Genomic variation in human health and disease

Davis James McCarthy
Balliol College
University of Oxford

A thesis submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

Trinity 2015

To my parents

Abstract

Understanding the structure and function of genomic variation within and between individuals will be crucial for the translation of genomics into improved health and clinical outcomes. This thesis addresses current issues around the study of genomic variation in that context.

Variant annotation is a vital step in the analysis of whole-genome and whole-exome sequence data. I compared variant annotations for 80 million variants from a clinically-focused whole-genome sequencing study, obtaining annotations with two different sets of transcripts and two different software tools. I found that choice of transcripts and choice of software both have a large effect on variant annotation. The extent of discrepancy in annotations has implications for all research that relies on variant annotation, especially as we try to use whole-genome sequencing in the clinic.

Type 2 diabetes (T2D) is a common, complex genetic disease imposing a large global health burden. Although over 80 genomic loci have been associated with increased risk for T2D, many questions remain about the genomic architecture of the disease. I used 11 million rare, low-frequency and common single-nucleotide variants obtained from whole-genome sequence data from 2,657 individuals with and without T2D to assess the contributions of different classes of genomic variation to T2D susceptibility. Using linear mixed model methods and variance partitioning approaches I characterised contributions from variants in different allele frequency classes. Partitioning variance into different functional classes revealed significant four-fold enrichment ($P < 0.01$) for variants in enhancer regions identified in pancreatic islet cells and significant depletion for variants without any functional annotation ($P < 0.01$).

Single-cell RNA-sequencing (scRNA-seq) technologies are rapidly gaining traction to interrogate transcriptomic heterogeneity across individual cells. However, raw scRNA-seq data require a large amount of processing to obtain a clean, tidy dataset ready for statistical modeling. I have developed a new, self-contained R software package, SCATER, to fill the niche between raw scRNA-seq data and downstream analysis. The package streamlines the pre-processing, quality control and data normalisation procedures while enabling flexible ways to visualise data and integration with other scRNA-seq analysis tools.

Keywords: *genomic variation, variant annotation, type 2 diabetes, heritability, linear mixed models, variance partitioning, enhancers, single-cell, RNA-seq, gene expression, quality control, software*

Acknowledgements

Thanks to my supervisor Professor Peter Donnelly who persuaded me to come to Oxford and whose generous support made possible all of the work presented here. I have learnt a great deal working with him. Many thanks to Professor Gerton Lunter and Dr Rory Bowden who examined me for Transfer of Status and Confirmation of Status. Their thoughtful feedback was much appreciated and helped to shape the direction of my research.

My DPhil studies were supported by a General Sir John Monash Scholarship. Many thanks to the General Sir John Monash Foundation, Australia, for providing me with the opportunity to study in Oxford. Thanks, in particular, to Dr Peter Binks, a fantastic mentor who provided consistently sound and thoughtful advice, including the encouragement to think seriously about studying at Oxford instead of in the United States. It has been a pleasure and a privilege to be part of the family of Monash Scholars.

Many thanks to Kate Distin-Harvey for her organisational skills and forbearance in the face of overloaded schedules and hectic travel arrangements. Without her, this thesis would have taken even longer to bring to completion.

My heartfelt thanks to Dr Loukas Moutsianas and Dr Quin Wills, with whom I worked in close collaboration over the course of my studies. Working with both was invigorating and enjoyable and they contributed a great deal to the work in this thesis.

Thanks to all of the members of the Donnelly Group, members of the WGS500 and GoT2D projects and my office mates on the link corridor at the Wellcome Trust Centre for Human Genetics. I have been lucky to work with excellent colleagues and I have benefited from many scientific and non-scientific conversations with them. In particular, thanks to my fellow Australian DPhil student Hilary Martin whose energy and drive enlivened the environment and provided an excellent example.

Thanks to the patients and their families for participating in the studies used in this thesis and for permitting their data to be used for scientific research.

Thanks to all of my friends in Oxford, many of whom I met in the Balliol College Middle Common Room (MCR). The MCR was an important part of my life in Oxford and it was an honour to lead the MCR as President for the second year of my DPhil. I met many wonderful people in Oxford, in the MCR and beyond, whose friendship I value highly and who made the experience of studying in Oxford such a good one. The positive effects of a word of encouragement, a random chat, a game of cricket or, occasionally, a vent cannot be underestimated.

It has been particularly special to share the experience of living and studying in the UK with some of my closest friends from Melbourne. Congratulations to Jimmy Hillis who came to Oxford on the same plane from Melbourne as I did and successfully completed his DPhil this year. Thanks also to my friends in Melbourne and around the world whose support from afar I appreciate deeply.

Many thanks to all who read and provided comments on drafts of this thesis: Julie Waters, John Waters, Gavan McCarthy, Elizabeth James, Quin Wills, Loukas Moutsianas, Florian Buettner and Verena Zuber. Their feedback immeasurably improved the manuscript. Any mistakes remaining are my own.

My deepest thanks and love to my family, without whom this could never have happened. Thanks to my parents-in-law John and Julie Waters, who provided extraordinary support and encouragement, far beyond what any son-in-law could reasonably expect. Thanks to my sister, Arlie; I'm very lucky to have such a wonderful sibling who shares so many scientific and other interests. Thanks to my parents, for everything. And finally thanks to Isabel Waters, my greatest source of support. Issy, you are the best. Here's to life after the DPhil!

Contents

List of Figures

List of Tables

Chapter 1

From Mendel to genomics

1.1 Introduction

When Gregor Mendel was breeding pea plants in an Austrian monastery in the mid-nineteenth century he could scarcely have imagined the impact that genetics would have on biology and medicine over the next 150 years. From abstract investigations into the nature of heritable traits grew a vast scientific enterprise, multi-billion dollar industries and clinical treatments based on an improved understanding of the genome. As we seek to exploit remarkable technological developments to improve disease outcomes and advance human health, we must further our understanding of genomic variation. This thesis addresses some current problems in understanding the relationships between genomic variation and human health and disease. Specifically, the thesis discusses issues surrounding: (1) the functional annotation of genomic variants; (2) the relative contributions of different classes of genomic variation to variability in susceptibility to type 2 diabetes; and (3) software for pre-processing, quality control and visualisation of single-cell RNA-sequencing data.

Around one hundred years ago, theoretical and experimental breakthroughs revolutionised physics and presaged a century of extraordinary discoveries. From theories of general relativity and quantum mechanics through harnessing nuclear fission to the development of the "standard model" and recent discovery of the Higgs boson (Chatrchyan et al., 2012), remarkable intellectual and technological progress has derived from breakthroughs in the first decades of the twentieth century. Physics enthusiasts might look back at that time with a mixture of admiration and envy: admiration for the individuals who developed the field, especially with the knowledge of where fundamental theories and particle acceleration would lead; envy for not having been present during one of science's most exciting epochs.

The fields of genetics and genomics currently find themselves in a similarly revolutionary era. Staggering advancements over the past two decades have fundamentally changed

the practice of genetics and biomedical research. High-throughput technologies have drastically reduced the cost of DNA sequencing, ushering in the "genomic era". The genomic revolution builds on the foundations of population, quantitative and disease genetics first laid by Mendel, combining genetics with molecular biology, statistics and computer science (Figure 1.1). The ability to assay all of an organism's genetic material (the "genome") at once instead of one gene at a time, achieved within the last twenty years, has expanded the horizons of genetic research. As population-scale cohorts are being characterised with genomic assays, genomic technologies are being married to molecular cell biology to link genomic and cellular variation at hitherto implausibly fine-grained resolution. Biomedical research today generates vast quantities of data, resulting in deep collaboration with statisticians, mathematicians and computer scientists to make sense of the data deluge. Bioinformatics, a term rarely heard (if at all) twenty years ago, is now a crucial aspect of any undertaking in genomics and biomedical research.

Whereas Mendel may not have suspected where research in genetics could lead, today we have grand ambitions for the field. Questions that five years ago would have been impossible, or impossibly expensive, can now be addressed. The hope for the twenty-first century is that the genomics revolution can revolutionise health care, with "precision medicine" delivering individually-tailored genomic solutions for drug treatment, clinical care and preventative measures to improve human health. Nevertheless, we are only at the beginning of the path that we hope will lead to the genomic transformation of disease treatment and understanding of healthy biology. The early years of genomics have demonstrated that human health and disease are incredibly complex. Understanding the extent of genomic variation within and between individuals, and how this variation relates to health and disease, is an imposing challenge, but also an exciting opportunity that underpins all of the work described in this thesis.

Genetics and DNA have infiltrated the contemporary zeitgeist. Mainstream novels, movies and television programmes regularly incorporate analysis of DNA in their plots, from identifying criminals to creating theme parks with live dinosaurs. The term "gene" is ubiquitous today. More importantly, the term is central to biomedical research and, increasingly, clinical medicine. With the rise of genetics in the public consciousness has risen the level of expectation about what benefits the "genomic revolution" can provide to society.

1.2 Historical overview: the path to genomics

In 1866, Gregor Mendel published a paper on plant hybridisation, revealing that certain characteristics differing in parents, such as height and flower colour, are not "blended"

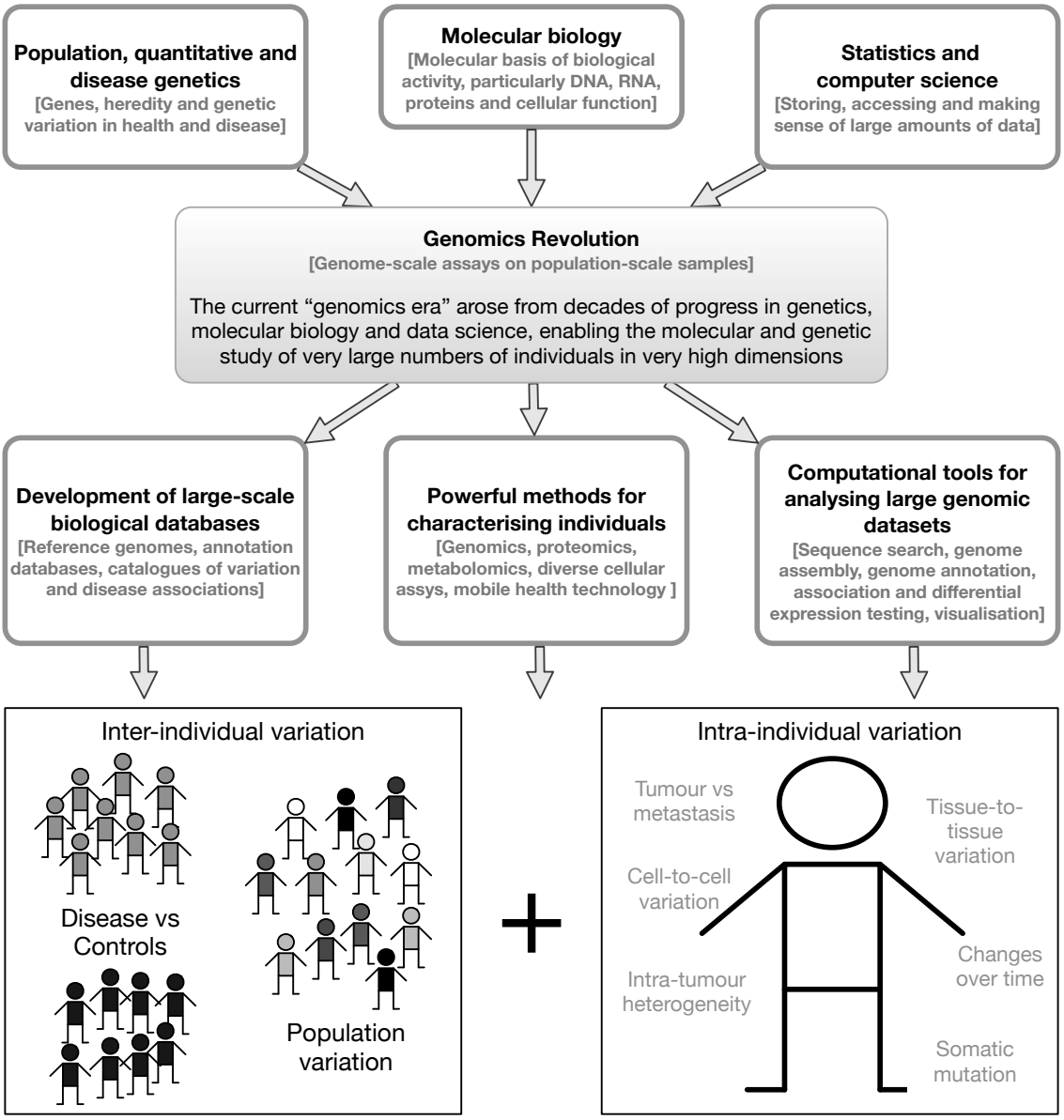


Figure 1.1: *From Mendel to genomics:* a brief and naïve schematic of the history of genomics. The diagram maps out the key inputs to and outcomes from the genomics revolution.

in the offspring, but are passed on as distinct, discrete entities. Mendel's paper is generally regarded as the moment that genetics was born. It was ignored for approximately thirty years. In 1900, Carl Correns, Erich von Tschermak-Seysenegg, and Hugo De Vries repeated and rediscovered Mendel's work (Tschermak, 1900; de Vries, 1900; Rheinberger, 1995). From this point the study of inheritance as the field of "genetics", a term first used by William Bateson in 1905 in a letter to Adam Sedgewick (Bateson, 2002), gathered steam. The term "gene" itself was first used by Wilhelm Johannsen in 1909 (Johannsen, 1909), deriving from the term "pangene", itself referring to "pangenesis", Charles Darwin's proposed mechanism of heredity (Heimans, 1962).

In the early years, Mendelian geneticists were frequently at odds with Darwinian evolutionists, but by 1932, R. A. Fisher, J. B. S. Haldane and Sewall Wright had combined Mendelian genetics and Darwin's theory of evolution, to develop a theory of genetic change that is still broadly accepted today (Mayr & Provine, 1998). This "evolutionary synthesis" resolved differences among geneticists, but many naturalists, concerned with speciation and the origin of organismal diversity, remained unaware of the advances in genetics. The divide between the population geneticists and naturalists was eventually bridged by Dobzhanksy in 1937 with the publication of his *Genetics and the Origin of Species* (Dobzhansky, 1937). Dobzhansky's newly synthesised, integrated evolutionary theory was widely accepted by 1947 (Mayr & Provine, 1998).

While Mendel was breeding his peas, researchers began to think about how heritable traits are transmitted. Ernst Haeckel proposed in 1866 that the cell nucleus contains the factors responsible for the transmission of heritable traits. Chromosomes in the nucleus were first described by Walther Flemming in 1879, who went on to examine their behaviour during cell division (Flemming, 1882). In 1902, Theodor Boveri and Walter Sutton postulated that the heredity units are located on chromosomes, as they observed that the segregation pattern of chromosomes during meiosis matched the segregation pattern of Mendel's genes (Sutton, 1903).

In the early twentieth century, Thomas Hunt Morgan's laboratory at Columbia University used fruit flies (*Drosophila*) as a model organism to study heredity and found the first mutant (*white*) with white eyes. In 1910–11, Morgan and his students showed that genes, strung like "beads on a string" on chromosomes, are the units of heredity (Morgan et al., 1915). They showed that chromosomes carry genes, discovered genetic linkage (the fact that genes are arranged on linear chromosomes) and described chromosome recombination (Morgan et al., 1915). In 1913 (during his PhD), Alfred Sturtevant, with Morgan, produced the first genetic linkage map, for the *Drosophila* fruit fly (Sturtevant, 1913b,a).

Researchers tried to establish the molecule of heredity. In 1923, Frederick Griffith postulated that a "transforming principle" permits properties from one type of bacteria (heat-inactivated virulent *Streptococcus pneumoniae*) to be transferred to another (live non-virulent

S. pneumoniae) (Griffith, 1923). In 1944, Oswald Avery, Colin MacLeod and Maclyn McCarty demonstrated that Griffith's "transforming principle" is not a protein, but in fact deoxyribonucleic acid (DNA), suggesting that DNA may function as the genetic material (Avery et al., 1944). DNA was confirmed as the genetic material in 1952, when Alfred Hershey and Martha Chase demonstrated that during infection with bacteriophage T2, viral DNA enters the bacterium, while the viral proteins do not, and that this DNA can be found in progeny virus particles (Hershey & Chase, 1952).

In 1869, Friedrich Miescher isolated a peculiar substance from the nucleus of white blood cells, which he called "nuclein" (Dahm, 2005). With the benefit of hindsight, it is clear that Miescher had obtained the first crude purification of DNA. The building blocks of DNA, the four bases adenine (A), cytosine (C), guanine (G) and thymine (T), were identified within 50 years. Establishing its structure took longer. Phoebus Levene wrongly predicted a constantly repeating four-base sequence, so he is not remembered as fondly as he might be (Levene, 1919). In 1943, however, William Astbury obtained the first X-ray diffraction pattern of DNA, which revealed that DNA must have a regular, periodic structure (Astbury, 1947). At the close of the decade, Erwin Chagaff found that the bases in DNA are always present in fixed ratios: the same number of As as Ts and the same number of Cs as Gs (Chargaff, 1950). In 1952, Rosalind Franklin and Maurice Wilkins used X-ray analyses to demonstrate that DNA has a regularly repeating helical structure, before James Watson and Francis Crick published the molecular structure of DNA: a double helix in which A always pairs with T, and C always with G (Watson et al., 1953). It was clear that this structure could copy and transmit genetic information. Connecting the abstract "gene" to a physical model was a major advance, and inextricably linked genetics with molecular biology.

Understanding of the relationship between DNA and proteins progressed quickly, with Francis Crick proposing the "central dogma" that information in DNA is translated into proteins through ribonucleic acid (RNA). He speculated that three-base sequences in DNA ("codons") always specify one amino acid in a protein (Crick, 1970). Sydney Brenner, Francois Jacob and Matthew Meselson discovered in 1961 that messenger RNA (mRNA) is the molecule that ferries information from DNA in the nucleus to the protein-making machinery in a cell's cytoplasm (Brenner et al., 1961). During the first half of the 1960s, Robert Holley, Har Khorana, Heinrich Matthaei, Marshall Nirenberg and colleagues cracked the "genetic code", describing how DNA sequences encode protein sequences (Nirenberg et al., 1963, 1965a). The genetic code is remarkably elegant, but discoveries like that of "introns", sequences in genes and mRNA that are not translated into protein sequences, began to reveal the deeper complexity of the genome (Berget et al., 1977; Berk & Sharp, 1977; Chow et al., 1977).

After the structure of DNA had been established, a major goal was to find ways to read the information it encodes—to sequence DNA. In 1955, Arthur Kornberg discovered DNA polymerase, an enzyme that replicates DNA. DNA polymerase has since been used for all kinds of recombinant DNA techniques and DNA sequencing (Kornberg, 1974). From the late 1960s, restriction enzymes were first used to cut DNA in specific places (Meselson & Yuan, 1968; Smith & Wilcox, 1970; Linn & Arber, 1968; Arber & Linn, 1969; Smith & Nathans, 1973; Danna & Nathans, 1971; Danna et al., 1973), which paved the way for Paul Berg to use restriction enzymes to create the first piece of recombinant DNA in 1972 (Jackson et al., 1972; Cohen et al., 1973). The following year, the first animal gene was cloned. Researchers fused a segment of DNA containing a gene from the African clawed frog *Xenopus* with DNA from the bacterium *E. coli* and placed the resulting DNA back into an *E. coli* cell, where the frog DNA was copied and a specific frog protein was produced from the gene it contained (Morrow et al., 1974). In 1977, Frederick Sanger, Allan Maxam and Walter Gilbert developed methods to sequence DNA (Sanger & Coulson, 1975; Sanger et al., 1977; Maxam & Gilbert, 1977), work that has profoundly shaped the trajectory of genetics and genomics to this day.

In the early 1980s, the first transgenic mice and fruit flies were produced (Gordon & Ruddle, 1981; Hogan & Williams, 1981; Costantini & Lacy, 1981; Rubin & Spradling, 1982). The GenBank databases were formed, receiving and sharing DNA sequence data (Benson et al., 2005), a precursor to the vast public databases of genomic data available today. Kary Mullis invented PCR (polymerase chain reaction) in 1983 as a method for amplifying DNA in vitro (Saiki et al., 1988), which has played a crucial role in sequencing ever since.

Marking the start of human disease genetics, Archibald Garrod observed the orderly inheritance of disease in 1902. He noted that the disease aklaptonuria is inherited according to Mendelian rules, and involves a rare recessive mutation (Garrod, 1902). The gene responsible for alkaptonuria was eventually mapped ninety years later (Pollak et al., 1993). A major disease-related breakthrough came in 1956, when sickle-cell anaemia was traced to a specific chemical alteration in a haemoglobin protein (Ingram, 1956). Chromosome abnormalities were first linked to disease in 1959, with publications on Down's syndrome (Lejeune et al., 1959), Turner's syndrome (Ford et al., 1959), and intersexuality (Jacobs & Strong, 1959). Two years later, Robert Guthrie developed the first screen for detecting a metabolic defect in newborns (Guthrie, 1961).

In 1982, Genentech released to market the first drug based on recombinant DNA, human insulin, which replaced animal insulin for treating diabetes. The following year, the first disease gene was mapped. Utilising DNA polymorphisms, researchers linked a genetic marker on chromosome 4 to Huntington disease, making Huntington disease the first disease to be genetically mapped (Gusella et al., 1983). In 1986 a gene for chronic granulomatous disease was the first human disease gene identified by positional cloning

(Royer-Pokora et al., 1986). The first comprehensive genetic map of human chromosomes was produced in 1987, based on 400 restriction fragment length polymorphisms (RFLPs), which are variations in DNA sequence that can be observed by digesting DNA with restriction enzymes (Donis-Keller et al., 1987). Genetic maps have been used for finding disease genes, utilising genetic markers such as microsatellites (Weber & May, 1989) and sequence-tagged sites (Olson et al., 1989). The identification of markers in genomic DNA was to prove crucial both for sequencing the human genome and finding associations between regions of the genome and diseases.

In 1990, the Human Genome Project (HGP) (Olson, 1993) was launched and the era of genomics began.

1.3 The genomic revolution

Dating back to the start of the HGP, the genomic revolution has transformed biomedical science. Remarkable developments in high-throughput technologies have continuously, and drastically, increased capabilities for genetics research and are gradually being translated into improvements in disease prevention and clinical care. The genomics era can be split into two roughly equal time periods: the time before the completion of the HGP in 2003 and the time since. In this section I will describe the path to completion of the HGP, and in the next section I will discuss the "contemporary" approaches in genomics—essentially methods developed after (and to a large extent, out of) the HGP—that are relevant to current research practices.

A major waypoint on the route to determining the sequence of all 3.2 billion bases of the human genome was to complete a physical map of the whole human genome with a marker every 100,000 base pairs. Steadily improving physical maps were produced first with microsatellites (Weissenbach et al., 1992; Murray et al., 1994) and later with sequence-tagged sites (Hudson et al., 1995), aided by the development of bacterial artificial chromosomes (BACs) for cloning longer human DNA sequences (Shizuya et al., 1992). To complement the physical map of the genome, researchers began to characterise genes using expressed-sequence tags (ESTs), stretches of DNA sequence made by copying a portion of an mRNA molecule. Thus, ESTs replicate sequences from genes. First proposed as a useful way to find genes in the genome in 1991 (Adams et al., 1991), a set of 260,000 ESTs was published in 1996 (Hillier et al., 1996), and two years later a human gene map was released that contained 30,000 human genes (Deloukas et al., 1998). At the time, this was thought to represent about one third of all human genes. In actual fact, there may be fewer than 20,000 human protein-coding genes if current estimates are correct (Ezkurdia et al., 2014).

Whole-genome sequencing had its first successes with much simpler organisms than humans. The very first complete sequence of the genome of a free-living organism was

published in 1995 when researchers sequenced the genome of the bacterium *Haemophilus influenzae* (Fleischmann et al., 1995), which causes respiratory and other infections, as well as flu. Sequencing its 1,803,137 base-pair genome (a little over 5% of the size of the human genome) aided development of efficient methods for sequencing. Sequencing of the smallest known genome, that of the bacterium *Mycoplasma genitalium*, was completed a few months later (Fraser et al., 1995). The complete genome sequence of the *Saccharomyces cerevisiae* (yeast used for brewing and baking) was obtained the following year (Mewes et al., 1997). This was the first eukaryotic organism to be completely sequenced and the first archaea genome was sequenced in 1996 (Bult et al., 1996), confirming the existence of a third branch of life on Earth. In 1998 the first complete genome sequence was obtained for a multicellular organism, the nematode worm *Caenorhabditis elegans* (C. elegans Sequencing Consortium, 1998), paving the way for full-scale human genome sequencing.

In May 1998, the private company Celera Genomics announced its plan to sequence the human genome within three years (Venter et al., 1998). Celera positioned itself as a direct competitor to the international HGP effort, with a novel "shotgun sequencing" method, in which the entire genome is fragmented and random segments are sequenced and then put in order. This strategy contrasted with the HGP's approach of building detailed maps before sequening defined regions. Driven by the intense competition between the public and private sequencing efforts, the HGP completed the first "finished", full-length sequence of a human chromosome (the relatively small chromosome 22) in December 1999 (Dunham et al., 1999).

The first draft of the whole human genome sequence was released, to great fanfare, in 2001. The HGP international consortium published its first draft and initial analysis of the human genome sequence (Lander et al., 2001) at the same time as J. Craig Venter and colleagues at Celera Genomics published another version of the human genome sequence (Venter et al., 2001). The draft sequence covered more than 90 percent of the human genome and was immediately and freely released into the public domain, a philosophy in the field that has underpinned much of the extraordinary growth in genomics since. In 2003, a "finished" version of the human genome was released, covering 99 percent of the genome with an accuracy of 99.99 percent, completing the HGP two years ahead of schedule and under budget (International Human Genome Sequencing Consortium, 2004). Immediately on the heels of the completion of the HGP, the Encyclopedia of DNA Elements (ENCODE) project began (ENCODE Project Consortium et al., 2012), one of many efforts around the globe involved in the ongoing endeavour to understand function in the genome.

A goal of the HGP had been to develop DNA sequencing technology and encourage commercial investment in genomics. What followed must surely be beyond the wildest hopes of members of the HGP. New technologies rapidly appeared that decreased the

cost of sequencing DNA by approximately four orders of magnitude in about seven years (National Human Genome Research Institute, 2015). In 2015, population-scale sequencing technologies have broken the almost-mystical barrier of the "$1,000 genome". It is now possible to obtain a high-quality, individual human genome sequence for less than the cost of many routine medical tests.

The staggering reduction in DNA sequencing costs along with rapid development of many different types of genomic assay have driven massive increases in data volume in biomedicine. Traditionally, biology and genetics had been "small data" enterprises, where data analysis could feasibly be done by researchers without particular statistical expertise and computing and data storage needs were negligible. Today, genomics is truly a "big data" field. As such, statistics and computing are vital cogs in the wheel of genomics. Research into specialised computational and statistical methods for genomics is a substantial field in its own right, and old divides between data analysis and biology are diminishing, accelerated by increasing the interdisciplinarity of individual researchers and collaborations.

The "open data" philosophy from the HGP, permeating both publicly and privately generated data, must not be underestimated as a force driving the genomic revolution. Incredible value for public and private research efforts has been gained from access to a public human reference genome, and reference genomes for the major model organisms, and many non-model organisms. A virtuous cycle has arisen, with a culture in the genomics field of depositing experimental data in public databases that make the data, often processed or otherwise organised to increase its utility, available for use by other researchers. As discussed below, the field now has a range of powerful tools to continue research into understanding the relationships between genomic variation and disease, and understanding function in the genome.

1.3.1 Studying genomic variation in health and disease

Genomic methods can be used for numerous distinct, but connected, areas in biomedical research:

- population genetics, quantitative genetics and evolution;
- disease genetics/genomics, discovering associations between genomic variation and disease states; and
- elucidating the path from genotype to phenotype, studying function in the genome.

These three broad areas are developing simultaneously, in concert, as advances in any one area assists progress in the other areas.

We are living in a "golden age of human population genetics" with unprecedented opportunities to reconstruct the entire genealogical and mutational history of humans (Przeworski, 2011). The affordability of DNA sequencing has made it possible to sequence the genomes of many individuals from distinct populations across the globe. Comparing sequences allows inferences to be made about population history and past demographic events (for example Schiffels & Durbin, 2014; Hellenthal et al., 2014; Haak et al., 2015; Bryc et al., 2015). Recent sequencing of ancient DNA has revealed the fascinating history of humans interbreeding with Neanderthals (Fu et al., 2015; Sankararaman et al., 2014; Prüfer et al., 2014).

Population genetics is interconnected with quantitative genetics, which focuses on the inheritance of traits, both continuous (like milk yield of dairy cattle) and binary (like pea-flower colour or disease status). A great deal of progress in quantitative genetics has come out of agricultural and livestock genetics, where, for example, selective breeding programmes have developed much of the theory and methods around the study of heritability of traits and trait prediction from genotypes (discussed in detail in Chapter 3).

Population and quantitative genetics, underpinned by evolutionary concepts, are vital foundations for disease genetics. Evolutionary and genetic forces have shaped every region of the genome, to a greater or lesser extent, so population genetics can help to answer questions like why disease mutations are present in human populations and what the sequence of demographic events was that led to the colonisation of the globe by modern humans (Przeworski, 2011). Understanding the evolutionary and genetic forces at work, therefore, greatly informs disease-focused studies, which can easily be confounded by population structure effects. Similarly, characterising natural variation in human populations is very important for building the tools for assaying genomic variation for disease studies and also for interpreting results of disease studies. For example, the first International HapMap project (International HapMap Consortium, 2005) catalogued single nucleotide polymorphisms in the population and quantified correlation between genetic variants, which enabled genome-wide association studies (GWAS) as a relatively small number of "tag" single nucleotide polymorphisms (SNPs) can provide most of the information on the pattern of genetic variation in a given region. The International HapMap 3 project integrated common and rare genetic variation in diverse human populations (International HapMap 3 Consortium et al., 2010), further expanding the range of GWAS in terms of the allele frequencies of variants that could be assayed and the variety of human populations that could be studied.

Large-scale whole-genome sequencing (for example the 1000 Genomes Project: 1000 Genomes Project Consortium et al., 2012) characterising rare variation (minor allele frequency (MAF) less than 0.5%) is increasing our ability to probe the effects of rare variation

on common disease genetics. Debate continues about the extent of the role that rare variation may play in common disease (Witte et al., 2014), as the genome-wide genotyping arrays that have been used for GWAS to this point are unable to assay rare variation. Thus, large cohort studies for common disease have been limited to exploring the effects of common (MAF greater than 5%) and low-frequency (MAF between 0.5% and 5%) variation. The Genetics of Type 2 Diabetes Project (discussed in Chapters 3 & 4), the first whole-genome sequencing study for complex, commmon disease, and similar projects such as the UK10K (`http://www.uk10k.org/`), along with large-scale cancer projects (see Weinstein et al., 2013, for example), are drastically increasing efforts to understand the role of rare variation in complex disease, which has proven so important for solving rare, Mendelian diseases.

The first genetic marker linked to disease was only discovered in 1983, when a marker on chromosome 4 was linked to Huntington disease using DNA plymorphisms (Gusella et al., 1983). Available technologies had limited the identification of causal, or markers for causal, variants. In the 1980s, however, family-based linkage and candidate gene assocation studies became mainstream techniques (see Lander & Schork, 1994, for an overview). Since then, discovery of causal genes for disease has followed three main waves (McCarthy, 2010):

1. Family-based linkage analyses;
2. Tests of association for candidate genes; and
3. Systematic large-scale surveys of association between common DNA variants and disease (following the advent of the GWAS).

We now appear to be cresting a fourth wave in which relatively inexpensive whole-genome and whole-exome sequencing enables association studies on the full catalogue of genetic variation (or of coding variation in the case of whole-exome sequencing), extending association studies to now include low-frequency, rare and structural variation.

Genetic linkage, first conceived by Alfred Surtevant and Thomas Hunt Morgan, has had a very important role in disease genetics. For Mendelian (typically rare) diseases, family-based linkage studies have been very successful for mapping disease genes (Jimenez-Sanchez et al., 2001). As technologies have steadily improved, the resolution for linkage studies has increased, along with abilities to fine-map disease-associated loci to identify causal variants. Literally thousands of genetic variants have been associated with Mendelian diseases using linkage methods, as recorded in the "Online Mendelian Inheritance in Man" database (McKusick-Nathans Institute of Genetic Medicine, 2015). Surprisingly, given the success of linkage studies, genomics has revolutionised human Mendelian genetics, making it much easier to "solve" the genetic cause of rare diseases (Gibbs, 2011; Boycott et al., 2013). Affordable whole-genome and whole-exome sequencing, particularly

using trio (sequencing both parents and an affected proband) or other family-based experimental designs, have proven to be powerful tools for Mendelian disease gene discovery and for diagnosing previously undiagnosed rare genetic disorders (Bamshad et al., 2011; Jacob et al., 2013; Taylor et al., 2015).

It is now clear that the characteristics of linkage studies mean that they are not well suited to finding genetic variants associated with common, complex diseases (Hirschhorn & Daly, 2005; McCarthy, 2010). The major issues affecting the efficacy of linkage studies are "incomplete penetrance" and "locus heterogeneity". For linkage analysis to succeed, markers that flank disease genes must segregate with the disease in families (that is, affected individuals must possess the disease alleles and unaffected individuals the non-disease alleles.) This brings us to the concept of "penetrance": the proportion of individuals with a specific genotype who manifest the genotype at the phenotype level. For example, if all individuals with a specific disease genotype show the disease phenotype, then the disease is said to be "completely penetrant". Thus, a disease is said to be "incompletely penetrant" if not all individuals with the disease genotype manifest the disease phenotype. Linkage analysis is much less powerful for detecting common alleles that have low penetrance, which (we now know) is typical for common, complex diseases such as type 2 diabetes. Locus heterogeneity, the situation in which a single disorder or trait is caused by variation in genes at different chromosomal loci, is also challenging for linkage analysis. When locus heterogeneity is present, clinical use of linkage analysis can be problematic because it is often not possible to determine the locus at which mutations are occurring in a given family. A further challenge for pedigree-based analyses lies in mapping *de novo* causal mutations, as they cannot be analysed with traditional linkage methods (Hu et al., 2014), a task with which association testing using whole-genome or whole-exome sequencing can assist.

An alternative approach to linkage studies, with many advantages over linkage analysis but also its own disadvantages, is to test for association between genetic variants and disease status in unrelated individuals with and without a given disease (Risch & Merikangas, 1996). These tests of association are intrinsically more powerful than linkage studies, but signals of association are limited to variants that are directly assayed, or tagged by (correlated with) variants assayed (McCarthy, 2010). Depending on the technology used, assayed and tagged variants could include just a single locus or a substantial fraction of the genome. Prior to the cheap chip-based genotyping that enabled GWAS, researchers were limited to testing association for specific candidate variants or genes of interest, based on prior knowledge from biological and pharmacological studies of protein function, animal models, monogenic or syndromic forms of the disease and positional information from linkage studies.

By the end of the HGP, candidate-gene association studies had identified some genes that contribute to susceptibility to common disease (Cardon & Bell, 2001; Tabor et al., 2002).

However, candidate-gene studies rely on having predicted the identity of the correct gene or genes, usually on the basis of biological hypotheses or the location of the candidate within a previously determined region of linkage. Even if these hypotheses are broad (for example, involving the testing of all genes in the glycolysis pathway), they will, at best, identify only a fraction of genetic risk factors, even for diseases in which the pathophysiology is relatively well understood. When the fundamental physiological defects of a disease are unknown, the candidate-gene approach will clearly be inadequate to fully explain the genetic basis of the disease (Hirschhorn & Daly, 2005). On the whole, candidate-based tests of association were not very successful for complex disease. They were hampered by lack of power (driven both by small sample size and lower than expected effect sizes) and confounding, or focused on inappropriate candidates (Hattersley & McCarthy, 2005). A meta-analysis revealed that much fewer than half of the reported candidate genetic associations were correct (Lohmueller et al., 2003). It is now clear that we were generally not very good at choosing candidate genes. A new approach was needed and genome-wide association studies arrived, subsequently uncovering many completely unexpected genes associated with disease.

In a review of association study designs in 2001, Cardon & Bell had predicted that the discovery of large numbers of genetic markers coupled with the development of better tools for genotyping would lead to the inevitable proliferation of GWAS. The approach for GWAS is the inverse of that in candidate gene studies. Instead of starting with a gene of interest, GWAS begin with a large sample of individuals with and without a disease of interest (or, alternatively, varying in a quantitative trait like height), and in a relatively unbiased fashion look across the entire genome to find regions associated with the trait (e.g. higher or lower risk of disease; increased or decreased height).

The events of the last fifteen years have profoundly vindicated that prediction (Figure 1.2). In 2007, GWAS came of age in a major way, profoundly changing the direction and focus of genetics research (WTCCC, 2007; Sladek et al., 2007; Diabetes Genetics Initiative, 2007; Scott et al., 2007; Steinthorsdottir et al., 2007; Zeggini et al., 2007). Since then, GWAS has become a standard approach for studying the genetics of complex diseases. With genotype data from SNP microarrays, GWAS have been successful in identifying over 9,000 genome-wide associations with over 1,300 traits (Burdett et al., 2015; Welter et al., 2014b). Now, instead of working on model organisms and trying, generally unsuccessfully, to leverage discoveries into insight on specific genetic associations with human disease, the focus has turned to uncovering the mechanisms underlying the robust genetic associations found directly in humans. In the vast majority of cases, the causal mechanisms for GWAS associations are currently not known, with the search for function taking us to the limits of our abilities to understand and interpret the genome.

Figure 1.2: *Diagram of the catalogue of loci robustly associated with diseases and traits using genome-wide association studies*. The diagram shows the human chromosomes 1–22 (the "autosomes") and chromosomes X and Y (the sex chromosomes) as banded rectangles. Overlaid on the chromosomes are coloured dots representing loci associated with diseases and traits through GWAS, with lines indicating roughly where they are located on the various chromosomes. The different colours represent different categories of disease or trait (though distinguishing between them is not a focus here), such as metabolic disease, immune system disease, cancer and so on. Almost all of these thousands of associations have been discovered since 2006. The catalogue is manually curated from the published literature by the National Human Genome Research Institute (NHGRI) and the European Bioinformatics Institute (EBI) with strict quality control. It contains several thousand associations for over one thousand diseases and traits (Welter et al., 2014a). This figure is reproduced with permission from the NHGRI-EBI GWAS catalogue website (`www.ebi.ac.uk/gwas/`).

The ten thousand-fold decrease in sequencing costs in the last decade now means that association studies covering (very nearly) the full catalogue of genomic variation are possible. Progress is being made on understanding the genomics of common disease, and genomic methods have been very successful for finding underlying genetic causes of rare, Mendelian diseases. Ending the expensive and emotionally fraught "diagnostic odyssey" for patients and their families is justification enough for widespread use of genomics in the clinic, but the next step is to turn diagnoses into successful treatments. In the current era, we can practise "genomic epidemiology", using the multitudinous genomic assays available to better understand human population health and disease. A major challenge, and aim, for the next phase of the genomics era is to improve our understanding of function in the genome so that we can make sense of disease associations and gain insight into how diseases can be better prevented and treated.

1.3.2 Understanding function in the genome

Understanding function in the genome is fundamental for basic science and disease genomics. The overarching goal of genetic association studies has been, and remains, the discovery of relationships between genomic and disease variation that can be used to understand disease aetiologies and potential treatments. Studying variation in DNA is a powerful approach for advancing genetics and disease research, but there is much more to genomic variation than variation in the sequence of genes and other genomic elements encoded in DNA. To make the most of discovered associations, the function of disease-associated genomic variation must be understood. There are now several genomic subfields that can be integrated with genomic DNA variation to better understand function in the genome. Variation in gene expression (transcriptomic variation), protein expression, the epigenome, and the metabolome can be assayed across time, tissues and biological conditions using high-throughput, genome-scale technologies (Table 1.1).

Following Crick's "central dogma" of molecular biology (Crick, 1970), we often study RNA levels as a proxy for gene and protein expression. This approach was proved successful over 15 years ago with the advent and wide adoption of gene expression microarrays, arguably the first genome-wide "genomic" assay. In the last 8 years or so we have seen microarrays largely superseded by RNA-sequencing (RNA-seq). RNA-seq takes advantage of our ability to reverse-transcribe RNA into complementary DNA (cDNA), which we can sequence on the usual high-throughput DNA sequencing platforms. As sequencing cost has decreased, RNA-seq has become cheaper and more accessible, and risen in popularity due to its reduced bias and increased flexibility in comparison to microarrays. RNA-seq has much higher dynamic range (can measure a wider range of expression values) and is not reliant on pre-defined "probes" in order to measure a gene's expression, and so is less dependent on prior knowledge of transcript isoforms and known genes. Assaying gene expression genome- or transcriptome-wide enables deeper understanding of functional effects of changes to DNA or the expression of other genes or regulatory mechanisms. With microarrays and then bulk-tissue RNA-sequencing it became cost-effective to study the whole transcriptome in hundreds of samples.

