.

[64] J. X. Chong, K. J. Buckingham, S. N. Jhangiani, C. Boehm, N. Sobreira, J. D. Smith, T. M. Harrell, M. J. McMillin, W. Wiszniewski, T. Gambin, Z. H. Coban Akdemir, K. Doheny, A. F. Scott, D. Avramopoulos, A. Chakravarti, J. Hoover-Fong, D. Mathews, P. D. Witmer, H. Ling, K. Hetrick, L. Watkins, K. E. Patterson, F. Reinier, E. Blue, D. Muzny, M. Kircher, K. Bilguvar, F. López-Giráldez, V. R. Sutton, H. K. Tabor, S. M.

Leal, M. Gunel, S. Mane, R. A. Gibbs, E. Boerwinkle, A. Hamosh, J. Shendure, J. R. Lupski, R. P. Lifton, D. Valle, D. A. Nickerson, Centers for Mendelian Genomics, and M. J. Bamshad. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet*, 97(2):199–215, August 2015.

[65] S. J. Clark, S. A. Smallwood, H. J. Lee, F. Krueger, W. Reik, and G. Kelsey. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat Protoc*, 12(3):534–547, March 2017.

[66] M. B. Colović, D. Z. Krstić, T. D. Lazarević-Pašti, A. M. Bondžić, and V. M. Vasić. Acetylcholinesterase inhibitors: pharmacology and toxicology. *Curr Neuropharmacol*, 11(3):315–335, May 2013.

[67] A. G. Comuzzie, S. A. Cole, S. L. Laston, V. S. Voruganti, K. Haack, R. A. Gibbs, and N. F. Butte. Novel genetic loci identified for the pathophysiology of childhood obesity in the hispanic population. *PLoS ONE*, 7(12):e51954, December 2012.

[68] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biol*, 17(1):13, January 2016.

[69] A. Contreras-Ferrat, S. Lavandero, E. Jaimovich, and A. Klip. Calcium signaling in insulin action on striated muscle. *Cell Calcium*, 56(5):390–396, November 2014.

[70] S. L. Corcoran, P. G. Wylie, N. V. Hayes, A. J. Baines, and H. M. Thomas. Characterisation of spectrin isoforms associated with GLUT4. *Biochem Soc Trans*, 25(3):483S, August 1997.

[71] J. Coulombe-Huntington, K. C. L. Lam, C. Dias, and J. Majewski. Fine-scale variation and genetic determinants of alternative splicing across individuals. *PLoS Genet*, 5(12): e1000766, December 2009.

[72] J. Craft. Genes and genetics: the language of scientific discovery. http://public.oed.com/aspects-of-english/shapers-of-english/genes-and-genetics-the-language-of-scientific-discovery, 2017. Oxford English Dictionary Online Accessed: 2017-09-12.

[73] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1859.

[74] S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, and C. Fuchsberger. Next-generation genotype imputation service and methods. *Nat Genet*, 48(10):1284–1287, October 2016.

[75] J. K. Davies. genos. In S. Hornblower and A. Spawforth, editors, *The Oxford classical dictionary*. Oxford University Press, Oxford, 3$^{rd}$ ed. rev. edition, 2005.

[76] B. De Las Casas and D. Waddell. FGGY carbohydrate kinase domain containing is upregulated during neurogenic skeletal muscle atrophy. *The FASEB journal*, January 2016.

[77] A. M. Deaton and A. Bird. CpG islands and the regulation of transcription. *Genes Dev*, 25(10):1010–1022, May 2011.

[78] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks. Evaluation of the infinium methylation 450k technology. *Epigenomics*, 3(6):771–784, December 2011.

[79] S. Dedeurwaerder, M. Defrance, M. Bizet, E. Calonne, G. Bontempi, and F. Fuks. A comprehensive overview of Infinium HumanMethylation450 data processing. *Brief Bioinformatics*, 15(6):929–941, November 2014.

[80] O. Delaneau, J. Marchini, and J. Zagury. A linear complexity phasing method for thousands of genomes. *Nat Methods*, 9(2):179–181, December 2011.

[81] O. Delaneau, H. Ongen, A. A. Brown, A. Fort, N. I. Panousis, and E. T. Dermitzakis. A complete tool set for molecular QTL discovery and analysis. *Nat Commun*, 8:15452, May 2017.

[82] Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, Novartis Institutes of BioMedical Research, R. Saxena, B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. W. de Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn, M. J. Daly, T. E. Hughes, L. Groop, D. Altshuler, P. Almgren, J. C. Florez, J. Meyer, K. Ardlie, K. Bengtsson Boström, B. Isomaa, G. Lettre, U. Lindblad, H. N. Lyon, O. Melander, C. Newton-Cheh, P. Nilsson, M. Orho-Melander, L. Raastam, E. K. Speliotes, M. Taskinen, T. Tuomi, C. Guiducci, A. Berglund, J. Carlson, L. Gianniny, R. Hackett, L. Hall, J. Holmkvist, E. Laurila, M. Sjögren, M. Sterner, A. Surti, M. Svensson, M. Svensson, R. Tewhey, B. Blumenstiel, M. Parkin, M. Defelice, R. Barry, W. Brodeur, J. Camarata, N. Chia, M. Fava, J. Gibbons, B. Handsaker, C. Healy, K. Nguyen, C. Gates, C. Sougnez, D. Gage, M. Nizzari, S. B. Gabriel, G.-W. Chirn, Q. Ma, H. Parikh, D. Richardson, D. Ricke, and S. Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–1336, June 2007.

[83] DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Mexican American Type 2 Diabetes (MAT2D) Consortium, Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium, A. Mahajan, M. J. Go, W. Zhang, J. E. Below, K. J. Gaulton, T. Ferreira, M. Horikoshi, A. D. Johnson, M. C. Y. Ng, I. Prokopenko, D. Saleheen, X. Wang, E. Zeggini, G. R. Abecasis, L. S. Adair, P. Almgren, M. Atalay, T. Aung, D. Baldassarre, B. Balkau, Y. Bao, A. H. Barnett, I. Barroso, A. Basit, L. F. Been, J. Beilby, G. I. Bell, R. Benediktsson, R. N. Bergman, B. O. Boehm, E. Boerwinkle, L. L. Bonnycastle, N. Burtt, Q. Cai, H. Campbell, J. Carey, S. Cauchi, M. Caulfield, J. C. N. Chan, L. Chang, T. Chang, Y. Chang, G. Charpentier, C. Chen, H. Chen, Y. Chen, K. Chia, M. Chidambaram, P. S. Chines, N. H. Cho, Y. M. Cho, L. Chuang, F. S. Collins, M. C. Cornelis, D. J. Couper, A. T. Crenshaw, R. M. van Dam, J. Danesh, D. Das, U. de Faire, G. Dedoussis, P. Deloukas, A. S. Dimas, C. Dina, A. S. Doney, P. J. Donnelly, M. Dorkhan, C. van Duijn, J. Dupuis, S. Edkins, P. Elliott, V. Emilsson, R. Erbel, J. G. Eriksson, J. Escobedo, T. Esko, E. Eury, J. C. Florez, P. Fontanillas, N. G. Forouhi, T. Forsen,

C. Fox, R. M. Fraser, T. M. Frayling, P. Froguel, P. Frossard, Y. Gao, K. Gertow, C. Gieger, B. Gigante, H. Grallert, G. B. Grant, L. C. Grrop, C. J. Groves, E. Grundberg, C. Guiducci, A. Hamsten, B. Han, K. Hara, N. Hassanali, A. T. Hattersley, C. Hayward, A. K. Hedman, C. Herder, A. Hofman, O. L. Holmen, K. Hovingh, A. B. Hreidarsson, C. Hu, F. B. Hu, J. Hui, S. E. Humphries, S. E. Hunt, D. J. Hunter, K. Hveem, Z. I. Hydrie, H. Ikegami, T. Illig, E. Ingelsson, M. Islam, B. Isomaa, A. U. Jackson, T. Jafar, A. James, W. Jia, K. Jöckel, A. Jonsson, J. B. M. Jowett, T. Kadowaki, H. M. Kang, S. Kanoni, W. H. L. Kao, S. Kathiresan, N. Kato, P. Katulanda, K. M. Keinanen-Kiukaanniemi, A. M. Kelly, H. Khan, K. Khaw, C. Khor, H. Kim, S. Kim, Y. J. Kim, L. Kinnunen, N. Klopp, A. Kong, E. Korpi-Hyövälti, S. Kowlessur, P. Kraft, J. Kravic, M. M. Kristensen, S. Krithika, A. Kumar, J. Kumate, J. Kuusisto, S. H. Kwak, M. Laakso, V. Lagou, T. A. Lakka, C. Langenberg, C. Langford, R. Lawrence, K. Leander, J. Lee, N. R. Lee, M. Li, X. Li, Y. Li, J. Liang, S. Liju, W. Lim, L. Lind, C. M. Lindgren, E. Lindholm, C. Liu, J. J. Liu, S. Lobbens, J. Long, R. J. F. Loos, W. Lu, J. Luan, V. Lyssenko, R. C. W. Ma, S. Maeda, R. Mägi, S. Männisto, D. R. Matthews, J. B. Meigs, O. Melander, A. Metspalu, J. Meyer, G. Mirza, E. Mihailov, S. Moebus, V. Mohan, K. L. Mohlke, A. D. Morris, T. W. Mühleisen, M. Müller-Nurasyid, B. Musk, J. Nakamura, E. Nakashima, P. Navarro, P. Ng, A. C. Nica, P. M. Nilsson, I. Njolstad, M. M. Nöthen, K. Ohnaka, T. H. Ong, K. R. Owen, C. N. A. Palmer, J. S. Pankow, K. S. Park, M. Parkin, S. Pechlivanis, N. L. Pedersen, L. Peltonen, J. R. B. Perry, A. Peters, J. M. Pinidiyapathirage, C. G. Platou, S. Potter, J. F. Price, L. Qi, V. Radha, L. Rallidis, A. Rasheed, W. Rathman, R. Rauramaa, S. Raychaudhuri, N. W. Rayner, S. D. Rees, E. Rehnberg, S. Ripatti, N. Robertson, M. Roden, E. J. Rossin, I. Rudan, D. Rybin, T. E. Saaristo, V. Salomaa, J. Saltevo, M. Samuel, D. K. Sanghera, J. Saramies, J. Scott, L. J. Scott, R. A. Scott, A. V. Segrè, J. Sehmi, B. Sennblad, N. Shah, S. Shah, A. S. Shera, X. O. Shu, A. R. Shuldiner, G. Sigurđsson, E. Sijbrands, A. Silveira, X. Sim, S. Sivapalaratnam, K. S. Small, W. Y. So, A. Stančáková, K. Stefansson, G. Steinbach, V. Steinthorsdottir, K. Stirrups, R. J. Strawbridge, H. M. Stringham, Q. Sun, C. Suo, A. Syvänen, R. Takayanagi, F. Takeuchi, W. T. Tay, T. M. Teslovich, B. Thorand, G. Thorleifsson, U. Thorsteinsdottir, E. Tikkanen, J. Trakalo, E. Tremoli, M. D. Trip, F. J. Tsai, T. Tuomi, J. Tuomilehto, A. G. Uitterlinden, A. Valladares-Salgado, S. Vedantam, F. Veglia, B. F. Voight, C. Wang, N. J. Wareham, R. Wennauer, A. R. Wickremasinghe, T. Wilsgaard, J. F. Wilson, S. Wiltshire, W. Winckler, T. Y. Wong, A. R. Wood, J. Wu, Y. Wu, K. Yamamoto, T. Yamauchi, M. Yang, L. Yengo, M. Yokota, R. Young, D. Zabaneh, F. Zhang, R. Zhang, W. Zheng, P. Z. Zimmet, D. Altshuler, D. W. Bowden, Y. S. Cho, N. J. Cox, M. Cruz, C. L. Hanis, J. Kooner, J. Lee, M. Seielstad, Y. Y. Teo, M. Boehnke, E. J. Parra, J. C. Chambers, E. S. Tai, M. I. McCarthy, and A. P. Morris. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet*, 46(3):234–244, March 2014.

[84] Z. Ding, Y. Ni, S. W. Timmer, B. Lee, A. Battenhouse, S. Louzada, F. Yang, I. Dunham, G. E. Crawford, J. D. Lieb, R. Durbin, V. R. Iyer, and E. Birney. Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet*, 10(11):e1004798, November 2014.

[85] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[86] H. Donis-Keller, P. Green, C. Helms, S. Cartinhour, B. Weiffenbach, K. Stephens, T. P. Keith, D. W. Bowden, D. R. Smith, and E. S. Lander. A genetic linkage map of the human genome. *Cell*, 51(2):319–337, October 1987.

[87] P. Du, X. Zhang, C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11:587, November 2010.