To understand the regulation of gene expression, it is necessary to study the epigenetics, the layer of modifiable chemical markers added to DNA and associated molecules such as histones that act as signals for genomic regulation. Thus, epigenomics covers DNA methylation, histone modifications, transcription factor binding, chromatin acessibility, and more. Unlike DNA, which under normal circumstances is identical in all cells in an organism, epigenetic marks can vary greatly across time (e.g. development), tissue and environment (Zhu et al., 2013, among many others). Thus, there is great interest in

conducting epigenomic assays across many individuals, tissues, timepoints and other conditions, and large consortia projects like ENCODE and the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al., 2015) are building database resources with rich epigenomic resources available for use by other researchers.

Quantitative RNA expression assays, such as those described above, are typically used as a proxy for protein expression. In many settings, it is of greater interest to know the quantities of different types of proteins under different conditions or times, than of the mRNA per se. However, proteomic assays have traditionally been noisy and lacked coverage across the whole proteome. Nevertheless, recent developments with mass spectrometry are enabling high-throughput, multi-dimensional proteomic assays. Current technologies enable the measurement of multiple properties for thousands of proteins (see Larance & Lamond, 2015, for an overview of the field). Supported by complementary data analysis, integration and visualisation tools, mass spectrometry-based proteomics can assay protein abundance, isoform expression, turnover rate, sub-cellular localisation, post-translational modification and interactions, adding valuable information about cell biology to layer onto other types of "omics" data.

Beyond transcriptomics, epigenomics and proteomics there is a large set of other omics approaches with varying degrees of maturity and utility, from established sub-fields like metabolomics (studying small molecules involved in the metabolism; see Zamboni et al., 2015) to currently less widely adopted approaches. As further technologies develop we can expect to see more omics approaches to add to the genomics toolbox for biomedical research.

The many different omics technologies for exploring genomic function are feeding back into directly understanding the function of DNA-sequence variation. Large consortium projects, for example ENCODE (ENCODE Project Consortium et al., 2012), FANTOM (Ravasi et al., 2010; Andersson et al., 2014; Arner et al., 2015), GTEx (GTEx Consortium, 2013) and Geuvadis (Lappalainen et al., 2013), have generated very large datasets of functional information that are now added to bioinformatic databases. Adding to genome annotations in this way provides ever-improving possibilities for interpreting discovered genetic associations in terms of likely function. Despite becoming something of a buzzword, however, genomic data integration remains challenging in practice in spite of the large potential it has to add value to present and future research.

1.3.3 Single-cell genomics: from inter-individual variation to intra-individual variation:

Bulk-tissue analyses have revealed much about the effects of functional genomic variation (for example Lappalainen et al., 2013), and how genomic DNA, epigenomic marks, gene expression and protein expression vary between individuals. Major questions, however,

Field	Assay
Traditional genomics	
Genomic DNA	SNP-chips
	whole-exome sequencing
	whole-genome sequencing
Transcriptomics	
gene/transcript/exon expression	expression microarrays
	RNA-sequencing (RNA-seq)
Epigenomics	
methylation	DNA methylation arrays
	bisulfite chip and sequencing (BS-seq)
	methyl-CpG binding domain (MBD) protein-enriched genome sequencing (MBD-seq)
	methylated DNA immunoprecipitation (MeDIP) chip and sequencing (MeDIP-seq)
histone modifications, transcription factor binding sites, open chromatin/nucleosomes	chromatin immuno-precipitation sequencing (ChIP-seq)
	DNAse 1 hypersensitivity with sequencing (DNAse-seq)
	assay for transposase-accessible chromatin with sequencing (ATAC-seq)
	micrococcal nuclease (MNase) sequencing (MNase-seq)
Proteomics	
multiple properties of proteins	mass spectrometry

Table 1.1: Major sub-fields of genomics and some of the major assays used to investigate the genome and its function.

remain about the role of genomic variation within individuals. Heterogeneity of genomic variation (broadly conceived) and its effects between individual cells within an individual could be important for developing better understanding of complex traits and diseases, and causal variation with heterogeneous effects across different people. With recent developments in cell handling and genomics technologies, researchers can now apply many of the most common genomics approaches at the level of individual cells. Genomics methods are now available at unprecedented resolution to study single-cell dynamics and genomic heterogeneity across biological systems, treatment conditions or throughout processes like cell differentiation.

Single-cell genomics (encompassing DNA-sequencing, RNA-sequencing, epigenomic and proteomic assays at the individual-cell level) offers new perspectives on cell biology and will revolutionise whole-organism science (Shapiro et al., 2013). They have already illuminated dynamic localised hormonal control of cellular variation (Shalek et al., 2014)

and have been used to study the dynamics of genomic clones in breast cancer (Eirew et al., 2015). Eventually, single-cell RNA, DNA and bisulfite sequencing, sequencing of accessible chromatin, mass spectrometry and proteomic technologies will help us understand the extent, basis and function of intercellular variation in genomic DNA, gene expression and regulation and protein expression. However, these are early days for single-cell genomics technologies, and as a field we are still coming to terms with the inherent biases and failure modes of the experimental protocols and the data produced. Much further work is required, but in the near future we should be able to explore the relationships between inter-individual genomic variation and intra-individual genomic variation and thus come to a more detailed and comprehensive understanding of genomic function to continue the extraordinary arc of progress in biomedicine driven by the genomic revolution.

1.4 Looking ahead

As discussed above, the major outcomes of the genomics revolution include:

- the development of large-scale biological databases (such as the human genome sequence);
- powerful methods for characterising patients (such as proteomics, transcriptomics, genomics, diverse cellular assays, and even mobile health technology);
- statistical and computational tools for analyzing large sets of data.

The extraordinary outcomes of the genomics revolution are allowing us to study genomic (in the broadest sense) variation both between and within individuals. The field now has many genomic assays available for studying variation within and between individuals.

A potential opportunity for genomics research in the twenty-first century is to use genomic technology and discoveries to deliver precise, individually-tailored approaches to improve disease prevention and treatment. One possible direction for this is so-called "precision medicine", which like many promising avenues in science has been hyped too much too soon. At risk of becoming just another buzzword, precision medicine is nevertheless a strategic priority for the United States government and many major biomedical research institutions and funding bodies (Collins & Varmus, 2015). Truly population-scale sequencing is now on the doorstep, with endeavours like the 100,000 Genomes Project (`http://www.genomicsengland.co.uk/`) in the UK's National Health Service aiming to bring genomics into the clinic and linked to electronic health records in the context of real-world healthcare systems. Currently, the emphasis for precision medicine is on rare diseases and cancer, but complex diseases will surely follow soon.

Cystic fibrosis provides an example of the hopes for precision medicine. The story of the cystic fibrosis gene, *CFTR*, traces a fascinating arc (Tsui & Dorfman, 2013). Cystic fibrosis was identified as a recessive disorder in the 1950s and linkage analysis and positional cloning were used to locate the gene on chromosome 7 in the 1980s. Confirmation of numerous disease-causing mutations was achieved in the 1990s and recently the cystic fibrosis treatment "kalydeco", which specifically targets the effects of certain *CFTR* mutations, has been made available (Vertex Pharmaceuticals Incorporated, 2015). Instead of facing an inevitable early death, patients with one of nine particular *CFTR* mutations may be able to treat the symptoms of cystic fibrosis with daily tablets and, it is hoped, live longer, fuller lives.

Aside from some very promising case studies in precision medicine, all of the research in genomics to this point confirms that the human genome is incredibly complex and entangled. Delivering precision medicine to all people is thus a huge challenge. Understanding the genome and its function, through better characterisation of genomic variation and its many roles in human health and disease, will be vital to the success of precision medicine. This thesis addresses some small, specific questions in the context of this large, international endeavour.

1.5 Outline

In this thesis I discuss three distinct facets of studying the structure and function of genomic variation.

Chapter 2 discusses the problem of variant annotation, a vital step in the analysis of whole-genome and whole-exome sequence data. I compare variant annotations for 80 million variants from a clinically-focused whole-genome sequencing study, obtaining annotations with two different sets of transcripts and two different software tools. I found that choice of transcripts and choice of software both have a large effect on variant annotation. The extent of discrepancy in annotations has implications for all research that relies on variant annotation, especially as we try to use whole-genome sequencing in the clinic. The work described in Chapter 2 was published in McCarthy et al. (2014). In addition, the analysis in that chapter guided variant annotation approaches used for the WGS500 Project (Taylor et al., 2015).

In **Chapters 3 & 4**, I discuss the use of whole-genome sequence data to investigate the genomic architecture of type 2 diabetes. Chapter 3 presents the estimation of heritability for susceptibility to type 2 diabetes using whole-genome sequence data. Chapter 4 describes partitioning variance in liability for type 2 diabetes to assess the relative contributions of different classes of genomic variation to variability in susceptibility to type 2 diabetes. I use linear mixed model methods to analyse genomic variation in 2,657 individuals with

and without type 2 diabetes. This dataset from the Genetics of Type 2 Diabetes (GoT2D) project was, at the time of writing, one of the largest samples for whole-genome sequencing for a complex disease. It provides a novel opportunity to apply methods developed for chip genotype data to a very dense set of variants from whole-genome sequence data. I estimate the narrow-sense heritability of type 2 diabetes using these data and characterise the relative contributions to the heritability of type 2 diabetes from variants of different allele frequency classes. I also analyse the contributions to susceptibility to type 2 diabetes from variants in different functional classes, finding enrichment for variants in enhancer regions in pancreatic islet cells. Throughout, I assess and comment on the performance of the models used. The GoT2D project has recently submitted two papers, one focusing on the analysis of whole-exome data (Teslovich et al., 2015) and one focusing on whole-genome data (Flannick et al., 2015). Some of the work described in Chapters 3 & 4 appears in the GoT2D Genomes paper (Flannick et al., 2015).

Chapter 5 presents software and methods for the pre-processing, quality control and normalisation of single-cell RNA-seq data. Single-cell RNA-seq is rapidly gaining traction as a tool for investigating transcriptomic profiles and variation in individual cells. As a relatively new technology, however, there is much yet to be understood about experimental protocols, biases, failure modes, and how analysis could be affected by artifacts. Many statistical methods for analysis have already been proposed, but all assume a clean, tidy dataset ready to analyse. The reality is that raw single-cell RNA-seq data requires a large amount of processing to prepare it for analysis. I have developed an R software package to fill the niche between raw single-cell RNA-seq data and downstream analysis methods, focusing on streamlining the pre-processing and quality control procedure while enabling flexible ways to visualise the data.

Across the three distinct projects runs the theme of understanding the function of genomic variation as it relates to human health and, in particular, complex human disease. This thesis is my own work, but it was conducted in collaboration with several other individuals as described in the individual chapters.

Chapter 2

Choice of transcripts and software has a large effect on variant annotation

2.1 Background and introduction

The advent of accessible and relatively inexpensive high-throughput sequencing technology has resulted in extensive sequencing of whole human genomes or exomes in a research setting and seems likely to lead to an explosion of genomic sequencing in a clinical context. While there remain challenges in unambiguously determining an individual's genome or exome sequence (Green et al., 2011; Schrijver et al., 2012), the focus here is on the downstream interpretation of that sequence. Let us take as a starting point a specified list of positions, assumed to be correct, at which the nucleotides in the individual's sequence differ from the human reference sequence. I will restrict the scope here to single nucleotide variants and short indels. A crucial step in linking sequence variants with changes in phenotype is variant annotation. This chapter describes a comparison of variant annotation methods. The material presented here expands on results in McCarthy et al. (2014). I carried out this analysis and wrote the paper, but benefited greatly from conversations with Peter Humburg, Alexander Kanapin, Kyle Gaulton, Manny Rivas and Jean-Baptiste Cazier.

Variant annotation is the process of assigning functional information to DNA variants. There are many different types of information that could be associated with variants, from measures of sequence conservation (Cooper et al., 2005) to predictions about the effect of a variant on protein structure and function (Kumar et al., 2009; Adzhubei et al., 2010; Schwarz et al., 2010). Here I focus on the most fundamental level of variant annotation, which is categorising each variant based on its relationship to coding sequences in the genome and how it may change the coding sequence and affect the gene product.

The coding sequences of the genome are, broadly speaking, the genes: "gene" has come to refer principally to a genomic region producing (through transcription) polyadenylated mRNAs that encode a protein (Gingeras, 2007). I refer to these polyadenylated mRNAs as "transcripts", although the term transcript can refer to any RNAs produced from the

transcription of genomic DNA sequence. Thus, there are non-coding transcripts that do not encode a protein, but nevertheless can have a function, for example in regulation. When considering transcripts in the context of genomic DNA sequence, a transcript is defined by its exons, introns and untranslated regions (UTRs), and their locations (Figure 2.1). Many separate transcripts may overlap any given position in the genome, and it is not uncommon for genes to have many different transcripts (or "isoforms"), of which they can express many simultaneously (Djebali et al., 2012).

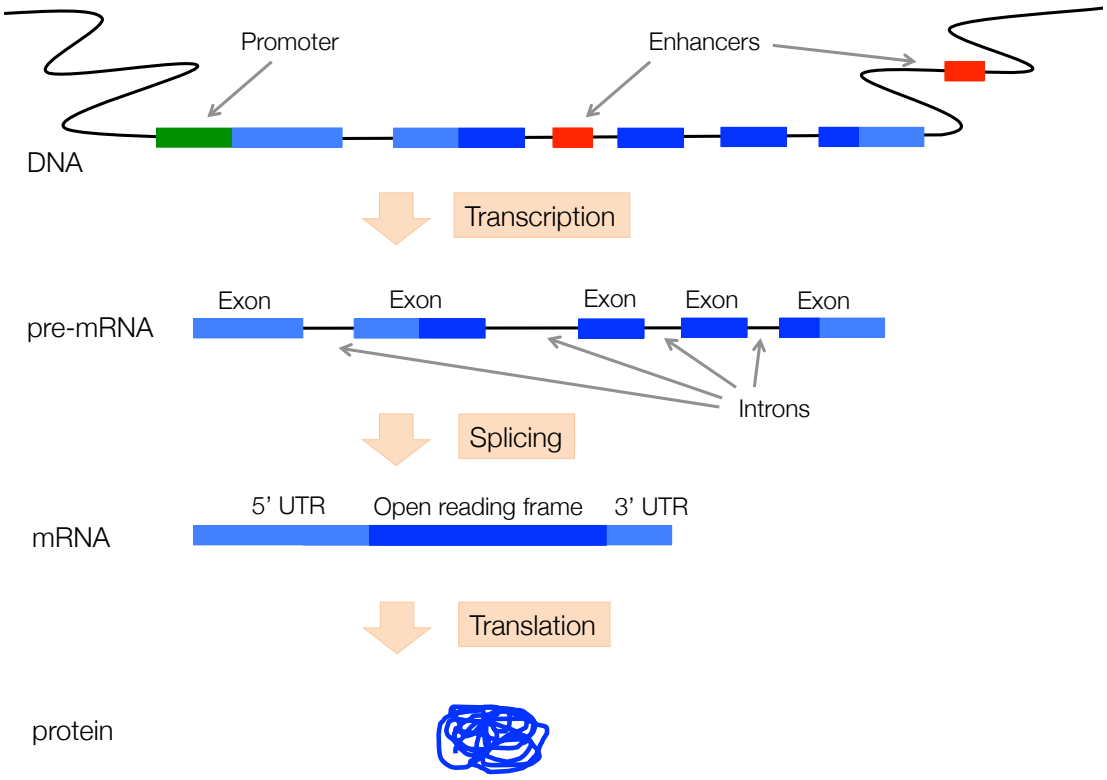


Figure 2.1: **Transcript structure.** A cartoon showing a simplified schematic of the structure of a transcript in genomic DNA and how it is modified through transcription and splicing before being translated into a protein. A single gene typically has multiple transcript isoforms that may yield modified proteins.

Our understanding of the protein-coding sequences in the genome is summarised in the set of transcripts believed to exist. In addition to protein-coding sequences, there are also transcribed regions of the genome that are relevant for gene expression regulation (non-coding transcripts). Thus, variant annotation depends on the set of transcripts used as the basis for annotation. The widely-used annotation databases and browsers—ENSEMBL (Flicek et al., 2012), REFSEQ (Pruitt et al., 2012) and UCSC (Fujita et al., 2011)—contain sets of transcripts that can be used for variant annotation, as well as a wealth of information of many other kinds as well, such as ENCODE (ENCODE Project Consortium, 2012) data about the function of non-coding regions of the genome. Current transcript sets

for variant annotation usually consist solely of protein-coding and non-coding transcripts, but in the future we may see the definition of a transcript set expanded to enable annotation of regulatory variants. As we improve our understanding of enhancer, promoter and transcription factor binding regions of the genome that affect the regulation of the expression of transcripts, we should see such annotations become part of the standard set of annotation terms. For the time being, however, the focus remains firmly on transcribed regions.

To annotate DNA variants we therefore require a set of transcripts that summarises our understanding of the genome. In addition, we need a software tool to determine the likely effect of each variant based on the transcripts (or other genomic features) that overlap the variant's position. One or more possible annotations for the variant can then be reported. An individual's genome will typically differ from the reference genome at over three million positions (Levy et al., 2007). Annotation pipelines for whole-genome and whole-exome sequencing studies thus need to process many millions of variants efficiently. Both annotation algorithms and particular software implementations are designed with this need for speed in mind, so software tools use shortcuts and heuristics where possible to enable rapid annotation of variants. Differences in how tools apply heuristics and algorithms—as well as deeper conceptual differences in how they approach variant annotation—result in differences in annotations for the same variants. Substantial downstream processing of variant annotation output is usually required to prioritise variants for follow-up investigation.

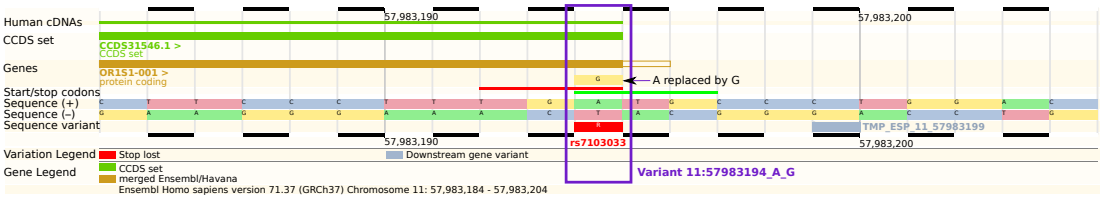
Let us start with a straightforward example of variant annotation, looking at the variant NC_000011.9:g.57983194A>G. This variant is here written in Human Genome Variation Society (HGVS) nomenclature, a commonly-used syntax for representing sequence variation in scientific literature, clinical reports and databases of variation. HGVS variant descriptions are always made relative to a reference sequence: either genomic DNA, coding DNA, mitochondrial, non-coding RNA, RNA or protein sequence. For this variant, the reference sequence is NC_000011.9, which refers to homo sapiens chromosome 11, version 9 from version 37 of the human reference genome (GRCh37). The "g" indicates that the reference sequence is genomic DNA (we use "c" for coding DNA, "m" for mitochondrial, "n" for non-coding RNA, "r" for RNA and "p" for protein sequence). The number 57983194 provides the position of the variant on this reference sequence (on chromosome 11), and A>G indicates that there is a substitution at this position, with the reference allele A replaced by G. HGVS nomenclature provides a compact, human-readable way to represent and communicate genomic variation.

Although the HGVS recommendations (Taschner & den Dunnen, 2011) are widely endorsed by journals and professional societies, including the HGVS itself, using HGVS nomenclature can be challenging. HGVS nomenclature was developed before the advent

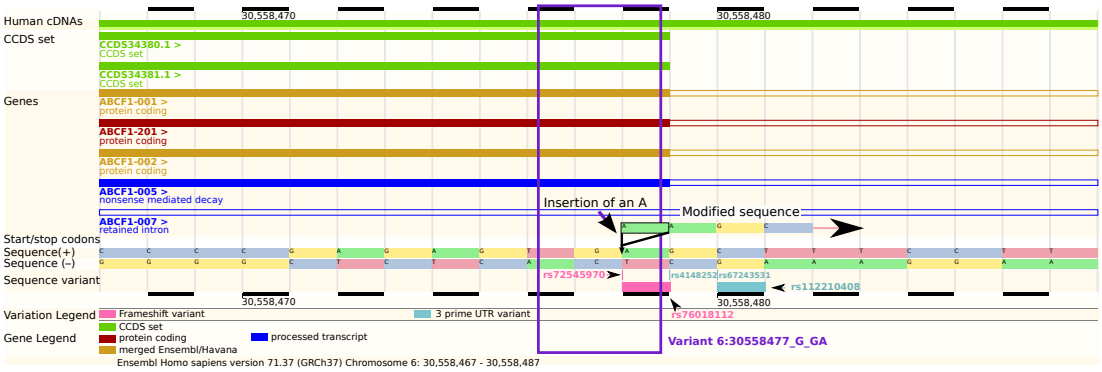
of high-throughput sequencing, and guidelines for the nomenclature continue to evolve. This makes the nomenclature difficult to understand and to use for experts and non-experts alike, even though tools have been developed to represent variants correctly with HGVS nomenclature (Hart et al., 2015). Given these difficulties, variants are often represented more simply in terms of genomic coordinates. When representing a variant in terms of genomic coordinates, a common format is Chromosome:Position_ReferenceAllele_ Alternate Allele. Thus, one would write NC_000011.9:g.57983194A>G simply as 11:57983194_A_G. Genomic coordinates refer to a specific "build" (version) of the reference genome, which is not included in the variant description (as is the case for HGVS nomenclature). Therefore, the genome build needs to be made clear when expressing variants in terms of genomic coordinates. In this chapter, I use genomic coordinates as well as HGVS notation to represent variants. In all cases I use the GRCh37 build of the human reference genome (Genome Reference Consortium, 2015).

Returning to the task of annotating the variant NC_000011.9:g.57983194A>G, we find that only two transcripts in the ENSEMBL transcript set, a Consensus Coding Sequence (CCDS) (Pruitt et al., 2009a) transcript and a merged ENSEMBL/HAVANA (GENCODE) transcript (Searle et al., 2010; Harrow et al., 2012), overlap the variant. In this case, the annotation of the transcript is the same regardless of which transcript is used (Figure 2.2a). This variant is unambiguously a stop-loss variant, as the final codon is changed from TGA (stop codon) to TGG (tryptophan) (Nirenberg et al., 1965b; Flicek et al., 2012). Both of the software tools that I use for the present study correctly annotate this variant as "stop-loss". With the stop-codon "lost", the cellular machinery will mistakenly continue to translate the RNA sequence which will disrupt the amino acid sequence of the resulting protein, meaning that the mRNA could be subject to nonstop-mediated decay (Frischmeyer et al., 2002; van Hoof et al., 2002; Maquat, 2002), or otherwise disrupt the protein's function.

Frequently, however, variant annotation is more complex. Typical pipelines are not well-suited to handling a variant that could have one consequence for one transcript and a different consequence for a different transcript. Even where a gene is relatively well defined and does not overlap other genes, there may be many transcripts (isoforms of the gene) to choose from, often supported by varying levels of evidence for their existence and structure. It is common for a gene to have multiple transcripts overlapping a given position in the genome, so given a set of transcripts a software tool has to choose which one to use. If it provides an annotation for the variant for each transcript, then the question becomes what annotation to report. If the software reports all possible annotations from all possible transcripts then the user must decide how to prioritise different, competing annotations, or how to integrate them into downstream analysis. This issue is further exacerbated in the uncommon case of a single variant affecting multiple genes (each of



(a) Straightforward annotation example



(b) More complex annotation example

Figure 2.2: **Annotation Examples.** These screenshots from the ENSEMBL web browser (Flicek et al., 2013) shows a straightforward example (top) and a more complex example of annotation (bottom). A browser image consists of several "tracks", each of which provides certain information about the DNA sequence at each base position. Two tracks, labeled "Sequence (+)" and "Sequence (-)", show the DNA sequence on the forward and reverse strands, respectively. Above these, there is a track that shows start and stop codons, and above that, there are several tracks indicating the presence and structure of different transcripts (labeled on the left as "Genes" and "CCDS set"; transcripts are read, in both examples here, from left to right). The "hollowed-out" parts of transcripts indicate that the sequence is non-coding at those positions. Below the DNA sequence, a track labeled "Sequence variant" shows known sequence variants from dbSNP, the NCBI database of genetic variation (Sherry et al., 2001) and the 1000 Genomes Project (1000G) (The 1000 Genomes Project Consortium, 2010). Appropriately, the "Variation Legend" and "Gene Legend" provide more information about the features referred to by different colours in the browser. The variant NC_000011.9:g.57983194A>G (rs7103033) (Figure 2.2a) is the final base of the final exon in both transcripts available at this position (here there is a CCDS transcript (green), and a "merged" ENSEMBL/HAVANA (GENCODE) transcript (gold)). The final codon is changed from TGA (stop codon) to TGG (tryptophan). Thus the variant is unambiguously a stop-loss variant. Indeed, using the ENSEMBL transcript set, both ANNOVAR and VEP correctly annotate this variant as "stop-loss". The variant NC_000006.11:g.30558477_30558478insA (rs72545970) (Figure 2.2b) is an example of a variant that is more complex to annotate. The variant is the penultimate base of the exon (so in the stop codon) for all but one of the transcripts shown. On one hand, the variant is a single-base insertion, so could be annotated as a frameshift variant. On the other hand, the variant is an insertion in a stop codon, so we might expect this to be a stop-loss variant. In fact, the final codon, TGA (stop codon), remains TGA with this variant (the insertion of a single base A), so it is actually a synonymous variant. One of the software tools used for this study (ANNOVAR) reports a "frameshift insertion" annotation and the other (VEP) a "stop-loss" annotation for this variant, when using ENSEMBL transcripts.

which likely has multiple transcripts). Current annotation tools vary in approaches to reporting consequences of a variant in multiple genes at once.

Choice of the underlying set of transcripts used for annotation can give the user more control over transcript use. Transcript sets from different sources can have different char-

acteristics. For example, both ENSEMBL and REFSEQ contain transcripts established from experimental evidence utilising automated annotation pipelines and manual curation, but their precise requirements for inclusion of transcripts differ. The result is that the ENSEMBL transcript set is larger than the REFSEQ set (see Section 2.2.2), but the REFSEQ transcript set is not simply a subset of the ENSEMBL transcript set.

Both ENSEMBL and REFSEQ contain transcripts established from experimental evidence. REFSEQ transcripts are constructed from sequence data submitted to the International Nucleotide Sequence Database Collaboration (INSDC) (Pruitt et al., 2002 [Updated 2012 Apr 6]). Similarly, all ENSEMBL transcripts are based on experimental evidence, namely mRNAs and protein sequences deposited in public databases (Ensembl, 2013b). Both transcript sets contain transcripts produced by automated pipelines run on the database sequences and manually-curated transcripts.

A portion of the REFSEQ dataset is manually curated by National Center for Biotechnology Information (NCBI) staff, with the remainder produced from the automated annotation pipeline. External information is incorporated entirely into REFSEQ's organizing framework. In contrast, ENSEMBL more explicitly imports information from external sources, such as the Consensus Coding Sequence (CCDS) set and HAVANA gene models. The ENSEMBL and HAVANA transcript models are compared where manual curation is available, and are merged when they agree on the same coding sequence (Ensembl, 2013b). This combined, merged geneset is the default gene set from the GENCODE project.

Although their approaches to building transcript sets are broadly similar, the resulting transcript sets for REFSEQ and ENSEMBL are substantially different. Unfortunately, detailed information about the automated pipelines and manual curation processes is not generally available. The quality assessment processes and inclusion requirements for both transcript sets are necessarily extensive and complicated. Furthermore, the transcript sets change regularly according to planned release schedules. As a result, detailed information on the breakdown of levels of support (evidence) for the human transcripts included in each set (e.g. proportion of transcripts from an automated pipeline versus the proportion that are manually curated) is not available, preventing us from a thorough characterisation of why certain transcripts may appear in the REFSEQ set and not the ENSEMBL set and vice versa. However, the REFSEQ frequently asked questions website (Pruitt & Brown, 2013) notes that, due to curation decisions made for REFSEQ transcript curation, alternately spliced gene products are underrepresented in the REFSEQ collection. The difference in the treatment of alternately spliced gene products is likely to contribute to the substantially greater number of transcripts in the ENSEMBL set than in the REFSEQ set. More information on the REFSEQ and ENSEMBL datasets is available from the NCBI Handbook (Pruitt et al., 2002 [Updated 2012 Apr 6]), and the REFSEQ and ENSEMBL websites (National Center for Biotechnology Information, 2013; Ensembl, 2013a).

Related to the issue of a variant frequently overlapping multiple transcripts and sometimes even multiple genes is the fact that any given variant can have several plausible annotations, even when considering just a single transcript as the basis for annotation. Choosing the "best" annotation is frequently not clear-cut, as in the case of the variant NC_000006.11:g.30558477_30558478insA, a single-base insertion at the end of an exon (Figure 2.2b). This variant could be annotated as a frame-shift insertion in a coding sequence (which it is), or as a stop-loss variant (as it falls in a stop codon). In fact, the correct annotation is that this is a synonymous variant. In many cases we would be misled into thinking that the variant is a frameshift or stop-loss variant, and therefore be likely to assume it to be of functional effect and include it in any list of variants of interest for further investigation. Indeed, one of the software tools used for this study reports a "frameshift insertion" annotation and the other a "stop-loss" annotation for this variant, when using ENSEMBL transcripts. In this example there seems to be a single "best" annotation, but many cases are more ambiguous, with several equally valid possible annotations. The software tool must make some sort of choice in such cases as to which annotation to report for the variant (and transcript used). Often, the software tool will apply a prioritisation rule, which typically prefers annotation terms perceived to have more severe consequence.

There are many other annotation tools available (for example, Mutalyzer 2 (Wildeman et al., 2008), VAT (Habegger et al., 2012), VAAST 2.0 (Hu et al., 2013), GATK VariantAnnotator (McKenna et al., 2010) and SnpEff (Cingolani et al., 2012)). Different tools will have better or worse performance for certain variants, but here I want to make the more general point, using two very widely used annotation tools, that there is a large degree of discrepancy between any two annotation tools, and researchers need to be aware of this when choosing a tool and conducting analysis.

Another major issue complicating variant annotation is the question of how to deal with genes and pseudogenes. There are widely varying levels of information available for different genes. Should we treat variants in well characterised genes in the same way as those in pseudogenes or non-genic regions of the genome? There is not currently a clear solution to this issue, although distinctions are usually made between annotations given from coding and non-coding transcripts. Again, careful choice of transcript set used for annotation can help.

Although there are many complications for variant annotation, we identify two major components:

1. Transcript set: a summary of information about genomic features, particularly the structure of transcripts (sequence and locations of exons, introns, UTRs and regulatory regions), used as the basis for determining the likely functional consequence of a variant.

2. Software tool: a piece of software that when given a particular variant can query a transcript set and return the functional annotation (or possibly annotations) of that variant. An annotation tool uses a particular algorithm applied to a given set of transcripts for annotating variants.

I examine the effects of fixing one component and then the other on a set of over 80 million single nucleotide variants (SNVs) and short indels from a large clinical sequencing project (see Section 2.2). ANNOVAR (Wang et al., 2010) is a popular annotation software tool, so I compare the results from ANNOVAR when used with the REFSEQ and ENSEMBL transcript sets. I also compare the annotation results from ANNOVAR and another popular annotation tool, VEP (McLaren et al., 2010) , the "Variant Effect Predictor" tool from ENSEMBL, when using the ENSEMBL transcript set and characterise the sorts of differences in annotation between the two tools and the apparent errors that ANNOVAR and VEP tend to make in annotation. Beyond issues specific to these particular transcript sets and software tools, I consider good practice for whole-genome and whole-exome variant annotation and problems that are yet to be solved.

2.2 Methods

2.2.1 Data generation

The data used in this study come from the "WGS500 Project", a collaboration between the University of Oxford and Illumina to sequence 500 genomes of clinical relevance. Samples were sought from patients where positive findings would have immediate clinical translational relevance in terms of clinical diagnosis, prognosis, genetics counselling and reproductive options, or treatment selection. The large umbrella project consists of many smaller sub-projects, each focusing on a particular patient cohort, as seen in some of the published studies that participated in the WGS500 project (Palles et al., 2013; Sharma et al., 2013; Cossins et al., 2013; Babbs et al., 2013; Lise et al., 2012; Martin et al., 2014; Ceroni et al., 2014). The research conformed to the Helsinki Declaration and to local legislation. Each individual project's ethics process was reviewed by the WGS500 Steering Committee and deemed to be sufficient for the WGS500 programme, including whole-genome analysis. As this was primarily a clinical study, clinical informed consent was considered sufficient for most samples. A small number of samples were part of research projects with separate ethical approval (references: MREC/06/Q1702/99; Riverside REC/09/H0706/20; Oxfordshire REC/06/Q1605/3; REC/09/H0606/74; and REC/09/H1204/3).

I focus here on whole genomes of 276 individuals sequenced as part of the WGS500 project. Only 276 whole genomes were used (not 500) as this was the number of non-tumour samples available at the time this work was carried out (this was "freeze 5" of

the WGS500 data). The samples included 80 patients with immune disease, 151 individuals from Mendelian disease studies (primarily parent-child trios) and 45 germline DNA samples from cancer patients. The sequencing was conducted using 100bp paired-end protocols on either the Illumina HiSeq 2000 instrument (Illumina, Inc, 2013a) or the Illumina HiSeq 2500 in standard mode (Illumina, Inc, 2013b), with a mixture of v2.5 and v3.0 chemistries, to at least 25x average coverage. Sequence reads were generated using the Illumina Off-Line Basecaller (v1.9.3) (Illumina, Inc, 2013c) and mapped to the human reference genome GRCh37d5/hg19d5 using Stampy, predominantly versions 1.0.12_(r975) and 1.0.13_(r1160) (Lunter & Goodson, 2011). Picard (picard-tools v1.67) was used to merge data and de-duplicate merged BAM files (Broad Institute, 2013). Variants were called from the aligned sequence reads using Platypus, version 0.1.9 (Rimmer et al., 2014). The raw data for annotation are variant call format (VCF) files (Danecek et al., 2011) containing information about the called variants.

In total, 80,995,744 unique variant calls were obtained from 276 individual genomes in the fifth "freeze" of the project's data, and merged into a preliminary "union" file. I compare functional annotations for 80,981,575 variants from the preliminary union file using different genome annotation databases and different annotation software tools, restricting ourselves to the set of variants for which an annotation was obtained using at least one transcript set or software tool.

2.2.2 Variant annotations

Variant annotations were obtained using the software ANNOVAR (version "2013Feb21"), using both the REFSEQ (release 57, January 2013) and ENSEMBL (version 69, October 2012) transcript sets (Pruitt et al., 2012; Flicek et al., 2013). I used the default transcript sets from REFSEQ and ENSEMBL. REFSEQ records are selected and curated from public sequence archives, so a REFSEQ record represents a synthesis, by a person or group, reducing the redundancy in the database. The REFSEQ database does not contain all possible (or even all observed) transcripts or gene models, but those that it does annotate feature strong evidence for their existence, structure and (possibly) function. Of a total of 105,258 human transcripts in REFSEQ release 57, 41,501 were used by ANNOVAR in the reported annotations for the variants in this study. Not all transcripts are used in reported annotations, because ANNOVAR reports only the most severe consequence of the variant, which corresponds to a subset of transcripts overlapping the variant. Typically, only one of many transcripts is reported, and so only about 40% of all REFSEQ transcripts are used for reported annotations across all 80 million variants.

Similarly, ENSEMBL, provides genome resources for chordate genomes with a particular focus on human genome data. ENSEMBL makes available substantial and diverse transcript information, including the Consensus Coding Sequence (CCDS) (Pruitt et al., 2009a;

Harte et al., 2012), Human and Vertebrate Analysis and Annotation (HAVANA) (Wellcome Trust Sanger Institute, 2012), Vertebrate Genome Annotation (Vega) (Ashurst et al., 2005), ENCODE data (ENCODE Project Consortium, 2012) and the GENCODE gene and transcript sets (Harrow et al., 2012). There are 208,677 transcripts in ENSEMBL version 69, of which 115,901 were used in reported annotations for this comparison. As for the REFSEQ transcripts, ANNOVAR's behaviour of reporting only one annotation from (typically) one of many possible transcripts for each variant results in only a subset of all ENSEMBL transcripts being used in reported annotations.

A broad interpretation of "splicing" regions was used for ANNOVAR annotations, so that all variants within 6 bases of an intron/exon boundary would fall into ANNOVAR's "splicing" annotation category. ANNOVAR returns a single annotation for each variant. If there are several relevant transcripts for a particular variant, then ANNOVAR will return the annotation with the most severe consequence according to its rules of precedence.

Variant annotations were also obtained using version 2.7 of ENSEMBL's VEP (Variant Effect Predictor), based on the ENSEMBL version 69 transcript set. As VEP returns all possible annotations for each variant (given the transcripts present at each variant's location in the genome), I prioritised annotations using a common-sense ranking of the "severity" of the consequence of the variant (Table 2.1) to make the VEP annotation results directly comparable with those from ANNOVAR. This prioritisation for consequences from VEP is just one possible way to prioritise variants and this subjectivity could affect the extent of matching between annotations from ANNOVAR and VEP. The most severe consequence for each variant was reported and compared to the ANNOVAR results.

2.2.3 Comparisons of variant annotations

I compare results across all annotation categories for the REFSEQ/ENSEMBL comparison. A comparison table was produced with a custom `Perl` (The Perl Foundation, 2013) script from VCF files containing ANNOVAR annotations when using REFSEQ and ENSEMBL transcripts and gene information for the transcript(s) used for each annotation. ANNOVAR reports only the "most damaging" annotation, but can return transcript information for all transcripts that would give the annotation reported. Subsequent statistical analysis was done in `R` version 2.15.0 (R Core Team, 2013).

For the comparison of ANNOVAR and VEP I focus on exonic variants (and especially loss-of-function (LoF) and nonsynonymous variants) for the ANNOVAR/VEP comparison as these are currently of the greatest interest in the majority of annotation applications in WGS studies. A VCF file containing all variants for comparison with annotations from ANNOVAR was processed to obtain VEP annotations, and the results were processed with a custom `Python` (Python Software Foundation, 2013) script to create a table of variants. The table provides annotation results obtained using ENSEMBL transcripts with ANNOVAR and

VEP, and information on transcripts used. The table was then analysed with R. I used the ENSEMBL Web Browser (archive version of ENSEMBL 69) (Flicek et al., 2013) and the UCSC Genome Browser (Rhead et al., 2010a) to inspect sets of variants identified to be of particular interest by comparing annotations using the DNA sequence and other information available in the browser. Files containing variant lists with annotations and source code for the analyses described here have been made available online in the figshare repository (http://dx.doi.org/10.6084/m9.figshare.798828), along with a "README" file that provides more details about the data and source code files, enabling these results to be reproducible.

2.2.4 Categories of variant annotations

To present, explain and discuss the results of these comparisons I need to introduce the different types of annotations produced by the different annotation tools. I define three categories of variants that are of particular interest for many functional studies:

1. **Putative loss-of-function variants (LoF):** variants that are likely to cause a gene product to be subject to nonsense-mediated decay and result in lost (or impaired) function of the gene. Included in this category are frameshift deletions, frameshift insertions, stop-gain, stop-loss and (most) splicing variants. Where finer resolution splicing categories are available (for example from VEP and some other annotation tools), we classify variants in splice acceptor and splice donor sites as LoF and other splicing variants as generically "exonic" (defined below). ANNOVAR does not provide subcategories of splicing variants. For this study I include all ANNOVAR splicing variants in the LoF category, but only include "splice donor" and "splice acceptor" annotations from VEP as LoF. While certain LoF variants have well-established causal roles in severe Mendelian diseases such as cystic fybrosis (Kerem et al., 1989), not all putative LoF variants seriously disrupt gene function. If a stop-gain or frameshift variant occurs towards the end of a transcript, the gene product may have residual or slightly modified function. Similarly, depending on surrounding sequence context and other factors, splicing and stop-loss variants may not result in serious disruption of gene function. MacArthur et al. (2012) present a systematic survey and discussion of loss-of-function variants, which explores these ideas in detail. Annotation as a LoF variant does not guarantee that the variant is of functional importance, but does flag it as likely to be of greater interest than other variants. As such, the LoF category is commonly used in discussions of genomic variation with the acceptance that annotation as a LoF variant does not always correspond to the variant having a genuine (or even if genuine, a serious) functional impact in a biological system.

Table 2.1: Common-sense precedence values used to prioritise VEP annotations for comparison with ANNOVAR annotations. I defined this prioritisation for this study based on on general expectations of the expected strength of effect of a typical variant with the given consequence. Higher values indicate higher precedence for the consequence, so if a variant receives more than one annotation the consequence with the higher precedence value is reported for the comparison.

VEP Consequence	Precedence
transcript_ablation	100
splice_donor_variant	87
splice_acceptor_variant	86
stop_gained	99
frameshift_variant	85
stop_lost	95
initiator_codon_variant	75
inframe_insertion	71
inframe_deletion	70
missense_variant	65
transcript_amplification	60
splice_region_variant	63
incomplete_terminal_codon_variant	50
synonymous_variant	40
stop_retained_variant	45
coding_sequence_variant	35
mature_miRNA_variant	30
UTR5_prime_UTR_variant	26
UTR3_prime_UTR_variant	25
intron_variant	24
NMD_transcript_variant	21
non_coding_exon_variant	20
nc_transcript_variant	19
upstream_gene_variant	18
downstream_gene_variant	17
TFBS_ablation	15
TFBS_amplification	16
TF_binding_site_variant	14
regulatory_region_variant	11
regulatory_region_ablation	12
regulatory_region_amplification	13
feature_elongation	2
feature_truncation	1
intergenic_variant	0

2. **Nonsynonymous/Missense variants:** variants in exons that change the amino-acid sequence encoded by the gene (but are not LoF), including single-base changes and nonframeshift indels. For this study I include VEP's "splice_region_variant" annotation in the missense category as this reflects the fact that general splice region variants are usually of a similar level of interest as canonical missense variants.

3. **Synonymous variants:** variants located in exons that do not change the translated amino acid sequence.

4. **Exonic variants:** variants that fall anywhere in exons or splicing regions, so this includes all variants in the LoF, nonsynonymous and synonymous categories above.

The exact terms used to denote annotation categories differ between ANNOVAR and VEP, but the correspondence in terms is almost always clear (Table 2.2). There are three exonic categories used by VEP ("initiator codon variant", "stop retained variant" and "other coding") for which there is no direct equivalent among the ANNOVAR categories.

Table 2.2: ANNOVAR and VEP terms in LoF, nonsynonymous/missense and exonic categories. In some figures and tables, frameshift insertions and deletions are combined into one "frameshift" category, with similar treatment for "nonframeshift" variants.

Category	ANNOVAR Terms	VEP Terms
Loss-of-function	"frameshift_deletion" "frameshift_insertion" "splicing" "stopgain_SNV" "stoploss_SNV"	"frameshift_variant" "splice_donor_variant" "splice_acceptor_variant" "stop_gained" "stop_lost" "transcript_ablation"
Missense/Nonsynonymous	"nonframeshift_deletion" "nonframeshift_insertion" "nonsynonymous_SNV"	"inframe_insertion" "inframe_deletion" "splice_region_variant" "initiator_codon_variant"
Synonymous	"synonymous_SNV"	"synonymous_variant" "stop_retained_variant"
Exonic	All of the above	All of the above plus "coding_sequence_variant" "incomplete_terminal_codon_variant"

2.3 Results

2.3.1 Same annotation tool, different transcript sets

The comparisons below of annotation results from ANNOVAR using either the REFSEQ or ENSEMBL transcript sets shows that the choice of transcript set has a large effect on the ultimate variant annotations. To summarise results, I look at the matching annotation rate, the percentage of variants that receive the same annotation with the two transcript sets. When looking at variants in particular classes (for example LoF variants), I define the match rate to be the percentage of all variants that receive an annotation in that class from *either* REFSEQ *or* ENSEMBL transcript sets (or both) that receive the same annotation from both REFSEQ and ENSEMBL transcripts. Across all 80 million variants there is an overall match rate of 85%. However, the matching annotation rate is 44% for LoF variants, the set of variants of most interest for biological and medical studies. The match rate is also substantially lower than the overall match rate for variants in non-coding RNA and UTR

regions, but there is better agreement for exonic and intronic variants. This observation accords with what one would expect: in areas of the genome where more is known about the protein-coding structure of the sequence the annotations agree more closely when using the two different transcript sets.

There are 590,893 variants given exonic annotations by ANNOVAR using REFSEQ or ENSEMBL (or both), of which 488,113 (83%) had precisely matching annotations when using the two different transcript sets (Table 2.3). The breakdown of matching variants by annotation reveals annotation categories showing greater and lesser difference when using REFSEQ or ENSEMBL. The extent of annotation matching is also summarised by category: "LoF", "LoF and missense (nonsynonymous)", "exonic" and "all annotated".

Visual comparison of transcript sets using REFSEQ- and ENSEMBL-normalized counts of variants with each combination of annotation terms from the two transcript sets highlights patterns in the differences in annotations provided by REFSEQ and ENSEMBL (Figures 2.3 & 2.4). By "REFSEQ-normalized", I mean that for each annotation term we consider all of the variants given that annotation using REFSEQ across all annotations using ENSEMBL and then normalize the count for each ENSEMBL annotation within the REFSEQ annotation by subtracting the mean number of counts per ENSEMBL annotation and dividing by the standard deviation. I do this independently for each REFSEQ annotation term. To obtain "ENSEMBL-normalized" values precisely the same thing is done, but the roles of the ENSEMBL and REFSEQ annotations are exchanged. Thus, for a given annotation term for a given transcript set, we can see the relative breakdown of annotations obtained when using the other transcript set. The REFSEQ-normalized values (Figure 2.3) show good agreement for indels (frameshift and nonframeshift), stop-gain, stop-loss and nonsynonymous variants, that is, a large proportion of variants given a particular annotation when using REFSEQ also get that annotation when using ENSEMBL. The agreement is not as good for synonymous and splicing variants, but we observe that variants given an exonic annotation when using REFSEQ usually get the same annotation when using ENSEMBL. Looking at ENSEMBL-normalized values (Figure 2.4), one sees generally lower matching rates. Agreement is good for variants called stop-gain, nonframeshift, nonsynonymous and synonymous by ENSEMBL, but variants annotated as frameshift, stop-loss and splicing are frequently given a different annotation when using REFSEQ.

Table 2.3: **Same software, different transcript sets:** This table summarises the number of annotations that match between the RefSeq and Ensembl results for each category of annotation. The table shows the number of variants given each type of annotation when using either RefSeq or Ensembl ("REF+ENS"; union), by RefSeq ("REF") and Ensembl ("ENS"), the number of variants that have matching annotations (i.e. the same annotation when using both transcript sets; intersection) and the match rate for each transcript set, which expresses the proportion of matching annotations for an annotation term relative to the total number of annotations in the category from the particular transcript set, as a percentage. The final column shows the "Overall match rate", which is the percentage of the variants with an annotation when using either RefSeq or Ensembl ("REF+ENS") that have a matching annotation when using the two transcript sets. Categories are loosely ordered by the severity of effect, with LoF annotations listed before nonsynonymous, synonymous, non-exonic categories and so on. Within each loose group, categories are sorted in descending order of overall matching rate. The bottom four rows show the total degree of matching across, respectively, all putative loss-of-function (LoF) categories, all LoF and missense categories, all exonic categories and, finally, all categories.

	REF+ENS	REF	ENS	Match	REF Match Rate (%)	ENS Match Rate (%)	Overall Match Rate (%)
stopgain_SNV	15835	14183	14960	13308	93.83	88.96	84.04
frameshift_insertion	6980	5298	6495	4813	90.85	74.10	68.95
frameshift_deletion	7491	4547	7380	4436	97.56	60.11	59.22
stoploss_SNV	946	503	906	463	92.05	51.10	48.94
splicing	47878	14154	45839	12115	85.59	26.43	25.30
frameshift_substitution	1960	195	1947	182	93.33	9.35	9.29
nonsynonymous_SNV	321669	291898	315592	285821	97.92	90.57	88.86
nonframeshift_insertion	3506	2888	2844	2226	77.08	78.27	63.49
nonframeshift_deletion	5136	3321	4963	3148	94.79	63.43	61.29
nonframeshift_substitution	933	226	843	136	60.18	16.13	14.58
synonymous_SNV	178559	167561	172463	161465	96.36	93.62	90.43
UTR3	724802	574255	622441	471894	82.17	75.81	65.11
UTR5	177832	94545	162684	79397	83.98	48.80	44.65
UTR5_UTR3	2183	292	2092	201	68.84	9.61	9.21
ncRNA_intronic	8992009	2113428	8244441	1365860	64.63	16.57	15.19
ncRNA_exonic	654098	140303	597947	84152	59.98	14.07	12.87
ncRNA_UTR3	53379	10712	47133	4466	41.69	9.48	8.37
ncRNA_UTR5	10683	1989	9444	750	37.71	7.94	7.02
ncRNA_splicing	13931	1051	13562	682	64.89	5.03	4.90
ncRNA_UTR5_ncRNA_UTR3	107	1	106	0	0.00	0.00	0.00
intronic	29289037	26805864	27743749	25260576	94.24	91.05	86.25
intergenic	50305202	49797113	41307708	40799619	81.93	98.77	81.10
downstream	991811	474684	840376	323249	68.10	38.46	32.59
upstream	910818	440728	762664	292574	66.38	38.36	32.12
upstream_downstream	53608	15621	47293	9306	59.57	19.68	17.36
unknown	11205	6215	5703	713	11.47	12.50	6.36
ALL LOF	81090	38880	77527	35317	90.84	45.55	43.55
ALL LOF and MISSENSE	412334	337213	401769	326648	96.87	81.30	79.22
ALL EXONIC	590893	504774	574232	488113	96.70	85.00	82.61
ALL	80981575	80981575	80981575	69181552	85.43	85.43	85.43

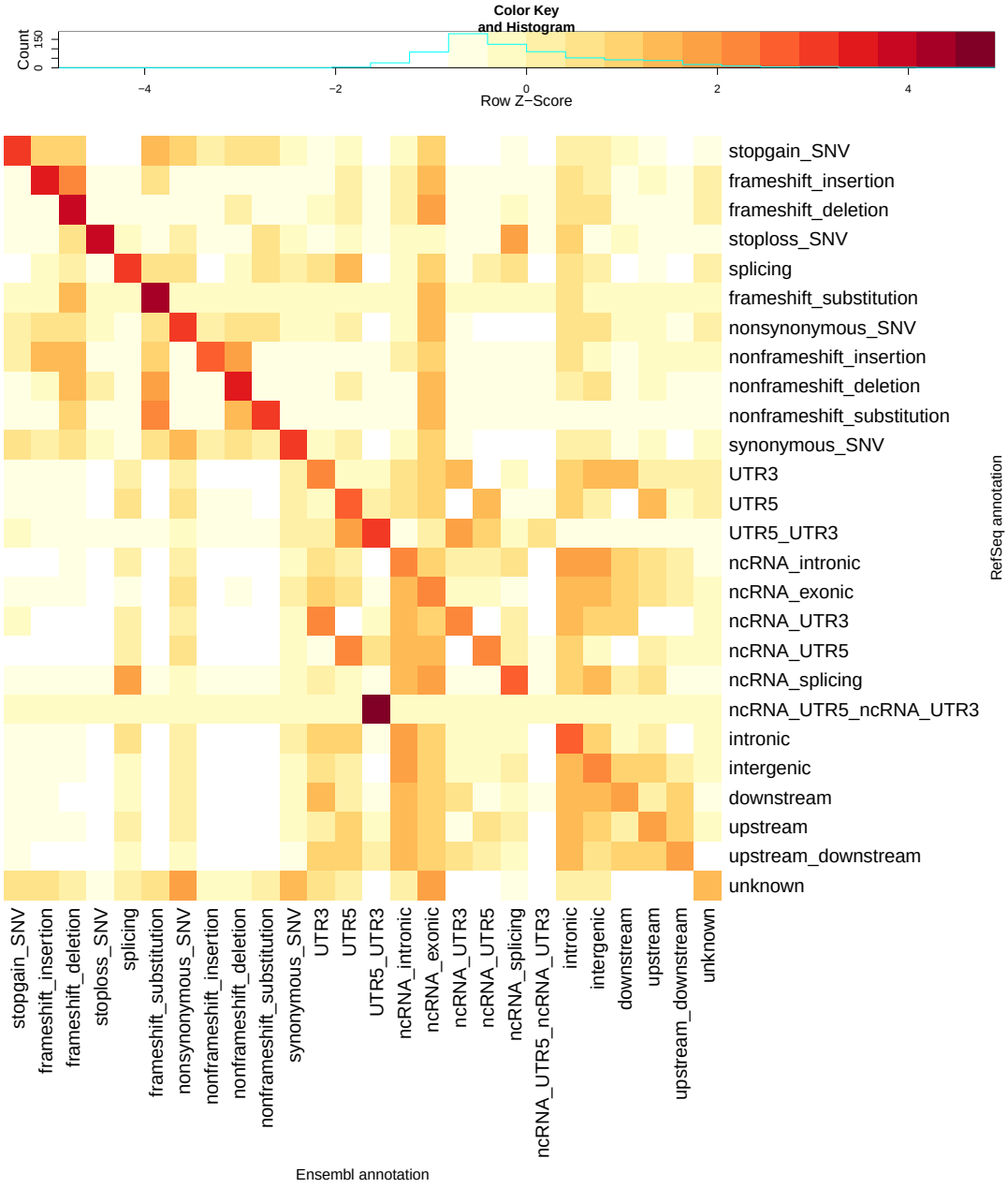


Figure 2.3: **RefSeq-normalized heatmap:** This heatmap shows scaled numbers of variants (with log10 transformation with offset of 1 applied) for all different combinations of ANNOVAR categories of annotations when using the ENSEMBL transcript set (columns) and REFSEQ transcript set (rows). Values are Z-scaled (mean-centred, divided by standard deviation) by row (each row is scaled separately; contrast with Figure 2.4). The key above the heatmap shows the values indicated by different colours. This row-normalized heatmap allows us to see which categories of annotation are overrepresented (relative to the total number of variants in the column/category) in the ENSEMBL annotations for each category (i.e. row) of REFSEQ annotation. Ideally, all of the dark red squares would lie on the diagonal, with white squares on the off-diagonals, indicating complete agreement in the annotations from the two databases. Compare with Table 2.4, which provides the numbers used for this heatmap. Categories are ordered as per Table 2.3.

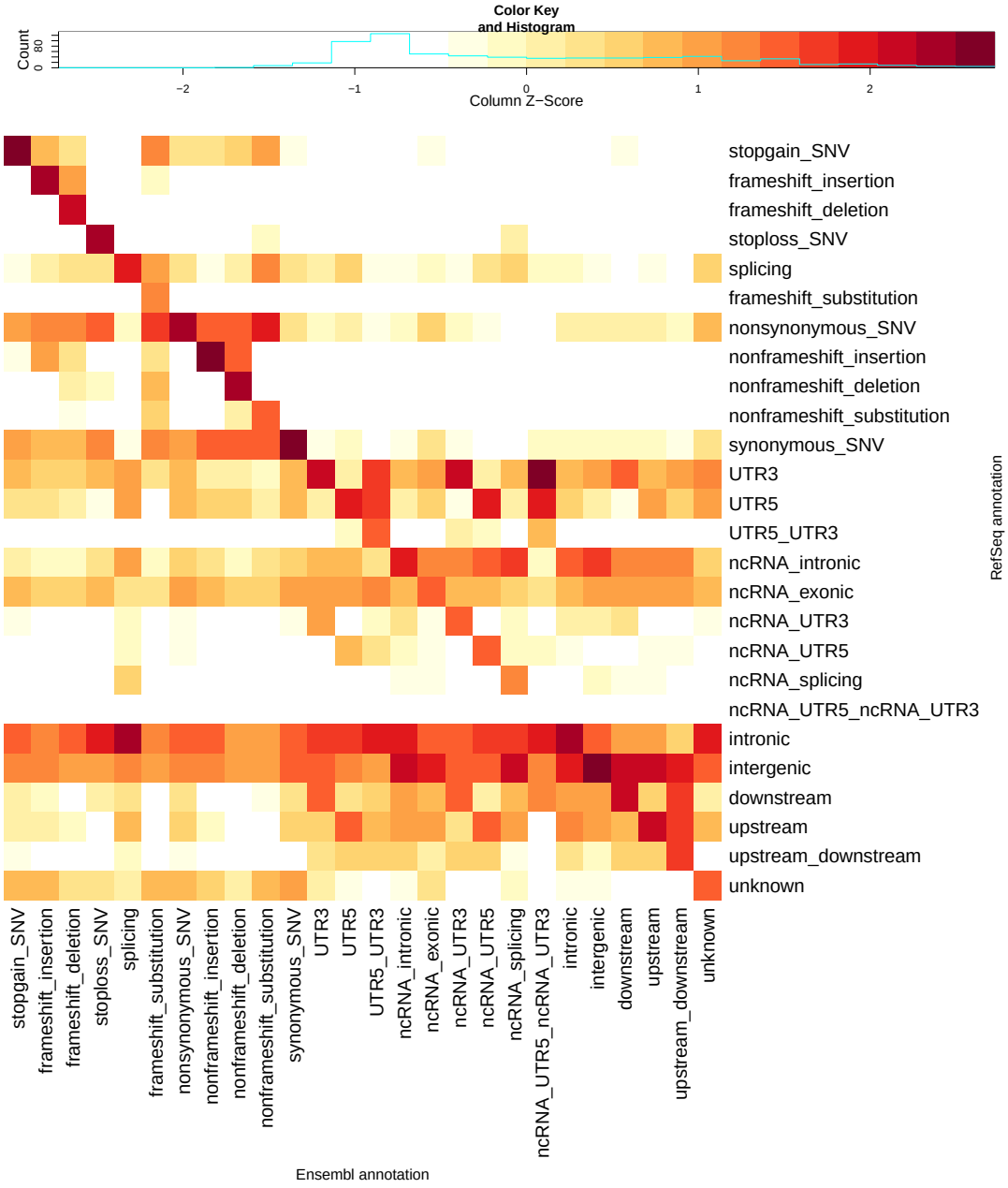


Figure 2.4: **ENSEMBL-normalized heatmap:** This heatmap shows scaled numbers of variants (with log10 transformation with offset of 1 applied) for all different combinations of ANNOVAR categories of annotations when using the ENSEMBL transcript set (columns) and REFSEQ transcript set (rows). Values are Z-scaled (mean-centred, divided by standard deviation) by column (each column is scaled separately; contrast with Figure 2.3). The key above the heatmap shows the values indicated by different colours. The column-normalized heatmap allows us to see which categories of annotation are overrepresented (relative to the total number of variants in the column/category) in the REFSEQ annotations for each category (i.e. column) of ENSEMBL annotation. Ideally, all of the dark red squares would lie on the diagonal, with white squares on the off-diagonals, indicating complete agreement in the annotations when using the two transcript sets. Compare with Table 2.4, which provides the numbers used for this heatmap. Categories are ordered as per Table 2.3.

Table 2.4: The count for each combination of REFSEQ annotation and ENSEMBL annotation for the 80971319 variants that are annotated by either REFSEQ or ENSEMBL. The abbreviated column names are the same as the rownames, which indicate the different categories of annotation. The entries along the diagonal should be equal (identical categories). Ideally, all of the off-diagonal entries should be zero.

	DN	FD	FI	FS	IG	IN	nc_EX	nc_IN	nc_splicing	nc_UTR3	nc_UTR5	nc_UTR5_nc_UTR3	NFD	NFI	NFS	NS	SP	SG	SL	SY	UN	UP	UP_DN	UTR3	UTR5	UTR5_UTR3
downstream	323249	5	10	0	24905	39579	7199	36584	152	2066	15	7	1	0	2	196	49	11	5	79	28	609	9493	30145	270	25
frameshift_deletion	0	4436	0	0	7	8	84	2	1	0	0	0	2	0	0	0	0	0	0	0	3	0	0	1	3	0
frameshift_insertion	0	373	4813	7	4	14	76	3	0	0	0	0	0	0	0	0	0	0	0	0	4	1	0	0	3	0
frameshift_substitution	0	5	0	182	0	2	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
intergenic	481779	304	253	139	40799619	2004467	327992	5637168	7657	3544	930	9	112	94	55	8345	1371	351	37	3529	923	448820	18240	42210	9087	78
intronic	3255	866	436	220	105303	25260576	108884	1152225	2226	3178	1874	21	179	136	54	11053	29953	571	189	4209	2764	4385	192	62935	49352	828
ncRNA_exonic	4173	122	70	30	14473	10635	84152	9340	117	294	161	2	52	34	16	2161	66	120	18	862	150	2819	643	7308	2365	120
ncRNA_intronic	9639	25	10	10	329371	353656	35082	1365860	2079	1581	739	1	6	9	12	575	810	18	8	255	73	9023	1228	2083	1261	14
ncRNA_splicing	8	0	0	0	58	21	101	47	682	0	0	0	0	0	0	3	107	0	0	2	0	15	0	4	3	0
ncRNA_UTR3	195	0	0	0	324	349	189	649	13	4466	0	0	0	0	0	20	18	3	0	9	6	0	0	4465	0	6
ncRNA_UTR5	0	0	0	0	3	35	138	78	9	0	750	1	0	0	0	14	12	0	0	5	3	12	4	1	909	15
ncRNA_UTR5_ncRNA_UTR3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
nonframeshift_deletion	0	46	1	67	7	3	42	0	0	0	0	0	3148	0	0	0	0	0	2	0	0	1	0	0	4	0
nonframeshift_insertion	0	72	150	35	3	12	29	6	0	0	0	0	346	2226	0	0	0	4	0	0	3	1	1	0	0	0
nonframeshift_substitution	0	9	0	49	0	0	16	0	0	0	0	0	16	0	136	0	0	0	0	0	0	0	0	0	0	0
nonsynonymous_SNV	62	495	283	532	297	308	2455	74	0	11	3	0	484	150	257	285821	17	167	83	63	95	59	15	42	122	3
splicing	1	68	21	140	35	113	217	20	110	6	35	1	17	2	83	172	12115	5	8	73	63	7	1	149	689	3
stopgain_SNV	6	79	95	203	15	17	172	8	3	0	0	0	37	11	53	158	0	13308	0	6	3	2	0	2	5	0
stoploss_SNV	1	4	0	0	0	6	1	1	19	0	0	0	0	0	3	2	1	0	463	1	0	0	0	0	1	0
synonymous_SNV	20	212	139	235	119	114	1152	30	2	7	0	1	494	131	135	2862	8	205	60	161465	44	35	6	21	64	0
unknown	2	65	91	71	46	41	2151	47	9	1	1	0	18	24	28	1777	36	72	7	920	713	0	0	67	28	0
upstream	863	27	18	3	13181	57988	11313	35689	275	41	1122	0	1	3	0	435	349	17	0	189	113	292574	7380	741	18343	63
upstream_downstream	603	0	1	0	141	1364	635	1850	7	102	68	0	0	0	0	15	11	3	0	3	1	585	9306	290	607	29
UTR3	16513	126	64	22	19497	12027	13081	4392	175	31795	23	39	13	8	4	1159	463	72	25	447	460	784	667	471894	140	365
UTR5	7	41	40	2	300	2414	2778	368	25	4	3715	19	37	16	5	823	453	32	1	343	254	2932	117	81	79397	341
UTR5_UTR3	0	0	0	0	0	0	2	0	1	37	8	5	0	0	0	1	0	1	0	3	0	0	0	2	31	201

The asymmetry in the differences in annotations between REFSEQ and ENSEMBL is striking. There are many more exonic annotations, across all LoF, nonsynonymous and synonymous categories, when using ENSEMBL transcripts (Table 2.3 and Supplementary Table 1). There are several thousand variants that are called exonic by ENSEMBL and yet are called as intergenic, intronic or in a non-coding RNA by REFSEQ. Conversely, there are only a few hundred exonic variants from REFSEQ that are annotated as intergenic, intronic or in non-coding RNA according to ENSEMBL. Using ENSEMBL here would gain over 2000 frameshift indels and over 1000 stop-gain/stop-loss variants compared with using REFSEQ, all LoF variants of substantial interest for follow-up. This asymmetry is not surprising when one considers the composition of the two transcript sets. The REFSEQ set contains 105,258 human transcripts in release 57, for which the protein-coding sequences cover approximately 1.07% of the genome (34Mb). ANNOVAR actively used 41,501 of these transcripts for annotation of this set of variants. The ENSEMBL version 69 set contains 208,677 transcripts (192,635 on chromosomes 1–22, X and Y, excluding patches and alternate loci), covering approximately 28% of the genome (892Mb), including introns. The protein-coding sequences in the ENSEMBL transcript set cover approximately 1.12% of the genome (35Mb). Of these transcripts, 115,091 were actively used for the annotation of this set of variants, including the set of 92,776 transcripts containing protein-coding sequences.

This extent of discrepancy in annotations can be partially explained by the fact that a high proportion of REFSEQ transcripts have an equivalent or highly similar transcript in ENSEMBL, but in the other direction there are many transcripts in ENSEMBL that do not appear to have a similar transcript in REFSEQ. ANNOVAR reports the most severe consequence for a variant across all transcripts present at that position in the genome, so with more transcripts available when using ENSEMBL there is an elevated chance of finding a more severe consequence for one of the ENSEMBL transcripts. There were no significant differences in annotation agreement rates across different variant frequencies (Table 2.5).

MAF Range	Total	Match	Match Rate (%)
0–1%	67313352	57701490	85.72
1–5%	5802063	4882133	84.14
5–10%	1656880	1388842	83.82
10+%	6209280	5209087	83.89
ALL	80981575	69181552	85.43

Table 2.5: **Same software, different transcript sets:** This table summarises the number of annotations that match between the REFSEQ and ENSEMBL results for different ranges of minor allele frequency (MAF). The match rates across the different ranges of MAF are not substantially different from the match rate across all variants.

2.3.1.1 Examples of variants with differing annotations

Studying examples of variants with striking differences in annotation helps to characterise the sorts of differences seen when using REFSEQ and ENSEMBL transcripts.

The variant 1:26879920_T_C is annotated as synonymous when using REFSEQ transcripts, but as stop-loss when using ENSEMBL transcripts (Figure 2.5). It looks like this difference in annotation is caused by there being more ENSEMBL transcripts available to use for annotation, and thus a higher chance of seeing a more severe consequence. ANNOVAR reports the most severe consequence, so there is a better chance of getting a "stop-loss" annotation when using ENSEMBL transcripts than REFSEQ transcripts. In the case of this variant, there are many ENSEMBL transcripts that would give the same annotation as from the REFSEQ transcript, but the one used is the one that gives the most severe consequence (stop-loss). Noteworthy is the fact that the ENSEMBL transcript used is labelled as being subject to nonsense-mediated decay (NMD). So although it is a possible choice of transcript, it may not be the best possible choice, as any changes to this transcript would likely be irrelevant as the transcript is believed to be subject to NDM. This case highlights the value of careful curation of transcript sets for annotation. More transcripts is not necessarily better. If transcripts that are of poor quality or in some way of lesser interest are used as the basis for annotation, then we risk diluting the set of truly interesting variants for follow-up study with variants that are of substantially lesser interest.

The variant 15:44093914_T_C provides an example of how the use of transcripts with different structures can give rise to different annotations (Figure 2.6). The variant is annotated as synonymous when using ENSEMBL transcripts, but stop-loss when using REFSEQ. In this case there are many ENSEMBL transcripts that look similar to one of the REFSEQ transcripts, but one REFSEQ transcript is noticeably different from the others, and this one transcript yields a stop-loss annotation. Following ANNOVAR's precedence rules, the stop-loss annotation is reported in preference to the synonymous annotation that would follow from using one of the other transcripts. Given the transcripts used, both annotations look valid here.

The variant 11:70279766_C_T (annotated as synonymous using REFSEQ and stop-gain with ENSEMBL transcripts) gives another example of differing annotations caused by the use of transcripts with different structures. In this case there are many ENSEMBL transcripts that look similar to one of the REFSEQ transcripts, but the ENSEMBL transcript used to give the annotation reported is noticeably different from the others, and again noted to be subject to NMD. Given the transcripts used, both annotations look "correct" here. I note that a difference in the reading frames for the transcripts used accounts for the variant being a "synonymous" variant for many of the transcripts shown and "stop-gain" for the ENSEMBL transcript used.

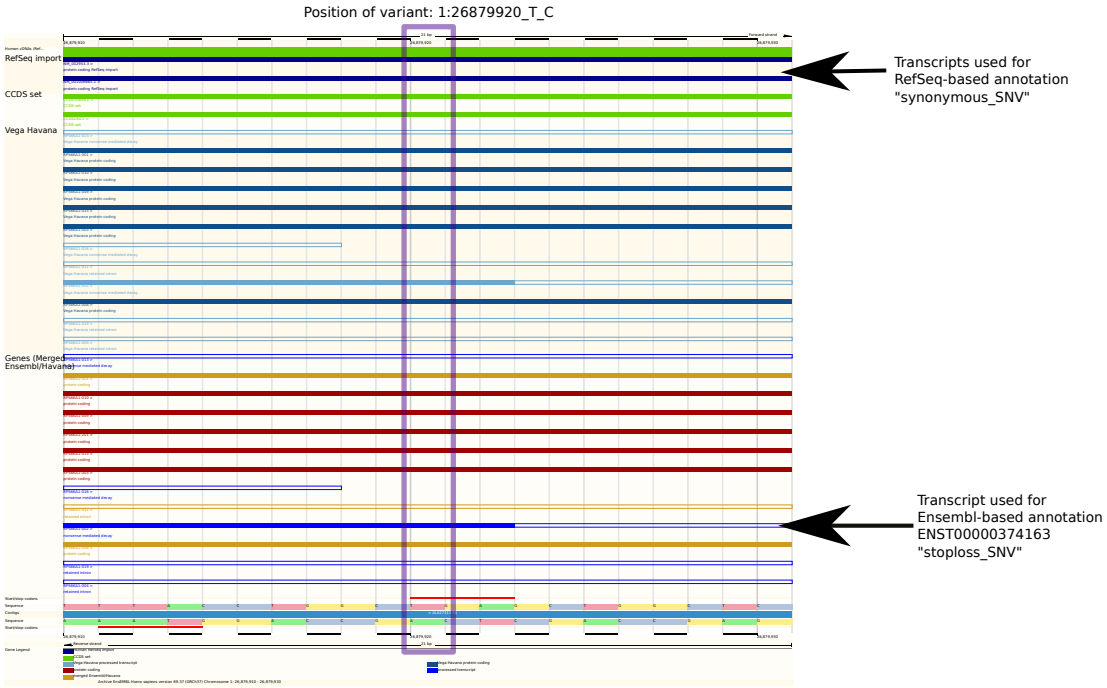


Figure 2.5: **Browser Image: RefSeq synonymous, Ensembl stoploss.** For this variant, 1:26879920_T_C, the different annotations from RefSeq and Ensembl are due to the use of transcripts with substantially different structures. There are many Ensembl transcripts that would give the same annotation as from the RefSeq transcript, but the one used is the one that gives the most severe consequence, in this case a stoploss variant. Noteworthy is the fact that the Ensembl transcript used is labelled as being subject to nonsense-mediated decay (NMD). So although it is a possible choice of transcript, it may not be the best possible choice, as any changes to this transcript would likely be irrelevant as the transcript is subject to NDM anyway. See Figure 2.2 for an explanation of the elements of Ensembl Web Browser images.

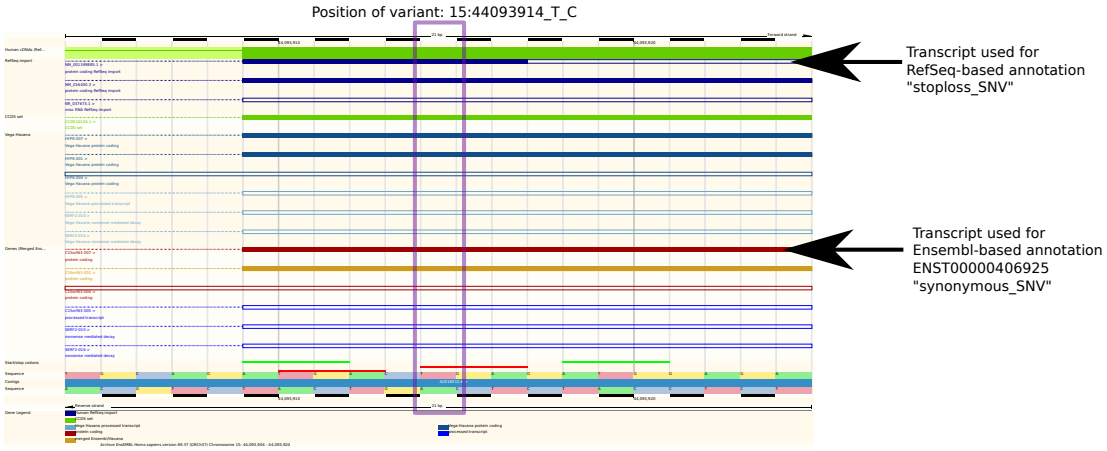


Figure 2.6: **Browser Image: RefSeq stoploss, Ensembl synonymous.** For this variant, 15:44093914_T_C, the different annotations from RefSeq and Ensembl are due to the use of transcripts with different structures (as seen in Figure 2.5). In this case there are many Ensembl transcripts that look similar to one of the RefSeq transcripts, but one RefSeq transcript is noticeably different from the others. Given the transcripts used, both annotations look "correct" here. See Figure 2.2 for an explanation of the elements of Ensembl Web Browser images.

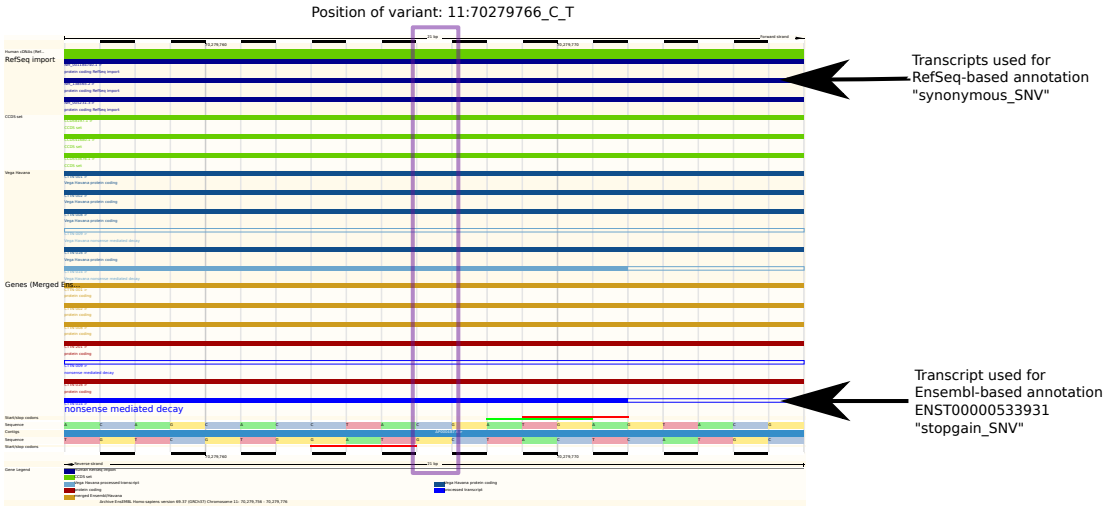


Figure 2.7: **Browser Image: REFSEQ synonymous, ENSEMBL stopgain.** For this variant, 11:70279766_C_T, the different annotations from REFSEQ and ENSEMBL are again due to the use of transcripts with different structures (as seen in Figure 2.5). In this case there are many ENSEMBL transcripts that look similar to one of the REFSEQ transcripts, but ENSEMBL used to give the annotation reported is noticeably different from the others, and again noted to be subject to NMD. Given the transcripts used, both annotations look "correct" here, and we note that a difference in the reading frames for the transcripts used accounts for the variant being a "synonymous" variant for many of the transcripts shown and "stopgain" for the ENSEMBL transcript used. See Figure 2.2 for an explanation of the elements of Ensembl Web Browser images.

The variant 16:4745030_C_T is annotated as stop-gain using REFSEQ transcripts and as synonymous using ENSEMBL transcripts (Figure 2.8). Again, this discrepancy is due to the use of transcripts with different structures. Looking at the transcripts used from REFSEQ and ENSEMBL, both annotations look correct. Again, a difference in the reading frames for the transcripts used accounts for the variant annotated as synonymous for many of the transcripts shown and as stop-gain for the REFSEQ transcript used. It is not clear why ANNOVAR used the ENSEMBL transcript that it did, which returns the synonymous annotation, and not the shorter red transcript below it. It appears that if ANNOVAR used that red transcript it would yield a stop-gain annotation like the blue REFSEQ transcript, which it seems to match. While not common, one does see instances like this—where the software tool shows behaviour that is not immediately explicable to the end user—across all software tools. In some cases, inexplicable behaviour can be attributed to complicated annotation decisions in which it is not straight-forward to "eye-ball" the variant and determine what the correct annotation should be. In other cases, inexplicable behaviour can arise due to programming bugs in the software. Out of necessity, current variant annotation tools are complex pieces of software. Despite the best efforts of their developers, a certain number of bugs (or other unexpected "features") in these software tools is inevitable. Understanding how bugs can arise and affect variant annotations is therefore useful for users of annotation software.

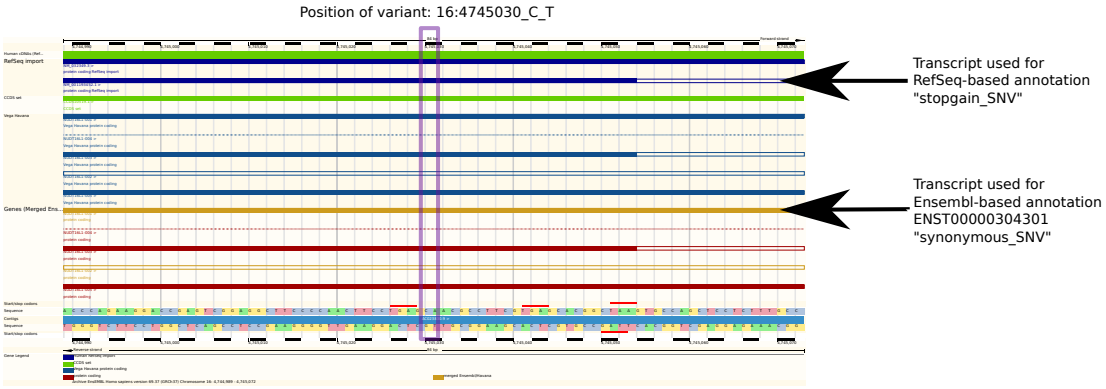


Figure 2.8: **Browser Image: REFSEQ stopgain, ENSEMBL synonymous.** For this variant, 16:4745030_C_T, the different annotations from REFSEQ and ENSEMBL are again due to the use of transcripts with different structures (as seen in Figure 2.7). Given the transcripts used, both annotations look "correct" here, and again it is noted that a difference in the reading frames for the transcripts used accounts for the variant being a "synonymous" variant for many of the transcripts shown and "stopgain" for the REFSEQ transcript used. What is unclear is why ANNOVAR used the ENSEMBL transcript that it did, returning a "synonymous" annotation, rather than the shorter red transcript below it, which (it appears) would yield a "stopgain" annotation like the blue REFSEQ transcript, which it seems to match. See Figure 2.2 for an explanation of the elements of Ensembl Web Browser images.

Inspecting the variant 11:8149801_CAT_T highlights the large number of isoforms a gene can have (Figure 2.9). As the browser image shows, there are eleven "merged ENSEMBL" transcripts, nine Vega/HAVANA transcripts, four CCDS transcripts and five REFSEQ transcripts overlapping this position. Although the annotations (intronic from REFSEQ and frameshift from ENSEMBL) are straight-forward to understand here, it is easy to imagine how such large numbers of transcripts and gene isoforms lead to great complexity in variant annotation. At this position, all of the multiple REFSEQ transcripts and an even larger number of ENSEMBL transcripts have an intron. There is one ENSEMBL transcript that has coding sequence at this position, and this is the one used by ANNOVAR for the ENSEMBL-based annotation. Given the transcripts used, both annotations look "correct" here.