[88] J. Dupuis, C. Langenberg, I. Prokopenko, R. Saxena, N. Soranzo, A. U. Jackson, E. Wheeler, N. L. Glazer, N. Bouatia-Naji, A. L. Gloyn, C. M. Lindgren, R. Mägi, A. P. Morris, J. Randall, T. Johnson, P. Elliott, D. Rybin, G. Thorleifsson, V. Steinthorsdottir, P. Henneman, H. Grallert, A. Dehghan, J. J. Hottenga, C. S. Franklin, P. Navarro, K. Song, A. Goel, J. R. B. Perry, J. M. Egan, T. Lajunen, N. Grarup, T. Sparsø, A. Doney, B. F. Voight, H. M. Stringham, M. Li, S. Kanoni, P. Shrader, C. Cavalcanti-Proença, M. Kumari, L. Qi, N. J. Timpson, C. Gieger, C. Zabena, G. Rocheleau, E. Ingelsson, P. An, J. O'Connell, J. Luan, A. Elliott, S. A. McCarroll, F. Payne, R. M. Roccasecca, F. Pattou, P. Sethupathy, K. Ardlie, Y. Ariyurek, B. Balkau, P. Barter, J. P. Beilby, Y. Ben-Shlomo, R. Benediktsson, A. J. Bennett, S. Bergmann, M. Bochud, E. Boerwinkle, A. Bonnefond, L. L. Bonnycastle, K. Borch-Johnsen, Y. Böttcher, E. Brunner, S. J. Bumpstead, G. Charpentier, Y.-D. I. Chen, P. Chines, R. Clarke, L. J. M. Coin, M. N. Cooper, M. Cornelis, G. Crawford, L. Crisponi, I. N. M. Day, E. J. C. de Geus, J. Delplanque, C. Dina, M. R. Erdos, A. C. Fedson, A. Fischer-Rosinsky, N. G. Forouhi, C. S. Fox, R. Frants, M. G. Franzosi, P. Galan, M. O. Goodarzi, J. Graessler, C. J. Groves, S. Grundy, R. Gwilliam, U. Gyllensten, S. Hadjadj, G. Hallmans, N. Hammond, X. Han, A.-L. Hartikainen, N. Hassanali, C. Hayward, S. C. Heath, S. Hercberg, C. Herder, A. A. Hicks, D. R. Hillman, A. D. Hingorani, A. Hofman, J. Hui, J. Hung, B. Isomaa, P. R. V. Johnson, T. Jorgensen, A. Jula, M. Kaakinen, J. Kaprio, Y. A. Kesaniemi, M. Kivimaki, B. Knight, S. Koskinen, P. Kovacs, K. O. Kyvik, G. M. Lathrop, D. A. Lawlor, O. Le Bacquer, C. Lecoeur, Y. Li, V. Lyssenko, R. Mahley, M. Mangino, A. K. Manning, M. T. Martínez-Larrad, J. B. McAteer, L. J. McCulloch, R. McPherson, C. Meisinger, D. Melzer, D. Meyre, B. D. Mitchell, M. A. Morken, S. Mukherjee, S. Naitza, N. Narisu, M. J. Neville, B. A. Oostra, M. Orrù, R. Pakyz, C. N. A. Palmer, G. Paolisso, C. Pattaro, D. Pearson, J. F. Peden, N. L. Pedersen, M. Perola, A. F. H. Pfeiffer, I. Pichler, O. Polasek, D. Posthuma, S. C. Potter, A. Pouta, M. A. Province, B. M. Psaty, W. Rathmann, N. W. Rayner, K. Rice, S. Ripatti, F. Rivadeneira, M. Roden, O. Rolandsson, A. Sandbaek, M. Sandhu, S. Sanna, A. A. Sayer, P. Scheet, L. J. Scott, U. Seedorf, S. J. Sharp, B. Shields, G. Sigurethsson, E. J. G. Sijbrands, A. Silveira, L. Simpson, A. Singleton, N. L. Smith, U. Sovio, A. Swift, H. Syddall, A. Syvänen, T. Tanaka, B. Thorand, J. Tichet, A. Tönjes, T. Tuomi, A. G. Uitterlinden, K. W. van Dijk, M. van Hoek, D. Varma, S. Visvikis-Siest, V. Vitart, N. Vogelzangs, G. Waeber, P. J. Wagner, A. Walley, G. B. Walters, K. L. Ward, H. Watkins, M. N. Weedon, S. H. Wild, G. Willemsen, J. C. M. Witteman, J. W. G. Yarnell, E. Zeggini, D. Zelenika, B. Zethelius, G. Zhai, J. H. Zhao, M. C. Zillikens, DIAGRAM Consortium, GIANT Consortium, Global BPgen Consortium, I. B. Borecki, R. J. F. Loos, P. Meneton, P. K. E. Magnusson, D. M. Nathan, G. H. Williams, A. T. Hattersley, K. Silander, V. Salomaa, G. D. Smith, S. R. Bornstein, P. Schwarz, J. Spranger, F. Karpe, A. R. Shuldiner, C. Cooper, G. V. Dedoussis, M. Serrano-Ríos, A. D. Morris, L. Lind, L. J. Palmer, F. B. Hu, P. W. Franks, S. Ebrahim, M. Marmot, W. H. L. Kao, J. S. Pankow, M. J. Sampson, J. Kuu-

sisto, M. Laakso, T. Hansen, O. Pedersen, P. P. Pramstaller, H. E. Wichmann, T. Illig, I. Rudan, A. F. Wright, M. Stumvoll, H. Campbell, J. F. Wilson, A. Hamsten, Procardis Consortium, MAGIC investigators, R. N. Bergman, T. A. Buchanan, F. S. Collins, K. L. Mohlke, J. Tuomilehto, T. T. Valle, D. Altshuler, J. I. Rotter, D. S. Siscovick, B. W. J. H. Penninx, D. I. Boomsma, P. Deloukas, T. D. Spector, T. M. Frayling, L. Ferrucci, A. Kong, U. Thorsteinsdottir, K. Stefansson, C. M. van Duijn, Y. S. Aulchenko, A. Cao, A. Scuteri, D. Schlessinger, M. Uda, A. Ruokonen, M. Jarvelin, D. M. Waterworth, P. Vollenweider, L. Peltonen, V. Mooser, G. R. Abecasis, N. J. Wareham, R. Sladek, P. Froguel, R. M. Watanabe, J. B. Meigs, L. Groop, M. Boehnke, M. I. McCarthy, J. C. Florez, and I. Barroso. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*, 42(2):105–116, February 2010.

[89] T. Elsir, A. Smits, M. S. Lindström, and M. Nistér. Transcription factor PROX1: its role in development and cancer. *Cancer Metastasis Rev*, 31(3-4):793–805, December 2012.

[90] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[91] J. Ermann and L. H. Glimcher. After GWAS: mice to the rescue? *Curr Opin Immunol*, 24(5):564–570, October 2012.

[92] J. Ernst and M. Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, 28(8):817–825, August 2010.

[93] J. Ernst and M. Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods*, 9(3):215–216, March 2012.

[94] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, May 2011.

[95] J. Fadista, P. Vikman, E. O. Laakso, I. G. Mollet, J. L. Esguerra, J. Taneera, P. Storm, P. Osmark, C. Ladenvall, R. B. Prasad, K. B. Hansson, F. Finotello, K. Uvebrant, J. K. Ofori, B. Di Camillo, U. Krus, C. M. Cilio, O. Hansson, L. Eliasson, A. H. Rosengren, E. Renström, C. B. Wollheim, and L. Groop. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci U S A*, 111(38):13924–13929, September 2014.

[96] B. P. Fairfax, P. Humburg, S. Makino, V. Naranbhai, D. Wong, E. Lau, L. Jostins, K. Plant, R. Andrews, C. McGee, and J. C. Knight. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, 343(6175): 1246949, March 2014.

[97] G. Felsenfeld. A brief history of epigenetics. *Cold Spring Harb Perspect Biol*, 6(1), January 2014.

[98] Y. Field, E. A. Boyle, N. Telis, Z. Gao, K. J. Gaulton, D. Golan, L. Yengo, G. Rocheleau, P. Froguel, M. I. McCarthy, and J. K. Pritchard. Detection of human adaptation during the past 2000 years. *Science*, 354(6313):760–764, November 2016.

[99] R. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1918.

[100] T. P. Flagg, D. Enkvetchakul, J. C. Koster, and C. G. Nichols. Muscle KATP channels: recent insights to energy sensing and myoprotection. *Physiol Rev*, 90(3):799–829, July 2010.

[101] J. Flannick, N. L. Beer, A. G. Bick, V. Agarwala, J. Molnes, N. Gupta, N. P. Burtt, J. C. Florez, J. B. Meigs, H. Taylor, V. Lyssenko, H. Irgens, E. Fox, F. Burslem, S. Johansson, M. J. Brosnan, J. K. Trimmer, C. Newton-Cheh, T. Tuomi, A. Molven, J. G. Wilson, C. J. O'Donnell, S. Kathiresan, J. N. Hirschhorn, P. R. Njolstad, T. Rolph, J. G. Seidman, S. Gabriel, D. R. Cox, C. E. Seidman, L. Groop, and D. Altshuler. Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat Genet*, 45(11):1380–1385, November 2013.

[102] J. Flannick, G. Thorleifsson, N. L. Beer, S. B. R. Jacobs, N. Grarup, N. P. Burtt, A. Mahajan, C. Fuchsberger, G. Atzmon, R. Benediktsson, J. Blangero, D. W. Bowden, I. Brandslund, J. Brosnan, F. Burslem, J. Chambers, Y. S. Cho, C. Christensen, D. A. Douglas, R. Duggirala, Z. Dymek, Y. Farjoun, T. Fennell, P. Fontanillas, T. Forsén, S. Gabriel, B. Glaser, D. F. Gudbjartsson, C. Hanis, T. Hansen, A. B. Hreidarsson, K. Hveem, E. Ingelsson, B. Isomaa, S. Johansson, T. Jorgensen, M. E. Jorgensen, S. Kathiresan, A. Kong, J. Kooner, J. Kravic, M. Laakso, J. Lee, L. Lind, C. M. Lindgren, A. Linneberg, G. Masson, T. Meitinger, K. L. Mohlke, A. Molven, A. P. Morris, S. Potluri, R. Rauramaa, R. Ribel-Madsen, A. Richard, T. Rolph, V. Salomaa, A. V. Segrè, H. Skärstrand, V. Steinthorsdottir, H. M. Stringham, P. Sulem, E. S. Tai, Y. Y. Teo, T. Teslovich, U. Thorsteinsdottir, J. K. Trimmer, T. Tuomi, J. Tuomilehto, F. Vaziri-Sani, B. F. Voight, J. G. Wilson, M. Boehnke, M. I. McCarthy, P. R. Njolstad, O. Pedersen, Go-T2D Consortium, T2D-GENES Consortium, L. Groop, D. R. Cox, K. Stefansson, and D. Altshuler. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet*, 46(4):357–363, April 2014.

[103] J. Flint and T. F. C. Mackay. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research*, 19(5):723–733, May 2009.

[104] T. Flutre, X. Wen, J. Pritchard, and M. Stephens. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet*, 9(5):e1003486, May 2013.

[105] M. P. Fogarty, T. M. Panhuis, S. Vadlamudi, M. L. Buchkovich, and K. L. Mohlke. Allele-specific transcriptional activity at type 2 diabetes-associated single nucleotide polymorphisms in regions of pancreatic islet open chromatin at the JAZF1 locus. *Diabetes*, 62(5):1756–1762, May 2013.

[106] J. Fortin and K. D. Hansen. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol*, 16(1):180, August 2015.

[107] J. Fortin, E. Fertig, and K. Hansen. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. *F1000Res*, 3:175, July 2014.

[108] P. W. Franks and M. I. McCarthy. Exposing the exposures responsible for type 2 diabetes and obesity. *Science*, 354(6308):69–73, October 2016.

[109] W. M. Freeman, S. J. Walker, and K. E. Vrana. Quantitative RT-PCR: pitfalls and potential. *BioTechniques*, 26(1):112–22, 124, January 1999.

[110] W. T. Friedewald, R. I. Levy, and D. S. Fredrickson. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem*, 18(6):499–502, June 1972.

[111] C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, M. A. Rivas, J. R. B. Perry, X. Sim, T. W. Blackwell, N. R. Robertson, N. W. Rayner, P. Cingolani, A. E. Locke, J. F. Tajes, H. M. Highland, J. Dupuis, P. S. Chines, C. M. Lindgren, C. Hartl, A. U. Jackson, H. Chen, J. R. Huyghe, M. van de Bunt, R. D. Pearson, A. Kumar, M. Müller-Nurasyid, N. Grarup, H. M. Stringham, E. R. Gamazon, J. Lee, Y. Chen, R. A. Scott, J. E. Below, P. Chen, J. Huang, M. J. Go, M. L. Stitzel, D. Pasko, S. C. J. Parker, T. V. Varga, T. Green, N. L. Beer, A. G. Day-Williams, T. Ferreira, T. Fingerlin, M. Horikoshi, C. Hu, I. Huh, M. K. Ikram, B. Kim, Y. Kim, Y. J. Kim, M.-S. Kwon, J. Lee, S. Lee, K. Lin, T. J. Maxwell, Y. Nagai, X. Wang, R. P. Welch, J. Yoon, W. Zhang, N. Barzilai, B. F. Voight, B. Han, C. P. Jenkinson, T. Kuulasmaa, J. Kuusisto, A. Manning, M. C. Y. Ng, N. D. Palmer, B. Balkau, A. Stančáková, H. E. Abboud, H. Boeing, V. Giedraitis, D. Prabhakaran, O. Gottesman, J. Scott, J. Carey, P. Kwan, G. Grant, J. D. Smith, B. M. Neale, S. Purcell, A. S. Butterworth, J. M. M. Howson, H. M. Lee, Y. Lu, S. Kwak, W. Zhao, J. Danesh, V. K. L. Lam, K. S. Park, D. Saleheen, W. Y. So, C. H. T. Tam, U. Afzal, D. Aguilar, R. Arya, T. Aung, E. Chan, C. Navarro, C. Cheng, D. Palli, A. Correa, J. E. Curran, D. Rybin, V. S. Farook, S. P. Fowler, B. I. Freedman, M. Griswold, D. E. Hale, P. J. Hicks, C. Khor, S. Kumar, B. Lehne, D. Thuillier, W. Y. Lim, J. Liu, Y. T. van der Schouw, M. Loh, S. K. Musani, S. Puppala, W. R. Scott, L. Yengo, S. Tan, H. A. Taylor, F. Thameem, G. Wilson, T. Y. Wong, P. R. Njolstad, J. C. Levy, M. Mangino, L. L. Bonnycastle, T. Schwarzmayr, J. Fadista, G. L. Surdulescu, C. Herder, C. J. Groves, T. Wieland, J. Bork-Jensen, I. Brandslund, C. Christensen, H. A. Koistinen, A. S. F. Doney, L. Kinnunen, T. Esko, A. J. Farmer, L. Hakaste, D. Hodgkiss, J. Kravic, V. Lyssenko, M. Hollensted, M. E. Jorgensen, T. Jorgensen, C. Ladenvall, J. M. Justesen, A. Käräjämäki, J. Kriebel, W. Rathmann, L. Lannfelt, T. Lauritzen, N. Narisu, A. Linneberg, O. Melander, L. Milani, M. Neville, M. Orho-Melander, L. Qi, Q. Qi, M. Roden, O. Rolandsson, A. Swift, A. H. Rosengren, K. Stirrups, A. R. Wood, E. Mihailov, C. Blancher, M. O. Carneiro, J. Maguire, R. Poplin, K. Shakir, T. Fennell, M. DePristo, M. H. de Angelis, P. Deloukas, A. P. Gjesing, G. Jun, P. Nilsson, J. Murphy, R. Onofrio, B. Thorand, T. Hansen, C. Meisinger, F. B. Hu, B. Isomaa, F. Karpe, L. Liang, A. Peters, C. Huth, S. P. O'Rahilly, C. N. A. Palmer, O. Pedersen, R. Rauramaa, J. Tuomilehto, V. Salomaa, R. M. Watanabe, A. Syvänen, R. N. Bergman, D. Bharadwaj, E. P. Bottinger, Y. S. Cho, G. R. Chandak, J. C. N. Chan, K. S. Chia, M. J. Daly, S. B. Ebrahim, C. Langenberg, P. Elliott, K. A. Jablonski, D. M. Lehman, W. Jia, R. C. W. Ma, T. I. Pollin, M. Sandhu, N. Tandon, P. Froguel, I. Barroso, Y. Y. Teo, E. Zeggini, R. J. F. Loos, K. S. Small, J. S. Ried, R. A. DeFronzo, H. Grallert, B. Glaser, A. Metspalu, N. J. Wareham, M. Walker, E. Banks, C. Gieger, E. Ingelsson, H. K. Im, T. Illig, P. W. Franks, G. Buck, J. Trakalo, D. Buck, I. Prokopenko, R. Mägi, L. Lind, Y. Farjoun, K. R. Owen, A. L.