The position 1:28785729 provides a peculiar example of annotation behaviour from ANNOVAR. There are two variants at this position, 1:28785729_GA_G and 1:28785729_G_GA. If there is a transcript with coding sequence at this location, then one would expect the first variant to be annotated as "frameshift deletion" and the second to be annotated as "frameshift insertion". However, ANNOVAR behaves oddly here as the two variants are *both* annotated as "frameshift insertion" when REFSEQ transcripts are used and *both* annotated as "frameshift deletion" when ENSEMBL transcripts are used. There are 373 variants in the dataset for which this or similar behaviour occurs, but it is not at all clear why ANNOVAR would behave in this way for these variants. As mentioned above, a certain number of bugs are inevitable in complex software. This is a particularly inexplicable ex-

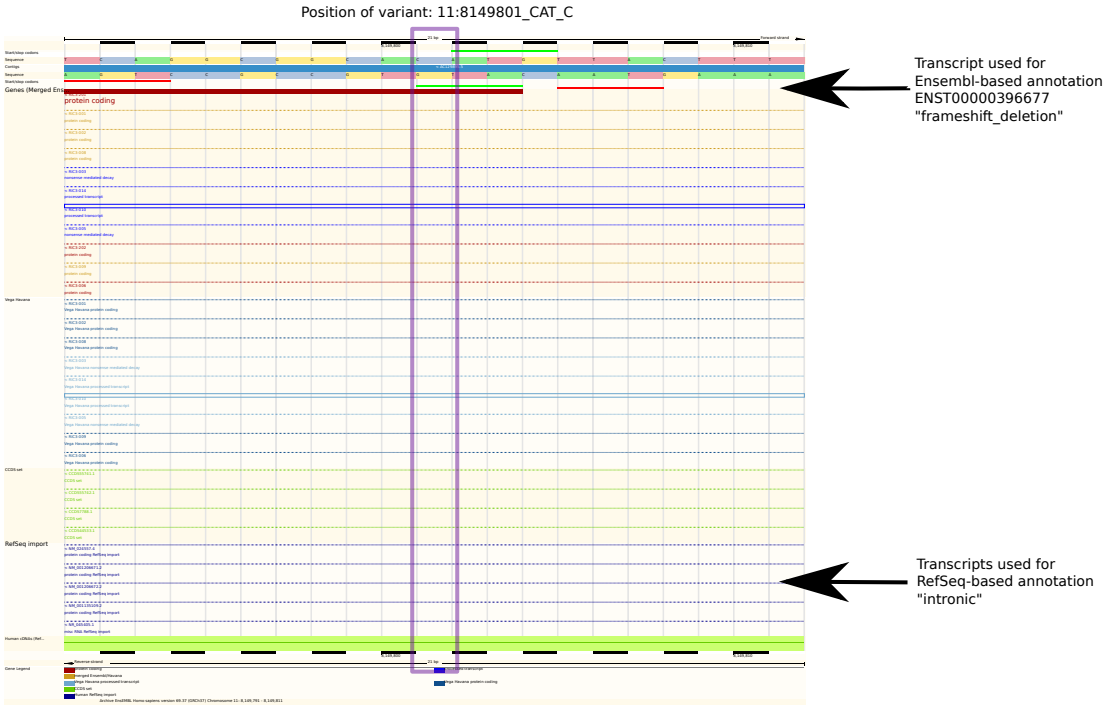


Figure 2.9: **Browser Image: RefSeq intronic, Ensembl frameshift deletion.** For this variant, 11:8149801_CAT_T, the different annotations from RefSeq and Ensembl are again due to the use of transcripts with different structures. At this position in the genome there are multiple RefSeq transcripts and even more Ensembl transcripts that have an intron. There is one Ensembl transcript that has coding sequence at this position, and this is the one used by Annovar for the Ensembl-based annotation. Given the transcripts used, both annotations look "correct" here. This image shows how many isoforms (i.e., transcripts) a gene can have. See Figure 2.2 for an explanation of the elements of Ensembl Web Browser images.

ample of behaviour from Annovar, but thankfully the 373 variants where this sort of behaviour occurs represent only a very small proportion of the roughly 15,000 variants that receive either a frameshift insertion or frameshift deletion annotation when using either RefSeq or Ensembl transcripts. Overall, the match rates for frameshift insertions (69%) and frameshift deletions (59%) are resonably good (Table 2.3), although there is much higher concordance for RefSeq annotations than Ensembl annotations. The particular bug that causes the peculiar behaviour for this variant does not seem to have a very large effect on frameshift annotations overall.

The variant 11:104761100_T_C (nonsynonymous using RefSeq and splicing using Ensembl) highlights how difficult it can sometimes be to assess the validity of variant annotations by eye. Here, the reason for the different annotations from RefSeq and Ensembl is not apparent from looking at the Ensembl Web Browser image (Figure 2.11, top). At this position in the genome there are multiple Ensembl transcripts with the same structure and one RefSeq transcript that matches them (at least in this region shown). The "splicing" annotation given by the Ensembl transcript looks correct. Looking at the RefSeq transcript, the nonsynonymous annotation looks incorrect. However, if this variant

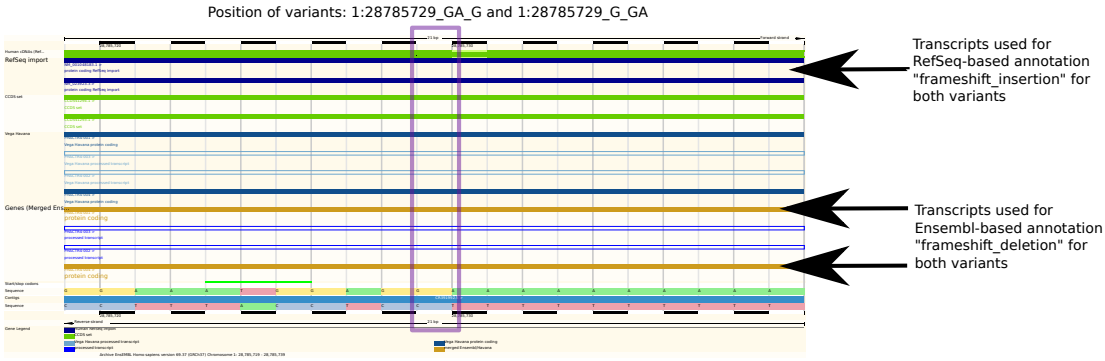


Figure 2.10: **Browser Image: REFSEQ frameshift insertion, ENSEMBL frameshift deletion.** This browser image shows quite a peculiar example of annotations. At this location two variants are observed, 1:28785729_GA_G and 1:28785729_G_GA. If there is a transcript with coding sequence at this location, then one would expect the first variant to be annotated as "frameshift deletion" and the second to be annotated as "frameshift insertion". However, ANNOVAR behaves oddly here as the two variants are *both* annotated as "frameshift insertion" when REFSEQ transcripts are used and *both* annotated as "frameshift deletion" when ENSEMBL transcripts are used. There are 373 variants in the dataset for which this or similar behaviour occurs, but it is not at all clear why ANNOVAR would behave in this way for these variants. See Figure 2.2 for an explanation of the elements of Ensembl Web Browser images.

is inspected in the UCSC Web Browser (Figure 2.11, bottom), then one sees that there are five REFSEQ transcripts available at this position that, if used, would give an annotation of "nonsynonymous". Thus it looks like both annotations are reasonable based on the transcripts used, although this is not clear from looking only at the (archived version 69) Ensembl Web Browser.

The variant 17:38064469_T_C provides a final example of different transcript structures giving rise to different annotations. Here, the variant is annotated as intronic when using REFSEQ transcripts and as synonymous when using ENSEMBL transcripts. At this position in the genome there are two REFSEQ transcripts and several more ENSEMBL transcripts that share structure and have an intron. There is one ENSEMBL transcript that has coding sequence at this position, and this is the one used by ANNOVAR for the ENSEMBL-based annotation. Given the transcripts used, both annotations look valid.

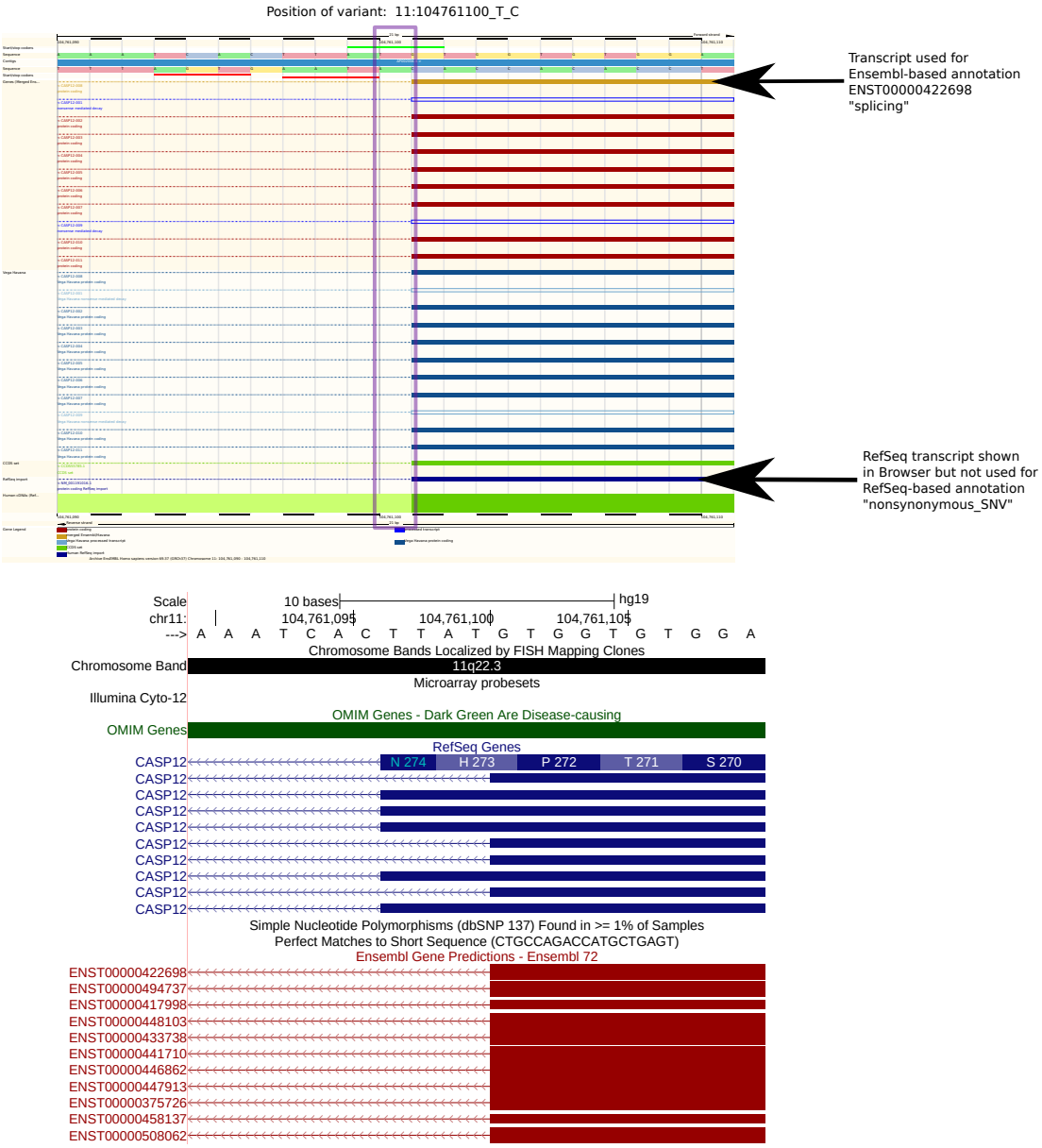


Figure 2.11: **Browser Image: RefSeq nonsynonymous, Ensembl splicing.** For this variant, 11:104761100_T_C, the reason for the different annotations from RefSeq and Ensembl is not apparent from looking at the Ensembl Web Browser image (top). At this position in the genome there are multiple Ensembl transcripts with the same structure and one RefSeq transcript that matches them (at least in this region shown). The "splicing" annotation given by the Ensembl transcript looks correct. Looking at the RefSeq transcript, the nonsynonymous annotation looks incorrect. However, inspecting this variant in the UCSC Web Browser (bottom) reveals that there are five RefSeq transcripts available at this position that, if used, would give an annotation of "nonsynonymous". Thus it looks like both annotations are reasonable based on the transcripts used, although this is not clear from looking only at the (archived version 69) Ensembl Web Browser. See Figure 2.2 for an explanation of the elements of Ensembl Web Browser images.

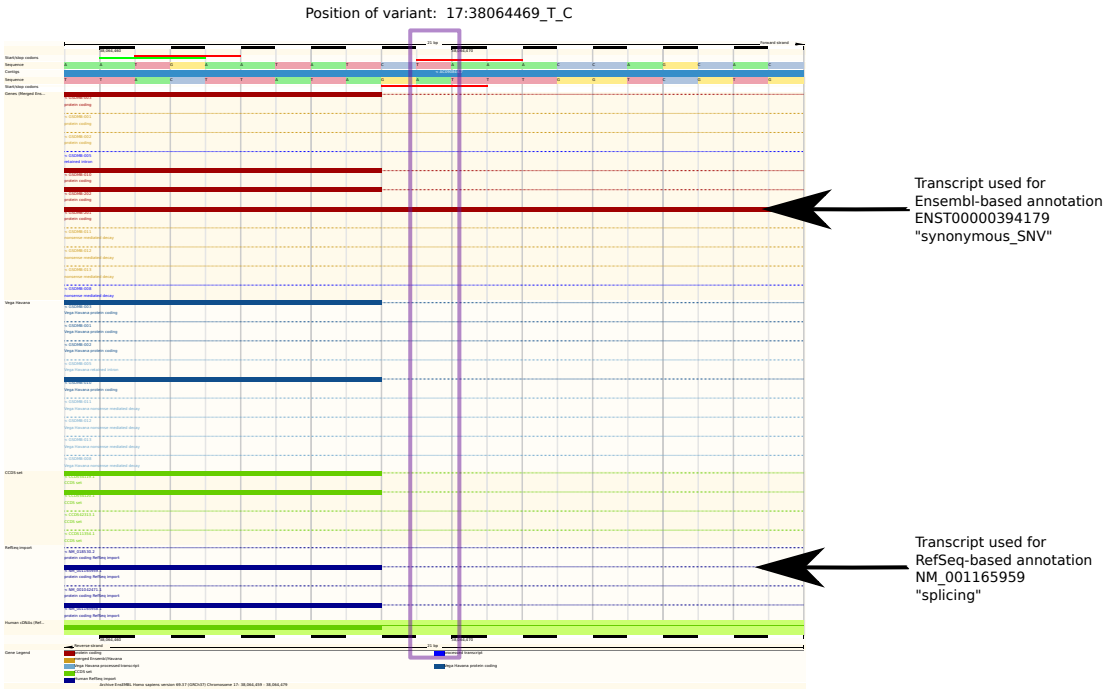


Figure 2.12: **Browser Image: REFSEQ intronic, ENSEMBL frameshift deletion.** For this variant, 17:38064469_T_C, the different annotations from REFSEQ and ENSEMBL are again due to the use of transcripts with different structures. At this position in the genome there are two REFSEQ transcripts and several more ENSEMBL transcripts that share structure and have an intron. There is one ENSEMBL transcript that has coding sequence at this position, and this is the one used by ANNOVAR for the ENSEMBL-based annotation. Given the transcripts used, both annotations look "correct" here. See Figure 2.2 for an explanation of the elements of Ensembl Web Browser images.

2.3.2 Same transcript set, different annotation tools

I also investigate the extent to which using different software tools influences the final annotations. Here I compare annotations from ANNOVAR and VEP using the ENSEMBL transcript set, focusing on exonic annotation categories. I look at the rate of "exactly matching" annotations and the rate of "category matching" annotations. I refer to an exact match when the annotations from both software tools are exactly equivalent given the annotation terms used by the two tools, for example both tools annotate a variant as "frameshift". By "category match", I mean that annotations from both software tools are in the same higher-level category of "LoF", "Missense" or "Synonymous and other coding" (with categories defined in Table 2.2). So if a variant received an annotation of "frameshift" from one tool and "stop gain" from the other, I would designate this as a "category match" as both are LoF annotations. Overall, there is only a small difference in matching rates when considering category matches as opposed to exact matches, with category matching rates approximately 1% higher than exact matching rates (Table 2.6).

In total, 637,841 variants were given exonic annotations by either ANNOVAR or VEP (Table 2.6). Of these, 551,983 (86.5%) had exactly matching annotations from the two tools and 556,387 (87.2%) have category matching annotations. However, the match rate is substantially lower (65% for exact matches, 66% for category matches) for LoF annotations (Table 2.6). We observe that 89% of exonic variants from VEP get an exactly matching annotation from ANNOVAR and 96% of exonic variants according to ANNOVAR get an exactly matching annotation from VEP. These percentages of agreement should not be taken to show that ANNOVAR is "more accurate" than VEP—the difference between the tools for exonic variants is driven by the larger number of "splicing" annotations from VEP, which is due to a difference in the definition of a splicing variant used by the two tools.

Considering all annotation categories for VEP and ANNOVAR annotations shows a substantial amount of disagreement in annotations from the two tools, even when using the same transcripts (Figures 2.13 & 2.14). The heatmaps in Figures 2.13 & 2.14 represent the normalized counts for each combination of VEP and ANNOVAR annotation. The ANNOVAR- and VEP-normalized counts were computed in an analogous fashion to the computation of the REFSEQ- and ENSEMBL-normalized counts in the transcript set comparison above. For an annotation term category under consideration for one software tool, counts across all categories from the other software tool are mean-centered and divided by the standard deviation, giving normalized counts that indicate, for a given annotation term for a given software tool, the relative breakdown of annotations from the other software tool. Ideally, one would like to see, for example, that all variants called "synonymous" by VEP are also annotated as "synonymous" by ANNOVAR and vice-versa. We would like to see agreement in annotations across all categories. ANNOVAR-normalized values (Figure 2.13)

Table 2.6: **Same transcript set, different software:** This table summarises the number of annotations that match between the ANNOVAR and VEP results for each exonic category of annotation. Columns one to five show the number of variants given each type of annotation by either ANNOVAR or VEP ("ANV+VEP"; union), by ANNOVAR ("ANV") and VEP ("VEP"), the number of variants that have exact matching annotations (i.e. exactly the same annotation from both tools; intersection), and category-matching annotations (i.e. annotations from the two tools in the same high-level category—LoF, Missense, Synonymous and Other Coding—even if not an exact match). Columns six and seven show the match rate for each tool, which gives the percentage of matching annotations for an annotation term from ANNOVAR and VEP, respectively, relative to the total number of annotations in the category from the particular software tool. Column eight gives the percentage of variants with annotations from ANNOVAR and VEP in the same high-level category ("Overall Category Match Rate"). Column nine shows the "Overall Exact Match Rate", which is the percentage of the variants with an annotation from either ANNOVAR or VEP ("ANV+VEP") that have an exactly matching annotation from the two tools. Here, the specific annotations from equivalent terms for ANNOVAR and VEP have been aggregated to enable the comparison (see Supplementary Table 4). The final three rows of the table show aggregate counts and match rates for all loss-of-function categories, all LoF and missense categories and all exonic categories, respectively. Note that the "all splicing" category for VEP includes 5,011 "splice acceptor" variants, 8,544 "splice donor" variants and 49,298 more general "splice region" variants. ANNOVAR, in contrast, only has one general "splicing" category, and does not distinguish between acceptor, donor and other splicing variants.

	ANV + VEP	ANV	VEP	Exact Match	Category Match	ANV Match Rate (%)	VEP Match Rate (%)	Overall Category Match Rate (%)	Overall Exact Match Rate (%)
LOF Total	104915	77527	96761	68284	69373	88.08	70.57	66.12	65.09
frameshift	19021	15822	16685	13486	-	85.24	80.83	-	70.90
stop gained	16758	14960	16146	14348	-	95.91	88.86	-	85.62
stop lost	1113	906	1077	870	-	96.03	80.78	-	78.17
all splicing	69112	45839	62853	39580	-	86.35	62.97	-	57.27
MISSENSE Total	350806	324242	347752	318056	321188	98.09	91.46	91.56	90.66
inframe indel	9455	8650	6600	5795	-	66.99	87.80	-	61.29
missense	343284	315592	339953	312261	-	98.94	91.85	-	90.96
initiator codon	1199	0	1199	0	-	-	0.00	-	0.00
SYNONYMOUS and OTHER CODING Total	182120	172463	175483	165643	165826	96.05	94.39	91.05	90.95
synonymous	181873	172463	175053	165643	-	96.05	94.62	-	91.08
stop retained	203	0	203	0	-	-	0.00	-	0.00
other coding	227	0	227	0	-	-	0.00	-	0.00
ALL LOF	104915	77527	96761	68284	69373	88.08	70.57	66.12	65.09
ALL LOF and MISSENSE	455721	401769	444513	386340	390561	96.16	86.91	85.70	84.78
ALL EXONIC	637841	574232	619996	551983	556387	96.13	89.03	87.23	86.54

indicate generally good agreement of ANNOVAR annotations with VEP annotations. Nevertheless, there are substantial numbers of variants receiving differing annotations from the two tools across all categories of variants. The VEP-normalized values (Figure 2.14) confirm this view.

Relatively lower concordance is observed for intergenic, intronic, miRNA and splicing variants. Even in well-defined categories such as nonsynonymous (missense) and

frameshift, there is a large amount of disagreement in annotations between the two tools. There were no significant differences in annotation agreement rates across different variant frequencies (Table 2.7).

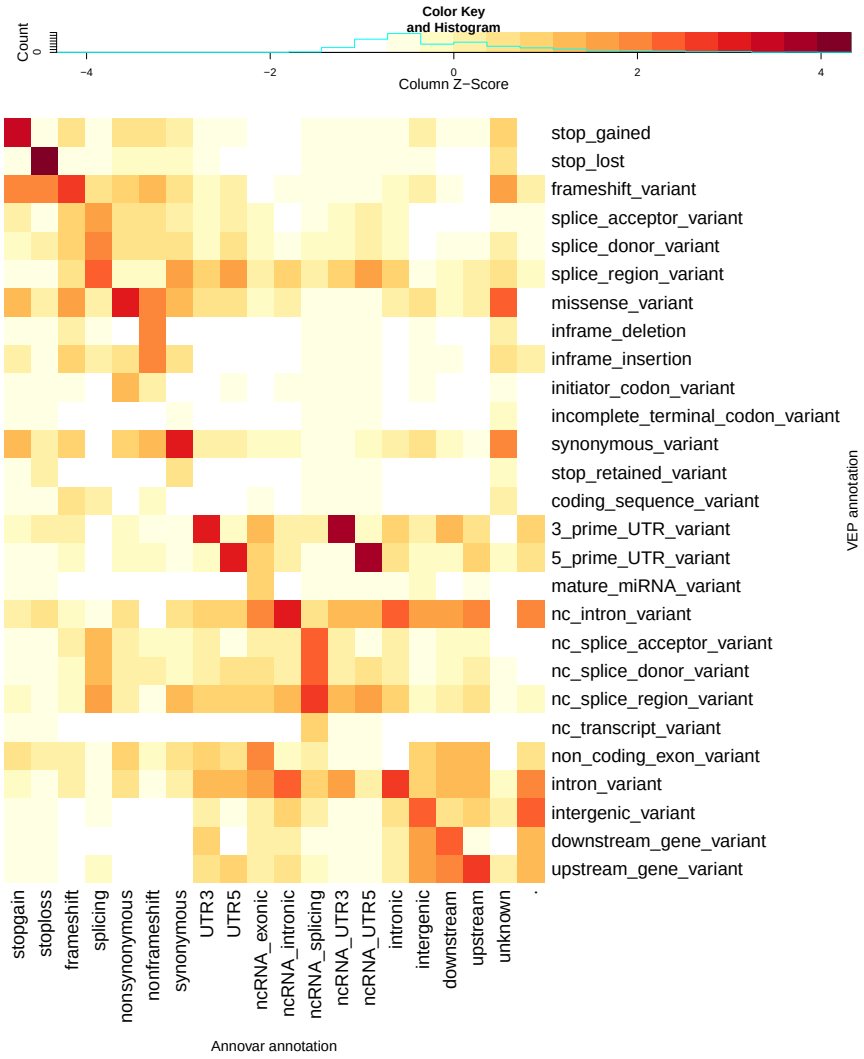


Figure 2.13: **ANNOVAR-normalized heatmap:** This heatmap shows scaled numbers of variants for all different combinations of categories of annotations when using the VEP annotation tool (rows) and the ANNOVAR software (columns), with the ENSEMBL transcript set. The values (log10 of the count of variants with that combination of annotations from the two tools, with an offset of 1 applied; see Supplementary Table 2 for raw counts) are Z-scaled (mean-centred, divided by standard deviation) by column (i.e. standardising ANNOVAR annotations). The key above the heatmap shows the values indicated by different colours. ANNOVAR annotation categories are ordered similarly to Table 2.6, but with all categories of annotation represented in loosely decreasing order of severity, and VEP categories are ordered to correspond (as far as possible) with their matching ANNOVAR categories. This column-normalized heatmap allows us to see which categories of annotation are overrepresented (relative to the total number of variants in the column/category) in the VEP annotations for each category (i.e. column) of ANNOVAR annotation. Contrast this figure with Figure 2.14 and compare with Supplementary Table 2, which provides the counts used for this heatmap.

To characterise the sorts of apparent errors or inconsistencies that commonly emerge in

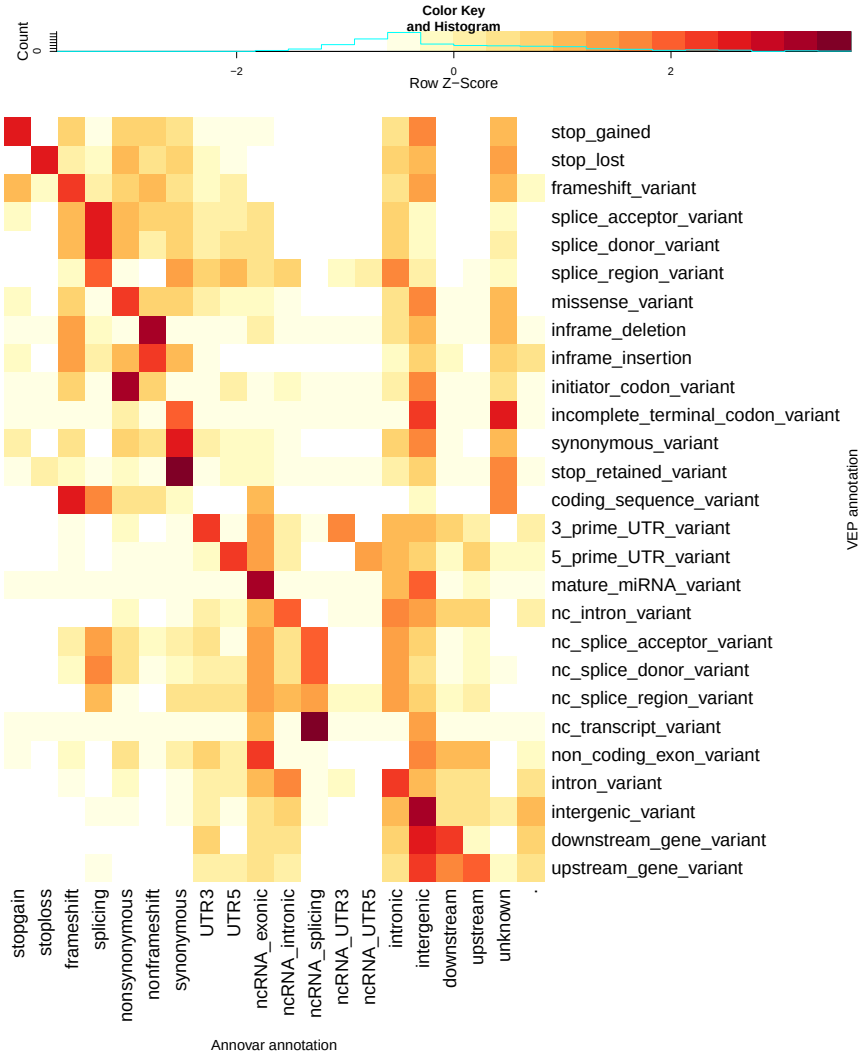


Figure 2.14: **VEP-normalized heatmap:** This heatmap shows scaled numbers of variants for all different combinations of categories of annotations when using the VEP annotation tool (rows) and the ANNOVAR software (columns), with the ENSEMBL database. The values (log10 of the count of variants with that combination of annotations from the two tools, with an offset of 1 applied; see Supplementary Table 2 for raw counts) are Z-scaled (mean-centred, divided by standard deviation) by row (i.e. standardising VEP annotations). The key above the heatmap shows the values indicated by different colours. Categories are ordered as per Figure 4. This row-normalized heatmap allows us to see which categories of annotation are overrepresented (relative to the total number of variants in the row/category) in the ANNOVAR annotations for each category (i.e. row) of VEP annotation. Contrast this figure with Figure 4 and compare with Supplementary Table 2, which provides the counts used for this heatmap.

annotation by ANNOVAR and VEP, I investigate cases for which annotations from ANNOVAR and VEP disagree. Although it is counter-intuitive (since the annotations were based on the same set of transcripts), ANNOVAR and VEP do not always use the same transcript for the annotation of a variant. This is a result of the interaction of different annotation categories, different precedence rules and the fact (for this study) of reporting only one consequence for each variant.

MAF Range	Total	Matching	Match Rate
0–1%	67327521	61819362	91.82
1–5%	5802063	5308312	91.49
5–10%	1656880	1510709	91.18
10+%	6209280	5667996	91.28

Table 2.7: **Same transcript set, different software:** This table summarises the number of annotations that match between the ANNOVAR and VEP results for different ranges of minor allele frequency (MAF). The match rates across the different ranges of MAF are not substantially different from the match rate across all variants.

I focus on loss-of-function variants—frameshift, stop-gain, stop-loss and splicing—as they are currently of most interest in disease studies, and better than 90% agreement is observed between ANNOVAR and VEP annotations for nonsynonymous and synonymous variant categories (Table 2.6). Where possible (as in the case of splicing annotations), I discuss differences in annotation algorithms that are likely causes of differences in annotation, but detailed information on annotation algorithms is not available for ANNOVAR or VEP, even in online documentation (Wang, 2013; Ensembl, 2013c). The numbers of variants that are given a particular annotation by ANNOVAR and not VEP, and vice versa, are inspected, as are the categories that most frequently feature amongst the "disagreeing" annotations. Further, I examine (in the ENSEMBL Web Browser; browser results not shown) some specific examples of variants with disagreeing annotations to try to characterise the causes of some of the differences in annotations from ANNOVAR and VEP. When characterising differences and apparent errors in annotation, I focus particularly on variants for which it is known ANNOVAR and VEP used the same transcript as the basis for annotation. These cases are better able to give us insight into the different characteristics of the differing annotations.

2.3.2.1 Frameshift variants

ANNOVAR annotates 15,822 variants as "frameshift", of which 13,486 (85%) also receive a "frameshift" annotation from VEP (Table 2.6). This leaves 2,336 variants annotated as "frameshift" by ANNOVAR that get a different annotation from VEP. Of these, both tools used the same transcript for 2,015 variants and they used different transcripts for the remaining 321 variants (Table 2.8).

When matching transcripts were used, 905 of the 2,015 variants were annotated as "missense" by VEP, and 257 as "synonymous" (the two most common categories; Table 2.8a). Remarkably, 54% of these variants are single nucleotide variants (data not shown), which means that ANNOVAR annotates over 300 variants as frameshift despite them being SNVs, so the ANNOVAR annotation is unequivocally incorrect for these variants. For the majority

of these variants, however, it is not possible to say conclusively from manual inspection whether the ANNOVAR or VEP annotation is correct.

There are some cases where neither tool seems to get the annotation correct. For example, the variant 1:52499090_G_GGGTTCT is a 6bp insertion, so it seems that "inframe insertion" would be the best annotation, but it is annotated as "frameshift" by ANNOVAR and "missense" by VEP. From looking at specific examples it appears that only a tiny fraction of variants get an incorrect annotation from both software tools.

There are 16,685 variants that VEP annotates as "frameshift", of which 13,486 (81%) also receive a "frameshift" annotation from ANNOVAR (Table 2.6). This leaves 3,199 variants annotated as "frameshift" by VEP that get a different annotation from ANNOVAR. The two tools used the same transcript for 1,300 of these variants and different transcripts for the remaining 1,899 variants (Table 2.8b. All of the variants annotated as "frameshift" by VEP are indels (not single nucleotide variants) and none are a multiple of three bases, so VEP looks to be correctly identifying these variants as frameshift indels.

Several hundred variants get "nonframeshift", "nonsynonymous" and "synonymous" annotations from ANNOVAR, which are incompatible with the frameshift annotations from VEP. When matching transcripts are used, 394 variants of the 1,300 variants are annotated as "nonframeshift" by ANNOVAR, 437 as "stopgain" and 283 as "nonsynonymous" (Table 2.8b). The "nonframeshift" and "nonsynonymous" annotations from ANNOVAR are incompatible with the "frameshift" annotations from VEP and the "frameshift" annotations seem reasonable, so ANNOVAR seems to give incorrect annotations for these variants.

The several hundred variants annotated as "stop-gain" by ANNOVAR and frameshift by VEP provide an interesting case study. A stop-gain annotation is not necessarily incompatible with a frameshift annotation from VEP, as ANNOVAR inspects the transcript produced by the insertion/deletion and sometimes finds that a stop codon is introduced by the indel. Following its precedence rules, it then returns an annotation of stop-gain rather than frameshift. The disagreement between annotations for such variants is thus reasonable once one takes into account how the two tools report annotations. From looking at specific examples it appears that only a small fraction of variants get an incorrect annotation from both software tools.

2.3.2.2 Stop-gain variants

ANNOVAR annotates 14,960 variants as "stopgain" and 96% of these are annotated as "stopgain" by VEP too (Table 2.6). This leaves 612 variants that get different annotations from VEP. The two tools use the same transcript to annotate 570 of these variants, and different transcripts for the remaining 42 variants (Table 2.9).

When matching transcripts are used, 437 of the 570 variants with discrepant annotations are given "frameshift" annotations by VEP. It was noted above that ANNOVAR's

Table 2.8: Differences in annotations for variants annotated as frameshift by only one of ANNOVAR or VEP.

Consequence	Matching_Tx	Mismatching_Tx
3_prime_UTR_variant	18	7
5_prime_UTR_variant	10	2
coding_sequence_variant	102	2
inframe_deletion	37	18
inframe_insertion	217	12
initiator_codon_variant	8	1
intron_variant	9	33
missense_variant	905	39
nc_intron_variant	0	13
nc_splice_acceptor_variant	0	32
nc_splice_donor_variant	0	23
nc_splice_region_variant	5	11
non_coding_exon_variant	24	4
splice_acceptor_variant	113	38
splice_donor_variant	178	58
splice_region_variant	52	8
stop_gained	74	4
stop_lost	5	1
stop_retained_variant	1	0
synonymous_variant	257	13
upstream_gene_variant	0	2
TOTAL	2015	321

(a) VEP consequences for variants that were annotated as frameshift by ANNOVAR but not VEP.

Consequence	Matching_Tx	Mismatching_Tx
.	0	16
downstream	0	2
intergenic	0	1115
intronic	0	94
ncRNA_exonic	0	5
ncRNA_intronic	0	1
nonframeshift	394	27
nonsynonymous	283	12
splicing	28	40
stopgain	437	14
stoploss	25	0
synonymous	133	4
unknown	0	518
UTR3	0	18
UTR5	0	33
TOTAL	1300	1899

(b) ANNOVAR consequences for variants that were annotated as frameshift variants by VEP but not ANNOVAR.

precedence rules can lead it to give a "stopgain" annotation to an indel for which "frameshift" would otherwise be a reasonable annotation. Here too, all of the variants annotated as "frameshift" by VEP seem to be genuine frameshift variants (as they are indels that are not a multiple of 3bp in size). These discrepancies, therefore, reflect a difference in precedence for reporting annotations, rather than a true difference between the annotation algorithms, and the ANNOVAR annotation (assuming it correctly identifies introduced stop codons) adds information of interest. There is a much smaller number of variants given "missense" (77) and "synonymous" (39) annotations by VEP (Table 2.9a). Manual inspection in the ENSEMBL Genome Browser of ten of those discrepant variants on chromosome 1 shows that for eight of the ten "missense" (from VEP) variants, the VEP annotation looks correct (for two variants neither annotation looks correct). Of the two given "synonymous" annotations by VEP, the ANNOVAR annotation looks correct for one and the VEP annotation looks correct for the other. For other variants with discrepant annotations, manual inspection reveals that the VEP annotation looks correct more often than the ANNOVAR annotation (data not shown).

VEP annotates 16,146 variants as "stop-gain" and 89% of these are also annotated as "stopgain" by ANNOVAR (Table 2.6). Thus, 1,798 variants get different annotations from ANNOVAR, of which the two tools used the same transcript for 225 variants and different transcripts for 1,573 variants (with 91% of those getting an "intergenic" or "unknown" annotation from ANNOVAR; Table 2.9b).

When matching transcripts are used, the breakdown of the ANNOVAR annotations for

the 225 variants is 74 "frameshift", 52 "nonframeshift", 77 "nonsynonymous" and 22 "synonymous" (Table 2.9b). Approximately 20% (30) of the "frameshift" or "nonframeshift" (from ANNOVAR) variants are single nucleotide variants, so an annotation of "frameshift" or "nonframeshift" cannot be correct. I conclude that these ANNOVAR annotations must be a result of a software bug. For the remaining genuine indels, VEP finds a stop codon introduced where ANNOVAR does not—these variants are difficult to assess by eye to determine which annotation is correct. The "frameshift" and "nonframeshift" annotations from ANNOVAR are reasonable, but if the VEP annotations are correct then annotating the variants as "stop-gain" provides more useful information. Checking the 8 variants on chromosome 1 given "nonsynonymous" or "synonymous" annotations by ANNOVAR indicates that the VEP annotation is correct for 5 variants, and for the other 3 the correct annotation was not possible to determine by visual inspection of the variant in the ENSEMBL Genome Browser. For the remaining variants it is difficult to assess whether the ANNOVAR or VEP annotation is better. Even after taking into account the differences in annotation caused by different precedence rules, the stop-gain annotations from VEP look more reliable than those from ANNOVAR.

Table 2.9: Differences in annotations for variants labelled as stop-gain by either ANNOVAR or VEP.

Consequence	Matching_Tx	Mismatching_Tx
3_prime_UTR_variant	0	2
frameshift_variant	437	14
inframe_insertion	7	0
intron_variant	0	1
missense_variant	77	4
nc_intron_variant	0	6
nc_splice_region_variant	1	0
non_coding_exon_variant	6	4
splice_acceptor_variant	2	4
splice_donor_variant	1	1
synonymous_variant	39	6
TOTAL	570	42

(a) VEP consequences for variants annotated as stop-gain by ANNOVAR but not VEP.

Consequence	Matching_Tx	Mismatching_Tx
downstream	0	1
frameshift	74	4
intergenic	0	1271
intronic	0	31
ncRNA_exonic	0	2
nonframeshift	52	11
nonsynonymous	77	55
splicing	0	3
synonymous	22	33
unknown	0	154
upstream	0	1
UTR3	0	4
UTR5	0	3
TOTAL	225	1573

(b) ANNOVAR consequences for variants annotated as stop-gain by VEP but not ANNOVAR.

2.3.2.3 Stop-loss variants

There are only small numbers of variants that are annotated as "stop-loss" by ANNOVAR and not by VEP, but almost all of these are annotated as "frameshift" by VEP. ANNOVAR annotates 906 variants as "stop-loss" of which 870 are also annotated as "stop-loss" by VEP (Table 2.6). Therefore, only 36 variants get different annotations from VEP, 30 of which were annotated with the same transcript by both tools, and 6 of which were annotated with different transcripts.

When matching transcripts are used, 25 of the 30 variants are given "frameshift" annotations by VEP (Supplementary Table 2.10a). These "frameshift" variants are all indeed

indels that are not a multiple of three bases—as such, annotations of "frameshift" from VEP are reasonable. Across a selection of 22 of the 30 variants, it seems that ANNOVAR gives the best annotation for 8 variants, VEP gives the best annotation for 11 variants and neither appears to give the best possible annotation for the remaining 3 variants. Thus, for these variants there is a roughly even split between when the ANNOVAR and the VEP annotation look better.

There are only 16 variants that are annotated as "stop-loss" by VEP and as something else by ANNOVAR when the two tools use the same transcript for annotation (Table 2.10). VEP annotates 1,077 variants as "stop-loss", of which 870 (81%) are also annotated as "stop-loss" by ANNOVAR (Table 2.6). That leaves 207 variants that are given different anntoations by ANNOVAR. The two tools use the same transcript for only 16 of these variants, using different transcripts for the other 191 (Table 2.10b). When matching transcripts are used one therefore sees a negligibly small number of variants that are not given a "stop-loss" annotation by ANNOVAR when they are annotated as "stop-loss" by VEP. These results support the notion that VEP is doing an excellent job of annotating "stop-loss" variants.

Table 2.10: Differences in annotations for variants labelled as stop-loss by either ANNOVAR or VEP.

Consequence	Matching_Tx	Mismatching_Tx
3_prime_UTR_variant	1	0
frameshift_variant	25	0
missense_variant	1	1
nc_intron_variant	0	3
non_coding_exon_variant	1	0
splice_donor_variant	0	1
stop_retained_variant	2	0
synonymous_variant	0	1
TOTAL	30	6

(a) VEP consequences for variants annotated as stop-loss by ANNOVAR but not VEP.

Consequence	Matching_Tx	Mismatching_Tx
frameshift	5	1
intergenic	0	40
intronic	0	17
nonframeshift	5	4
nonsynonymous	4	32
splicing	0	2
synonymous	2	14
unknown	0	77
UTR3	0	3
UTR5	0	1
TOTAL	16	191

(b) ANNOVAR consequences for variants annotated as stop-loss by VEP but not ANNOVAR.

2.3.2.4 Splicing variants

The category (or categories) of splicing variants is a source of many differences in annotations from different annotation software tools. Unlike most other categories of annotation, there are still multiple notions in the field of what entails a "splicing" variant. ANNOVAR defines just one broad category, "splicing", for these variants: any variant within x-bp of a splicing junction receives the annotation "splicing". The value of x can be specified by the user of ANNOVAR, and for the annotations here a broad definition of splicing was used, by setting $x = 6$. In contrast, VEP uses three categories of splicing variant: (1) "splice donor variant", a splice variant that changes the 2-base region at the 5′ end of an intron; (2) "splice

acceptor variant", a splice variant that changes the 2-base region at the 3' end of an intron, and (3) "splice region variant", a sequence variant in which a change has occurred within the region of the splice site, either within 1–3 bases of the exon or 3–8 bases of the intron. VEP thus gives more useful information, through its subcategories of "splicing" variants, about the likely function of a variant.

Differences in annotation can also arise simply as a result of differing definitions of what a "splicing" variant is, rather than any truly substantial differences in the algorithms producing the annotations. I investigated these differences in annotation on variants where both tools used the same transcript for annotation, and annotations did not match, that is, a variant with a "splicing" annotation from ANNOVAR did not get an annotation of one of "splice donor variant", "splice acceptor variant" or "splice region variant", or the inverse.

ANNOVAR annotates 45,839 variants as "splicing", of which 86% receive a splicing annotation from VEP (Table 2.6). Of those 6,259 variants with differing annotations from VEP, the two tools used the same transcript for 119 variants and different transcripts for the remaining 6,140 variants. The major source of difference in splicing annotations is that the overwhelming proportion of ANNOVAR "splicing" variants that receive non-splicing annotations from VEP actually receive one of VEP's three splicing annotations, but reported as being in a non-coding transcript (Table 2.11a). This result suggests that VEP does a better job at reporting when the transcript it uses for annotation is non-coding, but that there may actually not be such a large degree of difference between splicing annotations as appears initially.

VEP annotates 62,853 variants in one of its three "splicing" categories, of which 63% are also annotated as "splicing" by ANNOVAR. Thus, there are 23,273 variants that are annotated as "splicing" by VEP and not by ANNOVAR. Of these, the two tools used the same transcript for 3,540 variants (15%) and different transcripts for the other 19,733 variants.

When matching transcripts are used, 88% of the 3,540 variants are given "synonymous" annotations by ANNOVAR (Table 2.11b). Looking closely in the ENSEMBL Web Browser at 20 of these "synonymous" (according to ANNOVAR) variants reveals them all to be annotated as "splice region variant" by VEP, and all are in an exon, either in the first 3 bases (5' end) or last 3 bases (3' end) of the exon. Thus, these annotation differences seem to be a systematic result of differences in the annotation algorithms used by ANNOVAR and VEP, and for these variants the VEP annotations look to be preferable.

These splicing variants highlight the combined effect of different definitions of splicing variants and precedence rules that result in a splicing variant found in one transcript being reported instead of a less "serious" variant seen in another transcript.

Table 2.11: Differences in annotations for variants labelled as splicing by either ANNOVAR or VEP.

Consequence	Matching_Tx	Mismatching_Tx
coding_sequence_variant	28	3
frameshift_variant	28	40
inframe_deletion	0	1
inframe_insertion	9	9
intergenic_variant	0	3
intron_variant	0	1
missense_variant	1	21
nc_intron_variant	0	2
nc_splice_acceptor_variant	11	630
nc_splice_donor_variant	6	1169
nc_splice_region_variant	36	4249
non_coding_exon_variant	0	2
stop_gained	0	3
stop_lost	0	2
upstream_gene_variant	0	5
TOTAL	119	6140

(a) VEP consequences for variants annotated as splicing by ANNOVAR but not VEP.

Consequence	Matching_Tx	Mismatching_Tx
.	0	6
downstream	0	7
frameshift	343	104
intergenic	0	153
intronic	0	13375
ncRNA_exonic	0	514
ncRNA_intronic	0	659
ncRNA_splicing	2	5
ncRNA_UTR3	0	51
ncRNA_UTR5	0	100
nonframeshift	74	49
nonsynonymous	15	386
stopgain	3	5
stoploss	0	1
synonymous	3103	1020
unknown	0	125
upstream	0	24
UTR3	0	884
UTR5	0	2265
TOTAL	3540	19733

(b) ANNOVAR consequences for variants annotated as splicing by VEP but not ANNOVAR.

2.4 Discussion

The results of the comparison of annotations obtained using REFSEQ and ENSEMBL transcript sets emphasise the importance of the choice of transcript set used for annotation. Applying the same annotation software with different transcript sets saw a matching rate of 44% for putative loss-of-function annotations. Through these results we see that the choice of transcript set has a large effect on variant annotations, especially on variants of most interest. In this study we used the "default" transcript sets provided by REFSEQ and ENSEMBL. More selective choice of transcripts would alter the comparison results, and could also improve annotation results in many practical settings.

Though not done here, transcript sets from REFSEQ and ENSEMBL (or other sources) can be restricted to a subset of transcripts to exclude low confidence annotations. For many applications, the cost of a false positive (following up an interesting variant in a transcript that proves not to exist, not to be expressed, or similar) is higher than missing a possibly-relevant variant. In such cases, one would want to be very selective with the choice of transcripts, providing more confidence in the relevance of variants that obtain LoF or missense annotations. For example, excluding transcripts expected to undergo nonsense-mediated decay would be sensible, since any consequence of a variant in such a transcript is unlikely to be of functional importance or particular clinical or biological interest. Making better use of information about canonical transcripts and consensus coding sequence, or demanding the highest-level of experimental support for a transcript to be included in the set used for annotation could further increase confidence in variant annotations. Projects like GENCODE aim to provide a carefully curated transcript set supported

by experimental evidence (Harrow et al., 2006; Coffey et al., 2011; Derrien et al., 2012; Harrow et al., 2012). As we improve the quality and precision of our transcript sets, so will variant annotations improve.

Where a specific tissue of interest is known, annotation could be restricted to use only the set of transcripts known to be expressed in that tissue. Defining a targeted set of transcripts will not always be easy, but for sequencing studies where the cost of false positives (e.g. through follow-up experiments) is high, and where information on the expression of specific transcripts exists, a set of high-confidence transcripts tailored to the study at hand may be preferable. Incorporating information about tissue-specific expression from, for example, the GTEx (Lonsdale et al., 2013) and ENCODE projects should allow us to build tissue-specific transcript sets, which may be of use for particular studies.

Through efforts such as GENCODE, GTEx, and ENCODE, as well as the ongoing work to constantly improve the REFSEQ, ENSEMBL and UCSC databases, we may see annotation results converge as (ideally tissue-specific) transcript sets align across different repositories. For the time being, though, large differences remain.

Variant annotation remains challenging for current software tools—differing choices made in annotation packages on how to analyse, categorise and prioritise annotations for a variant lead to differing annotations from different tools, even when using the same set of transcripts as the basis for annotation. Differences in annotations from different software tools (e.g. 64% overall agreement for LoF annotations) are not as large as those seen when using different transcript sets (44% overall agreement for LoF annotations), and are often caused by differences in the annotation categories defined by different tools. Nevertheless, the extent of the differences seen shows that, again, careful consideration must be given when choosing a software tool to make sure that it is well suited to the goals of the scientific investigation.

Standardising definitions of variants across the field, to reduce the scope for apparent differences in annotations returned by different software tools and to crystallise the (epistemic) meaning of terms used for annotations, could be of value. The Global Alliance for Genomics and Health (GA4GH) project (GA4GH Project, 2015) could help to achieve this goal. Although currently in its early stages, a Variant Annotation Task Team has been established. This team is already working on the standardisation of annotation terms and formats for reporting of annotations from software tools. The team may also undertake systematic benchmarking of variant annotation approaches, extending the work discussed here. In the results here, for example, differing definitions of splicing variants cause tens of thousands of annotation differences. The Sequence Ontology Project (Eilbeck et al., 2005) may help with this. Indeed, VEP and SNPEFF have now adopted SO terms for reporting annotations, and we may hope that other software tools will follow suit. The SO terms themselves likely need expansion and further refinement to ensure that the information

that we would like to express about variant annotations can be expressed. However, with interaction between the SO team and developers of annotation tools (possibly under the umbrella of GA4GH) we should see the suitability of SO terms for summarising variant annotations to continually improve.

It would be beneficial for phase information to be used in annotating variants in close proximity, given the extent of "rescue" of LoF variants by nearby variants (MacArthur et al., 2012). LoF rescue can occur when, for example, a frameshift insertion occurs and then, shortly downstream another, in-phase, insertion or deletion occurs that restores the transcript's proper reading frame. Without phase information, one may conclude that both variants are LoF variants. However, if when accounting for the phase information, it becomes apparent that the second variant "rescues" the first LoF variant, and the combined consequence of the two variants is more akin to that of a missense variant, and one would not see the function of the transcript lost. Variant phasing can be computationally expensive, but for some applications could provide very useful information.

Currently, annotation tools typically do not associate any measure of uncertainty with reported variant annotations. Such information could be useful for downstream analysis, especially for consideration when allocating resources for follow-up experiments on variants of interest. When a high level of certainty about the validity of an annotation is required, annotations could be obtained with two software tools and variants with differing annotations could be flagged to be treated with caution. Information about variant frequency is commonly used with variant annotations when prioritising interesting variants for follow-up study. Resources like the 1000 Genomes Project (1000 Genomes Project Consortium, 2010), Exome Variant Server (NHLBI GO Exome Sequencing Project (ESP), 2015) and, most recently, the Exome Aggregation Consortium (EXAC) (Exome Aggregation Consortium (ExAC), 2015) contain a great deal of information about variant frequency. It may be possible to incorporate frequency information more formally into variant annotations. EXAC may prove to be most useful for this purpose, given its explicit focus on linking variants to function at a very large scale.

In the comparison of annotation tools here, I restricted each tool to report only the annotation of most severe consequence for each variant. This was necessary to directly compare ANNOVAR and VEP and to avoid comparisons becoming too unwieldy. However, VEP and other annotation tools can (and often by default do) report annotations for all transcripts, providing extra information that is often valuable. Adding this extra information, as with utilising phase information or tissue-specific transcripts, increases the challenges of data processing and interpretation by adding complexity to the treatment of variant annotation, but with good reason—this added complexity reflects the underlying biology, so taking this information into account potentially adds significant value to analyses of DNA variants.

Our understanding of the human genome continues to improve rapidly as we gain a better appreciation of the genome's complexity. As a result, at some point we may see the variant annotations from different approaches converge. For the time being, though, we confront an epistemic challenge (determining the meaning or function of variants observed) because our ontological foundations (knowledge and understanding of what all sequences in the genome actually do) remain unresolved or unclear. Thus, the choices of transcript set and software tool can have substantial effects on the annotation results obtained, and from there, large effects on all downstream aspects of the analysis of whole-genome and whole-exome data. Variant annotation is not yet a "plug-and-play" procedure and should not be treated as such.

In addition to different variant annotation approaches (of which there are many more than I have compared here), there are different sequencing technologies, read mappers and variant callers. Each of these can potentially have substantial impact on the final variants and annotations obtained, but comparison of other sources of variation is beyond the scope of this work. Interested readers may refer to systematic comparisons of other aspects of the NGS pipeline, for example, comparisons of benchtop high-throughput sequencing technologies (Loman et al., 2012), short-read mappers (Hatem et al., 2013), variant callers (Yu & Sun, 2013) and variant-calling pipelines as a whole (Jason et al., 2013; Pabinger et al., 2014).

I have aimed to highlight the impact on final annotation results that can arise from two aspects of analyses of whole genome (or whole exome) sequence data, namely choice of transcripts and choice of annotation software. While I am not advocating any particular software or transcript set, I suggest researchers be aware of the impact of these choices and tailor their decisions to suit the particular requirements of any given study. I hope these comparisons may inform such decisions.

2.5 Conclusions

This study quantified the extent of disparity in variant annotation when different transcript sets and different software tools are used. This comparison of annotations for 80 million human DNA variants revealed many substantial differences between annotations based on different transcript sets and different software tools. The extent of differences in annotations was particularly large in annotation categories of most interest, namely putative loss-of-function and nonsynonymous variants. I found many more variants with annotations in interesting categories when using ENSEMBL transcripts compared with REFSEQ transcripts only. If it is important not to miss potential loss-of-function variants, then there are advantages to using ENSEMBL transcripts. If it is important to reduce false positives, then a carefully curated set of transcripts tailored to the study at hand may be preferred. Even when using the same transcript set, different annotation software packages can provide substantially different annotations.

There are variants with potentially severe effects that are identified with one method and not another. We require consistent, accurate and reliable annotation of variants to support the use of whole-genome and whole-exome sequencing in making diagnostic and treatment decisions. The dependence of current annotation results on the set of transcripts and software used can be managed, with sufficient care, in the research context. However, more work is required to improve variant annotation for clinical use. The differences in annotation due to choice of transcript set and software package quantified here should be given due consideration when undertaking variant annotation in practice. Careful thought needs to be given to the choice of transcript sets and software packages for variant annotation in sequencing studies. As discussed, making blanket recommendations is difficult, but from the comparisons of annotation approaches conducted in this chapter I would recommend the VEP software in preference to the ANNOVAR software, because: (1) detailed investigation of discrepant annotations from the two packages revealed VEP to provide a better annotation more often than ANNOVAR; and (2) VEP offers more refined variant annotation categories, particularly for splice variants, which is valuable in practice. Choice of transcript set should be more study-specific, but for general research purposes when annotating human genome data I would recommend the GENCODE set of transcripts (Harrow et al., 2012), as it is very carefully curated and I think it strikes a good balance between sensitivity (detecting potentially interesting annotations) and specificity (interesting annotations are "real" or relevant to actual biology). The GENCODE set of transcripts is available as a subset of the ENSEMBL transcripts.

Chapter 3

Estimating the heritability of type 2 diabetes susceptibility using whole-genome sequence data

"Man may be the captain of his fate, but he is also the victim of his blood sugar" — Wilfrid Oakley (Oakley, 1962)

3.1 Background and introduction

This chapter and the next investigate the heritabilty of type 2 diabetes (T2D) and the relative contributions of different classes of genetic variant to explaining variance in susceptibility to T2D. Linear mixed model (LMM) methods are applied to partition variance in T2D susceptibility using whole-genome sequence data on a set of 2,700 individuals with and without T2D. The chapter begins with an introduction to the study of T2D and an overview of research into the genetics of T2D. I introduce the concept of heritability and the use of LMMs in genetics and then discuss the datasets available for analysis. Specific methods used for estimating heritability using LMMs are described and then I discuss the implementation of the analysis, including quality control. In this chapter, I present the results for estimating the heritabiity of T2D with a single variance component, and then discuss the robustness of these results.

In the next chapter, I present the analyses using variance partitioning to investigate the contribution of different classes of genetic variation to T2D. I describe the approach taken for variant annotation for this study and introduce the methods that I use to assess the relative importance of different classes of genetic variation to T2D susceptibility. The results sections begin by presenting a streamlined version of the analysis that focuses on presenting the key findings. I then replicate the key findings using a second, larger cohort with imputed data. Following those results I present a section with detailed discussion of

the robustness of the results to changes to many different modeling assumptions, including permutation and other empirical testing results probing the behaviour of the models and the significance of the results. Finally I check corresponding results in sub-populations of the complete cohort.

The work presented in this and the next chapter was conducted as part of the Genetics of Type 2 Diabetes (GoT2D) project. Assigning appropriate credit in a large collaborative project like this can be complicated. The work presented here is primarily joint work between Loukas Moutsianas and myself, but utilises data from the GoT2D project prepared by others. Broadly, all credit for ascertaining, sequencing, and genotyping individuals, integrating data across platforms, calling variants, applying quality control and integrating genotypes into haplotypes goes to "the GoT2D project"—Loukas and I made use of the "Integrated Panel" data (after QC) for our analyses. Kyle Gaulton was responsible for the annotation of variants, as described in Section 4.2. He also provided us with the imputed data where variants from the 1000 Genomes Project (1000 Genomes Project Consortium, 2010; Consortium et al., 2012) were used as the reference panel.

Loukas Moutsianas and I planned, discussed and interpreted analyses and results together, but there was a broad split in our contributions. Loukas was primarily responsible for organising and managing the genotype data and generating lists of annotated variants that define the classes of variants that used across analyses. Loukas also generated the imputed data using GoT2D as the reference panel and imputing variants into a larger UK cohort (see Section 3.2.5). I was responsible for the statistical analyses, running the linear mixed model software, compiling results, deriving necessary theoretical results, computing enrichment scores, and plotting results and producing display items. I wrote up the chapters presented here, with advice from Loukas on structure and content. Loukas did, however, also run statistical analyses using only SNPs passing quality control in the first Wellcome Trust Case-Control Consortium study and using genotype dosages rather than hard calls. Conscious attempt is made below to give appropriate credit for contributions, especially when contributions differ from the schema described here. When the term "I" is used below, it should be taken to include contributions from Loukas as above. I will try to be specific when others deserve credit.

3.1.1 Introduction to the study of type 2 diabetes

Diabetes mellitus (diabetes) is a group of metabolic diseases characterised by inadequate production or utilisation of the hormone insulin resulting in high blood sugar levels (hyperglycaemia) persisting over a prolonged period. Insulin regulates how the body uses and stores glucose and fat. Crucially for diabetes, insulin signals to the body's cells for them to take in glucose and use it as energy (Figure 3.1). If insufficient insulin is produced, or if cells become resistant to insulin, glucose is not taken up from the blood resulting in

prolonged high blood glucose levels. In the final decade of the nineteenth century, Oskar Minkowski and Joseph von Mering identified the crucial role of the pancreas in diabetes, when they discovered that removing the pancreas caused diabetes in dogs (von Mering & Minkowski, 1889, 1890; Minkowski, 1893). In the early 1920s, Frederick Banting, Charles Best, John MacLeod and Bertram Collip discovered that insulin was the critical product of the pancreas by isolating the hormone and demonstrating its efficacy as a treatment on diabetic dogs and rabbits (Banting et al., 1922a,b, 1923; Banting & Gairns, 1924). Banting and MacLeod were awarded the 1923 Nobel Prize in Physiology and Medicine "for the discovery of insulin" (Nobel Media AB, 2014). Since the first successful insulin treatment for human diabetes, administered to 14-year-old Leonard Thompson in Toronto in 1922 (Banting et al., 1922a), the production of synthetic insulin and its use has saved the lives of millions of people with diabetes. Nevertheless, the full aetiologies of the major forms of diabetes confound us to this day.

The different types of diabetes differ with respect to how insulin production or sensitivity, or both, are compromised. There are three main types of diabetes (type 1, type 2 and gestational) and many rarer types, such as various forms of maturity-onset diabetes of the young (MODY) and other monogenic disorders. Type 1 diabetes (T1D) is caused by an absolute deficiency of insulin secretion, often identified by evidence of an autoimmune pathologic process occurring in the pancreatic islets (American Diabetes Association, 2010). Typically, T1D occurs in children or young adults, although it can affect people of any age (International Diabetes Federation, 2013). T2D, in contrast, features insulin resistance, when the cells of the body no longer respond properly to insulin, usually combined with relative (rather than absolute) insulin deficiency (American Diabetes Association, 2010). The specific aetiology of T2D is not fully understood, and there are probably many forms of this disease. Indeed, T2D is in a sense a catch-all category for diabetes that does not feature autoimmune destruction of β-cells (as in T1D), present during pregnancy (gestational diabetes) or have the genetic and phenotypic hallmarks of a form of MODY or monogenic diabetes. (American Diabetes Association, 2010). Gestational diabetes features insulin resistance and glucose intolerance similar to T2D, but arises in women during pregnancy and entails risks for both mother and child (International Diabetes Federation, 2013). MODY denotes many hereditary forms of diabetes, typically characterised by an autosomal dominant mode of inheritance (Fajans et al., 2001). As such, the forms of MODY are sometimes referred to as "monogenic" forms of diabetes.

Despite our knowledge for almost a century of the link between insulin and diabetes , the incidence of diabetes worldwide has increased dramatically over the last hundred years due to changes in diet and lifestyle (Zimmet et al., 2001). Today approximately 350 million people are afflicted with the disease (International Diabetes Federation (2013) estimates 382 million people; World Health Organisation (2014) estimates 347 million), with almost

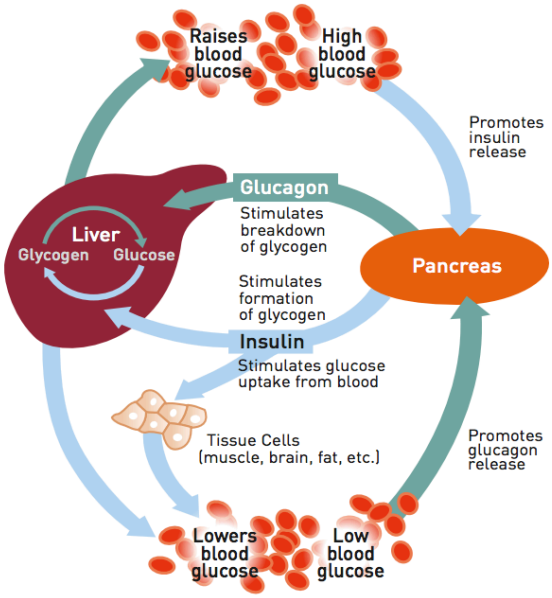


Figure 3.1: Insulin production and action. This figure provides an overview of metabolic interactions between the pancreas, liver and tissue cells that influence blood glucose levels. The pancreas produces the hormones insulin and glucagon. Both hormones affect formation and storage of glucose in the liver. Glucagon stimulates the breakdown of glycogen into glucose, raising blood glucose levels. High blood glucose then promotes insulin release in the pancreas. Insulin release stimulates formation of glycogen from glucose in the liver, and also stimulates glucose update from blood in tissue cells. Both of these processes lower blood glucose. Low blood glucose levels then promote glucagon release from the pancreas. If the pancreas produces insufficient insulin, or if tissue cells no longer respond properly to insulin and do not take up glucose appriopriately, then this can lead to chronic high blood glucose levels and diabetes. Reproduced with permission, courtesy International Diabetes Federation (2013).

half of these people remaining undiagnosed (International Diabetes Federation, 2013). All types of diabetes are increasing, T2D particularly, and the number of people living with diabetes is expected to rise to approximately 600 million by 2035 (International Diabetes Federation, 2013). In both human and financial terms, the burden of diabetes is immense. The International Diabetes Federation (2013) estimates that in 2013 diabetes cost US$548 billion in health spending (11% of the total spent worldwide) and caused 5.1 million deaths. Worldwide, 6% of deaths are caused by high blood glucose (Mathers et al., 2009). More than 80% of deaths from diabetes occur in low- and middle-income countries (Mathers & Loncar, 2006).

Of all the types of diabetes, T2D is by far the most prevalent, accounting for approximately 90% of all diabetes cases (World Health Organisation, 2014). Increasingly, across the world, children and adolescents are being diagnosed with T2D, previously considered a lifestyle disease of adults (Fazeli Farsani et al., 2013). T2D afflicts many people in developing countries and vulnerable populations, with particularly high prevalence rates among indigenous peoples of Australia, North and Central America and people in Mauritius, In-

dia, China and the island nations of the Western Pacific (International Diabetes Federation, 2013). A chronic disease, T2D can be managed effectively where healthcare services are adequate. Where they are not, T2D sufferers face a higher risk for the worst complications of the disease: cardiovascular illness, kidney disease, eye disease, nerve damage (which can lead to limb amputation), pregnancy complications, coma and death (International Diabetes Federation, 2013).

Tackling diabetes represents a daunting global health challenge, and our failure to understand the pathophysiology of T2D frustrates efforts to prevent and cure the disease (McCarthy, 2010). Genetics gives us another lens (along with epidemiology, physiology, molecular biology and clinical practice) through which to examine T2D. It is hoped that improving understanding of how genomic variation influences variation in T2D predisposition will provide clues to the processes involved in disease pathogenesis. Here, I address some of the challenges of developing a better understanding of the genetics of T2D.

3.1.2 The genetics of type 2 diabetes: an overview

T2D, in the public consciousness, is often seen as a lifestyle disease. The links between obesity, physical activity and T2D are well known and well publicised, while genetic influences on the disease are often overlooked. However, the simple observation that there are many obese people who do not have T2D and many non-obese people who do, shows that there is much more involved in the aetiology of T2D than just obesity, and allows for a substantial contribution from genetics (Frayling, 2007). Certainly, by the 1930s the idea that diabetes was at least partly heritable had been well established (Allan, 1933; Pincus & White, 1933, 1934a,b; Levit & Pessikova, 1934; White et al., 1934). This was driven by the observation of concurrence of diabetes in twins (Curtis, 1929, for example) and increased risk of diabetes in relatives of diabetic patients compared with a control population.

Available technologies limited the identification of causal, or markers for causal, variants until the 1980s when family-based linkage and candidate gene assocation studies became mainstream techniques (see Lander & Schork, 1994, for an overview). Since then, discovery of causal genes for diabetes has followed three main waves (McCarthy, 2010):

1. Family-based linkage analyses;
2. Tests of association for candidate genes; and
3. Systematic large-scale surveys of association between common DNA variants and disease (following the advent of the GWAS).

The field is now, possibly, entering a fourth wave in which relatively inexpensive whole-genome and whole-exome sequencing enables association studies on the full catalogue of genetic variation (or of coding variation in the case of whole-exome sequencing), extending GWAS to now include low-frequency, rare and structural variation.

Significant breakthroughs in diabetes genetics were made with family-based linkage analyses and candidate gene studies, which proved effective in identifying genes responsible for extreme forms of diabetes segregating as monogenic (Mendelian) disorders (McCarthy, 2010). For several distinct, familial forms of diabetes that did not have autoimmune characteristics, researchers characterised genes in which specific mutations are believed to cause the disease (see Waterfield & Gloyn, 2008). These discoveries yielded insights into processes for maintenance of normal blood glucose levels, energy balance, and the inner workings of the pancreatic β-cell and the hypothalamus. For many families, identification of the monogenic basis for disease improved their diagnostic and therapeutic options. However, attempts to apply similar approaches to families in which common forms of diabetes were segregating proved to be largely unrewarding (McCarthy, 2010).

An alternative approach to linkage studies, with different advantages and disadvantages, was to test association between genetic variants and disease status in unrelated individuals with and without diabetes. These tests of association are intrinsically more powerful than linkage studies, but signals of association are limited to variants that are directly assayed, or tagged by (correlated with) variants assayed (McCarthy, 2010). Depending on the technology used, assayed and tagged variants could include just a single locus or a substantial fraction of the genome. Prior to the cheap chip-based genotyping that enabled GWAS, researchers were limited to testing association for specific candidate variants or genes of interest.

The candidate gene approach, based on prior knowledge from biological and pharmacological studies of protein function, animal models, monogenic or syndromic forms of the disease and positional information from linkage studies, uncovered a small number of genuine susceptibility variants. The first unequivocal evidence for common variants involved in T2D (Frayling, 2007) came from candidate gene studies: the E23K variant in *KCNJ11* (Nielsen et al., 2003; Gloyn et al., 2003; Florez et al., 2004), the P12A variant in *PPARG* (Altshuler et al., 2000), and common variation in *TCF2* (Gudmundsson et al., 2007; Winckler et al., 2007) and *WFS1* (Sandhu et al., 2007). Each of these genes advanced understanding of known biology and drug treatments, as well as yielding new avenues of inquiry into biological processes underpinning T2D. For example, sulfonylureas and thiazolidinediones are classes of therapeutic agents widely and long-used in diabetes management. The prior knowledge of the targets of these agents led to the study of *KCNJ11*, which encodes a protein target for sulfonylureas, and *PPARG*, which encodes a protein target for thiazolidinediones, in candidate gene studies (Pacanowski et al., 2008). *KCNJ11* (potassium inwardly-rectifying channel, subfamily J, member 11) encodes the inwardly rectifying potassium channel (Kir6.2) subunit of the pancreatic β-cell ATP-sensitive potassium (K_{ATP}) channel (which controls insulin secretion). *PPARG* (peroxisome proliferator-activated receptor-γ) is a transcription factor involved in adipocyte differentiation. On the

whole, though, candidate-based tests of association were not very successful, hampered by lack of power (driven both by small sample size and lower than expected effect sizes) or focused on inappropriate candidates (Hattersley & McCarthy, 2005). A new approach was needed, which came in the form of genome-wide association studies.

A first proof of concept for the genome-wide association approach in T2D came through the identification of a T2D risk variant in *TCF7L2*. Grant et al. (2006) discovered this association by genotyping 228 microsatellite loci in Icelandic individuals with and without T2D, in a previously reported 10.5Mb linkage region on chromosome 10. This association provided indirect evidence that suggested GWAS could work and be valuable, because: (1) the variant found did not explain the linkage signal, indicating that association testing could find signals independent of those found by linkage approaches; and (2) the associated gene, *TCF7L2*, would not have been linked to T2D through candidate gene approaches. *TCF7L2* was a completely unexpected gene—it encodes a transcription factor expressed in the foetal pancreas and involved in the *WNT* signaling pathway—and no previous connection between the gene and T2D had been made, so it would not have been studied as a candidate gene (Frayling, 2007). The association in *TCF7L2* remains the strongest genome-wide signal for T2D even after many subsequent GWAS investigating T2D.

In 2007, the year following the *TCF7L2* discovery, six papers from five separate GWAS collectively identified six new loci associated with T2D (WTCCC, 2007; Sladek et al., 2007; Diabetes Genetics Initiative, 2007; Scott et al., 2007; Steinthorsdottir et al., 2007; Zeggini et al., 2007). In a study of obesity, Frayling et al. (2007) found that the *FTO* locus, which had just been associated with T2D, was also associated with body mass index (BMI). The six GWAS loci discovered in 2007 were (with descriptions from Frayling, 2007):

CDKAL1: *CDK5* regulatory subunit associated protein 1-like; highly expressed in human islets; shares sequence with *CDK5RAP1*, a known inhibitor of *CDK5* (implicated in reduced β-cell function) activation; association between *CDKAL1* risk allele and reduced insulin secretion;

CDKN2A-CDKN2B: cyclin-dependent kinase inhibitor genes; *CDKN2A* encodes p16INK, overexpression of which leads to decreased islet proliferation in ageing mice;

FTO: fat-mass and obesity associated gene; expressed in the hypothalamus;

HHEX-IDE: haematopoietically expressed homebox-insulin degrading enzyme; encodes a transcription factor with a key role in pancreatic development and a target of *TCF7L2*;

IGF2BP2: insulin-like growth factor 2 mRNA-binding protein 2; binds to IGFII and is expressed in the pancreatic islet

SLC30A8: a zinc transporter that is expressed in the β-cell.

These six loci cover a potentially wide range of physiology, with genes expressed from pancreatic islets (perhaps expected) to the hypothalamus (probably less so). Of these six

GWAS loci, only the *SLC30A8* variant (non-synonymous, changing an arginine to a tryptophan) has an obvious functional implication. That the remaining five fell outside coding regions is typical of what was to follow in subsequent GWAS across all traits. A GWAS SNP is likely just a tag for causative variants, so most associated loci do not have an obvious mechanism to affect protein function, necessitating fine-mapping of associated regions and substantial downstream functional work. Even with such work being done, most associated loci elude complete functional characterisation (Hindorff et al., 2009). Indeed, the example of *SLC30A8* itself neatly encapsulates the challenge. Previous functional studies, including mouse knockouts, were inconclusive in determining the physiology underpinning the assocation, and it took genotyping of loss-of-function mutations in 150,000 individuals across several populations to provide strong evidence that *SLC30A8* haploinsufficiency protects against T2D (Flannick et al., 2014).

About 80 loci have now been associated with T2D across the genome, from 39 studies recorded in the NHGRI GWAS Catalog (Welter et al., 2014a) up to the end of June 2013 (Morris et al., 2012; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2014). Whereas early GWAS studied individuals of European descent, subsequent studies broadened their reach to include people of Native American, Mexican, African, South and East Asian descent. Large consortia have succeeded in sharing data across studies and countries to conduct trans-ancestry meta-analyses, revealing shared and novel loci for T2D risk in different populations (see DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al., 2014, for the most recent, and largest, example). Details of the many recent T2D GWAS are too extensive to be presented here, but recent reviews by Kahn et al. (2014), Hivert et al. (2014), Hara et al. (2014), and Grarup et al. (2014) provide a survey for the interested reader of the current understanding of the genetic architecture and pathophysiology of T2D, as well as coverage of the now sizeable literature in this area.

The success of GWAS for T2D underlined the genetic complexity of the disease, but also highlighted how much remains to be discovered. One outstanding question for the genetic architecture of diabetes—as for most common traits—regards so-called "missing heritability" (Maher, 2008; Manolio et al., 2009; Eichler et al., 2010). Heritability estimates for T2D, which I will discuss further below, range from 30% to 70% in twin- and family-based studies (Poulsen et al., 1999; Almgren et al., 2011). However, the common genetic variants discovered through GWAS each impose only a modest risk increment for T2D and in total explain less than 10% of the variance in T2D risk. The question of where the heritability "missing" between estimates from GWAS loci and family studies resides has been a point of much discussion in the literature. Debate continues about the potential roles of rare variants, epistasis (interactions), parent-of-origin effects, and the collective effect of large numbers of common SNPs, and the best ways to attempt to resolve these

questions (Makowsky et al., 2011; Zuk et al., 2012, 2014; Hunt et al., 2013; Zaitlen & Kraft, 2012; Mott et al., 2014; Bloom et al., 2013). As an alternative to the family- and associated variant-based approaches to estimating heritability, Yang et al. (2010a) demonstrated the utility of variance components analysis using LMMs to obtain estimates of heritability from chip-genotype data in unrelated individuals.

The strength of the LMM approach comes from being able to consider all variants assayed as contributing to trait variance, rather than only computing variance explained from variants that reached very stringent thresholds for evidence of genome-wide assocation. The success of these methods, which I discuss in much greater detail below, motivates the use of variance components analysis with LMMs to investigate further the genetic architecture of T2D using the near-complete catalogue of genetic variation available from the GoT2D project. I will use these models to try to answer two questions:

1. How much do variants across the allele-frequency spectrum contribute to variance in T2D risk?
2. How much do variants in different functional classes contribute to variance in T2D risk?

Answers to these questions could have useful implications for the interpretation of associated variants and the design of future studies on the genetics of T2D.

3.1.3 Heritability and linear mixed models in genetics

The above summary of the study of genetics in T2D introduced us to a crucial concept in genetics: heritability. This concept is used (among other applications) to help understand the genetic component of risk to disease, independently of known environmental risk factors.

Heritability is the proportion of total variance in a population for a particular measurement (among comparable individuals) that is attributable to variation in genetic values (Visscher et al., 2008). In defining heritability, one refers to *additive* and *non-additive* genetic values. The additive genetic value (or breeding value) is defined as the sum of the average effects of parents' alleles that give rise to the mean genotypic value of their progeny (Visscher et al., 2008). The genotypic value for a polygenic trait is defined as the mean phenotypic value of all those individuals with that genotype in the population. Breeding values can be measured even when the average effects of individual loci cannot, a fact crucial for agricultural and livestock genetics. The individual average allelic effects (or additive effects) across loci are treated as independent of each other. The non-additive (or interaction) genetic value, in contrast, is an aggregate effect that is not simply the sum of the individual additive effects taken in isolation. Non-additive effects include dominance (interactions within a locus) and epistatic (interactions between loci) effects. Two easily confused types

of heritability are used: "narrow-sense" heritability (or just heritability, h^2), the proportion of total variance in a population for a particular measurement that is attributable to variation in additive genetic values, and "broad-sense" heritability (H^2), the proportion of total variance attributable to variation in total genetic values (including dominance and epistatic effects). To define these terms formally, I follow the treatment by Visscher et al. (2008).

According to biologically plausible nature-nurture models, observed phenotypes (P) of a trait of interest can be partitioned into a statistical model representing the contribution of the unobserved genotype (G) and unobserved environmental factors (E):

$$\text{Phenotype (P)} = \text{Genotype (G)} + \text{Environment (E)}. \tag{3.1}$$

This model assumes that genotype and environment are uncorrelated and do not interact. The observable phenotypic variance (σ_p^2), which usually excludes variation that is due to known fixed factors and covariates such as sex, age and cohort, can be expressed as a sum of unobserved underlying variances (σ_G^2 and σ_e^2):

$$\sigma_p^2 = \sigma_G^2 + \sigma_e^2. \tag{3.2}$$

Heritability is defined as a ratio of variances, expressing the proportion of the phenotypic variance that can be attributed to variance of genotypic values:

$$\text{Broad-sense heritability} = H^2 = \frac{\sigma_G^2}{\sigma_p^2}. \tag{3.3}$$

The genetic variance itself can be partitioned into the variance of additive genetic effects (also known as "breeding values"; σ_A^2), of dominance genetic effects (interactions between alleles at the same locus; σ_D^2), and of epistatic genetic effects (interactions between alleles at different loci; σ_I^2): $\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$. Thus we obtain:

$$\text{Narrow-sense heritability} = h^2 = \frac{\sigma_A^2}{\sigma_p^2}. \tag{3.4}$$

In general, σ_e^2 can be broken down into any number of identifiable, but random, contributing factors. However, here I use the simplest partitioning and let σ_e^2 represent the environmental residual variance, which includes individual stochastic error variance and measurement error. This partitioning of the phenotypic variance (Equation 3.2) assumes the absence of genotype by environment covariance ($\sigma_{G,e}$), and it ignores the interaction between the genotype and environment (G $*$ E). Although similar terms, G $*$ E covariation and G $*$ E interaction are different. G $*$ E interaction refers to different genotypes responding to environmental variation in different ways, whereas G $*$ E covariation is said to occur when exposure to environmental conditions depends on an individual's genotype, which can arise by both causal and non-causal mechanisms (Jaffee & Price, 2007). At a global scale there will surely be G $*$ E covariation, but both G and E covariation and G $*$ E interaction

are often ignored, especially in human genetic studies. This is usually because they cannot be estimated or one lacks power to study interactions at current sample sizes.

Studing G $*$ E interactions is underpowered with current sample sizes, because the required sample-size for finding G $*$ E interactions can be enormous. Testing for interaction terms in regression models drastically increases the multiple-testing burden, which is already high for testing marginal genetic effects in GWAS. A useful rule of thumb is that the detection of an interaction requires a sample size at least four times greater than that required for the detection of a main effect of comparable magnitude (Smith and Day, 1984). Thus, studies designed to have reasonable power to detect main genetic effects are unlikely to have adequate power to find interaction effects of a similar magnitude. Assuming that G $*$ E effects in human studies will be relatively modest (say interaction odds ratios of 0.8—1.2), then sample sizes in the hundreds of thousands will be required to detect them (Thomas, 2010). We are only just starting to see GWAS on this scale emerging through large consortium and meta-analysis projects (see Wood et al., 2014, for example). Following standard practice, then, G $*$ E covariation and interactions are ignored in analyses here, which means that if G and E covariation is present then estimates of σ^2_G will be inflated and if G $*$ E interaction is present σ^2_e estimates will be inflated. I will use the simple partitioning focusing on additive effects. Henceforth, I use σ^2_g to denote the variance of additive genetic effects (σ^2_A, above), and use "heritability" to refer to narrow-sense heritability.