Gloyn, K. Strauch, T. Tuomi, J. S. Kooner, J. Lee, T. Park, P. Donnelly, A. D. Morris, A. T. Hattersley, D. W. Bowden, F. S. Collins, G. Atzmon, J. C. Chambers, T. D. Spector, M. Laakso, T. M. Strom, G. I. Bell, J. Blangero, R. Duggirala, E. S. Tai, G. McVean, C. L. Hanis, J. G. Wilson, M. Seielstad, T. M. Frayling, J. B. Meigs, N. J. Cox, R. Sladek, E. S. Lander, S. Gabriel, N. P. Burtt, K. L. Mohlke, T. Meitinger, L. Groop, G. Abecasis, J. C. Florez, L. J. Scott, A. P. Morris, H. M. Kang, M. Boehnke, D. Altshuler, and M. I. McCarthy. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, August 2016.

[112] D. J. Gaffney. Global properties and functional complexity of human gene regulatory variation. *PLoS Genet*, 9(5):e1003501, May 2013.

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

[114] J. Gagneur, O. Stegle, C. Zhu, P. Jakob, M. M. Tekkedil, R. S. Aiyar, A. Schuon, D. Pe'er, and L. M. Steinmetz. Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. *PLoS Genet*, 9(9):e1003803, September 2013.

[115] I. Gallego Romero, A. A. Pai, J. Tung, and Y. Gilad. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol*, 12:42, May 2014.

[116] F. Galton. Hereditary talent and character. *Macmillan's magazine*, 12(157):318–327, 1865.

[117] F. Galton. The history of twins, as a criterion of the relative powers of nature and nurture. *The journal of the Anthropological Institute of Great Britain and Ireland*, 5: 391–406, 1876.

[118] F. Galton. Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45(273-279):135–145, January 1888.

[119] F. Galton. *Inquiries Into Human Faculty and Its Development*. Macmillan, London, 1883.

[120] F. Galton. The average contribution of each several ancestor to the total heritage of the offspring. *Proceedings of the Royal Society of London*, 61:401–413, 1897.

[121] C. C. Garcia, J. G. Potian, K. Hognason, B. Thyagarajan, L. G. Sultatos, N. Souayah, V. H. Routh, and J. J. McArdle. Acetylcholinesterase deficiency contributes to neuromuscular junction dysfunction in type 1 diabetic neuropathy. *Am J Physiol Endocrinol Metab*, 303(4):E551–61, August 2012.

[122] A. E. Garrod. The incidence of alkaptonuria: a study in chemical individuality. *Lancet*, 160(4137):1616–1620, December 1902.

[123] K. J. Gaulton, C. J. Willer, Y. Li, L. J. Scott, K. N. Conneely, A. U. Jackson, W. L. Duren, P. S. Chines, N. Narisu, L. L. Bonnycastle, J. Luo, M. Tong, A. G. Sprau, E. W. Pugh, K. F. Doheny, T. T. Valle, G. R. Abecasis, J. Tuomilehto, R. N. Bergman, F. S.

Collins, M. Boehnke, and K. L. Mohlke. Comprehensive association study of type 2 diabetes and related quantitative traits with 222 candidate genes. *Diabetes*, 57(11): 3136–3144, November 2008.

[124] J. Gayon. From Mendel to epigenetics: History of genetics. *C R Biol*, 339(7-8): 225–230, August 2016.

[125] S. Ghosh, R. M. Watanabe, E. R. Hauser, T. Valle, V. L. Magnuson, M. R. Erdos, C. D. Langefeld, J. Balow, D. S. Ally, K. Kohtamaki, P. Chines, G. Birznieks, H. S. Kaleta, A. Musick, C. Te, J. Tannenbaum, W. Eldridge, S. Shapiro, C. Martin, A. Witt, A. So, J. Chang, B. Shurtleff, R. Porter, K. Kudelko, A. Unni, L. Segal, R. Sharaf, J. Blaschak-Harvan, J. Eriksson, T. Tenkula, G. Vidgren, C. Ehnholm, E. Tuomilehto-Wolf, W. Hagopian, T. A. Buchanan, J. Tuomilehto, R. N. Bergman, F. S. Collins, and M. Boehnke. Type 2 diabetes: evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. *Proc Natl Acad Sci U S A*, 96(5):2198–2203, March 1999.

[126] S. Ghosh, R. M. Watanabe, T. T. Valle, E. R. Hauser, V. L. Magnuson, C. D. Langefeld, D. S. Ally, K. L. Mohlke, K. Silander, K. Kohtamäki, P. Chines, J. Balow Jr, G. Birznieks, J. Chang, W. Eldridge, M. R. Erdos, Z. E. Karanjawala, J. I. Knapp, K. Kudelko, C. Martin, A. Morales-Mena, A. Musick, T. Musick, C. Pfahl, R. Porter, and J. B. Rayman. The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. i. an autosomal genome scan for genes that predispose to type 2 diabetes. *Am J Hum Genet*, 67(5):1174–1185, November 2000.

[127] N. W. Gillham. The battle between the biometricians and the Mendelians: how Sir Francis Galton's work caused his disciples to reach conflicting conclusions about the hereditary mechanism. *Sci & Educ*, 24(1-2):61–75, January 2015.

[128] A. D. Goldberg, C. D. Allis, and E. Bernstein. Epigenetics: a landscape takes shape. *Cell*, 128(4):635–638, February 2007.

[129] T. Gong and J. D. Szustakowski. DeconRNAseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-seq data. *Bioinformatics*, 29 (8):1083–1085, April 2013.

[130] B. L. Goode and M. J. Eck. Mechanism and function of formins in the control of actin assembly. *Annu Rev Biochem*, 76:593–627, 2007.

[131] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6):333–351, May 2016.

[132] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol*, 29(7):644–652, May 2011.

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

[134] T. A. Greenwood, B. K. Rana, and N. J. Schork. Human haplotype block sizes are negatively correlated with recombination rates. *Genome Res*, 14(7):1358–1361, July 2004.

[135] E. Grundberg, V. Adoue, T. Kwan, B. Ge, Q. L. Duan, K. C. L. Lam, V. Koka, A. Kindmark, S. T. Weiss, K. Tantisira, H. Mallmin, B. A. Raby, O. Nilsson, and T. Pastinen. Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. *PLoS Genet*, 7(1):e1001279, January 2011.

[136] E. Grundberg, K. S. Small, A. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S.-Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, T. D. Spector, and Multiple Tissue Human Expression Resource (MuTHER) Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet*, 44(10):1084–1089, October 2012.

[137] GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat Genet*, 45(6): 580–585, June 2013.

[138] GTEx Consortium. Human genomics. the genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.

[139] H. Guo, P. Zhu, F. Guo, X. Li, X. Wu, X. Fan, L. Wen, and F. Tang. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nat Protoc*, 10(5):645–659, May 2015.

[140] J. U. Guo, Y. Su, J. H. Shin, J. Shin, H. Li, B. Xie, C. Zhong, S. Hu, T. Le, G. Fan, H. Zhu, Q. Chang, Y. Gao, G. Ming, and H. Song. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci*, 17 (2):215–222, February 2014.

[141] M. Gutierrez-Arcelus, T. Lappalainen, S. B. Montgomery, A. Buil, H. Ongen, A. Yurovsky, J. Bryois, T. Giger, L. Romano, A. Planchon, E. Falconnet, D. Bielser, M. Gagnebin, I. Padioleau, C. Borel, A. Letourneau, P. Makrythanasis, M. Guipponi, C. Gehrig, S. E. Antonarakis, and E. T. Dermitzakis. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife*, 2:e00523, June 2013.

[142] M. Gutierrez-Arcelus, H. Ongen, T. Lappalainen, S. B. Montgomery, A. Buil, A. Yurovsky, J. Bryois, I. Padioleau, L. Romano, A. Planchon, E. Falconnet, D. Bielser, M. Gagnebin, T. Giger, C. Borel, A. Letourneau, P. Makrythanasis, M. Guipponi, C. Gehrig, S. E. Antonarakis, and E. T. Dermitzakis. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet*, 11(1):e1004958, January 2015.

[143] K. S. C. Hamming, D. Soliman, L. C. Matemisz, O. Niazi, Y. Lang, A. L. Gloyn, and P. E. Light. Coexpression of the type 2 diabetes susceptibility gene variants KCNJ11 E23K and ABCC8 S1369A alter the ATP and sulfonylurea sensitivities of the ATP-sensitive K(+) channel. *Diabetes*, 58(10):2419–2424, October 2009.

[144] E. Hannon, H. Spiers, J. Viana, R. Pidsley, J. Burrage, T. M. Murphy, C. Troakes, G. Turecki, M. C. O'Donovan, L. C. Schalkwyk, N. J. Bray, and J. Mill. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci*, 19(1):48–54, January 2016.

[145] K. D. Hansen, B. Langmead, and R. A. Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, 13(10): R83, October 2012.

[146] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res*, 22(9):1760–1774, September 2012.

[147] S. W. Hartley and J. C. Mullikin. QoRTs: a comprehensive toolset for quality control and data processing of RNA-seq experiments. *BMC Bioinformatics*, 16(1):224, July 2015.

[148] P. C. Haycock, S. Burgess, K. H. Wade, J. Bowden, C. Relton, and G. Davey Smith. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am J Clin Nutr*, 103(4):965–978, April 2016.

[149] B. He, C. Chen, L. Teng, and K. Tan. Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci U S A*, 111(21):E2191–9, May 2014.

[150] M. J. Heller. DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng*, 4:129–153, March 2002.

[151] A. Hellman and A. Chess. Gene body-specific methylation on the active X chromosome. *Science*, 315(5815):1141–1143, February 2007.

[152] G. Hemani, J. Zheng, K. H. Wade, C. Laurin, B. Elsworth, S. Burgess, J. Bowden, R. Langdon, V. Tan, J. Yarmolinsky, H. A. Shihab, N. Timpson, D. M. Evans, C. Relton, R. M. Martin, G. D. Smith, T. R. Gaunt, P. C. Haycock, and The MR-Base Collaboration. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv*, December 2016.

[153] G. Hemani, K. Tilling, and G. D. Smith. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet*, 13:e1007081, November 2017.

[154] D. Hockemeyer and R. Jaenisch. Induced pluripotent stem cells meet genome editing. *Cell Stem Cell*, 18(5):573–586, May 2016.

[155] R. Holliday and J. E. Pugh. DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, January 1975.

[156] Y. Huang, W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu, and A. Rao. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, 5(1):e8888, January 2010.

[157] D. J. Hunter. Gene-environment interactions in human diseases. *Nat Rev Genet*, 6(4): 287–298, April 2005.

[158] M. Hussain, A. C. Wareham, and S. I. Head. Mechanism of action of a K+ channel activator BRL 38227 on ATP-sensitive K+ channels in mouse skeletal muscle fibres. *J Physiol (Lond)*, 478 Pt 3:523–532, August 1994.

[159] Y. Idaghdour, J. Quinlan, J. Goulet, J. Berghout, E. Gbeha, V. Bruat, T. de Malliard, J. Grenier, S. Gomez, P. Gros, M. C. Rahimy, A. Sanni, and P. Awadalla. Evidence for additive and interaction effects of host genotype and infection in malaria. *Proc Natl Acad Sci U S A*, 109(42):16786–16793, October 2012.

[160] N. Ignatiadis, B. Klaus, J. B. Zaugg, and W. Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods*, 13(7): 577–580, July 2016.

[161] International Diabetes Federation. IDF diabetes atlas. Technical report, International Diabetes Federation, Brussels, Belgium, January 2015.

[162] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, October 2005.

[163] International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. Tsui, W. Mak, Y. Q. Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro,

Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, October 2007.

[164] T. Iskratsch, S. Lange, J. Dwyer, A. L. Kho, C. dos Remedios, and E. Ehler. Formin follows function: a muscle-specific isoform of FHOD3 is regulated by CK2 phosphorylation and promotes myofibril maintenance. *J Cell Biol*, 191(6):1159–1172, December 2010.

[165] T. Iskratsch, S. Reijntjes, J. Dwyer, P. Toselli, I. R. Dégano, I. Dominguez, and E. Ehler. Two distinct phosphorylation events govern the function of muscle FHOD3. *Cell Mol Life Sci*, 70(5):893–908, March 2013.

[166] H. Iwakawa and Y. Tomari. The functions of microRNAs: mRNA decay and translational repression. *Trends Cell Biol*, 25(11):651–665, November 2015.

[167] A. E. Jaffe and R. A. Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*, 15(2):R31, February 2014.

[168] H. S. Jang, W. J. Shin, J. E. Lee, and J. T. Do. CpG and non-CpG methylation in epigenetic gene regulation and brain function. *Genes (Basel)*, 8(6):148, May 2017.

[169] W. Y. Jang, K. B. Bae, S. H. Kim, D. H. Yu, H. J. Kim, Y. R. Ji, S. J. Park, S. J. Park, M. Kang, J. I. Jeong, S. Park, S. G. Lee, I. Lee, M. O. Kim, D. Yoon, and Z. Y. Ryoo. Overexpression of JAZF1 reduces body weight gain and regulates lipid metabolism in high fat diet. *Biochem Biophys Res Commun*, 444(3):296–301, February 2014.

[170] R. C. Jansen and J. P. Nap. Genetical genomics: the added value from segregation. *Trends Genet*, 17(7):388–391, July 2001.

[171] L. Jiang, F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Res*, 21(9): 1543–1551, September 2011.

[172] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M.

Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, January 2013.

[173] P. A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 13(7):484–492, May 2012.

[174] P. A. Jones and G. Liang. Rethinking how DNA methylation patterns are maintained. *Nat Rev Genet*, 10(11):805–811, September 2009.

[175] G. Jun, M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny, G. R. Abecasis, M. Boehnke, and H. M. Kang. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet*, 91(5): 839–848, November 2012.

[176] S. E. Kahn, R. L. Prigeon, D. K. McCulloch, E. J. Boyko, R. N. Bergman, M. W. Schwartz, J. L. Neifing, W. K. Ward, J. C. Beard, and J. P. Palmer. Quantification of the relationship between insulin sensitivity and beta-cell function in human subjects. evidence for a hyperbolic function. *Diabetes*, 42(11):1663–1672, November 1993.

[177] S. E. Kahn, R. L. Hull, and K. M. Utzschneider. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature*, 444(7121):840–846, December 2006.

[178] F. J. Kaiser, A. Osmanoric, A. Rakovic, A. Erogullari, N. Uflacker, D. Braunholz, T. Lohnau, S. Orolicki, M. Albrecht, G. Gillessen-Kaesbach, C. Klein, and K. Lohmann. The dystonia gene DYT1 is repressed by the transcription factor THAP1 (DYT6). *Ann Neurol*, 68(4):554–559, October 2010.

[179] M. Kan-O, R. Takeya, T. Abe, N. Kitajima, M. Nishida, R. Tominaga, H. Kurose, and H. Sumimoto. Mammalian formin FHOD3 plays an essential role in cardiogenesis by organizing myofibrillogenesis. *Biol Open*, 1(9):889–896, September 2012.

[180] H. Kanaya, R. Takeya, K. Takeuchi, N. Watanabe, N. Jing, and H. Sumimoto. Fhos2, a novel formin-related actin-organizing protein, probably associates with the nestin intermediate filament. *Genes Cells*, 10(7):665–678, July 2005.