Traditionally, heritability was estimated from simple and often balanced designs, such as simple functions of the regression of offspring on parental phenotypes, the correlation of full or half siblings, and the difference in the correlation of monozygotic and dizygotic twin pairs (Falconer & Mackay, 1996). However, when phenotypic measurements are available for individuals with a mixture of relationships, or in general when the design is unbalanced (as in GWAS), the most efficient way to estimate additive genetic variance and environmental components is to use an LMM.

LMMs are a well-established and widely-used element of the statistician's toolkit, even though, remarkably, variance components models were used by astronomers before they were known to statisticians (Airy, 1861; Chauvenet, 1863). An LMM is known as "mixed" because the dependent variable is a linear function of both fixed and random independent variables. Fixed effects are constant across the taking of repeated samples, whereas random effects are a sample from a distribution of effects. These models have a long history of use in genetics, in particular, tracing back to Fisher (1918) who introduced the term "variance" and implicitly used variance components in his classic paper. Often, aggregated genetic effects are treated as one or more random effects (sometimes called "polygenic" effects), as the individual genetic effects cannot be estimated, but the sampling distribution of the genetic effect(s) can be characterised.

LMMs play an important role in contemporary genetics. LMMs are primarily used in two ways: one in which the polygenic random effect is a nuisance parameter, and one in which it is of primary interest. The analysis of genome-wide association studies is an example of the first case, where LMMs have been shown to be effective in improving power and controlling for population structure and other technical artefacts by fitting a polygenic random effect that captures relatedness between individuals (Kang et al., 2008; Astle & Balding, 2009; Zhang et al., 2010; Kang et al., 2010; Pirinen et al., 2013; Lippert et al., 2011; Listgarten et al., 2012, 2013). However, in my analysis, I focus on the second type of case, in which the random effect itself is of primary interest. In this setting, I use the LMMs to estimate the proportion of variance in phenotype attributable to different types of genetic effects. The model of choice for this application was originally derived in the context of livestock genetics (Quaas & Pollak, 1980; Meuwissen et al., 2001; Goddard & Hayes, 2009), but Peter Visscher and colleagues have pioneered its use for quantifying the collective contribution of a set of genetic variants to variance in human traits and susceptibility to disease (Yang et al., 2010a, 2011b; Lee et al., 2010, 2011, 2012a,b). In this model, the focus is not on individual variant associations, but on fitting a random polygenic effect for each individual designed to capture aggregate additive effects of a set of variants. Genomic similarity is computed between all pairs of individuals in a way that accounts for additive genetic effects and is used to define the covariance structure for the random effect in the LMM.

More formally, given phenotypic values $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$, a typical form is

$$\mathrm{Var}(\boldsymbol{Y}) = \sigma_g^2 K + \sigma_e^2 I_N, \tag{3.5}$$

where K is a matrix of pairwise additive relationships between individuals, and I_N is the $N \times N$ identity matrix, which implicitly assumes independence across individuals of environmental effects and measurement error (Speed et al., 2012). In the LMM context, estimates $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ are typically obtained via residual (or restricted) maximum likelihood (REML) (Corbeil & Searle, 1976a). Under certain conditions, h^2 can be estimated by:

$$\hat{h}^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}. \tag{3.6}$$

We can derive Equation 3.5 by assuming a model in which all measured genetic markers contribute to the phenotype through (unknown) effects. Depending on the assumptions one makes about the effect sizes at the markers (or, equivalently, how one scales the observed genotypes at these markers), Equation 3.6 may not hold for estimating h^2. Specifically, if the genotypes have been normalised to have mean of zero and unit variance, then Equation 3.6 holds. This normalisation of genotypes implicitly assumes that the variance of genetic effect sizes increases as variant allele frequency decreases. If one makes any different assumption about the distribution of effect sizes (or transform the raw genotypes

in any different way), then h^2 cannot be estimated using Equation 3.6. In such cases, a more general form is available (Speed et al., 2012). The models I use for my analysis are discussed in much greater detail in Section 3.3.

Heritability is a widely-used but often misunderstood term. The definition of heritability means that it depends on the population, as both the variation in additive and non-additive genetic factors, and the environmental variance, are population-specific (Visscher et al., 2008). Only under specific conditions are the ratios of variance components estimated from linear mixed models interpretable as heritability estimates. Furthermore, estimates of heritability obtained from observed genotype data depend on the set of markers used and how well those markers tag the actual causal variants for the trait. Heritability estimates from SNP-chip genotype data are often referred to as "chip-heritability" estimates, to acknowledge this effect. Thus, it is often preferable to talk about the proportion of phenotypic variance explained (VE) by available genotypes. Later in this chapter, I explore possible formulations of mixed models, which make different assumptions about the effect sizes of the markers. I also discuss various considerations for and complications with applying these models to the analysis of the GoT2D data.

3.1.4 The Genetics of Type 2 Diabetes project

The work described in this chapter was conducted as part of the GoT2D project, a large-scale international collaboration. The consortium responsible for the project comprised over 80 researchers, primarily from the Broad Institute in Boston, US, the University of Michigan in Ann Arbor, US, the Helmholtz Zentrum München in Munich, Germany, Lund University in Lund, Sweden and the University of Oxford in Oxford, UK. Substantial funding was received from the National Institutes of Health and the National Human Genome Research Institute in the United States and the Wellcome Trust in the United Kingdom. The GoT2D project generated genomic data on a then unprecedented scale to gain further insight into the genetics of T2D.

The GoT2D project comprised three main studies:

1. Genomes Study: an investigation of the whole genome sequence of around 2,700 European individuals with and without T2D (Flannick et al., 2015).

2. Exomes Study: an investigation of the whole exome sequence of over 13,000 individuals and exome chip data on 44,000 individuals with and without T2D across five populations worldwide (Teslovich et al., 2015).

3. Pedigrees Study: an investigation of the genetics of T2D in large Mexican-American families with high incidence of T2D.

The work presented in this chapter and the following chapter was part of the GoT2D Genomes Study. That study aimed to extend our understanding of the as-yet-unexplained heritability and genetic architecture of T2D using the much more complete catalogue of variation captured by whole-genome sequencing. Some regions of the genome and many variants (particularly those at rare or low frequency) are tagged poorly by genotyping arrays. Thus, whole-genome sequencing greatly extends the catalogue of genomic variation available for study, particularly in the low-frequency and rare variant spectrum. The key areas of investigation for the GoT2D Genomes Study were therefore:

1. associations missed by common-variant GWAS, particularly associations with low-frequency (minor allele frequency 0.5–5%) and rare variants (minor allele frequency less than 0.5%);
2. causal genes and variants at T2D loci;
3. helping the design of future studies of T2D by expanding our knowledge of where causal variants could lie; and
4. models for the genetic architecture of T2D.

This chapter and the following chapter focus on addressing goals 3 and 4. The analysis presented focuses on investigating models for the genetic architecture of T2D. Here, I seek to estimate the heritability of susceptibility to T2D and in the next chapter I assess and present the contribution to the variance in T2D risk from DNA variants across the minor-allele frequency spectrum and in different functional classes.

3.1.5 Areas of focus

Common-variant GWAS have hugely advanced the study of the genetics of T2D (as outlined in Section 3.1.2). It is well established from common-variant GWAS and variance-partitioning analyses that common variants explain a large proportion of the variance in susceptibility to T2D. Gusev et al. (2013) report genome-wide heritability estimates of between 30% and 60% for T2D depending on the method used for the analysis. These analyses have used LMMs to measure the collective contribution of (almost exclusively) common SNPs, and so we are moving from a general definition of heritability (introduced in Section 3.1.3) to one which is specific to LMMs and SNP genotype data. Having access to whole-genome sequence data in the GoT2D Integrated Panel dataset (described in detail below in Section 3.2.1) allows us to investigate a much broader catalogue of genetic variation. We have assayed millions of variants that did not feature on previous GWAS genotyping chips, and so can ask: what have we missed in GWAS so far? Or, framed more positively: what more can we discover when using the near-complete catalogue of genomic variation in the GoT2D data?

The field has begun to investigate the contribution of rarer variants to T2D risk. One hypothesis (Goldstein, 2009) states that common-variant associations may in fact be "synthetic associations" due to tagging by a common SNP of multiple rare variants of larger effect. Under this model, numerous as yet undiscovered rare variants of large effect may, in aggregate, contribute substantially to complex disease risk. The GoT2D project (Flannick et al., 2015) ascertained 98% of all variants with MAF $>$ 0.1% in the 2,657 studied samples in the Integrated Panel and propagated variants via imputation into 44,414 additional European samples. This gave the association meta-analyis 80% power to identify a low-frequency (0.5%) variant in the dataset with relative risk $>$ 5. The GoT2D study produced no associations for rare (0.1–0.5%) or low-frequency (0.5–5%) variants, from which bounds on the effect size distributions for rare and low-frequency T2D risk variants can be inferred. These results suggest that the majority of unexplained heritability in T2D is not due to low frequency variants of large effect, nor due to variants not tagged by GWAS arrays.

However, the possible hypothesis remains that rare and low-frequency variants in aggregate could explain a substantial proportion of variance in T2D risk, even if effect sizes for rare and low-frequency variants are modest and individual variants cannot be significantly associated with T2D. For the analysis here, 3,123,033 rare variants and 3,550,861 low-frequency variants are used. Thus, even if on average the rare and low-frequency variants individually contribute a very small amount to T2D risk, in aggregate they could explain a substantial proportion of the overall genetic risk for T2D. A large number of rare and low-frequency variants are included in the estimates of heritability here, probing the contribution to heritability from lower frequency variants to an extent not previously possible.

This chapter introduces the datasets and LMM methods that used for heritability analyses. I outline the quality control procedures undertaken on the data, and report estimates for heritability using a single-variance component model under a range of settings and modeling assumptions. The next chapter discusses the partitioning of heritability onto different classes of genetic variant.

3.2 Data

To study the as-yet-unexplained heritability of T2D, the GoT2D project analyzed whole-genome sequence data in a large number of individuals with and without T2D (Flannick et al., 2015). This dataset represents the largest cohort for a complex disease sequenced to this point.

3.2.1 GoT2D Integrated Panel data

The primary dataset used for this analysis is known as the "GoT2D Integrated Panel" dataset. To generate this dataset, 2,874 individuals were selected from Northern and Central Europe, comprising approximately 50% cases who had early-onset T2D, were lean and/or had a familial form of the disease, and 50% normoglycaemic controls who were overweight. This extreme-phenotypes study design has been demonstrated to increase statistical power to detect associations (Voight et al., 2010). The genome of each individual was characterised through a combination of low-coverage whole-genome sequencing ($\sim$5x), deep-exome sequencing ($\sim$100x) (Teslovich et al., 2015), and dense genotyping (2.5M SNPs on the HumanOmni2.5 array (Illumina, Inc, 2014b). SNPs, short indels, and large deletions were identified, genotyped, and phased as haplotypes (Flannick et al., 2015). In total, 26.6 million variants were identified with an estimated heterozygous genotype accuracy of 99.1%. The 26.6 million variants comprise 25.1 million SNPs, 1.5 million short indels and 8,900 large deletions. This set is estimated to offer near-complete ascertainment of low-frequency and common variants in the GoT2D sample, since it contains 98.2% of variants observed with an allele count greater than five in study participants (MAF $>$ 0.1%).

3.2.2 Variant calling

The GoT2D project processed whole genome sequence reads across the 2,772 studied individuals using two SNP calling pipelines: `GotCloud` (`http://genome.sph.umich.edu/wiki/GotCloud`) and `GATK` (DePristo et al., 2011). Unfiltered SNP calls were merged across the two call sets. The merged site list was then processed through the SVM and VQSR filtering algorithms implemented by those pipelines, respectively. SNPs that failed both filtering algorithms were removed before genotyping and haplotype integration. The `GATK UnifiedGenotyper` was used to call SNPs in whole exome sequence data.

The `GATK UnifiedGenotyper` was also used to call short insertions and deletions (less than 50kb) from the whole-genome sequence (WGS) and whole-exome sequence (WES) data. Short insertions and deletions are known to have high false positive rates due to systematic sequencing and alignment errors (1000 Genomes Project Consortium et al., 2012). To counter this, used stringent filtering critera in SVM and VQSR were applied and variants that failed either algorithm were excluded. Illumina's `GenomeStudio` (Illumina, Inc, 2014a) software was used to call genotypes from the Omni2.5 data, with the default clustering provided by Illumina.

3.2.3 Haplotype integration

To produce a set of high-quality variants to take forward into downstream analyses the GoT2D project integrated variant calls from the WGS, WES and chip data into haplotypes.

The following strategy was used to avoid sample mismatch across the three sequencing and genotyping platforms. Sample pairs with less than 98% genotype concordance from the exome sequence and Omni2.5 genotypes were identified. For whole genome sequence and other platforms, `verifyBamID` (Jun et al., 2012) was used, setting exome or Omni2.5 genotypes as known. As `verifyBamID` had already been used for DNA contamination detection, it was possible to identify samples as matched or mismatched. If a sample showed evidence of mismatch in any of these three comparisons it was excluded from downstream analysis. Samples with exome sequence data only were also excluded due to lack of genome-wide coverage.

Genotype likelihoods were computed across all sites separately for each platform after merging SNPs discovered from the three different platforms into a single site list. Likelihoods were calculated across the genome combining the genome and exome sequence data, utilising the substantial off-target coverage of exome sequencing. Genotype likelihoods were calculated using `GotCloud` for genome sequence and `GATK UnifiedGenotyper` for exome sequence. Omni2.5 hard genotype calls were converted into genotype likelihoods assuming uniform small error rates (10^{-6}).

The GoT2D project integrated the genotypes from the three platforms by calculating combined genotype likelihoods across each of the 2,874 individuals as the product of the corresponding genome, exome and Omni2.5 likelihoods, assuming independence across platforms. The integrated genotype likelihood data were phased into a single haplotype map using `Beagle` (Browning & Browning, 2007), with 10,000 variants in each chunk and 1,000 overlapping variants between consecutive chunks. The phased sequences were refined to improve genotype and haplotype quality using `Thunder` (Li et al., 2011) as implemented in `GotCloud` with 400 states.

3.2.4 Using dosages instead of hard genotype calls

Most variance partitioning analyses using the approaches described above use hard genotype calls (or "best guess" genotype calls). Hard genotype calls provide the minor allele count for each variant, and so are 0, 1, or 2. The genotype assigned is the one determined to be most likely (i.e., the best guess) according to the variant calling algorithm. A genotype dosage, in contrast, represents the predicted dosage of the minor allele given the data available (posterior mean genotype), and so gives a value for the genotype on a continuous scale between 0 and 2 (Li et al., 2009; Howie et al., 2011). Thus, a genotype dosage incorporates information about the uncertainty in the genotype call in the reported genotype.

When imputation is used, the dosage value indicates how well the genotype is supported by the imputation method. For example, a dosage of 1.95 implies very high confidence that the true genotype is homozygous alternative. A dosage of 1.5, halfway between the genotype for heterozygous and homozygous alternative, indicates that the calling algorithm has equal confidence in the heterozygous or homozygous alternative calls.

The models described above for hard genotype calls can be used directly for the analysis of dosage data. The only differences are that when one mean-centres the genotypes one could assume that the mean was $2p_k$, where p_k is the observed minor allele frequency for the marker k. However, one cannot make this assumption for the dosages, so the sample mean of the dosages for marker k is used to mean-centre the dosage genotypes. Similarly, for hard genotypes one has the variance of the genotypes as $2p_k(1-p_k)$, but for dosages one does not have such a functional form, so must use the sample variance for marker k when this variance is required for weighting.

3.2.5 Imputed data for a UK cohort

Later in this chapter I attempt to replicate the key results from the analysis of the Integrated Panel data in a second dataset. Through the GoT2D consortium I was able to access genotype data from the Wellcome Trust Case Control Consortium (Wellcome Trust Case Control Consortium, 2007). Two separate imputed datasets from these samples were generated: Kyle Gaulton imputed 1000 Genomes variants into these samples (1000G-Imputed) and Loukas Moutsianas imputed GoT2D variants into these samples (GoT2D-Imputed). In both cases, IMPUTE2 (Howie et al., 2009) was used.

Loukas removed individuals who were sequenced for GoT2D so that there was no overlap of individuals between the Integrated Panel dataset and the imputed datasets. Loukas followed QC approaches taken by Gusev et al. (2014), filtering out samples with missing genotype rates greater than 5% and filtering out variants with missingness greater than 5% and imputation "info score" less than 0.5. For partitioning analyses, I also filtered out variants in the regions that were excluded for the Integrated Panel analyses (see 3.4.1.2). After this filtering, I retained 4,525 individuals (1,587 cases and 2,938 controls) for analysis.

3.3 Methods

This section presents the theory and methods underpinning the use of linear mixed models for the analysis of the contribution of different classes of genetic variant to variance in T2D susceptibility.

The material presented here is primarily an exposition and synthesis of previous work. In particular, I draw heavily on the work of the team responsible for the Genome-wide Complex Trait Analysis (GCTA) software presented in the following publications: Hayes

et al. (2009); Yang et al. (2010a); Lee et al. (2010); Yang et al. (2011b,a); Lee et al. (2011, 2012b,a, 2013). Ideas from these papers underpin Sections 3.3.1, 3.3.2, 3.3.3 & 3.3.7. Section 3.3.4 presents an alternative effect-size model first described explicitly for this application in Speed et al. (2012). Section 4.3.1, looking at models with multiple variance components, presents material presented in the GCTA team's publications listed above (most technical detail in Yang et al. (2011a)), but follows the treatment by Gusev et al. (2014), which focuses (as I do), on partitioning variance by functional class. Likewise, Section 4.3.2 on enrichment scores is from Gusev et al. (2014).

To the best of my knowledge I have made original contributions in Section 3.3.5, Section 4.3.1 and Section 4.3.2.1. Section 3.3.5 is a synthesis of the preceding sections, putting the models in Sections 3.3.3 & 3.3.4 into a unified framework. Although a simple extension of previous work, I have not seen a general framework that contains the more specific models previously described in the literature. Section 4.3.1 describes a model for non-hierarchical partitioning of variants. The suggestion was made to me by Hilary Finucane (personal communication), but I specify the details of this particular model, which I have not seen described elsewhere. Section 4.3.2.1, applying the delta method to obtain the required results for standard errors for enrichment estimates, is hardly a groundbreaking result, but has been independently derived for the setting here (Gusev et al. (2014) suggest this approach, but provide no details or results).

In the remaining sections (and, in fact, across all sections), I have tried to give appropriate credit in the text. The novelty of the work overall is primarily in the application of these methods and theory to the whole-genome sequence data at hand (as far as I know, a first), rather than, with the exception of the original contributions described above, the development of new theory and methods for these applications.

3.3.1 A linear mixed model framework for estimating heritability

Let there be n individuals with observed phenotypes, genotyped at m markers. Consider a linear model relating phenotypes $\boldsymbol{Y}$ to genotypes X through additive genetic effects of the genotyped markers:

$$\boldsymbol{Y} = \mathbf{1}_n \mu + X\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{3.7}$$

where

$$\boldsymbol{\epsilon} \sim \mathrm{MVN}_n(\mathbf{0}, I_n \sigma_e^2),$$

and:

- X is an $n \times m$ matrix of observed genotypes or genotype dosages. I assume that the raw genotypes are coded as 0, 1 and 2 copies of a reference allele at each marker (or are values between 0 and 2 if dosages are used), and that the columns of X are mean-centred, but not variance-standardised.

- $\boldsymbol{Y}$ is an n-vector of phenotypes measured at n individuals.
- $\boldsymbol{u}$ is an m-vector of (unknown) genetic marker effects.
- $\mathbf{1}_n$ is an n-vector of 1s.
- μ is a scalar representing the phenotype mean.
- $\boldsymbol{\epsilon}$ is an n-vector of independent error terms that have variance σ_e^2.
- MVN_n denotes the n-dimensional multivariate normal distribution.

One could extend this model to allow more flexible modelling of fixed effects by replacing $\mathbf{1}_n\mu$ by $Z\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of fixed effects and Z the corresponding design matrix.

Let us consider the genotypes as random variables. In this model, we let $\mathfrak{X}_k^{\text{raw}}$ denote the number of copies in an individual of the minor allele for marker k. If we let p_k denote the minor allele frequency (MAF) of marker k in the population and assume Hardy-Weinberg equilibrium (Hardy, 1908; Weinberg, 1908) then it follows that $\mathfrak{X}_k^{\text{raw}}$, takes the value 0 with probability $(1-p_k)^2$, 1 with probability $2p_k(1-p_k)$ and 2 with probability p_k^2. Thus, $\mathfrak{X}_k^{\text{raw}}$ has expected value $2p_k$ and variance $2p_k(1-p_k)$. We then define the mean-centred genotype $\mathfrak{X}_k = \mathfrak{X}_k^{\text{raw}} - 2p_k$, which has mean 0 and the same variance as the raw genotype, $2p_k(1-p_k)$. Since we do not usually know p_k we estimate it from the sample with the sample mean $\hat{p}_k$.

In our sample, we make an observation of $\mathfrak{X}_k$ for each of the n individuals, represented as $X^{(k)}$, the kth column of the observed (mean-centred) genotype matrix. Let $X_i^{(k)}$ denote the observed mean-centred genotype for marker k for individual i. We estimate the allele frequencies from our samples as $\hat{p}_k = 1/n \sum_{i=1}^{n} X_i^{(k)}$, and the variance of $\mathfrak{X}_k$ as $\hat{\text{var}}(\mathfrak{X}_k) = 2\hat{p}_k(1-\hat{p}_k)$.

In the type of genomic datasets that we are considering, we have m (e.g. $>$ 1 million) much larger than n (<100,000), so genetic variant effects, $\boldsymbol{u}$, cannot be estimated individually. Instead, we need to make further modelling assumptions to get estimates of the proportion of variance in T2D risk explained by genetic effects.

The LMM approach from Yang et al. (2010a) is equivalent to the assumption that genetic effect sizes are normally distributed. Indeed, different specifications for an LMM to partition phenotypic variance onto different categories of genetic variants can be defined through different assumptions about the effect size for each marker. Zhou et al. (2013) provide a comprehensive overview of various effect size distributions that have been used for similar applications as that considered here (often in animal breeding) and their corresponding models.

Two particular models are considered below in more detail:

1. A model where the effect size depends on the reference allele frequency of the marker. This the "allele-frequency dependent" or AFD model.

2. A model where effect size does not depend on allele frequency. This is the "constant effect size" or CES model.

The AFD model is the standard (default) model for such LMM analyses, so it is the obvious starting point. Under this model, estimates of the proportion of variance explained can be interpreted as estimates of the narrow-sense heritability. This has driven their widespread use in the investigation of "missing heritability" in chip-genotype data. Implicit in the model are the assumptions that all variants, on average, contribute equally to variance explained, and that effect-size increases as MAF decreases, which suggests a link to selection. The CES model, on the other hand, assumes a constant effect size for all variants, regardless of MAF. This has the effect of assuming that rarer variants contribute, on average, less to variance explained than more common variants. The decoupling of effect size from allele frequency is a plausible architecture for a disease with onset typically after reproductive years.

At first glance, the assumption that all variants make a non-zero contribution to the phenotype seems unrealistic. Indeed, it probably is not the case that all variants contribute to the phenotype, but the LMM approach has been shown to be robust to this assumption being violated (Speed et al., 2012).

Both models were studied to see how they perform in different settings, and see which findings are consistent, regardless of the effect-size distribution assumed. Below, I also describe a general framework for assumptions on marker effect sizes that can include marker weights. This general framework easily describes the two models above, and many more sophisticated weighting schemes as well.

3.3.2 The Genetic Relatedness Matrix

This section focuses on the interpretation of linear mixed models from the perspective that each variant has an effect (Equation 3.7). In this framework, one makes assumptions about the distribution of effect sizes and these assumptions define the model. However, one could just as well reframe Equation 3.7 as an equivalent polygenic model.

Under a polygenic model, we assume a single "polygenic" random effect, g_i, for each individual. We then assume a covariance structure for the random effect vector $\boldsymbol{g}$ based on the genetic (or genomic) similarity between individuals. The genetic relatedness matrix (GRM), sometimes also called a *kinship matrix*, is defined to be an $n \times n$ matrix containing the pairwise relatedness values between individuals. Although perhaps not as interpretable as the effect-size focused interpretation of the LMM, the GRM has been an important feature of the study of heritability and phenotypic variance explained by genetic effects (Hayes et al., 2009; Yang et al., 2010a; Lee et al., 2011). Let K denote the GRM, and

define it as:

$$K = X \cdot \text{diag}(\boldsymbol{v}) \cdot X^T \tag{3.8}$$

where X is the matrix of mean-centred observed genotypes and $\boldsymbol{v}$ is a vector of variant-specific weights.

The off-diagonal entries of this matrix provide a measure of the genetic (or genomic) relatedness between two individuals i and j. We can see how relatedness for two individuals is defined explicitly by expanding out the weights term:

$$K_{ij} = \sum_{k=1}^{m} v_k X_{ik} X_{jk}, \tag{3.9}$$

where X_{ik} denotes the mean-centred genotype for marker k in individual i (the ikth entry of the genotype matrix). This definition of K shows the equivalence between the effect-size interpretation and the polygenic interpretation of the LMM: the effect-size distribution defines K, or, alternatively, a particular way of computing K imposes implicit assumptions about the effect-size distribution.

The diagonal entries of the GRM, K_{ii}, would naïvely be equal to 1, but Yang et al. (2011a) instead propose that the value for an individual's relatedness with itself take into account possible inbreeding. As such, the diagonal entries of K are computed as $1 + \hat{F}_i^{\text{III}}$ in the GCTA software, where $\hat{F}_i^{\text{III}}$ is an unbiased estimator of F, the inbreeding coefficient, defined in Yang et al. (2011a).

This gives us an expression for the phenotypic variance:

$$\text{var}(\boldsymbol{Y}) = E(K)\sigma_g^2 + I_n\sigma_e^2. \tag{3.10}$$

In practice, we assume that the genotypes are known (as measured) and take $E(K) = K$. In this case, the GRM K is called the "realised relatedness matrix" (Hayes et al., 2009), and we have

$$\text{var}(\boldsymbol{Y}) = K\sigma_g^2 + I_n\sigma_e^2, \tag{3.11}$$

and so

$$\boldsymbol{Y} \sim \text{MVN}_n(Z\boldsymbol{\beta}, K\sigma_g^2 + I_n\sigma_e^2). \tag{3.12}$$

In describing the models in more detail below, I will focus on the effect-size interpretation of the models, but also explicitly define the GRM K for each model. Both ways of thinking about the model can be useful for understanding results of the models and potential pitfalls in their use.

3.3.3 Default Model: Effect sizes depend on allele frequency

Consider the effect size u_k of marker k. The default model in GCTA, which has become the model most widely used in the field, assumes that the variance of the distribution of effect sizes increases with decreasing MAF. I will refer to this model as the "allele-frequency dependent" (AFD) model.

In this setting, we assume the following effect-size distribution:

$$u_k \sim \mathrm{N}\left(0, \frac{\sigma_g^2}{m \cdot 2\hat{p}_k(1-\hat{p}_k)}\right) \tag{3.13}$$

where σ_g^2 is the component of the phenotypic variance attributable to additive genetic effects at the genotyped markers. We can then consider the distribution of $\boldsymbol{Y}$ under this model. This model allows each marker the same contribution to the VE, regardless of allele frequency.

Under this model, the GRM, K^{AFD}, is:

$$K_{ij}^{\mathrm{AFD}} = \frac{1}{m}\sum_{k=1}^{m} \frac{X_{ik}\, X_{jk}}{2\hat{p}_k(1-\hat{p}_k)}. \tag{3.14}$$

From these results it follows that

$$\boldsymbol{Y} \sim \mathrm{MVN}_n\left(\mathbf{1}_n\mu, K^{\mathrm{AFD}}\sigma_g^2 + I_n\sigma_e^2\right). \tag{3.15}$$

Apart from some adjustments to the diagonal entries of K^{AFD} to account for inbreeding (Yang et al., 2011a), as described above, this is the default model used in the GCTA software. Hence we see that this Default Model implicitly models genetic effect sizes that increase as the variant minor allele frequency decreases.

We can also consider $E(K^{\mathrm{AFD}})$ through:

$$\begin{aligned} E(K_{ij}^{\mathrm{AFD}}) &= \frac{1}{m}\sum_{m=1}^{m} \frac{E(X_{im}X_{jm})}{2p_m(1-p_m)} \\ &= \frac{1}{m}\sum_{m=1}^{m} \frac{\hat{\mathrm{cov}}(X_{im}, X_{jm})}{\sqrt{\hat{\mathrm{var}}(X_{im})}\cdot\sqrt{\hat{\mathrm{var}}(X_{jm})}} \\ &= \frac{1}{m}\sum_{m=1}^{m} \hat{\rho}_{ij}^{(m)}, \end{aligned}$$

using $\hat{\rho}_{ij}^{(m)}$ to denote the sample correlation in genotypes between individuals i and j at marker m, $\hat{\mathrm{corr}}(X_{im}, X_{jm})$. Thus, the relatedness between two individuals can be interpreted as the average sample genotypic correlation across all measured variants.

3.3.4 Alternative Model: Effect sizes do not depend on allele frequency

This section considers a model in which effect sizes do not depend on allele frequency, so that effect size is constant across the allele frequency range. This results in different model behaviour from the Default Model described above, in which it assumed that effect size is larger for rarer variants. I will refer to this alternative model as the "constant effect-size" (CES) model.

The distribution of effect sizes under this model is then:

$$u_k \sim \mathrm{N}\left(0, \frac{1}{\sum_{k=1}^{m} 2\hat{p}_k(1-\hat{p}_k)}\sigma_g^2\right), \tag{3.16}$$

and the GRM, K^{CES}, is:

$$K_{ij}^{\mathrm{CES}} = \frac{1}{\sum_{k=1}^{m} 2\hat{p}_k(1-\hat{p}_k)} \sum_{k=1}^{m} X_{ik} X_{jk}. \tag{3.17}$$

Just as above, it follows that

$$\boldsymbol{Y} \sim \mathrm{MVN}_n\left(\mathbf{1}_n\mu, K^{\mathrm{CES}}\sigma_g^2 + I_n\sigma_e^2\right). \tag{3.18}$$

This alternative model can be computed in more recent versions of the GCTA software. The assumption of constant genetic effect sizes across the allele frequency spectrum means that, on average, individual variants that are more common will explain more phenotypic variance than individual variants that are rarer, because the genotypic variance is larger for more common variants while the effect-size variance is the same regardless of allele frequency. In comparison to the AFD model, then, the CES model will give less weight to rare or low-frequency variants, and these variants will contribute comparatively less in aggregate to phenotypic variance explained than in the AFD model.

We can also consider $E(K^{\mathrm{CES}})$ as we did $E(K^{\mathrm{AFD}})$:

$$\begin{aligned} E(K_{ij}^{\mathrm{CES}}) &= \frac{1}{\sum_{k=1}^{m} 2\hat{p}_k(1-\hat{p}_k)} \sum_{m=1}^{m} E(X_{im}X_{jm}) \\ &= \frac{1}{\sum_{k=1}^{m} 2\hat{p}_k(1-\hat{p}_k)} \sum_{m=1}^{m} \hat{\mathrm{cov}}(X_{im}, X_{jm}). \end{aligned}$$

Thus the relatedness between two individuals can be interpreted as the average sample genotypic covariance across all measured variants.

3.3.5 A general LMM with marker weights

Next, I describe a general framework for estimating genetic variance explained using linear mixed moels. Let us reframe the LMM by taking a closer look at the distribution of effect sizes assumed for the genetic effects $\boldsymbol{u}$, which allows a more general description of such models. This section presents a model that is a synthesis of the two models described

above. I have not seen this general model described in the literature, so believe this formulation of a general LMM framework to be a new contribution.

Consider again Model 3.7, but allow each marker k to be assigned a weight w_k. Let us assume the following normal distribution for the effect sizes u_k:

$$u_k \sim \mathrm{N}\left(0, \frac{w_k}{\sum_{t=1}^{m} w_t \cdot \mathrm{var}(\mathfrak{X}_t)}\sigma_g^2\right). \tag{3.19}$$

The scaling of the individual marker weight w_k by the sum of the weighted genotype variances ensures that the ultimate estimate of σ_g^2 does not depend on the number of markers used. Replacing $\mathrm{var}(\mathfrak{X}_t)$ gives a convenient general form for variant effect sizes with weighting of variants:

$$u_k \sim \mathrm{N}\left(0, \frac{w_k}{\sum_{t=1}^{m} w_t \cdot 2\hat{p}_k(1-\hat{p}_k))}\sigma_g^2\right). \tag{3.20}$$

Crucial for the uses of this model in the current work is the variance attributable to a single variant. If we consider the marker k, then we have:

$$\begin{aligned}
\mathrm{var}(\mathfrak{X}_k u_k) &= E(\mathrm{var}(\mathfrak{X}_k u_k|\mathfrak{X}_k)) + \mathrm{var}(E(\mathfrak{X}_k u_k|\mathfrak{X}_k)) \\
&= E(\mathfrak{X}_k^2 \cdot \mathrm{var}(u_k)) + 0 \\
&= E(\mathfrak{X}_k^2) \cdot \frac{w_k}{\sum_{t=1}^{m} w_t \cdot 2\hat{p}_t(1-\hat{p}_t)}\sigma_g^2 \qquad [\text{as } E(\mathfrak{X}_k^2) = \mathrm{var}(\mathfrak{X}_k)] \\
&= 2p_k(1-p_k) \cdot \frac{w_k}{\sum_{t=1}^{m} w_t \cdot 2\hat{p}_t(1-\hat{p}_t)}\sigma_g^2.
\end{aligned}$$

If we plug in the estimate $2\hat{p}_k(1-\hat{p}_k)$ for the variance of $\mathfrak{X}_k$, then the phenotypic variance explained by a given marker in the sample is a fraction of the total additive genetic σ_g^2, and a function of the observed minor allele frequency and marker weight:

$$\mathrm{var}(\mathfrak{X}_k u_k) = \frac{w_k \cdot 2\hat{p}_k(1-\hat{p}_k)}{\sum_{t=1}^{m} w_t \cdot 2\hat{p}_t(1-\hat{p}_t)}\sigma_g^2. \tag{3.21}$$

We can apply this variant-weighting effect-size model back into the framework of Equation 3.7. This allows us to characterise the phenotypic variance, $\mathrm{var}(\boldsymbol{Y})$. Using the independence of the genetic and environmental (error) effects and the law of total variance, we find:

$$\begin{aligned}
\mathrm{var}(\boldsymbol{Y}) &= \mathrm{var}(Z\boldsymbol{\beta} + X\boldsymbol{u} + \boldsymbol{\epsilon}) \\
&= 0 + \mathrm{var}(X\boldsymbol{u}) + \mathrm{var}(\boldsymbol{\epsilon}) \\
&= E(\mathrm{var}(X\boldsymbol{u}|X)) + \mathrm{var}(E(X\boldsymbol{u}|X)) + I_n\sigma_e^2 \\
&= E(X \cdot \mathrm{var}(\boldsymbol{u}) \cdot X^T) + \mathrm{var}(XE(\boldsymbol{u})) + I_n\sigma_e^2
\end{aligned}$$

If we let $\boldsymbol{v} = \mathrm{var}(\boldsymbol{u})$ be a vector with entries $\frac{w_k}{\sum_{t=1}^{m} w_t \cdot 2\hat{p}_t(1-\hat{p}_t)}$ arising from the variant weights, then:

$$\mathrm{var}(\boldsymbol{Y}) = E(X \cdot \mathrm{diag}(\boldsymbol{v}) \cdot X^T)\sigma_g^2 + I_n\sigma_e^2.$$

Applying Equation 3.8 we write a form for the estimation of the relatedness for a pair of individuals i and j when we apply arbitrary marker weights:

$$K_{ij} = \frac{1}{\sum_{t=1}^{m} w_t \cdot 2\hat{p}_t(1-\hat{p}_t)} \sum_{k=1}^{m} w_k X_{ik} X_{jk}, \tag{3.22}$$

where X_{ik} denotes the mean-centred genotype for marker k in individual i (the ikth entry of the genotype matrix).

This framework is general, so it incorporates practically any variant-weighting scheme in which a weight is assigned to each variant. In practice, however, we focus our attention on a smaller number of models that have been shown to be useful in decomposing phenotypic variance into variance components.

I focus here on using weights of the form

$$w_k = [2\hat{p}_k(1-\hat{p}_k)]^s. \tag{3.23}$$

By setting s equal to an integer, most commonly -1 or 0, I can define convenient and interpretable effect-size distributions for the linear model. I discuss the AFD and CES models below and present variance-partitioning results for these models. Recently, other weighting approaches have been proposed attempting to address various aspects of variance partitioning using LMMs (Speed et al., 2012; Gusev et al., 2013).

This model is equivalent to a model with a polygenic term for each individual (rather than assuming an effect for each marker):

$$\boldsymbol{Y} = Z\boldsymbol{\beta} + \boldsymbol{g} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{g} \sim \mathrm{MVN}_n(\mathbf{0}, K\sigma_g^2)$$

with K defined as above. Indeed, when estimation of σ_g^2 and σ_e^2 is done using restricted maximum likelihood (REML), this polygenic alternative characterisation of the model is used, with the GRM K, which expresses the relatedness between individuals, defining the covariance structure of the random effect $\boldsymbol{g}$. Since the covariance structure of the model is exactly the same as that obtained when taking the individual variant effects model described above, the two models are completely equivalent.

It is typical to think about the polygenic model formulation when it comes to the actual model fitting with REML, but the individual variant effects model when thinking about contributions of different types of variant to phenotypic variance explained by the model. Switching between the two equivalent models can sometimes aid insights into how the models work.

3.3.6 Estimating variance components using residual maximum likelihood

Once covariance structures (GRMs in this setting) are defined for the polygenic random effect using the approaches outlined above we need a method for obtaining estimates of the variance components, σ^2_g and σ^2_e in the simplest case. Maximum likelihood (ML) is the obvious starting approach, but using ML to estimate variance components in LMMs results in estimates that depend on the fixed effects fitted in the model (see Quenouille, 1956; Harville, 1977, for example). To obtain variance component estimates that are independent of the fixed effects used, I use restricted (or residual) maximum likelihood (REML), proposed by Corbeil & Searle (1976a,b), as is standard.

The REML approach partitions the likelihood into two parts, one part being free of the fixed effects. Maximising this part of the likelihood yields the REML variance component estimators that retain the property of invariance under translation (Corbeil & Searle, 1976b). There are multiple algorithms available for the computation of REML estimates (see Section 3.4.2).

3.3.7 LMM heritability analysis of case-control data

The normal linear mixed models described above were originally designed for quantitative (continuous) phenotypes, such as height or blood glucose level. For continuous traits, the assumption that error terms should be normally distributed is reasonable. However, for binary traits, such as disease, it is not immediately obvious that normal linear mixed models are appropriate.

Lee et al. (2011) raise three particular issues that need to be addressed in order to estimate genetic variance without bias :

1. Scale — For disease traits, the phenotypes (case-control) are measured on the 0-1 scale, but heritability is most interpretable on a scale of liability (see Section 3.3.8 below).

2. Ascertainment — In case-control studies, the proportion of cases is usually (much) larger than the prevalence in the population. For example, in the GoT2D sample, the proportion of cases is approximately 50%, whereas the prevalence of T2D in European populations is approximately 8%. To make estimates of genetic variance interpretable, they should not be biased by this ascertainment.

3. Quality Control (QC) of SNPs — QC is more of a concern for case-control than quantitative traits.

Note that the binary trait data is analysed using the same models described above, not a logistic mixed model (which would be a reasonable expectation for 0-1 data). Logistic

mixed models are not used primarily for reasons of computational efficiency, but it has also been shown that (normal) linear mixed models perform well in terms of accuracy as well as computational efficiency when effect sizes are small (Pirinen et al., 2013). This is expected to be the case in genetic data, especially for a complex disease like T2D, for which even the strongest GWAS hits have relatively small odds-ratios (e.g. Morris et al., 2012).

Instead, variance components (e.g. σ^2_g and σ^2_e) and heritability are estimated on the "observed scale" using REML, and then transformed to the more interpretable liability scale. The relationship between observations on the observed risk scale and liabilities on the unobserved continuous scale can be modeled using a probit transformation to generate the classical liability threshold model (Falconer, 1965, 1967). Under the liability threshold model, it is assumed that each individual in a population has an unobserved liability for any binary trait (particularly, disease). Liability of disease is assumed to be the sum of environmental and additive genetic components from independent normal distributions, and so the liability distribution is assumed to be standard normal. All affected individuals (cases) have liability phenotypes exceeding some threshold t. REML estimates of the genetic variance on the observed scale can be transformed to obtain liability-scale estimates of the genetic variance and heritability as described in the next section. As Lee et al. (2011) state, working on the scale of liability has two major advantages:

1. Population parameters such as variance components and heritability are independent of prevalence, and

2. Statistical methods developed for quantitative traits can be applied to the trait liability.

The independence of variance components and heritability from prevalence enables us to compare variance partitioning results across populations and traits. This is important here, as I have the Integrated Panel data and imputed data wish to compare results between them.

3.3.8 Transforming variance estimates and heritability to the liability scale

Variance components estimates on the observed scale need to be transformed to the liability scale for case-control data. The transformation needs to account for the binary phenotype and, if present, ascertainment. For completeness and ease of understanding a brief summary is presented Lee et al. (2011)'s derivations of linear transformations of the heritability on the observed scale h^2_o to that on the liability scale h^2_l.

Here, K denotes the population prevalence of the disease trait, and d the density of the normal distribtution at the truncation threshold t. In a random sample from the population

(so that the proportion of cases in the sample is equal to the prevalence of disease in the population), then the required transformation is relatively straight-forward:

$$h_l^2 = h_o^2 K(1-K)/d^2. \tag{3.24}$$

This result was originally derived by Alan Robertson in the Appendix of Dempster & Lerner (1950). Briefly, it is obtained by applying the properties of truncated normal distributions and linking the observed and liability scales through a linear regression coefficient. The approximation is valid in samples of "unrelated" (in the conventional sense) individuals, such as those studied here, in which the non-additive component of the genetic variance is small relative to the additive component (Lee et al., 2011). However, this transformation does not take into account the inflated proportion of cases present in ascertained case-control samples.

To obtain a valid transformation in an ascertained sample, Lee et al. (2011) use the same liability model as above. However, the normality of liability is violated in a case-control study with ascertainment because we sample individuals from the tails of the liability distribution. Thus, the transformation also depends on, P, the proportion of cases in the sample. The derivation of the transformation to the liability scale is considerably more involved for case-control studies, but the crux is still the idea of regressing phenotype on the observed risk scale on genetic liability in the case-control study.

In this case, we estimate the genetic variance on the observed case-control scale ($\hat{\sigma}^2_{u_{cc}}$) and we can obtain $\hat{h}_{o^2_{cc}}$, an estimate of the heritability on the liability scaled in an ascertained case-control study. The heritability on the liability scale, h_l^2, is equal to the genetic variance on the liability, σ_g^2 (as the liability distribution is standard normal). After substantial algebraic manipulations, Lee et al. (2011) present (in Equation 23 of that paper) the transformation required to take an estimate of heritability from an ascertained case-control study to an estimate of heritability on the liability scale:

$$h_l^2 = \sigma_g^2 = \hat{\sigma}^2_{u_{cc}} \left[\frac{1}{d}\frac{K(1-K)}{P(1-P)}\right]^2 = h^2_{o_{cc}} \frac{K(1-K)}{d^2}\frac{K(1-K)}{P(1-P)}. \tag{3.25}$$

This simple result makes it easy to transform REML variance component estimates to liability-scale heritability estimates. It only depends on the prevalence of the disease (K) and P, the proportion of cases in the ascertained case-contol sample (the prevalence determines the value of d, as defined above). Importantly, for the focus on partitioning variance into multiple components defined by different classes of genetic variation, the transformation is linear in $h^2_{o_{cc}}$. Thus, when I look at the relative contributions of different variant classes (enrichment; in the next chapter), these remain unaffected by changing the prevalence value assumed.

Lee et al. (2011) apply a Taylor series expansion to derive the sampling variance of estimated heritability on the liability scale transformed from that on the observed scale:

$$\text{var}(h_l^2) \approx \left[\frac{d(h_l^2)}{d(h_{o_{cc}}^2)}\right]^2 \text{var}(h_{o_{cc}}^2) = \left[\frac{K(1-K)}{d^2}\frac{K(1-K)}{P(1-P)}\right]^2 \text{var}(h_{o_{cc}}^2). \tag{3.26}$$

Thus, from REML estimates of the heritability in ascertained case-control samples we can obtain liability-scale heritability estimates with estimates of the standard errors for those estimates.

3.4 Quality contol and implementation of analysis

Quality control is important for these analyses as LMM estimates of heritability can be sensitive to effects of artefacts in the data (Lee et al., 2011). This section describes standard quality control steps that were undertaken to exclude potentially problematic individuals and variants from the analysis. I also outline the implementation of the various analyses, including default parameter settings.

3.4.1 Quality control

In a sense, linear mixed models are a crude instrument for characterising the genetic contribution to a complex trait. There are many possible sources of biases that could distort the variance component estimates on which inferences are based. It is thus crucial to conduct quality control (QC) procedures on the input data to attempt to minimise such problems. This means that one needs to perform QC on both the variants and the samples included in the analysis.

3.4.1.1 Individual exclusions

For LMM analyses I used individuals that had passed GoT2D quality control. Individuals who were closely related, population outliers or structural variant outliers were excluded. After excluding these individuals the GoT2D project settled on a QC-passed set of 2,657 individuals as the Integrated Panel dataset consisting of 1,326 cases who had early-onset T2D, were lean and/or had a familial form of the disease, and 1,331 normoglycaemic controls who were overweight. Individuals came from five populations across Europe: Botnia (a homogeneous Swedish-speaking population in northern Finland with a shared Scandinavian and Finnish genetic background), Finland, Germany, Sweden and the United Kingdom (UK). There was an almost even split between cases and controls within each population (Table 3.1).

The LMM approach to estimating VE (and in the right settings, heritability) differs markedly from classical approaches to estimating these quantities by using "unrelated"

	Cases	Controls	N
GoT2D individuals after integration			**2,874**
Population outliers (excluded)			-43
Related individuals (excluded)			-44
Structural variant outliers (excluded)			-137
GoT2D sequenced QC+	**1,326**	**1,331**	**2,657**
Botnia	199	159	358
Finland	486	517	1,003
Germany	104	101	205
Sweden	222	227	447
UK	322	322	644

Table 3.1: GoT2D sequenced case-control cohort after QC.

individuals rather than closely related individuals (as exemplified by twin studies, the gold standard for estimating heritability). One possible criticism of the use of twin studies for estimating VE is that it can be difficult to tease apart the effects of shared environments from the genetic effects of interest. The approach to estimating VE used here can potentially avoid such problems, but only if care is taken to avoid using closely related individuals in the analysis. Typically, a threshold of "relatedness" between two individuals (the off-diagonal entries of the GRM) is set, often between 0.025 and 0.05 (Yang et al., 2010b; Lee et al., 2012a; Speed et al., 2012; Gusev et al., 2013), and a minimal set of individuals is removed such that the pairwise relatedness between individuals no longer exceeds the threshold.

For the GoT2D dataset, which contains individuals from five distinct groups (Botnia, Finland, Germany, Sweden and the UK), care needs to be taken when applying a threshold to remove closely related individuals. I observe that the "relatedness" measure used in a GRM is a relative term—relatedness is defined relative to the average level of genomic similarity in the group of individuals under consideration at a given time. Thus, one obtains different values for the relatedness of pairs of Botnian individuals if one computes GRMs using only Botnian samples, or if one includes other samples along with the Botnian individuals when computing GRMs.

Exploratory analysis of relatedness in the 2,657 individuals in the dataset showed that the individuals across the five countries divided broadly into two groups: the Botnian and Finnish samples on the one hand, and the German, Swedish and UK samples on the other hand. Thus I computed GRMs from the Botnian and Finnish samples and, separately, from the German, Swedish and UK samples. I found that the number of individuals to be removed from the analysis at a relatedness threshold of 0.025 or 0.05 differs according to the minimum MAF used when computing the GRMs, with higher minimum MAFs resulting in more conservative sample removal lists (that is, more samples removed from the analysis). By default, I use the sample inclusion lists produced using this approach with a minimum MAF of 5% to get a more conservative sample exclusion list. This approach retains 2,554

samples at a relatedness threshold of 0.05 and 2,261 at a relatedness threshold of 0.025. I used a relatedness threshold of 0.05 for our primary analyses as has become standard for these types of analyses.

3.4.1.2 Variant exclusions

Not all regions of the genome are suitable for inclusion when estimating variance explained using LMMs. I first obtained a list of complex regions of the genome (for example the major histocompatability region) that had been excluded from other analyses in the GoT2D project (Hyun Min Kang, personal communication). From the genome-focused analysis done by others in the GoT2D project, principal components had been computed from a carefully selected set of SNPs from the Omni chip to characterise population structure. Variants were selected to be informative for ancestry and reasonably independent of each other (correlation less than 0.1 between genotypes). Various numbers of these PCs (the number depending on the particular anlaysis) were fitted in statistical models to account for population structure in analyses of the association of SNPs with T2D phenotype.

Following standard practice for estimating phenotypic variance explained using LMMs (Yang et al., 2010a, 2011b; Lee et al., 2011; Janss et al., 2012; Speed et al., 2012), I planned to fit a number of these Omni PCs in the LMMs to account for (some of the) population structure effects. To identify regions with markers that could be problematic for inclusion in the LMMs, we sought to find marker loadings with respect to the PCs. Given the PCs and the genotype data, I used the software `shellfish` (version from January 2014) to compute marker loadings (Davison, 2009).

I inspected plots of marker loadings with respect to the first 30 PCs after variants in the first list of complex regions had been excluded. Ideally, one would like to see PC-loadings to be more-or-less constant across the genome. Regions with very high PC-loadings are likely to be problematic, as PCs are expected to identify population structure and other artefacts in the data. Variants highly correlated with these PCs are therefore likely to have more information about structure or effects other than the genetic effects of interest. Thus, I sought to identify such regions and exclude variants in them. We took a loading value of 6 from `shellfish` to indicate a "high" value as this value was the 99^{th} percentile loading value for the first principal component. I set 6 as the threshold value, and manually identified contiguous regions with substantial (greater than 100) numbers of variants with loads above the threshold. The exclusion list created was used to exclude variants, and GRMs were recomputed. I then repeated the procedure. After these two iterations of identifying exclusion regions I was satisfied that there were no substantially problematic regions remaining. This strategy excluded approximately 300,000 variants in potentially problematic regions from the analysis and left a total of 12,034,435 variants with MAF greater than 0.1%. These second-iteration exclusion lists were used for the analyses that follow.

I set a global minimum MAF threshold of 0.1% for this study because: (1) the LMM approaches I use are expected to break down when very rare variants are included (I observed this breakdown when including all variants, obtaining severely deflated heritability estimates; data not shown); and (2) rarer variants are also likely to be enriched for genotyping errors and artefacts (a variant with MAF $< 0.1\%$ would be observed fewer than five times in the Integrated Panel dataset). LMM methods are expected to break down when very rare variants are used, because they distort the estimated relatedness between individuals. Under the default model for computing relatedness between individuals (described in Section 3.3 below), sharing rarer variants is assigned much greater weight than sharing common variants. Thus the relatedness value for a pair of individuals could be dominated by sharing of a small number of very rare variants. However, there are a very large number of variants in the dataset with MAF $<0.1\%$ (over 12 million), almost all of which are not shared between any given individuals. As a result, relatedness values, which are in a sense averaged over all variants, are depressed by incorporating in the relatedness calculation a very large number of variants with very little information about genomic similarity between individuals. These two effects combined led to the decision to restrict all analyses to variants with MAF $>0.1\%$.

3.4.2 Software for variance partitioning analysis

Loukas and I were unable to get the LD_{res} method (Gusev et al., 2013) implemented in EIGENSOFT (Price et al., 2006; Patterson et al., 2006) to run on the Integrated Panel data, and we were not confident about the reliability of the results obtained using LDAK (Speed et al., 2012). Therefore, I only used the GCTA software for the variance partitioning analyses. Analysis with GCTA requires two main steps: computing GRMs (and merging if necessary) and obtaining variance component estimates from the LMM using the REML approach.

The GCTA software (Yang et al., 2011a) provides two options for computing GRMs, the AFD and CES models described above, and implements three different algorithms for computing REML estimates:

1. Average Information (AI)
2. Fisher-Scoring (FS)
3. Expecation Maximization (EM)

The default method is AI (Gilmour et al., 1995), as this is the fastest algorithm computationally. In preliminary investigations on our data, I found that the AI algorithm often gave VE estimates of exactly zero when fitting multiple variance components. For example, when obtaining estimates for each chromosome individually (fitting chromosomes jointly

in a single model), I found that the AI algorithm frequently constrained variance component estimates for several chromosomes to be zero. These contrained estimates arise when multiple successive iterations of the algorithm give a negative estimate for a variance component. To avoid returning (implausible) negative estimates, GCTA by default constrains these negative estimates to be exactly zero. When comparing the AI estimates to those obtained from Fisher-Scoring (Jennrich & Sampson, 1976) and EM (Dempster et al., 1977), I found that FS similarly constrained multiple estimates at zero, but that EM did not. Thus, I have used the EM algorithm for this study, which requires more time to compute but avoids these constrained zero estimates.

3.4.3 Code implementing the analysis

As is usual for a large computational and data analysis project, a significant amount of code was generated implementing the analyses presented in this chapter. There are three main code outputs from this project:

- The "GCTAtools" package, a lightweight Python module implementing a suite of tools useful for processing GCTA output files, computing enrichment scores and associated statistics and results, and returning results in a convenient form for downstream analysis and plotting.
- Many bash scripts for running GCTA jobs, R markdown documents, and R and Python scripts implementing analyses described here.
- IPython notebooks detailing the steps undertaken in the analysis (including calls used for scripts and dates run), providing an almost complete record of the analysis.

The GCTAtools package is publicly available on GitHub (`http://github.com/davismcc/GCTAtools`), as it might prove useful for others conducting similar analyses. I plan to submit the IPython notebooks recording the computations and other scripts to the figshare repository (`http://figshare.com/`) to aid the reproducibility of the analysis.

3.4.4 Default parameter settings for LMM analyses

The focus of the analyses is the contribution to VE from different classes of variants, and not on the influences of the various fixed effects that one can fit in LMMs. Thus, following the QC steps described above, I fix a set of default parameters that I use (unless specifically stated) in the results that follow.

I keep the fixed effects the same across models, fitting both Sex (factor with two levels), Batch (two levels), and Country (five levels) as categorial covariates and the first 10 Omni PCs (see above) as continuous covariates. I use the 2nd-iteration variant exclusion lists described above, so variants in these exclusion lists are not used for analysis regardless of

what MAF or functional class they may fall into, leaving a total of 11,892,083 variants with MAF $> 0.1\%$. I look at partitioning VE by MAF of variants (described in detail below), but only using variants with MAF $> 0.1\%$. By default, I present results using a relatedness threshold of 0.05 (as described in Section 3.4.1.1 above), having observed no substantially different results thresholds of 0.05, 0.025 and no threshold (data not shown).

When the GCTA software is used for the computation of GRMs and REML estimates of variance components, version 1.24 is used. I use the EM algorithm to obtain REML estimates. In all cases I report VE estimates on the liability scale so that they are comparable across models and with previous studies.

With these parameters all but fixed, I investigate the effects on partitioning VE by applying different MAF cutoffs, different LD-correction methods, different effect size models and separation of variants by functional class.

3.5 Results for estimating the heritability of type 2 diabetes

I first look at results from single-variance component models. That is, models that use all variants for given parameter settings in a single variance component in the mixed model. These are the standard models used in the computation of the "heritability" of a trait from chip-genotype data. As such, single-variance component results provide a baseline to which other results can be compared. Analyses will show, however, that the interpretation even of single-variance component results can be fraught.

This section presents the analysis of single variance component estimates of phenotypic variance explained by genetic effects. I begin here, as the single-variance component (VC) approach was the first approach taken to estimating narrow-sense heritability using LMMs. Multiple-VC approaches were proposed later as a solution to some of the objections raised against the inflexible, and highly parameterised single-VC approach. Noting the results obtained using the simplest LMM approach to understanding the proportion of T2D risk that could be explained by genetic effects provides a platform for investigating more complicated—and scientifically interesting—models for partitioning phenotypic variance in the next chapter.

I obtain estimates of the effect of all variants with MAF above a certain threshold. In such cases, I compute MAF-bin GRMs (useful for later analyses) and merge the appropriate GRMs together to obtain a single GRM that utilises variants with MAF above the threshold. Merging GRMs is a trivial operation, as one just needs to take a weighted average of the relatedness value across GRMs for each pair of individuals. Merging is performed easily with the GCTA software. The weighting is given either by the number of variants used to compute each GRM (default, allele-frequency dependent model) or the total sample genotypic variance of the variants used to compute the GRM (alternative, constant effect-size

model). I then fit the merged GRM as a single variance component in the model and estimate σ^2_g and σ^2_e to obtain an estimate of the phenotypic variance explained by additive genetic effects. In the context of the AFD model, the variance explained estimate is interpretable as a heritability estimate. I report estimates on the liability scale (as described above).

I compare results using thresholds of MAF $>$ 0.1% (i.e. rare, low-frequency and common variants), MAF $>$ 0.5% (i.e. low-freqency and common variants), MAF $>$ 1% and MAF $>$ 5% (common variants only). To begin with, I look at estimates using the Integrated Panel data from GoT2D. I then compare these results to those obtained using the same models from the two imputed datasets (Imputed-GoT2D and Imputed-1000G), which feature a larger UK cohort. This section focuses on presenting "baseline" results for the default allele-frequency dependent model. Section 3.6 explores the robustness of these baseline results to changes in modeling assumptions, approaches to accounting for population structure and linkage disequilibrium, and other considerations.

3.5.1 Single-variance component model using whole-genome sequence data

I first consider results for the default, allele-frequency dependent model when using a single genetic variance component to estimate liability-scale T2D heritability. Heritability estimates for the Integrated Panel data vary to a large degree when different minor allele frequency thresholds for variants used are applied (Table 3.2).

The heritability estimate increases as rarer variants are included in the model. Using only common variants (MAF greater than 5%; 5,360,541 variants) yields an estimate of 0.49 (s.e. 0.12). The point estimate increases slightly when low-frequency variants are added, as can be seen see when using all variants with MAF greater than 1% (0.50, s.e. 0.14; 7,753,493 variants) and variants with MAF greater than 0.5% (0.54, s.e. 0.15; 8,911,402 variants). These estimates are broadly concordant (given the uncertainty in the estimates) with recent estimates of liability-scale heritability for T2D from genome-wide SNPs (Gusev et al., 2013). A striking increase in the heritability occurs when using all variants with MAF greater than 0.1% (12,034,435 variants). When rare variants are included the estimate jumps to 0.68 (s.e. 0.19). In a sense, this is what one would expect to see. In the AFD model, each variant is assumed on average to make the same contribution to variance explained. Thus, this model assumes a substantial contribution from rare variants, and when one adds over three million rare variants (in addition to the variants with MAF greater than 0.5%) a marked increase in the heritability estimate is observed.

Uncertainty in these heritability estimates is large across the different minimum MAF thresholds. For all of the estimates in Table 3.2 the standard errors overlap, so one cannot make confident statements about the true differences in heritability from different sets of variants. This has little effect on the interpretation of results using variants with MAF

MAF Threshold	Number of variants	Heritability (s.e.)
MAF $> 0.1\%$	12,034,435	0.682 (0.193)
MAF $> 0.5\%$	8,911,402	0.536 (0.154)
MAF $> 1\%$	7,753,493	0.503 (0.140)
MAF $> 5\%$	5,360,541	0.490 (0.115)

Table 3.2: Estimates of liability-scale heritability (variance in risk for T2D explained) by single variance-components for the Integrated Panel data. The standard error (s.e.) for each estimate is shown in brackets. Estimates are shown for the default model in which variant effect-sizes are allele-frequency dependent (AFD model). Estimates were obtained for minimum MAF thresholds of 0.1%, 0.5%, 1% and 5% using the corresponding number of variants that passed quality control at that MAF threshold. Here, models with lower minimum MAF threshold contain all of the variants used in models with higher minimum MAF threshold.

greater than 5%, 1% and 0.5%, since the heritability estimates are so similar. However, it means that we cannot make strong statements about potentially the most interesting aspect of the analysis, namely the contribution to heritability of T2D from rare variants.

Another way to tease out the contribution of variants to T2D heritability using single-variance component models is to fit a variance component for rare variants, low-frequency variants and common variants in separate models. That is, I fit a model with a single genetic variance component that only uses rare variants (MAF 0.1–0.5%; 3,123,033 variants), fit a separate model with a single genetic variance component using low-frequency variants (MAF 0.5–5%; 3,550,861 variants), and a third separate model using only common variants (MAF greater than 5%; 5,360,541 variants). The results from these models show a large estimate for the liability-scale heritability from common variants (0.49, s.e. 0.12; exactly the model above with MAF greater than 5%), a small estimate from low-frequency variants (0.097, s.e. 0.15) and a substantial estimate of 0.29 (s.e. 0.17) from rare variants (Table 3.3).

MAF Variant Class	Number of variants	Heritability (s.e.)
Rare (MAF 0.1–0.5%)	3,123,033	0.286 (0.174)
Low-frequency (MAF 0.5–5%)	3,550,861	0.0972 (0.154)
Common (MAF $> 5\%$)	5,360,541	0.490 (0.115)

Table 3.3: Estimates of liability-scale heritability (variance in risk for T2D explained) by single variance-components for disjoint MAF ranges. The standard error (s.e.) for each estimate is shown in brackets. Estimates are shown for the default model in which variant effect-sizes are allele-frequency dependent (AFD model). Estimates were obtained separately for rare variants (variants with MAF 0.1–0.5%), low-frequency variants (MAF 0.5–5%) and common variants (MAF $> 5\%$). The number of variants in each class is shown.

I note that the sum of the estimates from the three separate models (0.87) is substantially larger than the estimate when I fit a single variance component in a model using all variants with MAF greater than 0.1% (0.68; cf. Table 3.2). This increase arises because variants included in a model can tag genetic variation not explicitly used in the computation of

the genetic relatedness matrix. When I use common variants only in the model, these common variants will, to a certain unknown extent, tag lower frequency variants with which they are correlated. Conversely, when I use only rare variants in the model, those rare variants will tag low-frequency and common variants to some extent. Low-frequency variants will tag some rare and some common variation. Thus, the heritability estimates for rare variants alone will not be strictly the variance explained by rare variants but the variance explained by those rare variants and other (predominantly higher frequency) variants that they tag. As a consequence, the estimates of the heritability from rare, low-frequency and common variants when fitted in separate models are slightly inflated by tagging variation not explicitly included in the model. This is not the case when I fit one variance component using all variants with MAF greater than 0.1%.

As a result, the sum of the separate rare, low-frequency and common variant estimates is substantially larger than the estimate from a single variance component using all variants with MAF greater than 0.1%. Nevertheless, even bearing in mind the inflation effects in a separate model fit, the point heritability estimates from rare variants estimate is large (0.29), and much larger than the standard error for the estimate. This suggests that there is a truly non-zero aggregate contribution of rare variants to heritability. The next chapter examines the use of models with multiple variance component models to fit variance components using variants in different allele frequency classes simultaneously. These models will allow us to probe the question of the contribution of rare variants more thoroughly.

Taken together these results for the default model for the Integrated Panel data show that variants identified with whole-genome sequence data (combined with chip and exome sequence data) can, indeed, be used to estimate the heritability of T2D using linear mixed models. The heritability estimates from common variants agree broadly with previously reported results using chip genotype data. With the near-complete catalogue of variation from the GoT2D Integrated Panel data we can begin to probe the contribution of rare variants to the heritability of T2D. The results leave open the possibility of a substantial contribution from rare variants, but there is considerable uncertainty in the heritability estimates obtained from the Integrated Panel data, so we cannot claim with confidence a large contribution from rare variants. I explore the robustness of these results in Section 3.6, below. In the next section I conduct the same analyses, using the same models, on two datasets consisting of genetic variants imputed into a larger cohort. I will assess how closely the results from whole-genome sequence data agree with results from data (with a larger sample size) of the kind that has been used to conduct this type of heritability analysis previously.

3.5.2 Single-variance component model using data imputed into a larger cohort

In this section I conduct analyses of single-variance component models, as above, using data imputed into a larger cohort. Through the GoT2D project I have access to a set of 4500 UK individuals who have been genotyped using a SNP array (see Section 3.2.5). For these individuals I have further genotypes imputed using the 1000 Genomes reference panel (Imputed-1000G data) and, separately, using the GoT2D Integrated Panel as a T2D-specific reference panel (Imputed-GoT2D data). I want an independent set of individuals for replication, so take the individuals in this dataset who did not feature in the Integrated Panel dataset. After removing individuals who were sequenced (and thus appear in the Integrated Panel dataset) and subsequent quality control, I have 4525 individuals (1587 cases and 2938 controls) in the UK imputation cohort. This gives roughly 300 more cases and more than twice as many controls in the imputation cohort as in the Integrated Panel. After removing closely related individuals, I use 4498 individuals with genome-wide relatedness less than 0.05, computed using variants with MAF greater than 5%.

I can compare the single-variance component results across MAF thresholds for the imputed data with those from the Integrated Panel data. Overall, the single-VC estimates are very similar for the Integrated Panel and imputed data, especially taking into account the standard errors of estimates (Table 3.4). The largest differences between Integrated Panel and imputed data results occur for MAF greater than 0.1% variants (8,26,421 variants for Imputed-GoT2D; 6,974,853 variants for Imputed-1000G) and the MAF greater than 5% variants (3,791,288 variants for Imputed-GoT2D; 3,316,072 variants for Imputed-1000G). When rare variants are included, as in the MAF greater than 0.1% model, the imputed results are larger than the Integrated Panel results, quite noticeably so in the case of the Imputed-GoT2D estimate. At the other extreme, for common variants, one sees that the estimates from imputed data are substantially lower than the estimate from the Integrated Panel data, with the discrepancy largest for the Imputed-1000G dataset. One would conclude that the estimates when using variants with MAF greater than 0.5% and MAF greater than 1% are highly concordant.

The estimates from the imputed data get progressively smaller relative to the Integrated Panel estimates as the MAF threshold increases. There are many fewer variants in the imputed datasets (approximately 30% fewer for the Imputed-GoT2D dataset and 40% fewer for the Imputed-1000G dataset than in the Integrated Panel; see Table 3.4) across all MAF thresholds, which could reduce the capacity for genomic similarity to explain phenotypic variance. Interestingly, though, this reduction effect is only apparent for higher MAF thresholds.

Estimates for the Imputed-1000G data are always lower than the estimates from the Imputed-GoT2D data. This difference must be due to the set of variants used for the anal-

	Data	Number of variants	Heritability (s.e.)
MAF >0.1%	Integrated Panel	12,034,435	0.682 (0.193)
	Imputed-GoT2D	8,261,421	0.744 (0.102)
	Imputed-1000G	6,974,853	0.692 (0.0963)
MAF >0.5%	Integrated Panel	8,911,402	0.536 (0.154)
	Imputed-GoT2D	6,100,925	0.552 (0.0807)
	Imputed-1000G	5,159,547	0.522 (0.0756)
MAF >1%	Integrated Panel	7,753,493	0.502 (0.140)
	Imputed-GoT2D	5,344,929	0.505 (0.0734)
	Imputed-1000G	4,587,286	0.471 (0.0697)
MAF >5%	Integrated Panel	5,360,541	0.490 (0.115)
	Imputed-GoT2D	3,791,288	0.429 (0.0611)
	Imputed-1000G	3,316,072	0.403 (0.0593)

Table 3.4: Single-variance component results for liability-scale heritability (proportion of variance explained) for the Integrated Panel, Imputed-1000G and Imputed-GoT2D datasets. The Imputed-1000G dataset refers to the imputation of variants from the 1000 Genomes project into the larger UK cohort, and Imputed-GoT2D refers to the imputation of GoT2D variants into this cohort. The proportion of cases in the imputed datasets is 35.1% compared with 50.4% in the integrated panel dataset. These liability-scale estimates were obtained using a prevalence value of 8%.

yses, because the cohort of individuals and the model parameters are the same for the two imputation analyses. The Imputed-GoT2D dataset contains roughly 15% more variants than the Imputed-1000G dataset (see Table 3.4). Using more variants in the model, all else being equal, thus appears to increase heritability estimates (though due to the overlapping standard errors one cannot claim this with great confidence). We have seen this to be the case when progressively including variants with lower MAF in the Integrated Panel data and both imputed datasets, but comparing the two imputed datasets shows this to be case even within fixed MAF ranges for the same set of individuals.

The estimates from the imputed datasets have smaller standard errors than those from the Integrated Panel data. We expect to obtain smaller standard errors with a larger sample size (such as for the imputed datasets). It is difficult to quantify the expected change in standard errors with increased sample size, however, because of the complexity of the form of the estimate of the variance of a REML variance component estimator (Corbeil & Searle, 1976b; Henderson, 1953). Standard errors for Imputed-1000G estimates are smaller than for the Imputed-GoT2D estimates in all cases. Since the same cohort of individuals and model parameters are used, the difference in the standard errors between the Imputed-GoT2D and Imputed-1000G estimates must arise from the difference in the set of variants used for the analysis. When more variants are used, standard errors are larger. This effect could contribute to the large standard errors for the Integrated Panel estimates, where a very large number of variants are used.

The LMM methods for estimating heritability were developed and shown to work effectively for chip-genotype and imputed data. Applying these methods to whole-genome sequence data, as done with the Integrated Panel data, is a new step. Obtaining single-variance component heritability results from two imputed datasets, which are comparable with those from the Integrated Panel data, provides us with reassurance that the LMM estimation methods can be used for whole-genome sequence data. These results provide some confidence that it will be possible to make meaningful comparisons when partitioning heritability and phenotypic variance onto more classes of genetic variation in the next chapter. However, the differences in the single-VC results at higher minimum MAF thresholds does indicate that differences in the sets of variants available between the two datasets could drive differences in the variance-component estimates. As will be seen in the next section, there are also issues with the interpretation of heritability estimates that need to borne in mind when drawing conclusions from the results, and indeed from all results that use a single variance component in an LMM to estimate heritability for a binary trait.

3.6 Robustness

This section explores the robustness of the single-variance component heritability estimates I have obtained in relation to:

1. Changing the effect-size model;
2. Effects of population structure;
3. Effects of linkage disequilibrium, and
4. Changing the assumed prevalence of T2D.

Each of these factors has the potential to affect the estimates of heritability and the interpretation of heritability results from LMMs, both in this study and more generally.

3.6.1 Changing the effect-size model

Section 3.3 described the linear mixed model framework that was used to estimate heritability in this context. In particular, we introduced the "default" model (Section 3.3.3) in which the effect-size distribution for individual variants is assumed to depend on the minor allele frequency of the variants. I refer to this model as the "allele-frequency dependent" (AFD) model. This default AFD model was used for the results presented in the preceding section. However, as described in Section 3.3.4, an alternative model for the effect size distribution is possible. Under this alternative model, I assume a constant effect size for all variants, thus removing the link between effect size and allele frequency. I refer to this alternative model as the "constant effect size" (CES) model.

MAF Threshold	Number of variants	AFD Model	CES Model
MAF $> 0.1\%$	12,034,435	0.682 (0.193)	0.466 (0.112)
MAF $> 0.5\%$	8,911,402	0.536 (0.154)	0.461 (0.111)
MAF $> 1\%$	7,753,493	0.502 (0.140)	0.458 (0.111)
MAF $> 5\%$	5,360,541	0.490 (0.115)	0.443 (0.106)
MAF $> 30\%$	928,498	0.228 (0.0724)	0.225 (0.0723)

Table 3.5: Estimates of liability-scale variance in risk for T2D explained (standard error in brackets) by single variance-components for the Integrated Panel data. Estimates were obtained using the default model in which variant effect-sizes are allele-frequency dependent (AFD model) and the alternative model in which variant effect-sizes are constant (CES model). Estimates were obtained for minimum MAF thresholds of 0.1%, 0.5%, 1%, 5% and 30%.

The assumed effect-size model has a large effect on heritability estimates obtained from an LMM with a single genetic variance component. The difference in heritability estimates between the AFD and CES models is larger when more variants with lower MAFs are included (Table 3.5 shows results for the Integrated Panel data).

To briefly recap the results from the AFD model discussed above, the highest heritability estimate arises when using variants with MAF greater than 0.1%, then there is a sharp drop when the minimum MAF threshold is raised to 0.5% and then relatively small decreases in the estimates for variants with MAF greater than 1% and MAF greater than 5%. When I raise the minimum MAF threshold to an extreme of MAF greater than 30% (928,498 variants) there is a large, significant decrease in the heritability estimate. These AFD model results suggest a large contribution to heritability from common variants (MAF greater than 5%), which is to be expected as previous chip-heritability estimates have shown this. The contribution from low-frequency variants with MAF in the range of 0.5–5% appears to be small. The large point estimate for the MAF greater than 0.1% suggests a potentially substantial aggregate contribution to heritability from rare variants (MAF 0.1–0.5%), but the uncertainty in the estimates prevents making any such claims with confidence.

A different picture emerges when looking at the results from the constant effect-size model. For the CES model, heritability estimates are almost constant across minimum MAF thresholds of 0.1%, 0.5%, 1% and 5%. There is only a very small increase in estimates when low-frequency and rare variants are included in the model (with differences not significant considering the standard errors). This is to be expected for the CES model. Under the CES model one expects variants, on average, to contribute to the heritability estimates in proportion to their genotypic variance. As the genotypic variance is taken to be $2p(1-p)$, where p is the variant's MAF, genotypic variance is, relatively speaking, high for common variants, low for low-freqency variants and very low for rare variants. The CES model thus gives low-frequency and rare variants less "opportunity" to contribute to the heritability estimate, and this is exactly what is observed in these results.

For a restricted MAF range one expects to see better agreement between the AFD and CES model results, as there is less scope for the different effect-size distribution assumptions to affect the heritability estimates. This is exactly what is observed when using only variants with MAF greater than 30% (928,498 variants). The AFD and CES model heritability estimates and standard errors are almost identical when only using variants in this restricted MAF range.

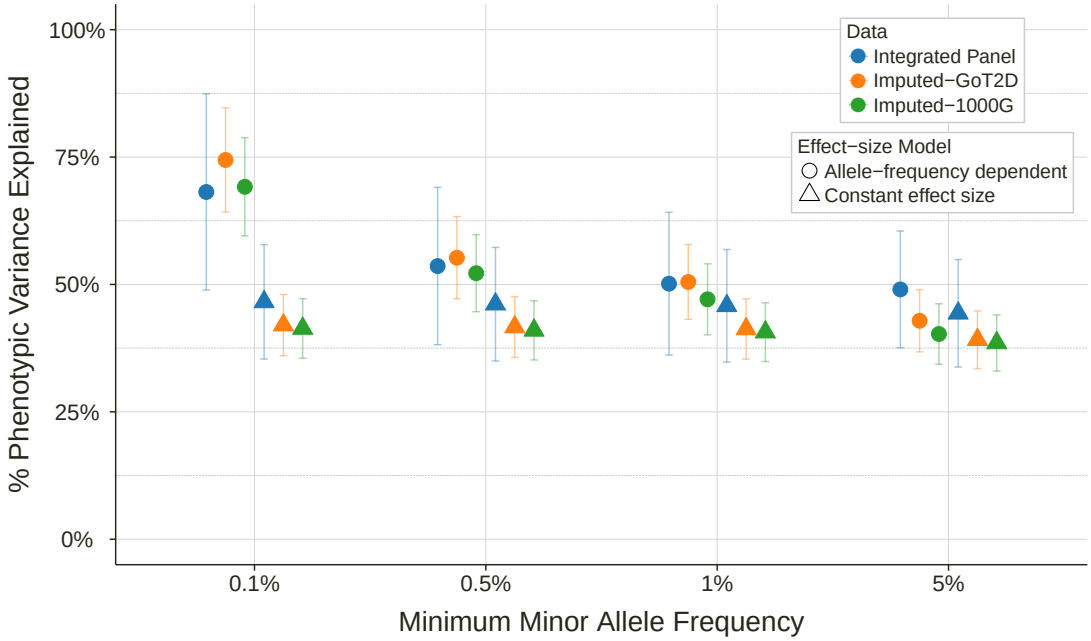


Figure 3.2: Liability-scale heritability (or variance explained) estimates showing the effect of changing assumptions about the effect-size distribution in the model. Estimates are obtained from a model with a single variance component for genetic effects. Results are shown for a range of minimum minor allele frequency (MAF) thresholds for the Integrated Panel, Imputed-GoT2D and Imputed-1000G datasets. Error bars represent ± 1 standard error.

The observations from the Integrated Panel data about the effects of changing the modeling assumptions hold true for the Imputed-GoT2D data and Imputed-1000G data as well (Figure 3.2). As was shown in the previous section, the AFD model heritability estimates from both imputed datasets are highly concordant with those from the Integrated Panel data. For the imputed data, the estimates decrease as the minimum MAF threshold increases, and we observe a large contribution from common variants, little from low-frequency variants and a potentially substantial contribution from rare variants. Unsurprisingly, then, given the agreement in results between imputed and Integrated Panel data for the AFD model, we also see highly concordant results for the CES model. Heritability estimates from the Imputed-GoT2D and Imputed-1000G data stay practically constant as the minimum MAF threshold decreases from 5% to 0.1%. For the imputed datasets, CES model estimates appear lower than the AFD model estimates across all of these minimum

MAF thresholds, with the most substantial difference between the models seen for MAF greater than 0.1% variants.

These results show that the modeling assumptions made about the effect-size distribution for variants have a large effect on heritability estimates obtained across varying MAF ranges. The results are highly concordant between the Integrated Panel, Imputed-GoT2D and Imputed-1000G datasets, demonstrating that this is an effect of the modeling assumptions, and not an effect particular to the data used. The difference in estimates between the constant effect-size model and the allele-frequency dependent model is largest, and substantial, when rare, low-frequency and common variants are all included. The field is currently interested in the possible contribution of rare variants to complex disease, so it is relevant to the interpretation of reported heritability results that the inferred contribution from rare variants could be large if using the AFD model but virtually non-existent if using the CES model.

When estimating heritability using a single variance component across a wide MAF range it seems to be a case of "seeing what you assume you will see" (c.f. confirmation bias) with regard to the contribution from rare variants. Put another way, the estimates of the collective contribution of rare variants are very sensitive to prior assumptions about their effect sizes. If we use a model that gives equivalent weight to rare and common variants (AFD model), then we see a substantial contribution from rare variants, whereas if we fit a model that implicitly gives little weight to rare variants (CES model) then we do not. A possible resolution, allowing us to probe the contribution from rare variants further, is suggested by the near-identical estimates obtained from the two models when using variants with MAF greater than 30%. This result shows that estimates from the two models agree if variants used are restricted to a narrow MAF range. The next chapter explores this further by fitting models with multiple genetic variance components partitioning phenotypic variance by allele-frequency class.

3.6.2 Accounting for and estimating effects of population structure

It is known that population structure can inflate heritability and variance explained estimates from LMMs (Browning & Browning, 2011; Goddard et al., 2011). Unfortunately, despite some recent suggested approaches (Janss et al., 2012), there is no established method to adequately account for population structure. The standard approach, first demonstrated by Yang et al. (2010a) and used in many subsequent studies (Yang et al., 2011b; Lee et al., 2011, 2012a, 2013; Gusev et al., 2013), is to fit a certain number of principal components (PCs; typically computed as the eigenvectors of the GRM) as fixed effects in the LMM. Typically between ten and twenty principal components are included in the model, and this is assumed to account for population structure sufficiently well. In theory there should be an ideal number of principal components to fit, so that the minimum number of PCs

to account for population structure are included, but not any unnecessary PCs that could be removing signal of interest from the results. In practice, there is no commonly-used approach in the field to select the ideal number of PCs to use (in spite of attempts in that direction by Janss et al. (2012)). Below, I look at the stability of heritability estimates to changing the number of PCs included in the model, and then explore two approaches to estimating the remaining population structure effects on the estimates when fitting ten PCs in the model.

3.6.2.1 Fitting principal components as fixed effects

Our default approach has been to fit ten principal components (PCs) as fixed effects in the linear mixed model. Here I assess the robustness of the heritability estimates to fitting different numbers of PCs in the model. Across all minimum MAF thresholds and both effect-size models the highest heritability estimate is obtained when no PCs are fitted (Figure 3.3). As more PCs are fitted the heritability estimates decrease steadily before appearing to stabilise once nine PCs are included. Virtually no change in heritability estimates is seen when between nine and 25 PCs are included in the model. The robustness of the heritability results shown here once at least nine PCs are used justifies the decision to fit ten PCs as the default approach for these analyses.

One cannot assume that fitting ten principal components in the model actually accounts for all of the population structure effects, but it is clear from Figure 3.3 that including more PCs has no further effect on removing any population structure (or other) effects from the heritability estimates. Having thus settled on including 10 PCs for these analyses, I can proceed, in the next section, to use two different approaches to try to quantify remaining effects of population structure on the heritability estimates.