[181] L. Kauppi, A. Sajantila, and A. J. Jeffreys. Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum Mol Genet*, 12(1):33–40, January 2003.

[182] J. C. Keen and H. M. Moore. The genotype-tissue expression (GTEx) project: Linking clinical data with molecular analysis to advance personalized medicine. *J Pers Med*, 5 (1):22–29, February 2015.

[183] B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, M. Buchwald, and L. C. Tsui. Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922):1073–1080, September 1989.

[184] P. Kheradpour and M. Kellis. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res*, 42(5):2976–2987, March 2014.

[185] H. Kilpinen, A. Goncalves, A. Leha, V. Afzal, K. Alasoo, S. Ashford, S. Bala, D. Bensaddek, F. P. Casale, O. J. Culley, P. Danecek, A. Faulconbridge, P. W. Harrison, A. Kathuria, D. McCarthy, S. A. McCarthy, R. Meleckyte, Y. Memari, N. Moens, F. Soares, A. Mann, I. Streeter, C. A. Agu, A. Alderton, R. Nelson, S. Harper, M. Patel, A. White, S. R. Patel, L. Clarke, R. Halai, C. M. Kirton, A. Kolb-Kokocinski, P. Beales, E. Birney, D. Danovi, A. I. Lamond, W. H. Ouwehand, L. Vallier, F. M. Watt, R. Durbin, O. Stegle, and D. J. Gaffney. Common genetic variation drives molecular heterogeneity in human ipscs. *Nature*, 546(7658):370–375, June 2017.

[186] R. Kivelä, I. Salmela, Y. H. Nguyen, T. V. Petrova, H. A. Koistinen, Z. Wiener, and K. Alitalo. The transcription factor PROX1 is essential for satellite cell differentiation and muscle fibre-type regulation. *Nat Commun*, 7:13124, October 2016.

[187] R. J. Klein, C. Zeiss, E. Y. Chew, J. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, C. Barnstable, and J. Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, April 2005.

[188] D. A. Knowles, J. R. Davis, H. Edgington, A. Raj, M. Favé, X. Zhu, J. B. Potash, M. M. Weissman, J. Shi, D. F. Levinson, P. Awadalla, S. Mostafavi, S. B. Montgomery, and A. Battle. Allele-specific expression reveals interactions between genetic variation and environment. *Nat Methods*, 14(7):699–702, May 2017.

[189] A. Kontrogianni-Konstantopoulos, E. M. Jones, D. B. Van Rossum, and R. J. Bloch. Obscurin is a ligand for small ankyrin 1 in skeletal muscle. *Mol Biol Cell*, 14(3): 1138–1148, March 2003.

[190] R. Kouki, U. Schwab, T. A. Lakka, M. Hassinen, K. Savonen, P. Komulainen, B. Krachler, and R. Rauramaa. Diet, fitness and metabolic syndrome–the DR's EXTRA study. *Nutr Metab Cardiovasc Dis*, 22(7):553–560, July 2012.

[191] S. Kriaucionis and N. Heintz. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, 324(5929):929–930, May 2009.

[192] J. E. Krolopp, S. M. Thornton, and M. J. Abbott. IL-15 activates the jak3/STAT3 signaling pathway to mediate glucose uptake in skeletal muscle cells. *Front Physiol*, 7:626, December 2016.

[193] L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nat Genet*, 27(3): 234–236, March 2001.

[194] L. Kruglyak. The road to genome-wide association studies. *Nat Rev Genet*, 9(4): 314–318, April 2008.

[195] Y. H. M. Krul-Poel, S. Westra, E. ten Boekel, M. M. ter Wee, N. M. van Schoor, H. van Wijland, F. Stam, P. T. A. M. Lips, and S. Simsek. Effect of vitamin D supplementation on glycemic control in patients with type 2 diabetes (SUNNY trial): A randomized placebo-controlled trial. *Diabetes Care*, 38(8):1420–1426, August 2015.

[196] N. Kumasaka, A. J. Knights, and D. J. Gaffney. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet*, 48(2):206–213, February 2016.

[197] E. S. Lander. The new genomics: global views of biology. *Science*, 274(5287): 536–539, October 1996.

[198] E. S. Lander and D. Botstein. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199, January 1989.

[199] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[200] C. R. Landry, J. Oh, D. L. Hartl, and D. Cavalieri. Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene*, 366(2):343–351, February 2006.

[201] T. Lappalainen, 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, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. J. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, Geuvadis Consortium, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A. Syvänen, G. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, and E. T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013.

[202] S. Lecompte, G. Pasquetti, X. Hermant, B. Grenier-Boley, M. Gonzalez-Gross, S. De Henauw, D. Molnar, P. Stehle, L. Béghin, L. A. Moreno, P. Amouyel, J. Dallongeville, and A. Meirhaeghe. Genetic and molecular insights into the role of PROX1 in glucose metabolism. *Diabetes*, 62(5):1738–1745, May 2013.

[203] M. S. LeDoux. The genetics of dystonias. *Adv Genet*, 79:35–85, 2012.

[204] M. N. Lee, C. Ye, A. Villani, T. Raj, W. Li, T. M. Eisenhaure, S. H. Imboywa, P. I. Chipendo, F. A. Ran, K. Slowikowski, L. D. Ward, K. Raddassi, C. McCabe, M. H. Lee, I. Y. Frohlich, D. A. Hafler, M. Kellis, S. Raychaudhuri, F. Zhang, B. E. Stranger, C. O. Benoist, P. L. De Jager, A. Regev, and N. Hacohen. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, 343(6175): 1246980, March 2014.

[205] S. Lee, J. Lee, K. Noh, W. Choi, S. Jeon, G. T. Oh, J. Kim-Ha, Y. Jin, S. Cho, and Y. Kim. Intragenic CpG islands play important roles in bivalent chromatin assembly of developmental genes. *Proc Natl Acad Sci U S A*, 114(10):E1885–E1894, March 2017.

[206] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–1735, September 2007.

[207] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11(10):733–739, September 2010.

[208] B. Lemon and R. Tjian. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev*, 14(20):2551–2569, October 2000.

[209] G. Lettre, C. Lange, and J. N. Hirschhorn. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol*, 31(4): 358–362, May 2007.

[210] G. Lev Maor, A. Yearim, and G. Ast. The alternative role of DNA methylation in splicing regulation. *Trends Genet*, 31(5):274–280, May 2015.

[211] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey. RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, February 2010.

[212] E. Li and Y. Zhang. DNA methylation in mammals. *Cold Spring Harb Perspect Biol*, 6(5):a019133, May 2014.

[213] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[214] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[215] L. Li, Y. Shi, X. Wang, W. Shi, and C. Jiang. Single nucleotide polymorphisms in K(ATP) channels: muscular impact on type 2 diabetes. *Diabetes*, 54(5):1592–1597, May 2005.

[216] S. Li, P. P. Łabaj, P. Zumbo, P. Sykacek, W. Shi, L. Shi, J. Phan, P. Wu, M. Wang, C. Wang, D. Thierry-Mieg, J. Thierry-Mieg, D. P. Kreil, and C. E. Mason. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol*, 32(9):888–895, September 2014.

[217] Y. Li, O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu, J. A. G. Riksen, E. Hazendonk, P. Prins, R. H. A. Plasterk, R. C. Jansen, R. Breitling, and J. E. Kammenga. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. *PLoS Genet*, 2(12):e222, December 2006.

[218] Y. I. Li, D. A. Knowles, and J. K. Pritchard. LeafCutter: Annotation-free quantification of RNA splicing. *bioRxiv*, March 2016.

[219] Y. I. Li, B. van de Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, and J. K. Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, April 2016.

[220] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, 34 (8):816–834, December 2010.

[221] M. E. Lindholm, M. Huss, B. W. Solnestam, S. Kjellqvist, J. Lundeberg, and C. J. Sundberg. The human skeletal muscle transcriptome: sex differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. *FASEB J*, 28(10): 4571–4581, October 2014.

[222] C. Lippert, F. P. Casale, B. Rakitsch, and O. Stegle. LIMIX: genetic analysis of multiple traits. *bioRxiv*, May 2014.

[223] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, November 2009.

[224] R. Lister, M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. A. Thomson, R. M. Evans, and J. R. Ecker. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471(7336):68–73, March 2011.

[225] R. Lister, E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, M. Yu, J. Tonti-Filippini, H. Heyn, S. Hu, J. C. Wu, A. Rao, M. Esteller, C. He, F. G. Haghighi, T. J. Sejnowski, M. M. Behrens, and J. R. Ecker. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146):1237905, August 2013.

[226] H. Liu, X. Liu, S. Zhang, J. Lv, S. Li, S. Shang, S. Jia, Y. Wei, F. Wang, J. Su, Q. Wu, and Y. Zhang. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucleic Acids Res*, 44(1): 75–94, January 2016.

[227] L. F. Lock, N. Takagi, and G. R. Martin. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell*, 48(1):39–46, January 1987.

[228] P. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef, H. K Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, R. Durbin, and A. L Price. Reference-based phasing using the haplotype reference consortium panel. *Nat Genet*, 48(11):1443–1448, November 2016.

[229] J. Lovén, H. A. Hoke, C. Y. Lin, A. Lau, D. A. Orlando, C. R. Vakoc, J. E. Bradner, T. I. Lee, and R. A. Young. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2):320–334, April 2013.

[230] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee. Transcriptomics technologies. *PLoS Comput Biol*, 13(5):e1005457, May 2017.

[231] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res*, 45(D1): D896–D901, January 2017.

[232] T. Macfarlan, S. Kutney, B. Altman, R. Montross, J. Yu, and D. Chakravarti. Human THAP7 is a chromatin-associated, histone tail-binding protein that represses transcription via recruitment of HDAC3 and nuclear hormone receptor corepressor. *J Biol Chem*, 280(8):7346–7358, February 2005.

[233] B. Machnicka, A. Czogalla, A. Hryniewicz-Jankowska, D. M. Bogusławska, R. Grochowalska, E. Heger, and A. F. Sikorski. Spectrins: a structural platform for stabilization and activation of membrane channels, receptors and transporters. *Biochim Biophys Acta*, 1838(2):620–634, February 2014.

[234] J. Majewski and T. Pastinen. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet*, 27(2):72–79, February 2011.

[235] L. M. Mangravite, B. E. Engelhardt, M. W. Medina, J. D. Smith, C. D. Brown, D. I. Chasman, B. H. Mecham, B. Howie, H. Shim, D. Naidoo, Q. Feng, M. J. Rieder, Y. I. Chen, J. I. Rotter, P. M. Ridker, J. C. Hopewell, S. Parish, J. Armitage, R. Collins, R. A. Wilke, D. A. Nickerson, M. Stephens, and R. M. Krauss. A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature*, 502 (7471):377–380, October 2013.

[236] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22): 2867–2873, November 2010.

[237] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[238] J. C. Maranville, F. Luca, M. Stephens, and A. Di Rienzo. Mapping gene-environment interactions at regulatory polymorphisms: insights into mechanisms of phenotypic variation. *Transcription*, 3(2):56–62, April 2012.

[239] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913, July 2007.

[240] J. A. Martin and Z. Wang. Next-generation transcriptome assembly. *Nat Rev Genet*, 12(10):671–682, September 2011.

[241] A. Mathelier, O. Fornes, D. J. Arenillas, C. Chen, G. Denay, J. Lee, W. Shi, C. Shyr, G. Tan, R. Worsley-Hunt, A. W. Zhang, F. Parcy, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 44(1):110–5, January 2016.

[242] F. M. Matschinsky. Assessing the potential of glucokinase activators in diabetes therapy. *Nat Rev Drug Discov*, 8(5):399–416, May 2009.

[243] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, September 2012.

[244] M. I. McCarthy. Painting a new picture of personalised medicine for diabetes. *Diabetologia*, 60(5):793–799, February 2017.

[245] M. I. McCarthy and D. G. MacArthur. Human disease genomics: from variants to biology. *Genome Biol*, 18(1):20, January 2017.

[246] S. McCarthy, S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood, A. Teumer, H. M. Kang, C. Fuchsberger, P. Danecek, K. Sharp, Y. Luo, C. Sidore, A. Kwong, N. Timpson, S. Koskinen, S. Vrieze, L. J. Scott, H. Zhang, A. Mahajan, J. Veldink, U. Peters, C. Pato, C. M. van Duijn, C. E. Gillies, I. Gandin, M. Mezzavilla, A. Gilly, M. Cocca, M. Traglia, A. Angius, J. C. Barrett, D. Boomsma, K. Branham, G. Breen, C. M. Brummett, F. Busonero, H. Campbell, A. Chan, S. Chen, E. Chew, F. S. Collins, L. J. Corbin, G. D. Smith, G. Dedoussis, M. Dorr, A. Farmaki, L. Ferrucci, L. Forer, R. M. Fraser, S. Gabriel, S. Levy, L. Groop, T. Harrison, A. Hattersley, O. L. Holmen, K. Hveem, M. Kretzler, J. C. Lee, M. McGue, T. Meitinger, D. Melzer, J. L. Min, K. L. Mohlke, J. B. Vincent, M. Nauck, D. Nickerson, A. Palotie, M. Pato, N. Pirastu, M. McInnis, J. B. Richards, C. Sala, V. Salomaa, D. Schlessinger, S. Schoenherr, P. E. Slagboom, K. Small, T. Spector, D. Stambolian, M. Tuke, J. Tuomilehto, L. H. Van den Berg, W. Van Rheenen, U. Volker, C. Wijmenga, D. Toniolo, E. Zeggini, P. Gasparini, M. G. Sampson, J. F. Wilson, T. Frayling, P. I. W. de Bakker, M. A. Swertz, S. McCarroll, C. Kooperberg, A. Dekker, D. Altshuler, C. Willer, W. Iacono, S. Ripatti, N. Soranzo, K. Walter, A. Swaroop, F. Cucca, C. A. Anderson, R. M. Myers, M. Boehnke, M. I. McCarthy, R. Durbin, and Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*, 48(10): 1279–1283, October 2016.

[247] D. L. McCartney, R. M. Walker, S. W. Morris, A. M. McIntosh, D. J. Porteous, and K. L. Evans. Identification of polymorphic and off-target probe binding sites on the Illumina infinium methylationepic BeadChip. *Genom Data*, 9:22–24, September 2016.

[248] K. McGregor, S. Bernatsky, I. Colmegna, M. Hudson, T. Pastinen, A. Labbe, and C. M. T. Greenwood. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biol*, 17(1):84, May 2016.

[249] G. Mendel. Versuche uber pflanzen-hybriden. *Verhandlungen des naturforschenden Vereins Brünn*, (4):3–47, 1866.

[250] E. M. Mendenhall, R. P. Koche, T. Truong, V. W. Zhou, B. Issac, A. S. Chi, M. Ku, and B. E. Bernstein. GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet*, 6(12):e1001244, December 2010.

[251] E. C. Merkhofer, P. Hu, and T. L. Johnson. Introduction to cotranscriptional RNA splicing. *Methods Mol Biol*, 1126:83–96, 2014.

[252] P. Michaely, D. R. Tomchick, M. Machius, and R. G. W. Anderson. Crystal structure of a 12 ANK repeat stack from human ankyrinR. *EMBO J*, 21(23):6387–6396, December 2002.

[253] T. S. Mikkelsen, Z. Xu, X. Zhang, L. Wang, J. M. Gimble, E. S. Lander, and E. D. Rosen. Comparative epigenomic analysis of murine and human adipogenesis. *Cell*, 143(1):156–169, October 2010.

[254] J. Millstein, B. Zhang, J. Zhu, and E. E. Schadt. Disentangling molecular relationships with a causal inference test. *BMC Genet*, 10:23, May 2009.

[255] J. Millstein, G. K. Chen, and C. V. Breton. cit: hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics*, 32(15):2364–2365, August 2016.

[256] G. Ming, X. Li, J. Yin, Y. Ai, D. Xu, X. Ma, Z. Liu, H. Liu, H. Zhou, and Z. Liu. JAZF1 regulates visfatin expression in adipocytes via PPARalpha and PPARbeta/delta signaling. *Metab Clin Exp*, 63(8):1012–1021, August 2014.

[257] R. L. Minster, N. L. Hawley, C. Su, G. Sun, E. E. Kershaw, H. Cheng, O. D. Buhule, J. Lin, M. S. Reupena, S. Viali, J. Tuitele, T. Naseri, Z. Urban, R. Deka, D. E. Weeks, and S. T. McGarvey. A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nat Genet*, 48(9):1049–1054, September 2016.

[258] I. Moltke, N. Grarup, M. E. Jorgensen, P. Bjerregaard, J. T. Treebak, M. Fumagalli, T. S. Korneliussen, M. A. Andersen, T. S. Nielsen, N. T. Krarup, A. P. Gjesing, J. R. Zierath, A. Linneberg, X. Wu, G. Sun, X. Jin, J. Al-Aama, J. Wang, K. Borch-Johnsen, O. Pedersen, R. Nielsen, A. Albrechtsen, and T. Hansen. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature*, 512 (7513):190–193, August 2014.

[259] S. A. Monks, A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, S. Edwards, J. W. Phillips, A. Sachs, and E. E. Schadt. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet*, 75(6):1094–1105, December 2004.

[260] M. J. Moore and N. J. Proudfoot. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell*, 136(4):688–700, February 2009.

[261] S. Moran, C. Arribas, and M. Esteller. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3):389–399, March 2016.

[262] T. H. Morgan. Random segregation versus coupling in Mendelian inheritance. *Science*, 34(873):384, September 1911.

[263] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–747, August 2004.

[264] A. P. Morris, B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segrè, V. Steinthorsdottir, R. J. Strawbridge, H. Khan, H. Grallert, A. Mahajan, I. Prokopenko, H. M. Kang, C. Dina, T. Esko, R. M. Fraser, S. Kanoni, A. Kumar, V. Lagou, C. Langenberg, J. Luan, C. M. Lindgren, M. Müller-Nurasyid, S. Pechlivanis, N. W. Rayner, L. J. Scott, S. Wiltshire, L. Yengo, L. Kinnunen, E. J. Rossin, S. Raychaudhuri, A. D. Johnson, A. S. Dimas, R. J. F. Loos, S. Vedantam, H. Chen, J. C. Florez, C. Fox, C. Liu, D. Rybin, D. J. Couper, W. H. L. Kao, M. Li, M. C. Cornelis, P. Kraft, Q. Sun, R. M. van Dam, H. M. Stringham, P. S. Chines, K. Fischer, P. Fontanillas, O. L. Holmen, S. E. Hunt, A. U. Jackson, A. Kong, R. Lawrence, J. Meyer, J. R. B. Perry, C. G. P. Platou, S. Potter, E. Rehnberg, N. Robertson, S. Sivapalaratnam,

A. Stančáková, K. Stirrups, G. Thorleifsson, E. Tikkanen, A. R. Wood, P. Almgren, M. Atalay, R. Benediktsson, L. L. Bonnycastle, N. Burtt, J. Carey, G. Charpentier, A. T. Crenshaw, A. S. F. Doney, M. Dorkhan, S. Edkins, V. Emilsson, E. Eury, T. Forsen, K. Gertow, B. Gigante, G. B. Grant, C. J. Groves, C. Guiducci, C. Herder, A. B. Hreidarsson, J. Hui, A. James, A. Jonsson, W. Rathmann, N. Klopp, J. Kravic, K. Krjutškov, C. Langford, K. Leander, E. Lindholm, S. Lobbens, S. Männistö, G. Mirza, T. W. Mühleisen, B. Musk, M. Parkin, L. Rallidis, J. Saramies, B. Sennblad, S. Shah, G. Sigurosson, A. Silveira, G. Steinbach, B. Thorand, J. Trakalo, F. Veglia, R. Wennauer, W. Winckler, D. Zabaneh, H. Campbell, C. van Duijn, A. G. Uitterlinden, A. Hofman, E. Sijbrands, G. R. Abecasis, K. R. Owen, E. Zeggini, M. D. Trip, N. G. Forouhi, A. Syvänen, J. G. Eriksson, L. Peltonen, M. M. Nöthen, B. Balkau, C. N. A. Palmer, V. Lyssenko, T. Tuomi, B. Isomaa, D. J. Hunter, L. Qi, Wellcome Trust Case Control Consortium, Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, A. R. Shuldiner, M. Roden, I. Barroso, T. Wilsgaard, J. Beilby, K. Hovingh, J. F. Price, J. F. Wilson, R. Rauramaa, T. A. Lakka, L. Lind, G. Dedoussis, I. Njolstad, N. L. Pedersen, K. Khaw, N. J. Wareham, S. M. Keinanen-Kiukaanniemi, T. E. Saaristo, E. Korpi-Hyövälti, J. Saltevo, M. Laakso, J. Kuusisto, A. Metspalu, F. S. Collins, K. L. Mohlke, R. N. Bergman, J. Tuomilehto, B. O. Boehm, C. Gieger, K. Hveem, S. Cauchi, P. Froguel, D. Baldassarre, E. Tremoli, S. E. Humphries, D. Saleheen, J. Danesh, E. Ingelsson, S. Ripatti, V. Salomaa, R. Erbel, K. Jöckel, S. Moebus, A. Peters, T. Illig, U. de Faire, A. Hamsten, A. D. Morris, P. J. Donnelly, T. M. Frayling, A. T. Hattersley, E. Boerwinkle, O. Melander, S. Kathiresan, P. M. Nilsson, P. Deloukas, U. Thorsteinsdottir, L. C. Groop, K. Stefansson, F. Hu, J. S. Pankow, J. Dupuis, J. B. Meigs, D. Altshuler, M. Boehnke, M. I. McCarthy, and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*, 44(9):981–990, September 2012.

[265] K. V. Morris and J. S. Mattick. The rise of regulatory RNA. *Nat Rev Genet*, 15(6): 423–437, June 2014.

[266] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods*, 5(7):621–628, July 2008.

[267] G. A. Moyerbrailean, A. L. Richards, D. Kurtz, C. A. Kalita, G. O. Davis, C. T. Harvey, A. Alazizi, D. Watza, Y. Sorokin, N. Hauff, X. Zhou, X. Wen, R. Pique-Regi, and F. Luca. High-throughput allele-specific expression across 250 environmental conditions. *Genome Res*, 26(12):1627–1638, December 2016.

[268] M. Mueckler. Insulin resistance and the disruption of GLUT4 trafficking in skeletal muscle. *J Clin Invest*, 107(10):1211–1213, May 2001.

[269] J. M. Murray, K. E. Davies, P. S. Harper, L. Meredith, C. R. Mueller, and R. Williamson. Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. *Nature*, 300(5887):69–71, November 1982.

[270] T. Nakajima, S. Fujino, G. Nakanishi, Y.-S. Kim, and A. M. Jetten. TIP27: a novel repressor of the nuclear orphan receptor TAK1/TR4. *Nucleic Acids Res*, 32(14): 4194–4204, August 2004.

[271] F. K. Ndiaye, A. Ortalli, M. Canouil, M. Huyvaert, C. Salazar-Cardozo, C. Lecoeur, M. Verbanck, V. Pawlowski, R. Boutry, E. Durand, I. Rabearivelo, O. Sand, L. Marselli, J. Kerr-Conte, V. Chandra, R. Scharfmann, O. Poulain-Godefroy, P. Marchetti, F. Pattou, A. Abderrahmani, P. Froguel, and A. Bonnefond. Expression and functional assessment of candidate type 2 diabetes susceptibility genes identify four new genes contributing to human insulin secretion. *Mol Metab*, 6(6):459–470, June 2017.

[272] F. Neri, D. Incarnato, A. Krepelova, C. Parlato, and S. Oliviero. Methylation-assisted bisulfite sequencing to simultaneously map 5fC and 5caC on a genome-wide scale for DNA demethylation analysis. *Nat Protoc*, 11(7):1191–1205, July 2016.

[273] B. Ng, C. C. White, H. Klein, S. K. Sieberts, C. McCabe, E. Patrick, J. Xu, L. Yu, C. Gaiteri, D. A. Bennett, S. Mostafavi, and P. L. De Jager. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat Neurosci*, September 2017.

[274] C. W. Ng, F. Yildirim, Y. S. Yap, S. Dalin, B. J. Matthews, P. J. Velez, A. Labadorf, D. E. Housman, and E. Fraenkel. Extensive changes in DNA methylation are associated with expression of mutant huntingtin. *Proc Natl Acad Sci U S A*, 110(6):2354–2359, February 2013.

[275] A. C. Nica and E. T. Dermitzakis. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond, B, Biol Sci*, 368(1620):20120362, May 2013.

[276] NIH/CEPH Collaborative Mapping Group. A comprehensive genetic linkage map of the human genome. *Science*, 258(5079):148–162, October 1992.

[277] H. O'Geen, Y.-H. Lin, X. Xu, L. Echipare, V. M. Komashko, D. He, S. Frietze, O. Tanabe, L. Shi, M. A. Sartor, J. D. Engel, and P. J. Farnham. Genome-wide binding of the orphan nuclear receptor TR4 suggests its general role in fundamental biological processes. *BMC Genomics*, 11:689, December 2010.

[278] K. Ohno, M. A. Rahman, M. Nazim, F. Nasrin, Y. Lin, J. Takeda, and A. Masuda. Splicing regulation and dysregulation of cholinergic genes expressed at the neuromuscular junction. *J Neurochem*, 142 Suppl 2:64–72, August 2017.

[279] L. J. Orozco, A. M. Buchleitner, G. Gimenez-Perez, M. Roqué I Figuls, B. Richter, and D. Mauricio. Exercise or exercise and diet for preventing type 2 diabetes mellitus. *Cochrane Database Syst Rev*, (3):CD003054, July 2008.

[280] J. Ott, J. Wang, and S. M. Leal. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet*, 16(5):275–284, May 2015.

[281] Oxford English Dictionary Online. chromosome, n. http://www.oed.com/view/Entry/32557, June 2017. Accessed: 2017-09-12.

[282] Oxford English Dictionary Online. epigenetic, adj. and n. http://www.oed.com/view/Entry/63355, June 2017. Accessed: 2017-09-12.

[283] Oxford English Dictionary Online. genetic, adj. http://www.oed.com/view/Entry/77550, June 2017. Accessed: 2017-09-12.

[284] Oxford English Dictionary Online. genetics, n. http://www.oed.com/view/Entry/268555, June 2017. Accessed: 2017-09-12.

[285] V. Papadopoulou, A. Postigo, E. Sánchez-Tilló, A. C. G. Porter, and S. D. Wagner. ZEB1 and CtBP form a repressive complex at a distal promoter element of the BCL6 locus. *Biochem J*, 427(3):541–550, April 2010.

[286] R. Parker and H. Song. The enzymes and control of eukaryotic mRNA turnover. *Nat Struct Mol Biol*, 11(2):121–127, February 2004.

[287] S. C. J. Parker, M. L. Stitzel, D. L. Taylor, J. M. Orozco, M. R. Erdos, J. A. Akiyama, K. L. van Bueren, P. S. Chines, N. Narisu, NISC Comparative Sequencing Program, B. L. Black, A. Visel, L. A. Pennacchio, and F. S. Collins. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A*, 110(44):17921–17926, October 2013.

[288] L. Pasquali, K. J. Gaulton, S. A. Rodríguez-Seguí, L. Mularoni, I. Miguel-Escalada, İ. Akerman, J. J. Tena, I. Morán, C. Gómez-Marín, M. van de Bunt, J. Ponsa-Cobas, N. Castro, T. Nammo, I. Cebola, J. García-Hurtado, M. A. Maestro, F. Pattou, L. Piemonti, T. Berney, A. L. Gloyn, P. Ravassard, J. L. G. Skarmeta, F. Müller, M. I. McCarthy, and J. Ferrer. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet*, 46(2):136–143, February 2014.

[289] V. Patil, R. L. Ward, and L. B. Hesson. The evidence for functional non-CpG methylation in mammalian cells. *Epigenetics*, 9(6):823–828, June 2014.

[290] A. S. Paul and T. D. Pollard. Review of the mechanism of processive actin filament elongation by formins. *Cell Motil Cytoskeleton*, 66(8):606–617, August 2009.

[291] K. Pearson and A. Lee. Mathematical contributions to the theory of evolution. VIII. on the inheritance of characters not capable of exact quantitative measurement. Part I. introductory. Part II. on the inheritance of coat-colour in horses. Part III. on the inheritance of eye-colour in man. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 195(262-273):79–150, January 1900.

[292] K. Pearson. *The life, letters and labours of Francis Galton*, volume II. Cambridge University Press, Cambridge, 1924.

[293] S. Perrotta, P. G. Gallagher, and N. Mohandas. Hereditary spherocytosis. *Lancet*, 372 (9647):1411–1426, October 2008.

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

[295] R. Pidsley, C. C. Y Wong, M. Volta, K. Lunnon, J. Mill, and L. C. Schalkwyk. A data-driven approach to preprocessing Illumina 450k methylation array data. *BMC Genomics*, 14:293, May 2013.

[296] R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Djik, B. Muhlhausler, C. Stirzaker, and S. J. Clark. Critical evaluation of the Illumina methylationepic BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol*, 17(1):208, October 2016.

[297] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res*, 21(3):447–455, March 2011.

[298] R. Plomin, C. M. A. Haworth, and O. S. P. Davis. Common disorders are quantitative traits. *Nat Rev Genet*, 10(12):872–878, December 2009.

[299] K. Popadin, M. Gutierrez-Arcelus, E. T. Dermitzakis, and S. E. Antonarakis. Genetic and epigenetic regulation of human lincRNA gene expression. *Am J Hum Genet*, 93 (6):1015–1026, December 2013.

[300] P. Poulsen, K. O. Kyvik, A. Vaag, and H. Beck-Nielsen. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42(2):139–145, February 1999.

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

[302] A. L. Price, M. E. Weale, N. Patterson, S. R. Myers, A. C. Need, K. V. Shianna, D. Ge, J. I. Rotter, E. Torres, K. D. Taylor, D. B. Goldstein, and D. Reich. Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet*, 83(1):132–135, July 2008.

[303] M. E. Price, A. M. Cotton, L. L. Lam, P. Farré, E. Emberly, C. J. Brown, W. P. Robinson, and M. S. Kobor. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin*, 6(1):4, March 2013.

[304] N. J. Proudfoot, A. Furger, and M. J. Dye. Integrating mRNA processing with transcription. *Cell*, 108(4):501–512, February 2002.

[305] E. L. Putiri and K. D. Robertson. Epigenetic mechanisms and genome stability. *Clin Epigenetics*, 2(2):299–314, August 2011.

[306] T. Raj, K. Rothamel, S. Mostafavi, C. Ye, M. N. Lee, J. M. Replogle, T. Feng, M. Lee, N. Asinovski, I. Frohlich, S. Imboywa, A. Von Korff, Y. Okada, N. A. Patsopoulos, S. Davis, C. McCabe, H. Paik, G. P. Srivastava, S. Raychaudhuri, D. A. Hafler, D. Koller, A. Regev, N. Hacohen, D. Mathis, C. Benoist, B. E. Stranger, and P. L. De Jager. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science*, 344(6183):519–523, May 2014.

[307] B. Rakitsch and O. Stegle. Modelling local gene networks increases power to detect trans-acting genetic effects on gene expression. *Genome Biol*, 17(1):33, February 2016.

[308] A. Regev, S. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Gottgens, N. Hacohen, M. Haniffa, M. Hemberg, S. K. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, J. Lundeberg, P. Majumder, J. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Philipakis, C. P. Ponting, S. R. Quake, W. Reik, O. Rozenblatt-Rosen, J. R. Sanes, R. Satija, T. Shumacher, A. K. Shalek, E. Shapiro, P. Sharma, J. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, A. van Oudenaarden, A. Wagner, F. M. Watt, J. S. Weissman, B. Wold, R. J. Xavier, N. Yosef, and H. C. A. M. Participants. The human cell atlas. *bioRxiv*, May 2017.

[309] R. D. Rende, R. Plomin, and S. G. Vandenberg. Who discovered the twin method? *Behav Genet*, 20(2):277–285, March 1990.

[310] M. J. Riedel and P. E. Light. Saturated and cis/trans unsaturated acyl CoA esters differentially regulate wild-type and polymorphic beta-cell ATP-sensitive K+ channels. *Diabetes*, 54(7):2070–2079, July 2005.

[311] M. J. Riedel, P. Boora, D. Steckley, G. de Vries, and P. E. Light. Kir6.2 polymorphisms sensitize beta-cell ATP-sensitive potassium channels to activation by acyl CoAs: a possible cellular mechanism for increased susceptibility to type 2 diabetes? *Diabetes*, 52(10):2630–2635, October 2003.

[312] A. Riggs. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet*, 14(1):9–25, 1975.

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

[314] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, September 1996.

[315] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*, 32(9):896–902, September 2014.

[316] Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. Farh, S. Feizi, R. Karlic, A. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall,

N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, February 2015.

[317] M. V. Rockman and L. Kruglyak. Genetics of global gene expression. *Nat Rev Genet*, 7(11):862–872, November 2006.

[318] R. Rodriguez-Diaz, R. Dando, M. C. Jacques-Silva, A. Fachado, J. Molina, M. H. Abdulreda, C. Ricordi, S. D. Roper, P.-O. Berggren, and A. Caicedo. Alpha cells secrete acetylcholine as a non-neuronal paracrine signal priming beta cell function in humans. *Nat Med*, 17(7):888–892, June 2011.

[319] C. E. Romanoski, S. Lee, M. J. Kim, L. Ingram-Drake, C. L. Plaisier, R. Yordanova, C. Tilford, B. Guan, A. He, P. S. Gargalovic, T. G. Kirchgessner, J. A. Berliner, and A. J. Lusis. Systems genetics analysis of gene-by-environment interactions in human cells. *Am J Hum Genet*, 86(3):399–410, March 2010.

[320] J. M. Rommens, M. C. Iannuzzi, B. Kerem, M. L. Drumm, G. Melmer, M. Dean, R. Rozmahel, J. L. Cole, D. Kennedy, and N. Hidaka. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 245(4922):1059–1065, September 1989.

[321] J. Ronald, R. B. Brem, J. Whittle, and L. Kruglyak. Local regulatory variation in Saccharomyces cerevisiae. *PLoS Genet*, 1(2):e25, August 2005.

[322] M. Rosado, C. F. Barber, C. Berciu, S. Feldman, S. J. Birren, D. Nicastro, and B. L. Goode. Critical roles for multiple formins during cardiac myofibril development and repair. *Mol Biol Cell*, 25(6):811–827, March 2014.

[323] A. J. Rose, B. Kiens, and E. A. Richter. Ca2+-calmodulin-dependent protein kinase expression and signalling in skeletal muscle during exercise. *J Physiol (Lond)*, 574(Pt 3):889–903, August 2006.

[324] Royal Horticultural Society. Hybridisation (the cross-breeding of genera or species), the cross-breeding of varieties, and general plant-breeding. In W. Wilks, editor, *Report of the Third International Conference 1906 on Genetics*, London, 1907. Royal Horticultural Society.

[325] D. E. Runcie, D. A. Garfield, C. C. Babbitt, J. A. Wygoda, S. Mukherjee, and G. A. Wray. Genetics of gene expression responses to temperature stress in a sea urchin gene network. *Mol Ecol*, 21(18):4547–4562, September 2012.

[326] V. E. A. Russo, R. A. Martienssen, and A. D. Riggs. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, Plainview, N.Y., 1996.

[327] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, D. Altshuler, and International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933, February 2001.

[328] N. J. Sakabe, I. Aneas, T. Shen, L. Shokri, S. Park, M. L. Bulyk, S. M. Evans, and M. A. Nobrega. Dual transcriptional activator and repressor roles of TBX20 regulate adult cardiac structure and function. *Hum Mol Genet*, 21(10):2194–2204, May 2012.

[329] A. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res*, 42(14):8845–8860, August 2014.

[330] D. Sambandan, M. A. Carbone, R. R. H. Anholt, and T. F. C. Mackay. Phenotypic plasticity and genotype by environment interaction for olfactory behavior in Drosophila melanogaster. *Genetics*, 179(2):1079–1088, June 2008.

[331] E. Sánchez-Tilló, A. Lázaro, R. Torrent, M. Cuatrecasas, E. C. Vaquero, A. Castells, P. Engel, and A. Postigo. ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1. *Oncogene*, 29(24):3490–3500, June 2010.

[332] M. S. Sandhu, M. N. Weedon, K. A. Fawcett, J. Wasson, S. L. Debenham, A. Daly, H. Lango, T. M. Frayling, R. J. Neumann, R. Sherva, I. Blech, P. D. Pharoah, C. N. A. Palmer, C. Kimber, R. Tavendale, A. D. Morris, M. I. McCarthy, M. Walker, G. Hitman, B. Glaser, M. A. Permutt, A. T. Hattersley, N. J. Wareham, and I. Barroso. Common variants in WFS1 confer risk of type 2 diabetes. *Nat Genet*, 39(8):951–953, August 2007.

[333] J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M. A. Pujana, M. Bibikova, and M. Esteller. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, 6(6):692–702, June 2011.

[334] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467, December 1977.

[335] S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*, 103(5):1412–1417, January 2006.

[336] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235): 467–470, October 1995.

[337] E. M. Schmidt, J. Zhang, W. Zhou, J. Chen, K. L. Mohlke, Y. E. Chen, and C. J. Willer. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics*, 31(16):2601–2606, August 2015.

[338] J. Schug, W.-P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol*, 6(4):R33, March 2005.

[339] A. Schulze and J. Downward. Navigating gene expression using microarrays–a technology review. *Nat Cell Biol*, 3(8):E190–5, August 2001.

[340] L. J. Scott, 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, L. Prokunina-Olsson, C. Ding, A. J. Swift, N. Narisu, T. Hu, R. Pruim, R. Xiao, X. Li, K. N. Conneely, N. L. Riebow, A. G. Sprau, M. Tong, P. P. White, K. N. Hetrick, M. W. Barnhart, C. W. Bark, J. L. Goldstein, L. Watkins, F. Xiang, J. Saramies, T. A. Buchanan, R. M. Watanabe, T. T. Valle, L. Kinnunen, G. R. Abecasis, E. W. Pugh, K. F. Doheny, R. N. Bergman, J. Tuomilehto, F. S. Collins, and M. Boehnke. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316(5829): 1341–1345, June 2007.

[341] L. J. Scott, M. R. Erdos, J. R. Huyghe, R. P. Welch, A. T. Beck, B. N. Wolford, P. S. Chines, J. P. Didion, N. Narisu, H. M. Stringham, D. L. Taylor, A. U. Jackson, S. Vadlamudi, L. L. Bonnycastle, L. Kinnunen, J. Saramies, J. Sundvall, R. D. Albanus, A. Kiseleva, J. Hensley, G. E. Crawford, H. Jiang, X. Wen, R. M. Watanabe, T. A. Lakka, K. L. Mohlke, M. Laakso, J. Tuomilehto, H. A. Koistinen, M. Boehnke, F. S. Collins, and S. C. J. Parker. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat Commun*, 7:11764, June 2016.

[342] R. Scott, L. Scott, R. Mägi, L. Marullo, K. Gaulton, M. Kaakinen, N. Pervjakova, T. Pers, A. Johnson, J. Eicher, A. Jackson, T. Ferreira, Y. Lee, C. Ma, V. Steinthorsdottir, G. Thorleifsson, L. Qi, N. Van Zuydam, A. Mahajan, H. Chen, P. Almgren, B. Voight, H. Grallert, M. Müller-Nurasyid, J. Ried, W. Rayner, N. Robertson, L. Karssen, E. van Leeuwen, S. Willems, C. Fuchsberger, P. Kwan, T. Teslovich, P. Chanda, M. Li, Y. Lu, C. Dina, D. Thuillier, L. Yengo, L. Jiang, T. Sparso, H. Kestler, H. Chheda, L. Eisele, S. Gustafsson, M. Fräanberg, R. Strawbridge, R. Benediktsson, A. Hreidarsson, A. Kong, G. Sigurðsson, N. Kerrison, J. Luan, L. Liang, T. Meitinger, M. Roden, B. Thorand, T. Esko, E. Mihailov, C. Fox, C. Liu, D. Rybin, B. Isomaa, V. Lyssenko, T. Tuomi, D. Couper, J. Pankow, N. Grarup, C. Have, M. Jorgensen, T. Jorgensen, A. Linneberg, M. Cornelis, R. van Dam, D. Hunter, P. Kraft, Q. Sun, S. Edkins, K. Owen, J. Perry, A. Wood, E. Zeggini, J. Tajes-Fernandes, G. Abecasis, L. Bonnycastle, P. Chines, H. Stringham, H. Koistinen, L. Kinnunen, B. Sennblad, T. Mühleisen, M. Nöthen, S. Pechlivanis, D. Baldassarre, K. Gertow, S. Humphries, E. Tremoli, N. Klopp, J. Meyer, G. Steinbach, R. Wennauer, J. Eriksson, S. Männistö, L. Peltonen, E. Tikkanen, G. Charpentier, E. Eury, S. Lobbens, B. Gigante, K. Leander, O. McLeod, E. Bottinger, O. Gottesman, D. Ruderfer, M. Blüher, P. Kovacs, A. Tonjes, N. Maruthur, C. Scapoli, R. Erbel, K. Jöckel, S. Moebus, U. de Faire, A. Hamsten, M. Stumvoll, P. Deloukas, P. Donnelly, T. Frayling, A. Hattersley, S. Ripatti, V. Salomaa, N. Pedersen, B. Boehm, R. Bergman, F. Collins, K. Mohlke, J. Tuomilehto, T. Hansen, O. Pedersen, I. Barroso, L. Lannfelt, E. Ingelsson, L. Lind, C. Lindgren, S. Cauchi, P. Froguel, R. Loos, B. Balkau, H. Boeing, P. Franks, A. Gurrea, D. Palli, Y. van der Schouw, D. Altshuler, L. Groop, C. Langenberg, N. Wareham, E. Sijbrands, C. van Duijn, J. Florez, J. Meigs, E. Boerwinkle, C. Gieger, K. Strauch, A. Metspalu,

A. Morris, C. Palmer, F. Hu, U. Thorsteinsdottir, K. Stefansson, J. Dupuis, A. Morris, M. Boehnke, M. McCarthy, I. Prokopenko, and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. An expanded genome-wide association study of type 2 diabetes in europeans. *Diabetes*, May 2017.

[343] SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat Biotechnol*, 32(9):903–914, September 2014.

[344] A. Seth, D. L. Stemple, and I. Barroso. The emerging use of zebrafish to model metabolic disease. *Dis Model Mech*, 6(5):1080–1088, September 2013.

[345] A. A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, May 2012.

[346] W. Shakespeare. *The Tempest*. Macmillian, London, 1864.

[347] O. Shalem, N. E. Sanjana, and F. Zhang. High-throughput functional genomics using CRISPR–Cas9. *Nat Rev Genet*, 16(5):299, May 2015.

[348] D. Shi, I. Ali, J. Tang, and W. Yang. New insights into 5hmC DNA modification: Generation, distribution and function. *Front Genet*, 8:100, July 2017.

[349] L. Sie, S. Loong, and E. K. Tan. Utility of lymphoblastoid cell lines. *J Neurosci Res*, 87(9):1953–1959, July 2009.

[350] SIGMA Type 2 Diabetes Consortium, K. Estrada, I. Aukrust, L. Bjorkhaug, N. P. Burtt, J. M. Mercader, H. García-Ortiz, A. Huerta-Chagoya, H. Moreno-Macías, G. Walford, J. Flannick, A. L. Williams, M. J. Gómez-Vázquez, J. C. Fernandez-Lopez, A. Martínez-Hernández, S. Jiménez-Morales, F. Centeno-Cruz, E. Mendoza-Caamal, C. Revilla-Monsalve, S. Islas-Andrade, E. J. Córdova, X. Soberón, M. E. González-Villalpando, E. Henderson, L. R. Wilkens, L. Le Marchand, O. Arellano-Campos, M. L. Ordóñez-Sánchez, M. Rodríguez-Torres, R. Rodríguez-Guillén, L. Riba, L. A. Najmi, S. B. R. Jacobs, T. Fennell, S. Gabriel, P. Fontanillas, C. L. Hanis, D. M. Lehman, C. P. Jenkinson, H. E. Abboud, G. I. Bell, M. L. Cortes, M. Boehnke, C. González-Villalpando, L. Orozco, C. A. Haiman, T. Tusié-Luna, C. A. Aguilar-Salinas, D. Altshuler, P. R. Njolstad, J. C. Florez, and D. G. MacArthur. Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA*, 311(22):2305–2314, June 2014.

[351] K. Silander, L. J. Scott, T. T. Valle, K. L. Mohlke, H. M. Stringham, K. R. Wiles, W. L. Duren, K. F. Doheny, E. W. Pugh, P. Chines, N. Narisu, P. P. White, T. E. Fingerlin, A. U. Jackson, C. Li, S. Ghosh, V. L. Magnuson, K. Colby, M. R. Erdos, J. E. Hill, P. Hollstein, K. M. Humphreys, R. A. Kasad, J. Lambert, K. N. Lazaridis, G. Lin, A. Morales-Mena, K. Patzkowski, C. Pfahl, R. Porter, D. Rha, L. Segal, Y. D. Suh, J. Tovar, A. Unni, C. Welch, J. A. Douglas, M. P. Epstein, E. R. Hauser, W. Hagopian, T. A. Buchanan, R. M. Watanabe, R. N. Bergman, J. Tuomilehto, F. S. Collins, and M. Boehnke. A large set of Finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. *Diabetes*, 53(3):821–829, March 2004.

[352] L. Siles, E. Sánchez-Tilló, J. Lim, D. S. Darling, K. L. Kroll, and A. Postigo. ZEB1 imposes a temporary stage-dependent inhibition of muscle gene expression and differentiation via CtBP-mediated transcriptional repression. *Mol Cell Biol*, 33(7): 1368–1382, April 2013.

[353] X. Sim, R. T. Ong, C. Suo, W. Tay, J. Liu, D. P. Ng, M. Boehnke, K. Chia, T. Wong, M. Seielstad, Y. Teo, and E.-S. Tai. Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS Genet*, 7(4):e1001363, April 2011.

[354] J. T. Simpson, R. E. Workman, P. C. Zuzarte, M. David, L. J. Dursi, and W. Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*, 14 (4):407–410, April 2017.

[355] P. Singh, J. C. Schimenti, and E. Bolcun-Filas. A mouse geneticist's practical guide to crispr applications. *Genetics*, 199(1):1–15, January 2015.

[356] D. A. Smirnov, M. Morley, E. Shin, R. S. Spielman, and V. G. Cheung. Genetic analysis of radiation-induced changes in human gene expression. *Nature*, 459(7246): 587–591, May 2009.

[357] A. K. Smith, V. Kilaru, M. Kocak, L. M. Almli, K. B. Mercer, K. J. Ressler, F. A. Tylavsky, and K. N. Conneely. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*, 15:145, February 2014.

[358] E. N. Smith and L. Kruglyak. Gene-environment interaction in yeast gene expression. *PLoS Biol*, 6(4):e83, April 2008.

[359] G. D. Smith and S. Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*, 32(1):1–22, February 2003.

[360] G. D. Smith, N. Timpson, and S. Ebrahim. Strengthening causal inference in cardiovascular epidemiology through Mendelian randomization. *Ann Med*, 40(7):524–541, 2008.

[361] Q. Song, B. Decato, E. E. Hong, M. Zhou, F. Fang, J. Qu, T. Garvin, M. Kessler, J. Zhou, and A. D. Smith. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS ONE*, 8(12):e81148, December 2013.

[362] W. W. Soon, M. Hariharan, and M. P. Snyder. High-throughput sequencing for biology and medicine. *Mol Syst Biol*, 9:640, 2013.

[363] H. Soreq and S. Seidman. Acetylcholinesterase–new roles for an old actor. *Nat Rev Neurosci*, 2(4):294–302, April 2001.

[364] M. B. Stadler, R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Schöler, E. van Nimwegen, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schübeler. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480 (7378):490–495, December 2011.

[365] A. Stancáková, T. Kuulasmaa, J. Paananen, A. U. Jackson, L. L. Bonnycastle, F. S. Collins, M. Boehnke, J. Kuusisto, and M. Laakso. Association of 18 confirmed susceptibility loci for type 2 diabetes with indices of insulin release, proinsulin conversion, and insulin sensitivity in 5,327 nondiabetic Finnish men. *Diabetes*, 58(9):2129–2136, September 2009.

[366] O. Stegle, L. Parts, R. Durbin, and J. Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol*, 6(5):e1000770, May 2010.

[367] O. Stegle, L. Parts, M. Piipari, J. Winn, and R. Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*, 7(3):500–507, February 2012.

[368] S. M. Stigler. Galton and identification by fingerprints. *Genetics*, 140(3):857–860, July 1995.

[369] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, August 2003.

[370] B. E. Stranger, M. S. Forrest, A. G. Clark, M. J. Minichiello, S. Deutsch, R. Lyle, S. Hunt, B. Kahl, S. E. Antonarakis, S. Tavaré, P. Deloukas, and E. T. Dermitzakis. Genome-wide associations of gene expression variation in humans. *PLoS Genet*, 1(6): e78, December 2005.

[371] B. E. Stranger, A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, S. Montgomery, S. Tavaré, P. Deloukas, and E. T. Dermitzakis. Population genomics of human gene expression. *Nat Genet*, 39(10): 1217–1224, October 2007.

[372] M. Stumvoll, B. J. Goldstein, and T. W. van Haeften. Type 2 diabetes: principles of pathogenesis and therapy. *Lancet*, 365(9467):1333–1346, April 2005.

[373] A. H. Sturtevant. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *J Exp Zool*, 14(1):43–59, January 1913.

[374] A. H. Sturtevant. *A History of Genetics*. CSHL Press, illustrated, reprint edition, 2001.

[375] F. Stutz and E. Izaurralde. The interplay of nuclear mRNP assembly, mRNA surveillance and export. *Trends Cell Biol*, 13(6):319–327, June 2003.

[376] L. Sun, K. Ma, H. Wang, F. Xiao, Y. Gao, W. Zhang, K. Wang, X. Gao, N. Ip, and Z. Wu. JAK1-STAT1-STAT3, a key pathway promoting proliferation and preventing premature differentiation of myoblasts. *J Cell Biol*, 179(1):129–138, October 2007.

[377] W. S. Sutton. On the morphology of the chromoso group in brachystola magna. *Biol Bull*, 4(1):24–39, December 1902.

[378] W. S. Sutton. The chromosomes in heredity. *Biol Bull*, 4(5):231–250, April 1903.

[379] D. I. Swerdlow. Mendelian randomization and type 2 diabetes. *Cardiovasc Drugs Ther*, 30(1):51–57, February 2016.

[380] A. Szanto, V. Narkar, Q. Shen, I. P. Uray, P. J. A. Davies, and L. Nagy. Retinoid x receptors: X-ploring their (patho)physiological functions. *Cell Death Differ*, 11 Suppl 2:S126–43, December 2004.

[381] P. A. C. 't Hoen, M. R. Friedländer, J. Almlöf, M. Sammeth, I. Pulyakhina, S. Y. Anvar, J. F. J. Laros, H. P. J. Buermans, O. Karlberg, M. Brännvall, GEUVADIS Consortium, J. T. den Dunnen, G. B. van Ommen, I. G. Gut, R. Guigó, X. Estivill, A. Syvänen, E. T. Dermitzakis, and T. Lappalainen. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol*, 31(11):1015–1022, November 2013.

[382] M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, and A. Rao. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324 (5929):930–935, May 2009.

[383] J. Taneera, S. Lang, A. Sharma, J. Fadista, Y. Zhou, E. Ahlqvist, A. Jonsson, V. Lyssenko, P. Vikman, O. Hansson, H. Parikh, O. Korsgren, A. Soni, U. Krus, E. Zhang, X. Jing, J. L. S. Esguerra, C. B. Wollheim, A. Salehi, A. Rosengren, E. Renström, and L. Groop. A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. *Cell Metab*, 16(1):122–134, July 2012.

[384] D. L. Taylor, D. A. Knowles, L. J. Scott, A. H. Ramirez, F. P. Casale, B. N. Wolford, L. Guan, A. Varshney, R. D. Albanus, S. C. Parker, N. Narisu, P. S. Chines, M. R. Erdos, R. P. Welch, L. Kinnunen, J. Saramies, J. Sundvall, T. A. Lakka, M. Laakso, J. Tuomilehto, H. A. Koistinen, O. Stegle, M. Boehnke, E. Birney, and F. S. Collins. Interactions between genetic variation and cellular environment in skeletal muscle gene expression. *bioRxiv*, February 2017.

[385] J. P. Thomson, P. J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A. R. W. Kerr, A. Deaton, R. Andrews, K. D. James, D. J. Turner, R. Illingworth, and A. Bird. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, 464 (7291):1082–1086, April 2010.

[386] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010.

[387] M. K. Trenerry, P. A. Della Gatta, and D. Cameron-Smith. JAK/STAT signaling and human in vitro MyoGenesis. *BMC Physiol*, 11:6, March 2011.

[388] T. Tsakiridis, M. Vranic, and A. Klip. Disassembly of the actin network inhibits insulin-dependent stimulation of glucose transport and prevents recruitment of glucose transporters to the plasma membrane. *J Biol Chem*, 269(47):29934–29942, November 1994.

[389] F. Urano. Wolfram syndrome: diagnosis, management, and treatment. *Curr Diab Rep*, 16(1):6, January 2016.

[390] S. Väätäinen, S. Keinänen-Kiukaanniemi, J. Saramies, H. Uusitalo, J. Tuomilehto, and J. Martikainen. Quality of life along the diabetes continuum: a cross-sectional view of health-related quality of life and general health status in middle-aged and older finns. *Qual Life Res*, 23(7):1935–1944, September 2014.

[391] T. Valle, J. Tuomilehto, R. N. Bergman, S. Ghosh, E. R. Hauser, J. Eriksson, S. J. Nylund, K. Kohtamäki, L. Toivanen, G. Vidgren, E. Tuomilehto-Wolf, C. Ehnholm, J. Blaschak, C. D. Langefeld, R. M. Watanabe, V. Magnuson, D. S. Ally, W. A. Hagopian, E. Ross, T. A. Buchanan, F. Collins, and M. Boehnke. Mapping genes for NIDDM. design of the Finland-United States investigation of NIDDM genetics (FUSION) study. *Diabetes Care*, 21(6):949–958, June 1998.

[392] M. van de Bunt and A. L. Gloyn. From genetic association to molecular mechanism. *Curr Diab Rep*, 10(6):452–466, December 2010.

[393] M. van de Bunt, J. E. Manning Fox, X. Dai, A. Barrett, C. Grey, L. Li, A. J. Bennett, P. R. Johnson, R. V. Rajotte, K. J. Gaulton, E. T. Dermitzakis, P. E. MacDonald, M. I. McCarthy, and A. L. Gloyn. Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. *PLoS Genet*, 11(12):e1005694, December 2015.

[394] J. G. Van Vranken, D. K. Bricker, N. Dephoure, S. P. Gygi, J. E. Cox, C. S. Thummel, and J. Rutter. SDHAF4 promotes mitochondrial succinate dehydrogenase activity and prevents neurodegeneration. *Cell Metab*, 20(2):241–252, August 2014.

[395] A. Varshney, L. J. Scott, R. P. Welch, M. R. Erdos, P. S. Chines, N. Narisu, R. D. Albanus, P. Orchard, B. N. Wolford, R. Kursawe, S. Vadlamudi, M. E. Cannon, J. P. Didion, J. Hensley, A. Kirilusha, NISC Comparative Sequencing Program, L. L. Bonnycastle, D. L. Taylor, R. Watanabe, K. L. Mohlke, M. Boehnke, F. S. Collins, S. C. J. Parker, and M. L. Stitzel. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc Natl Acad Sci U S A*, 114(9):2301–2306, February 2017.

[396] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu,

S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, February 2001.

[397] D. T. Villareal, J. C. Koster, H. Robertson, A. Akrouh, K. Miyake, G. I. Bell, B. W. Patterson, C. G. Nichols, and K. S. Polonsky. Kir6.2 variant E23K increases ATP-sensitive K+ channel activity and is associated with impaired insulin release and enhanced insulin sensitivity in adults with normal glucose tolerance. *Diabetes*, 58(8): 1869–1878, August 2009.

[398] E. Viré, C. Brenner, R. Deplus, L. Blanchon, M. Fraga, C. Didelot, L. Morey, A. Van Eynde, D. Bernard, J. Vanderwinden, M. Bollen, M. Esteller, L. Di Croce, Y. de Launoit, and F. Fuks. The polycomb group protein EZH2 directly controls DNA methylation. *Nature*, 439(7078):871–874, February 2006.

[399] P. M. Visscher and J. Yang. A plethora of pleiotropy across complex traits. *Nat Genet*, 48(7):707–708, June 2016.

[400] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *Am J Hum Genet*, 90(1):7–24, January 2012.

[401] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*, 101(1):5–22, July 2017.

[402] B. F. Voight, L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, R. P. Welch, E. Zeggini, C. Huth, Y. S. Aulchenko, G. Thorleifsson, L. J. McCulloch, T. Ferreira, H. Grallert, N. Amin, G. Wu, C. J. Willer, S. Raychaudhuri, S. A. McCarroll, C. Langenberg, O. M. Hofmann, J. Dupuis, L. Qi, A. V. Segrè, M. van Hoek, P. Navarro,

K. Ardlie, B. Balkau, R. Benediktsson, A. J. Bennett, R. Blagieva, E. Boerwinkle, L. L. Bonnycastle, K. Bengtsson Boström, B. Bravenboer, S. Bumpstead, N. P. Burtt, G. Charpentier, P. S. Chines, M. Cornelis, D. J. Couper, G. Crawford, A. S. F. Doney, K. S. Elliott, A. L. Elliott, M. R. Erdos, C. S. Fox, C. S. Franklin, M. Ganser, C. Gieger, N. Grarup, T. Green, S. Griffin, C. J. Groves, C. Guiducci, S. Hadjadj, N. Hassanali, C. Herder, B. Isomaa, A. U. Jackson, P. R. V. Johnson, T. Jorgensen, W. H. L. Kao, N. Klopp, A. Kong, P. Kraft, J. Kuusisto, T. Lauritzen, M. Li, A. Lieverse, C. M. Lindgren, V. Lyssenko, M. Marre, T. Meitinger, K. Midthjell, M. A. Morken, N. Narisu, P. Nilsson, K. R. Owen, F. Payne, J. R. B. Perry, A. Petersen, C. Platou, C. Proença, I. Prokopenko, W. Rathmann, N. W. Rayner, N. R. Robertson, G. Rocheleau, M. Roden, M. J. Sampson, R. Saxena, B. M. Shields, P. Shrader, G. Sigurdsson, T. Sparsø, K. Strassburger, H. M. Stringham, Q. Sun, A. J. Swift, B. Thorand, J. Tichet, T. Tuomi, R. M. van Dam, T. W. van Haeften, T. van Herpt, J. V. van Vliet-Ostaptchouk, G. B. Walters, M. N. Weedon, C. Wijmenga, J. Witteman, R. N. Bergman, S. Cauchi, F. S. Collins, A. L. Gloyn, U. Gyllensten, T. Hansen, W. A. Hide, G. A. Hitman, A. Hofman, D. J. Hunter, K. Hveem, M. Laakso, K. L. Mohlke, A. D. Morris, C. N. A. Palmer, P. P. Pramstaller, I. Rudan, E. Sijbrands, L. D. Stein, J. Tuomilehto, A. Uitterlinden, M. Walker, N. J. Wareham, R. M. Watanabe, G. R. Abecasis, B. O. Boehm, H. Campbell, M. J. Daly, A. T. Hattersley, F. B. Hu, J. B. Meigs, J. S. Pankow, O. Pedersen, H. Wichmann, I. Barroso, J. C. Florez, T. M. Frayling, L. Groop, R. Sladek, U. Thorsteinsdottir, J. F. Wilson, T. Illig, P. Froguel, C. M. van Duijn, K. Stefansson, D. Altshuler, M. Boehnke, M. I. McCarthy, MAGIC investigators, and GIANT Consortium. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet*, 42(7):579–589, July 2010.

[403] C. H. Waddington. The epigenotype. *Endeavor*, 1:18–20, 1942.

[404] C. H. Waddington. *The strategy of the genes: a discussion of some aspects of theoretical biology*. Allen & Unwin, London, 1957.

[405] G. P. Wagner, K. Kin, and V. J. Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*, 131 (4):281–285, December 2012.

[406] J. R. Wagner, S. Busche, B. Ge, T. Kwan, T. Pastinen, and M. Blanchette. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol*, 15(2):R37, February 2014.

[407] C. Wang, X. Zhan, L. Liang, G. R. Abecasis, and X. Lin. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *Am J Hum Genet*, 96(6):926–937, June 2015.

[408] Z. Wang, M. Gerstein, and M. Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, January 2009.

[409] L. D. Ward and M. Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol*, 30(11):1095–1106, November 2012.

[410] T. Wasada, T. Yano, M. Ohta, N. Yui, and Y. Iwamoto. ATP-sensitive potassium channels modulate glucose transport in cultured human skeletal muscle cells. *Endocr J*, 48(3):369–375, 2001.

[411] K. M. Waters, D. O. Stram, M. T. Hassanein, L. Le Marchand, L. R. Wilkens, G. Maskarinec, K. R. Monroe, L. N. Kolonel, D. Altshuler, B. E. Henderson, and C. A. Haiman. Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. *PLoS Genet*, 6(8):e1001078, August 2010.

[412] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[413] M. E. Weale. Quality control for genome-wide association studies. *Methods Mol Biol*, 628:341–372, 2010.

[414] J. L. Weber and P. E. May. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet*, 44(3):388–396, March 1989.

[415] M. N. Weedon and P. Light. From association to function: KCNJ11 and ABCC8. In J. C. Florez, editor, *The genetics of type 2 diabetes and related traits*, pages 363–377. Springer International Publishing, Cham, 2016.

[416] W. Wei, R. M. Cronin, H. Xu, T. A. Lasko, L. Bastarache, and J. C. Denny. Development and evaluation of an ensemble resource linking medications to their indications. *J Am Med Inform Assoc*, 20(5):954–961, October 2013.

[417] J. Weissenbach, G. Gyapay, C. Dib, A. Vignal, J. Morissette, P. Millasseau, G. Vaysseix, and M. Lathrop. A second-generation linkage map of the human genome. *Nature*, 359(6398):794–801, October 1992.

[418] W. F. R. Weldon. On the ambiguity of mendel's categories. *Biometrika*, 2(1):44–55, November 1902.

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

[420] U. Wellner, J. Schubert, U. C. Burk, O. Schmalhofer, F. Zhu, A. Sonntag, B. Waldvogel, C. Vannier, D. Darling, A. zur Hausen, V. G. Brunton, J. Morton, O. Sansom, J. Schüler, M. P. Stemmler, C. Herzberger, U. Hopt, T. Keck, S. Brabletz, and T. Brabletz. The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. *Nat Cell Biol*, 11(12):1487–1495, December 2009.

[421] W. A. Whyte, D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, April 2013.

[422] C. D. Willis, T. Oashi, B. Busby, A. D. Mackerell, and R. J. Bloch. Hydrophobic residues in small ankyrin 1 participate in binding to obscurin. *Mol Membr Biol*, 29(2): 36–51, March 2012.

[423] B. Wold and R. M. Myers. Sequence census methods for functional genomics. *Nat Methods*, 5(1):19–21, January 2008.

[424] E. C. Wooten, V. B. Hebl, M. J. Wolf, S. R. Greytak, N. M. Orr, I. Draper, J. E. Calvino, N. K. Kapur, M. S. Maron, I. J. Kullo, S. R. Ommen, J. M. Bos, M. J. Ackerman, and G. S. Huggins. Formin homology 2 domain containing 3 variants associated with hypertrophic cardiomyopathy. *Circ Cardiovasc Genet*, 6(1):10–18, February 2013.

[425] World Health Organization. *ATC classification index with DDDs*. WHO Collaborating Centre for Drug Statistics Methodology, Oslo, Norway, 2016.

[426] World Health Organization. The top 10 causes of death worldwide in 2015. http://www.who.int/mediacentre/factsheets/fs310/en, January 2017. Accessed: 2017-09-12.

[427] World Health Organization and International Diabetes Federation. *Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycaemia: Report of a WHO/IDF Consultation*. WHO Press, Geneva, Switzerland, 2006.

[428] X. Wu and Y. Zhang. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat Rev Genet*, 18(9):517–534, September 2017.

[429] Y. Yang, Y. Xu, W. Li, G. Wang, Y. Song, G. Yang, X. Han, Z. Du, L. Sun, and K. Ma. STAT3 induces muscle stem cell differentiation by interaction with myoD. *Cytokine*, 46(1):137–141, April 2009.

[430] C. J. Ye, T. Feng, H.-K. Kwon, T. Raj, M. T. Wilson, N. Asinovski, C. McCabe, M. H. Lee, I. Frohlich, H. Paik, N. Zaitlen, N. Hacohen, B. Stranger, P. De Jager, D. Mathis, A. Regev, and C. Benoist. Intersection of population variation and autoimmunity genetics in human t cell activation. *Science*, 345(6202):1254665, September 2014.

[431] Z. Ye, S. J. Sharp, S. Burgess, R. A. Scott, F. Imamura, C. Langenberg, N. J. Wareham, and N. G. Forouhi. Association between circulating 25-hydroxyvitamin D and incident type 2 diabetes: a Mendelian randomisation study. *The Lancet Diabetes & Endocrinology*, 3(1):35–42, January 2015.

[432] A. Yearim, S. Gelfman, R. Shayevitch, S. Melcer, O. Glaich, J. Mallm, M. Nissim-Rafinia, A. S. Cohen, K. Rippe, E. Meshorer, and G. Ast. HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. *Cell Rep*, 10(7):1122–1134, February 2015.

[433] W. Yong, F. Hsu, and P. Chen. Profiling genome-wide DNA methylation. *Epigenetics Chromatin*, 9(1):26, June 2016.

[434] M. Yu, G. C. Hon, K. E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, and C. He. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, 149(6):1368–1380, June 2012.

[435] L. Yuan, X. Luo, M. Zeng, Y. Zhang, M. Yang, L. Zhang, R. Liu, G. Boden, H. Liu, Z. A. Ma, L. Li, and G. Yang. Transcription factor TIP27 regulates glucose homeostasis and insulin sensitivity in a PI3-kinase/Akt-dependent manner in mice. *Int J Obes (Lond)*, 39(6):949–958, June 2015.

[436] K. Yuasa, N. Aoki, and T. Hijikata. JAZF1 promotes proliferation of C2C12 cells, but retards their MyoGenic differentiation through transcriptional repression of MEF2C and MRF4-implications for the role of JAZF1 variants in oncogenesis and type 2 diabetes. *Exp Cell Res*, 336(2):287–297, August 2015.

[437] G. Yule. Mendel's laws and their probable relations to intra-racial heredity. *New Phytologist*, 1(10):222–238, 1902.

[438] E. Zeggini, M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott, H. Lango, N. J. Timpson, J. R. B. Perry, N. W. Rayner, R. M. Freathy, J. C. Barrett, B. Shields, A. P. Morris, S. Ellard, C. J. Groves, L. W. Harries, J. L. Marchini, K. R. Owen, B. Knight, L. R. Cardon, M. Walker, G. A. Hitman, A. D. Morris, A. S. F. Doney, Wellcome Trust Case Control Consortium (WTCCC), M. I. McCarthy, and A. T. Hattersley. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341, June 2007.

[439] E. Zeggini, L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, T. Hu, P. I. W. de Bakker, G. R. Abecasis, P. Almgren, G. Andersen, K. Ardlie, K. B. Boström, R. N. Bergman, L. L. Bonnycastle, K. Borch-Johnsen, N. P. Burtt, H. Chen, P. S. Chines, M. J. Daly, P. Deodhar, C. Ding, A. S. F. Doney, W. L. Duren, K. S. Elliott, M. R. Erdos, T. M. Frayling, R. M. Freathy, L. Gianniny, H. Grallert, N. Grarup, C. J. Groves, C. Guiducci, T. Hansen, C. Herder, G. A. Hitman, T. E. Hughes, B. Isomaa, A. U. Jackson, T. Jorgensen, A. Kong, K. Kubalanza, F. G. Kuruvilla, J. Kuusisto, C. Langenberg, H. Lango, T. Lauritzen, Y. Li, C. M. Lindgren, V. Lyssenko, A. F. Marvelle, C. Meisinger, K. Midthjell, K. L. Mohlke, M. A. Morken, A. D. Morris, N. Narisu, P. Nilsson, K. R. Owen, C. N. A. Palmer, F. Payne, J. R. B. Perry, E. Pettersen, C. Platou, I. Prokopenko, L. Qi, L. Qin, N. W. Rayner, M. Rees, J. J. Roix, A. Sandbaek, B. Shields, M. Sjögren, V. Steinthorsdottir, H. M. Stringham, A. J. Swift, G. Thorleifsson, U. Thorsteinsdottir, N. J. Timpson, T. Tuomi, J. Tuomilehto, M. Walker, R. M. Watanabe, M. N. Weedon, C. J. Willer, Wellcome Trust Case Control Consortium, T. Illig, K. Hveem, F. B. Hu, M. Laakso, K. Stefansson, O. Pedersen, N. J. Wareham, I. Barroso, A. T. Hattersley, F. S. Collins, L. Groop, M. I. McCarthy, M. Boehnke, and D. Altshuler. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5):638–645, May 2008.

[440] B. Zhang, L. Yang, L. Yu, B. Lin, Y. Hou, J. Wu, Q. Huang, Y. Han, L. Guo, Q. Ouyang, B. Zhang, L. Lu, and X. Zhang. Acetylcholinesterase is associated with apoptosis in beta cells and contributes to insulin-dependent diabetes mellitus pathogenesis. *Acta Biochim Biophys Sin (Shanghai)*, 44(3):207–216, March 2012.

[441] W. Zhang, H. Tong, Z. Zhang, S. Shao, D. Liu, S. Li, and Y. Yan. Transcription factor EGR1 promotes differentiation of bovine skeletal muscle satellite cells by regulating MyoG gene expression. *J Cell Physiol*, March 2017.

[442] X. Zhang, W. Mu, and W. Zhang. On the analysis of the Illumina 450k array data: probes ambiguously mapped to the human genome. *Front Genet*, 3:73, May 2012.

[443] Y. Zhang, H. Liu, J. Lv, X. Xiao, J. Zhu, X. Liu, J. Su, X. Li, Q. Wu, F. Wang, and Y. Cui. QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res*, 39(9):e58, May 2011.

[444] D. V. Zhernakova, P. Deelen, M. Vermaat, M. van Iterson, M. van Galen, W. Arindrarto, P. van 't Hof, H. Mei, F. van Dijk, H. Westra, M. J. Bonder, J. van Rooij, M. Verkerk, P. M. Jhamai, M. Moed, S. M. Kielbasa, J. Bot, I. Nooren, R. Pool, J. van Dongen, J. J. Hottenga, C. D. A. Stehouwer, C. J. H. van der Kallen, C. G. Schalkwijk, A. Zhernakova, Y. Li, E. F. Tigchelaar, N. de Klein, M. Beekman, J. Deelen, D. van Heemst, L. H. van den Berg, A. Hofman, A. G. Uitterlinden, M. M. J. van Greevenbroek, J. H. Veldink, D. I. Boomsma, C. M. van Duijn, C. Wijmenga, P. E. Slagboom, M. A. Swertz, A. Isaacs, J. B. J. van Meurs, R. Jansen, B. T. Heijmans, P. A. C. 't Hoen, and L. Franke. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet*, 49(1):139–145, January 2017.

[445] H. Zhu, G. Wang, and J. Qian. Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet*, 17(9):551–565, August 2016.

[446] M. J. Ziller, F. Müller, J. Liao, Y. Zhang, H. Gu, C. Bock, P. Boyle, C. B. Epstein, B. E. Bernstein, T. Lengauer, A. Gnirke, and A. Meissner. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet*, 7(12):e1002389, December 2011.

# Appendix A

# Chromatin state information

| Label | Description | Source |
|---|---|---|
| Islets | islet tissue | [287] |
| SkeletalMuscle | skeletal muscle tissue | [316] |
| Adipose | floated adipocyte nuclei | [316] |
| GM12878 | lymphoblastoid cell line (blood) | [94] |
| AnteriorCaudate | anterior caudate tissue (brain) | [316] |
| CD34-PB | hematopoietic stem cell | [316] |
| CingulateGyrus | cingulate gyrus tissue (brain) | [316] |
| ColonicMucosa | colonic mucosal tissue (colon) | [316] |
| DuodenumMucosa | duodenum mucosal tissue (small intestine) | [316] |
| ES-HUES6 | embryonic stem cell line | [316] |
| ES-HUES64 | embryonic stem cell line | [316] |
| H1 | embryonic stem cell line | [94] |
| hASC-t1 | human adipose stromal cell differentiation to adipocytes, time point 1 | [253] |
| hASC-t2 | human adipose stromal cell differentiation to adipocytes, time point 2 | [253] |
| hASC-t3 | human adipose stromal cell differentiation to adipocytes, time point 3 | [253] |
| hASC-t4 | human adipose stromal cell differentiation to adipocytes, time point 4 | [253] |
| HepG2 | hepatocellular carcinoma cell line (liver cancer) | [94] |
| HippocampusMiddle | hippocampus tissue (brain) | [316] |
| HMEC | mammary epithelial cells (breast) | [94] |
| HSMM | skeletal muscle myoblasts | [94] |
| Huvec | umbilical vein endothelial cells (blood vessel) | [94] |
| InferiorTemporalLobe | inferior temporal lobe tissue (brain) | [316] |
| K562 | leukemia cell line (blood) | [94] |
| Liver | liver tissue | [316] |
| MidFrontalLobe | mid frontal lobe tissue (brain) | [316] |
| NHEK | epidermal keratinocytes (skin) | [94] |
| NHLF | lung fibroblasts | [94] |
| RectalMucosa | rectal mucosal tissue | [316] |
| RectalSmoothMuscle | rectal smooth muscle tissue | [316] |
| StomachSmoothMuscle | stomach smooth muscle tissue | [316] |
| SubstantiaNigra | substantia nigra tissue (brain) | [316] |

**Table A.1** Information on chromatin states from Varshney et al. [395]. Description of cell/tissue types and reference to source.