3.6.2.2 Estimating the effects of population structure on heritability estimates

The approach of fitting PCs as fixed effects is not entirely satisfactory, but having shown the robustness of heritability estimates to the number of PCs fitted for the data here, there are two diagnostic approaches available to probe the remaining influence of population structure on the obtained estimates of heritability. In all of these analyses I fit ten PCs as fixed effects, as well as our other default covariates (see Section 3.4.4).

Assessing population structure through individual chromosome heritability estimates Yang et al. (2011b) propose estimating heritability from variants on individual chromosomes as individual variance components jointly in one model ($h^2_c(\text{joint})$), and also separately as single variance components in separate models, one at a time ($h^2_c(\text{sep})$). The general idea is that population structure effects will manifest as correlation between variants on one chromosome with variants on other chromosomes. This correlation would lead to

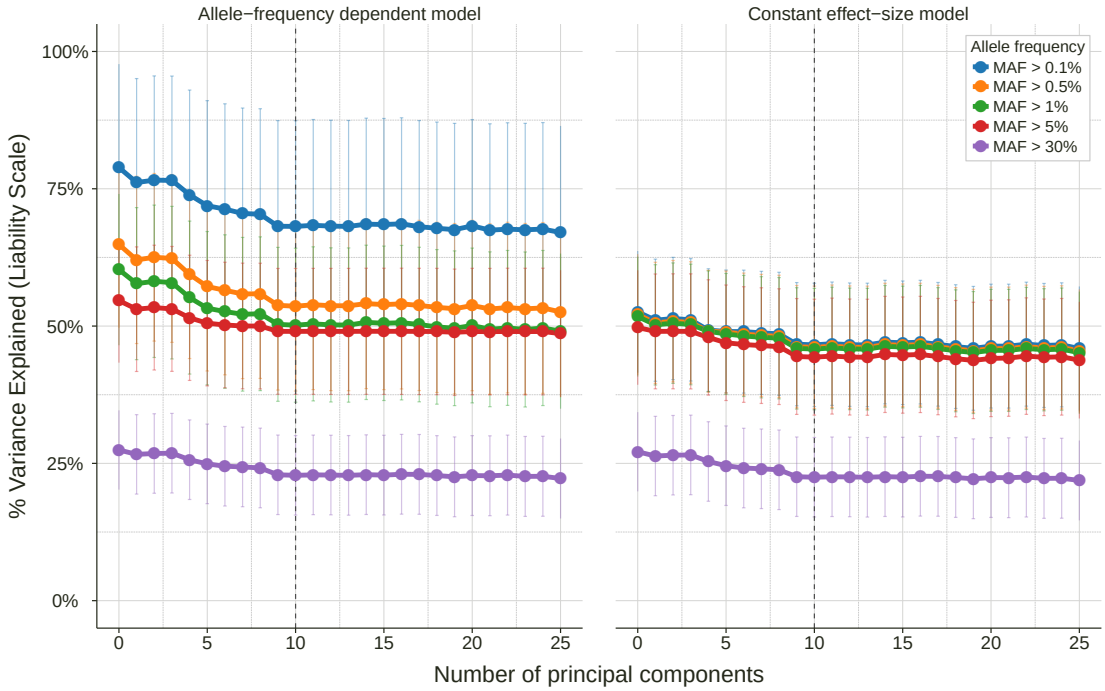


Figure 3.3: Liability-scale heritability (or percentage of phenotypic variance explained) estimates showing the effect of changing the number of principal components fitted as fixed effects in the linear mixed model. Estimates are obtained from a model with a single variance component for genetic effects. Results are shown for the Integrated Panel dataset with no LD-pruning of variants, for a range of minimum minor allele frequency (MAF) thresholds, distinguised by different colours. The left panel shows results when using the default effect-size model (allele-frequency dependent model) and the right panel shows results for the alternative model (constrant effect size model). Error bars represent ± 1 standard error.

overestimation of the genetic variance components in the separate analysis ($h^2_c(\text{sep})$). The difference between the heritability estimates for each chromosome from separate ($h^2_c(\text{sep})$) and joint fits ($h^2_c(\text{joint})$) are plotted against chromosome length. A non-zero intercept to a fitted linear regression line indicates inflation due to cryptic relatedness amongst samples, while a non-zero slope indicates other population structure affecting estimates (under the assumption that longer chromosomes track population structure better than smaller chromosomes). Thus, an ideal result would be to obtain a fitted line with intercept of zero and slope of zero, which would indicate no population structure effects on the heritability estimates.

I apply the Yang et al. (2011b) approach to the Integrated Panel for minimum MAF thresholds of 0.1%, 0.5%, 1% and 5% (as used above) and for three approaches to excluding individuals from the analysis according to their relatedness to other individuals in the cohort (Figure 3.4). I compare the default approach of filtering the individuals so that no pair of individuals has a pairwise-relatedness value greater than 0.05 (see Section 3.4.1.1) to applying a stricter relatedness threshold of 0.025 and to applying no relatedness threshold

(all individuals in the cohort included).

For the default relatedness threshold of 0.05, there is no significant slope to the fitted lines, so no evidence of "other" population structure effects on these estimates. For minimum MAF thresholds of 5% and 1% there is no significant evidence that the intercept of the fitted line is non-zero, so cryptic relatedness does not look to have an effect. For minimum MAF thresholds of 0.5% and 0.1%, however, the intercept looks to be non-zero, albeit small (less than 0.02). Taken together, this analysis suggests little effect of cryptic relatedness or other population structure on the heritability estimates.

Results when applying a relatedness threshold of 0.025 or no relatedness threshold tell a similar story. None of the fitted slopes are significantly different from zero, and only for a minimum MAF threshold of 0.1% does it look like there may be a significant effect for cryptic relatedness. Even in these cases, though, the 95% confidence interval spans approximately zero to less than 0.05, so even in the worst case scenarios here the cryptic relatedness effect is not large. However, there is substantial variability in the difference in per-chromosome estimates (particularly for chromosomes 10 and 11) across all models here, and even small amounts of inflation could become noticeable if I fit multiple variance components (as I will in the next chapter) that are all affected by these inflation effects. The greatest discrepancies in per-chromosome estimates arise for chromosomes 10 and 11. This effect appears to arise because there are relatively large genetic effects on chromosomes 10 and 11. *TCF7L2* and other GWAS-associated loci such as *HHEX*, *IDE* and *KIF11* lie on chromosome 10, and the *TCF7L2* association is the strongest genome-wide for T2D and explains the largest proportion of heritability of any variant or locus. The *MTNR1B* locus, another strong association, lies on chromosome 11. I repeated the analysis shown in Figure 3.4, but excluded variants in or within 1Mb of GWAS-associated loci (data not shown). Compared with the results shown including GWAS loci, the overall differences between jointly fitted and separately fitted estimates are substantially smaller, particularly for chromosomes 10 and 11, and chromosomes 10 and 11 do not appear to have outlying variance explained for their size.

Assessing population structure effects with half-genome heritability estimates Another approach, proposed by Speed et al. (2012), involves dividing the genome roughly in half (using odd chromosomes and even chromosomes) to form two variance components. Similar to the approach above, I then fit the two components jointly in a single model, and separately in two separate single-variance component models. I also compute the variance explained by a single variance component composed of all variants genome-wide. I then compare the sum of the odd and even chromosome components' estimates from the single-variance component models to the single-variance component whole-genome estimate and the sum of the joint fit estimates. If there is no inflation in estimates due to

Figure 3.4: Diagnostics for assessing effect of population structure following the approach of Yang et al. (2011b). This approach plots the difference of the per-chromosome estimate of heritability (h_c^2) when fit individually in separate models ($h_c^2(\text{sep})$) to that when all chromosomes are fit jointly in a single model ($h_c^2(\text{joint})$) against the length of the chromosome in megabases. The columns in the plot show results using all variants outside the second-iteration exclusion regions defined in Section 3.4.1.2 for minimum minor allele frequency (MAF) thresholds of 0.1% (far left, red), 0.5% (green), 1% (blue) and 5% (far right, orange). Rows show results when applying different relatedness thresholds to exclude individuals (see Section 3.4.1.1). Results are shown when no relatedness threshold is applied (top row), when the maximum relatedness between individuals is 0.025 (middle row) and 0.05 (bottom row; default approach). The chromosome number is given in the circle for each point plotted.

population structure, then one expects that the sum of estimates from the odd and even chromosomes in the separate fits will be equal to the sum from the whole genome. If there is a population structure effect, then this should be captured equally by the odd chromosome, even chromosome and whole-genome estimates (as the two halves of the genome should be large enough to capture the full effects of cryptic relatedness and population stratification). Thus, the difference in the sums gives an estimate of the effect of population structure on the heritability estimates.

I apply this approach to the heritability estimates from the Integrated Panel data for minimum MAF thresholds of 0.1%, 0.5%, 1% and 5% (Table 3.6). Unlike in the diagnostic analysis above, this approach estimates a substantial structure effect for the data when us-

ing variants with MAF>0.1% (0.098). The estimated structure effect is greatly reduced (is less than 0.05) as the the minimum MAF threshold is raised to 0.5% and reduces further for minimum MAF of 1% (0.044) and 5% (0.028). The sum of the odd and even chromosome estimates from the joint fit are slightly higher than the estimate from the single-variance component model. This suggests that a small degree of inflation in the total heritability estimate occurs when taking the sum of estimates from multiple variance components fitted jointly. This could be due either to upward bias in the individual REML variance component estimates or inflation due to correlation between the variance components, or both. These effects are important for the multiple variance component models explored in the next chapter, and are discussed in more detail there.

MAF	Component	Single VE Est (SE)	Joint VE est (SE)	Single Sum	Joint Sum	Single WG Est (SE)	Difference (Inflation due to structure)
MAF >0.1%	Odd Chr	0.347 (0.141)	0.299 (0.140)	0.780	0.688	0.682 (0.193)	0.098
	Even Chr	0.432 (0.140)	0.388 (0.139)				
MAF >0.5%	Odd Chr	0.254 (0.111)	0.244 (0.110)	0.585	0.570	0.536 (0.154)	0.049
	Even Chr	0.331 (0.110)	0.326 (0.110)				
MAF >1%	Odd Chr	0.226 (0.101)	0.216 (0.100)	0.546	0.531	0.502 (0.140)	0.044
	Even Chr	0.320 (0.101)	0.315 (0.100)				
MAF >5%	Odd Chr	0.225 (0.0830)	0.216 (0.0821)	0.518	0.504	0.490 (0.115)	0.028
	Even Chr	0.292 (0.0827)	0.287 (0.0822)				

Table 3.6: Half-genome variance explained (or heritability) estimates for T2D to assess the effects of population structure in the Integrated Panel dataset. Results are shown for four different minimum MAF thresholds and use individuals with relatedness less than 0.05 (Section 3.4.1.1). This table gives estimates of variance explained (VE) by variants on the odd- and even-numbered chromosomes when the variance components are fitted in separate models ("Single VE Est") and in a joint model ("Joint VE Est"). The sum of the estimates from the odd and even chromosomes ("Single Sum"/"'Joint Sum") is compared to the VE estimate from a single, whole-genome variance component. Therefore, we can estimate inflation by subtracting the whole-genome estimate ("Single WG Est") of VE from the sum of the estimates from odd and even chromosomes ("Single Sum"). The estimated inflation due to structure is shown in the column "Difference (Inflation due to structure)". The "joint" estimates show how inflation effects in individual components are reduced when two components are fit jointly in a model compared with when the same components are fit as single components in separate models.

These two diagnostic approaches both suggest that there is not a large effect of population structure on heritability estimates when using variants with MAF greater than 0.5%. The Yang et al. (2011b) per-chromosome approach suggested that the population structure effects are negligible, while the Speed et al. (2012) approach suggested that inflation from population structure should be less than 0.05 when using variants with MAF greater than 0.5%. This degree of inflation is not too large in the context of a liability-scale heritability estimate of approximately 50%. When rare variants are included in the model, however, the population structure effect seems to be substantial. The Yang et al. (2011b) approach hinted that structure effects could be present when using variants with MAF greater than 0.1%, while the Speed et al. (2012) approach indicated that the structure effect could be as large as 0.1 for these estimates. Overall, these analyses suggests that there is a population structure effect on the estimates, but that it is not too large when minimum MAF is greater than 0.5%, but is potentially large for variants with MAF greater than 0.1%.

These effects should be borne in mind when interpreting heritability estimates, and will be relevant when assessing inflation of total heritability estimates in the next chapter. In the analyses in the next chapter I do not explicitly account for the effects of structure identified here. Thus, heritability estimates from those models will be slightly inflated by the upward bias introduced by structure effects. One should, therefore, take the "raw" heritability estimates from these models (especially the total heritability estimate in multiple variance component models) with a grain of salt. However, the focus of subsequent analyses is to compare and assess the *relative* contributions of different classes of genetic variation to heritability. As such, even if the estimates of multiple variance components in a given model are all slightly upwardly biased one expects this effect will have little impact on the comparison of relative contributions to variance explained. Indeed, this expectation is borne out by extensive robustness checking of results in the second half of the next chapter.

3.6.3 Adressing linkage disequilibrium

In principle, and possibly in practice, correlations between markers, called linkage disequilibrium (LD) in the genetics context, can bias estimates of the true variance explained or heritability of a set of markers (Speed et al., 2012). LD between variants is often expressed using R^2 to quantify the strength of the correlation between the genotypes of a pair of variants. The conceptual model is that unknown and typically un-genotyped causal variants are "tagged" by genotyped "marker" variants. The problem introduced by LD is that the signal for causal variants across the genome can be tagged with differing efficiencies. Causal variants in LD with many marker variants could have their signal replicated by these correlated markers, giving them more influence than they should in the computation of a GRM. Conversely, causal variants that are only in relatively low LD with a small number of marker variants will not have their contribution measured appropriately. Of course, causal variants that are not in LD with genotyped markers are missed entirely in this analysis.

The LMM approach to estimating heritability grew in popularity with the rise of large sets of samples genotyped on chip arrays. The earlier generations of these genotyping arrays assayed a relatively sparse set of markers across the genome (say 200,000 markers or fewer). LD between these markers was typically low, so the issues caused by LD among marker variants were ignored. However, contemporary genotyping platforms assay much denser variants sets. For example, the Illumina Human Omni 2.5 chip (Illumina, Inc, 2014b) (one of the technologies used for the GoT2D project) contains 2.5 million markers, and the GoT2D Integrated Panel dataset takes this to an extreme with high-confidence genotype calls for over 26 million variants produced by combining Omni array, low-pass whole-genome sequence, and high-depth whole-exome sequence data. With these much higher density genotyping platforms, effects of LD among marker variants could become

problematic. Indeed, Speed et al. (2012) suggest that estimating the narrow-sense heritability for a trait using LMMs can be biased by differential LD-tagging of causal variants by marker variants.

For the single-variance component heritability estimates presented in this chapter, LD effects could influence the results. In the analyses in the next chapter, however, we are much more interested in the *relative* contributions of variants in different allele-frequency and functional classes to heritability than we are in estimates of the total heritability. So, the pertinent question there is whether or not LD is an issue for the variance partitioning analyses. In the paper in which they introduce the variance partitioning approach used in Chapter 4, Gusev et al. (2014) argue that pervasive LD across functional categories is handled by the joint estimation of the variance components in a single model. Joint estimation allows all components to compete for shared variation due to LD (Gusev et al., 2014). They argue that this LMM-based variance components analysis leverages the entire polygenic architecture of the trait and this feature of the LMM approach allows it to perform better than enrichment analyses of top GWAS hits (demonstrated in their analyses). Through simulations, Gusev et al. (2014) show that the LMM partitioning approach gives accurate genome-wide estimates of functional enrichment across varying genetic architectures.

In the Integrated Panel data the near complete catalogue of variation with MAF $> 0.1\%$ has actually directly genotyped. Thus, the causal as well as marker variants in this allele frequency spectrum have presumably been directly genotyped. Therefore, issues raised in the analysis of chip-genotype data (such as questions about tagging of causal variants) will not necessarily apply to genotype data from whole-genome sequencing. There could still be "double-counting" of the effects of some causal variants if they have many tagging variants, but no published work yet discusses LD issues in whole-genome sequence data such as that from the GoT2D study.

The arguments made by Gusev et al. (2014) provide confidence that the LMM partitioning approach in the next chapter should work appropriately on the dense variants in the GoT2D Integrated Panel dataset. As such, the "default" analysis approach includes all Integrated Panel variants (after quality control) with MAF greater than 0.1%. Nevertheless, one could apply "LD-pruning", that is thinning the set of variants used so that remaining variants have pairwise correlations below some R^2 threshold value. As a sanity check, I apply LD-pruning to the set of Integrated Panel variants to check how enrichment results are affected by LD. With a reasonable degree of LD-pruning the "whole-genome" variant set of the Integrated Panel can be reduced to a set that looks much more similar to an imputed dataset (for which Gusev et al. (2014) showed that the LMM partitioning methods perform well).

The obvious objection to LD-pruning genotyped variant sets is that it can be a self-defeating strategy. Either we prune so few variants that we make little change to the re-

sults, or we remove so many variants through stricter LD-pruning that we remove true signal along with "unwanted" correlations. Whatever set of variants we use to define GRMs will effectively "tag" variants that are not directly used in the model. This is, of course, how LMM methods have been able to explain a substantial proportion of "missing heritability" across traits: a relatively small set of directly genotyped SNPs (say 200,000) can tag a large amount of genome-wide variation. Thus, if we prune the set of variants so that remaining variants in the model have a maximum R^2 of, say, 0.8, then we would expect that a very large proportion of genome-wide variation would be captured by the remaining variants, and so we would expect to see little change in results. With stricter LD-pruning, say using an R^2 threshold of 0.3, much of the genome-wide variation will still be tagged by the remaining variants in the model, but we risk removing the signal that we are most interested in. Indeed, Gusev et al. (2014) obtain much better partitioning results when using (dense) imputed data rather than (relatively sparse) chip-genotype data, suggesting that we should not undertake LD-pruning for the partitioning of variance into functional classes.

There is one aspect of the investigation into the genetic architecture of T2D that can benefit from an approach originally designed to account for issues that LD causes in obtaining an absolute estimate of narrow-sense heritability. Lee et al. (2013) propose an approach to accounting for LD in response to the findings of Speed et al. (2012) that SNP-heritability estimates are sensitive to uneven LD between markers in dense genotype data. By "uneven LD", Speed et al. (2012) mean that across the genome local patterns of LD between variants will vary, and so causal variants in certain (unknown) regions will be better tagged by genotyped variants than causal variants in other (also unknown) regions. Speed et al. (2012) speculated that such "uneven LD" may bias heritability estimates and in response Lee et al. (2013) proposed a minor allele frequency (MAF)-stratified approach that gives heritability estimates that are robust to genotyping density and the underlying genetic architecture of the trait. This MAF-stratified, or MAF-binning, approach also provides the appealing possibility of providing insight into genetic architecture by dissecting out the heritability for variants in different MAF-classes.

Lee et al. (2013) claim that their "GCTA MAF-binning" approach breaks down the implicit relationship between SNP allele effects and heterozygosity (i.e. $2p(1-p)$ where p is the frequency of a given SNP), and that VE estimates with this approach are more robust to a range of underlying genetic architectures, different MAF-density distributions, and hence unequal tagging of causal variants. They also claim that the MAF-binning approach performs better than the LD-weighting of variants proposed by Speed et al. (2012), which can give upwardly-biased VE estimates because its weighting strategy is suboptimal for dense genotyping data and attributes too much weight to low-MAF variants.

As mentioned above, the primary focus (presented in the next chapter) is on partitioning phenotypic variance by variant class rather than obtaining an estimate of the narrow-sense heritability for T2D. However, we are interested in the contribution to variance explained from variants across the MAF spectrum. The MAF-binning approach from Lee et al. (2013) provides the means to do this, and also accounts for uneven LD between markers. The single-variance component heritability estimates presented in this chapter represent a standard, but more naive, approach to estimating variance explained by genetic effects. I present them in this chapter because they are a standard analysis conducted in this sort of setting, and provide a baseline against which to compare the partitioning results presented in the next chapter.

3.6.3.1 Single-variance component heritability estimates when LD-pruning variants

In spite of these arguments against LD-pruning outlined above, I undertook some analyses with LD-pruning as a sanity check. I used standard parameter settings in the `PLINK` (Purcell et al., 2007) software package to conduct LD-pruning with thresholds on R^2 between genotypes of 0.8, 0.5 and 0.3. LD-pruning substantially reduces the number of variants used for analysis, particularly in the common variant MAF range (Table 3.7). Here, I look at effects of LD-pruning just on single-variance component estimates for the Integrated Panel data. The next chapter explores effects of LD-pruning on results from multiple variance component partitioning analyses.

MAF Range	LD-pruning Approach	Number of Variants
Rare (MAF 0.1–0.5%)	No pruning	3,123,033
	$R^2 < 0.8$	2,110,658
	$R^2 < 0.5$	1,576,114
	$R^2 < 0.3$	1,221,178
Low-frequency (MAF 0.5–5%)	No pruning	3,550,861
	$R^2 < 0.8$	1,462,902
	$R^2 < 0.5$	964,908
	$R^2 < 0.3$	650,101
Common (MAF $>$ 5%)	No pruning	5,360,541
	$R^2 < 0.8$	936,797
	$R^2 < 0.5$	476,465
	$R^2 < 0.3$	264,721

Table 3.7: Number of Integrated Panel variants used when applying different LD-pruning approaches. When LD-pruning was undertaken, it was done using the default method in the `PLINK` software to prune variants to have a maximum correlation (R^2) between remaining variants of 0.8, 0.5 or 0.3. Here I show the number of variants after pruning for rare (MAF 0.1–0.5%), low-frequency (MAF 0.5–5%) and common (MAF $>$ 5%) variants.

The LD-pruning results for single-variance component heritability estimates on the Integrated Panel data suggest that LD-pruning is not the appropriate strategy here for deal-

ing with any worries about the effects of LD (Figure 3.5). For all models except the allele-frequency dependent model with minimum MAF threshold of 0.1%, the heritability estimates are larger with LD-pruning of variants (with R^2 threshold of 0.8, 0.5 or 0.3) than without any LD pruning. In the most extreme case (AFD model, MAF greater than 0.5%; number of variants 2,399,599, 1,441,373 and 914,822 variants for R^2 thresholds of 0.8, 0.5 and 0.3, respectively) some of the LD-pruned estimates are 50% larger than the no-pruning estimate. The LD-pruned estimates for the AFD model with MAF greater than 0.1% are more like what we might expect to see: the estimate for $R^2 < 0.8$ (4,510,257 variants) is very close to the estimate obtained without pruning, and when more variants are removed (3,017,486 variants used for $R^2 < 0.5$ and 2,136,000 variants for $R^2 < 0.3$) the heritability estimate decreases. It seems likely that these results are closer to what we expect, because LD-pruning has much less of an impact for rare variants than it does for low-frequency variants and (most extreme) common variants (Table 3.7). This accords with the results for the CES model, where the results for minimum MAF of 0.1% are very similar to those applying higher MAF thresholds. This is what is expected from the CES model, in which rare variants are given very little weight (as discussed above).

The standard errors for the LD-pruned estimates are substantially larger than those from the no-pruning estimates. This, again, is a somewhat surprising outcome, as for some models, such as when increasing the minimum MAF threshold, smaller standard errors are observed when fewer variants are used. Thus, when fitting far fewer variants in the model when applying LD-pruning than without LD-pruning, one might expect smaller standard errors. Results for the imputed datasets showed that generally smaller standard errors are obtained when fewer variants are used in the model, although that comparison is a little problematic as those smaller standard errors are likely also influenced by the larger sample size for the imputed data and different characteristics of that particular sample. Nevertheless, the fact that larger standard errors arise for the LD-pruned results when using fewer variants suggests that some aspect of LD-pruning upsets the heritability estimates, at least on the Integrated Panel data.

Overall, LD-pruning does not look like a good strategy for accounting for any LD effects for these data. Other methods exist that try to adjust heritability estimates to account for LD effects, but we did not have any success in applying the software implementations of those methods to the large Integrated Panel dataset (data not shown). On the basis of these results, it looks better to avoid LD-pruning for heritability analyses.

3.6.4 Effects of changing the disease prevalence value

Throughout all of the analyses I report heritability estimates on the liability scale as it is more interpretable and allows comparison of results between datasets and traits. As discussed in Section 3.3.8, transforming raw-scale variance component estimates to liability-

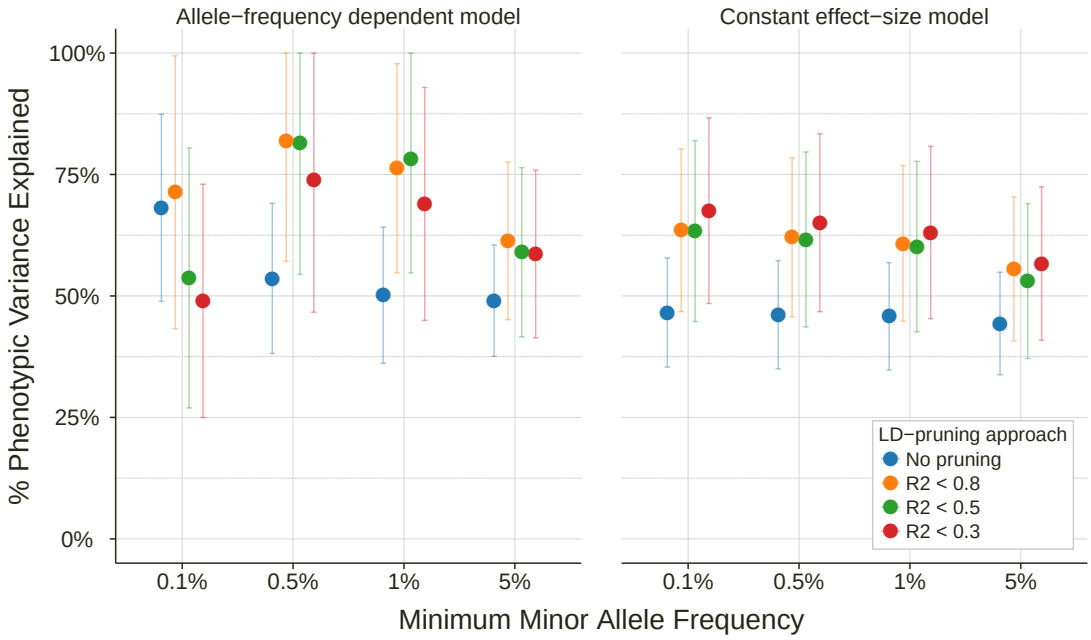


Figure 3.5: Liability-scale heritability (or percentage of phenotypic variance explained) estimates showing the effect of changing the approach to pruning variants to reduce linkage disequilibrium. Estimates are obtained from a model with a single variance component for genetic effects. Results are shown for the Integrated Panel dataset, for a range of minimum minor allele frequency (MAF) thresholds. Different colours distinguish between the LD-pruning approaches: no LD-pruning (our default), and pruning variants such that the maximum correlation (R^2) between variants used for the analysis is 0.8 ($R^2 < 0.8$), 0.5 ($R^2 < 0.5$) or 0.3 ($R^2 < 0.3$). The left panel shows results when using the default effect-size model (allele-frequency dependent model) and the right panel shows results for the alternative model (constrant effect size model). Error bars represent ± 1 standard error.

scale heritability estimates is straight-forward. However, the transformation strongly depends on the value assumed for the prevalence of the binary trait. In the default analyses, I have assumed the widely-used prevalence of T2D in European populations of 8%. However, since the liability-scale heritability estimates are sensitive to the prevalence value, I examine here the effects on our heritability estimates of using different prevalence values. After all, a prevalence value of 8% could be slightly misspecified (King et al., 1998), and we also face the possibility that a prevalence of 8% is not the appropriate value for the GoT2D cohort. After all, an opportunistic version of extreme-phenotype sampling was applied for the GoT2D project (see Section 3.2.1). Thus, the population-wide prevalence of 8% for T2D may be too high. It could be argued that a lower prevalence would be more appropriate when looking at more extreme T2D case-control phenotypes.

As expected from the theoretical results, changing the prevalence value has a large effect on liability-scale heritability estimates (Figure 3.6). Across the Integrated Panel, Imputed-GoT2D and Imputed-1000G datasets, doubling the assumed prevalence from 8%

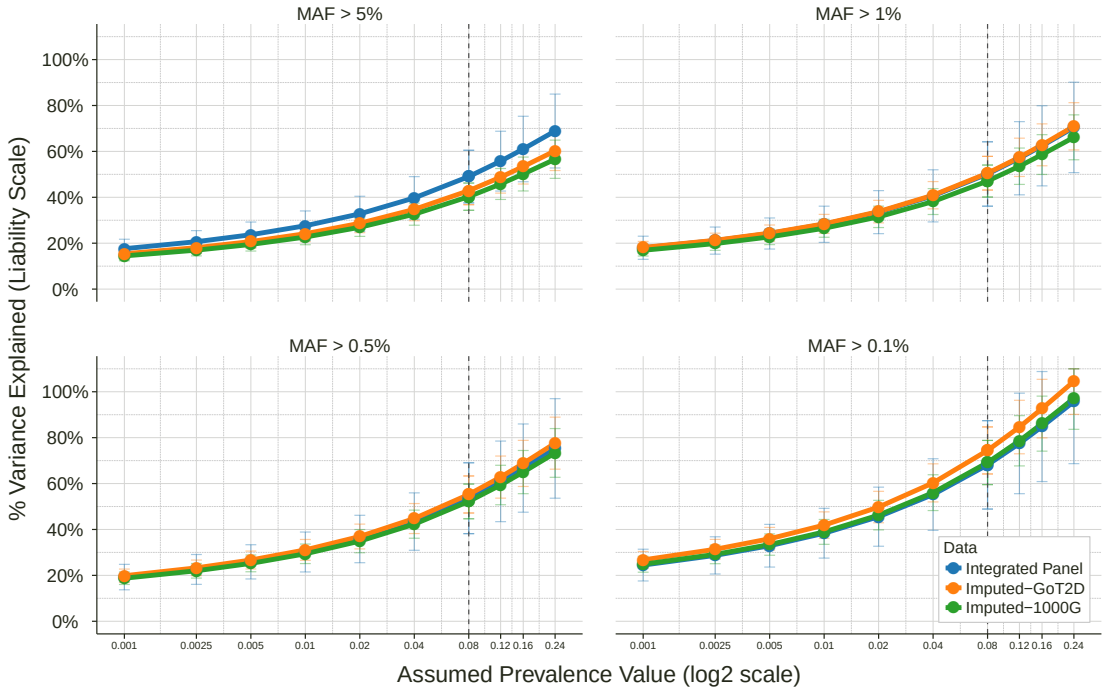


Figure 3.6: Liability-scale heritability (or percentage of phenotypic variance explained) estimates showing the effect of changing value assumed for the prevalence of the binary trait (here T2D case-control status). The x-axis shows possible assumed prevalence values on a log2-scale, and the y-axis shows the liability-scale heritability estimate corresponding to applying that prevalence value to transform observed-scale REML estimates to the liability scale. The vertical, black dotted line is placed at a prevalence of 0.08 (8%), which is the default value used for the analyses. Estimates are obtained from a model with a single variance component for genetic effects. Results are shown for the Integrated Panel, Imputed-GoT2D and Imputed-1000G datasets, for a range of minimum minor allele frequency (MAF) thresholds. Different colours distinguish between the different datasets. The four panels show results when using different minimum minor allele frequency (MAF) thresholds for the variants used in the analyses (MAF $>$ 5%, MAF $>$ 1%, MAF $>$ 0.5% and MAF $>$ 0.1%). Error bars represent ± 1 standard error.

to 16% would increase the heritability estimates by roughly 20%, while halving the prevalence to 4% would reduce heritability estimates by a similar amount. Completely misspecifying the prevalence can lead to nonsensical results: if one assumes a prevalence of 24% (grossly inflated for a contemporary European population, but actually a realistic prevalence for a country like Nauru with very high incidence of T2D), then when using variants with MAF greater than 0.1% one could actually obtain a liability-scale heritability estimate greater than 100%. At the other extreme, if one assumes a prevalence value of 0.1%, then one would get a liability-scale estimate of approximately 20% across all MAF thresholds and datasets.

The dependence of liability-scale heritability estimates on the assumed prevalence is strong. Thus, the assumed prevalence value has a large influence on our interpretation of liability-scale heritability estimates. If the assumed prevalence is lower than it should be

then we will underestimate the heritability, and if it is higher than it should be then we will overestimate the heritability. Given the extreme-sampling approach in the GoT2D project, it could be argued that a lower prevalence value would be more appropriate, although it is not clear exactly what a better prevalence value would be. Taking a prevalence value of 4% instead of 8%, which does not seem unreasonable, would reduce the heritability estimates by about 20%. Such a change would affect the interpretation of our heritability results and how they compare to heritability estimates from other studies. For example, a large estimate from these results of 68% (Integrated Panel, MAF greater than 0.1%) when using a prevalence of 8% would be less than 60% with a prevalence of 4%, and then a result more concordant with previous estimates of the heritability for T2D. If we do see heritability estimates that seem high, one possibility is that the prevalence assumed is higher than it should be.

3.7 Discussion

This chapter introduced the study of heritability and phenotypic variance explained for type 2 diabetes. It described the GoT2D project of which this work forms a part, and the Integrated Panel and imputed datasets that were used for the analyses. Here, I restricted the analyses to single-variance component estimates of heritability from linear mixed models, a standard approach in the field for obtaining heritability estimates from genome-wide variants in unrelated individuals in a case-control setting. To my knowledge, this is the first time that whole-genome sequence data has been used for these types of analyses, which have been previously been conducted on chip-genotype data (with and without imputation of larger numbers of variants).

The results show that common variants explain a large amount of the variance in liability for T2D. Using the default model, I obtained a liability-scale heritability estimate of 0.49 (s.e. 0.12) when using variants with MAF greater than 5%. The contribution from rare variants to susceptibility in T2D remains unclear. I obtained a high point estimate (0.68, se.e. 0.19) for the liability-scale heritability when using variants with MAF greater than 0.1%, but the uncertainty in the estimate is high. There is the possibility of a substantial contribution to T2D heritability from rare variants, but I cannot claim this with confidence. I also observed that the apparent contribution from rare variants depends greatly on the effect-size model assumed. When the default, allele-frequency dependent model is used, the contribution looks large, but when the alternative, constant effect size model is used, the contribution from rare variants appears negligible. Larger sample sizes will be required to characterise the contribution from rare variants more definitively.

Heritability estimates from the Integrated Panel dataset (incorporating whole-genome sequence data) were broadly concordant with estimates from two imputed datasets with

larger cohort size. I did observe, however, lower heritability estimates for common variants from the imputed datasets (0.43, s.e. 0.06 and 0.40, s.e. 0.06). The imputed dataset results showed the same patterns across different minimum MAF thresholds and the two effect size models as the Integrated Panel results. The agreement of the imputed data and Integrated Panel results provide evidence that the LMM-based approach to estimating heritability originally developed for sparse chip-genotype data does perform comparably when applied to very dense genotype data derived from whole-genome sequence, exome sequence and chip data.

The robustness of the heritability estimates was examined in various ways. The liability-scale heritability estimate is strongly affected by the assumed prevalence value for a binary trait. Prevalence of T2D can vary greatly between populations, from approximately 8% in western Europe to more than 30% in certain Pacific Island populations (Zimmet et al., 2001). The strong effect of the assumed prevalence value is important for the interpretation of the heritability results, since it is possible that the standard prevalence of 8% for T2D that I have applied is higher than it ought to be for the extreme-phenotype study design of GoT2D. The dependence of liability-scale estimates on prevalence is vital to bear in mind when interpreting heritability results for any binary trait. I saw that LD-pruning variants resulted in suspiciously high heritability estimates, and concluded that LD-pruning is not a good strategy for the types of analyses conducted here.

The heritability estimates that I have obtained here (ranging between 40% and 70% across minimum MAF thresholds and modeling assumptions) are broadly consistent with previously published estimates of the heritability of T2D. I note, however, that previously published estimates cover an even wider range than my results, from as low as 26% in a study from the Danish Twin Register (Poulsen et al., 1999) to as high as 69% in patients with age at onset 35–60 years in a family-based Botnian cohort (Almgren et al., 2011). From the point of view of study design (with an emphasis on early-onset T2D cases) the GoT2D cohort seems most similar to the cohort in the Almgren et al. (2011) study, so perhaps it is not unreasonable to see heritability estimates up to 68% in my analyses. More recent analyses on chip-genotype data (with and without imputation) which obtained estimates of T2D heritability between 30% and 60% depending on modeling assumptions and whether imputed variants were used or not (Gusev et al., 2013). Overall, then, while some of the heritability estimates appear large at first glance, they are eminently plausible in the context of other published estimates for the heritability of T2D.

These single-variance component estimates of the liability-scale heritability of T2D provide a useful baseline. The results look broadly sensible, and have been able to probe the robustness of the LMM-based heritability estimates in various ways. However, the results here are not conclusive regarding the contribution to T2D risk from rare and low-frequency variants, but the baseline that these results provide will be useful to bear in mind for the

analyses I conduct in the next chapter. Analyses in Chapter 4 partition the heritability onto variants in different allele-frequency and functional classes. The aim is to gain a finer-grained view of the contributions to heritability from different classes of genetic variation, probing further the contribution to disease liability from rare variants and, possibly, functional variants of different types identified from multiple cell types.

Chapter 4

Using variance partitioning to investigate the contribution of different classes of genetic variation to type 2 diabetes susceptibility

4.1 Introduction

This chapter extends the analyses presented in the previous chapter to investigate the relative contributions of different classes of genetic variation to T2D susceptibility. These investigations build on the baseline results provided by the single-variance component estimates of heritability. Using the same datasets I will partition phenotypic variance onto variants in different allele-frequency classes, and variants in different functional annotation classes. Linear mixed model methods continue to be used for estimating heritability and variance explained, but here I will fit models with multiple genetic variance components to partition heritability onto different classes of genetic variation in various ways.

There are two main thrusts to the variance-partitioning analysis of the GoT2D data:

1. **MAF classes models:** Variance components are computed using sets of variants stratified by minor allele frequency.

2. **Annotation classes models:** Sets of variants are defined based on annotation class (for example, "enhancer", "coding" and so on) and variance components are fitted using the variants for each annotation class.

In both cases, I fit multiple variance components in the LMM simultaneously and obtain REML estimates of VE for each class. The analysis of variance partitioning by functional class begins with high-level annotation categories (e.g. coding, UTR, promoter, and so on), but then homes in on enhancer variants, particularly those active in pancreatic islet cells. The focus is on enhancer variants as the analysis highlights their importance and they have also recently been shown to be particularly relevant for T2D (Pasquali et al., 2014).

In the previous chapter I used LMM methods to estimate the heritability of T2D from whole-genome sequence and imputed genotype data. Larger estimates were obtained for the estimates of T2D heritability when including rare and low-frequency variants along with common variants in a single-variance component model. However, results showed that modeling assumptions had a substantial effect on single-variance component estimates of heritability. This chapter, seeks to establish the contribution to T2D risk from rare and low-frequency variants in a finer-grained way by partitioning phenotypic variance into multiple allele frequency classes. Similar estimates were observed from the two effect-size models when restricting variants used to a narrow MAF range. Below, I fit multiple variance components covering in combination a wide MAF range, but each individually using variants in a narrow MAF range. From these results we will be able to assess agreement between the two effect-size models when using this partitioning approach. I will also attempt to separate contributions to heritability from variants in different MAF ranges, which may give us further insight into the collective importance of rare and low-frequency variance in explaining risk for T2D.

In addition, an aim is to identify what types of functional variants contribute more or less to T2D risk. As demonstrated in Chapter 2, great effort goes into assigning functional annotations to sequence variants (even though the resulting annotations remain imperfect). Since almost all thus-far T2D-associated loci fall outside genic regions (Morris et al., 2012), the debate continues about whether genic or non-genic variation is more important for the aetiology of the disease. It seems an excellent opportunity, especially given the vastly increased coverage of genomic variation afforded by the GoT2D Integrated Panel data, to add functional annotation information to augment the analyses.

The results of the GoT2D Genomes association analysis (Flannick et al., 2015) show that, in existing sample sizes, it is not possible to resolve most common variant signals to individual causal variants through genetic association alone. However, functional annotations can be used to add value to the interpretation of association signals. Schaub et al. (2012) proposed the intersection of variant sets with functional annotations as a complementary approach to pinpoint causal variants, as did Maurano et al. (2012) to pinpoint disease-relevant annotations. However, previous annotation-based analyses have primarily used association summary statistics (for example, *p*-values or estimated effect sizes) derived from incomplete variant catalogues (ENCODE Project Consortium et al., 2012; Trynka et al., 2013; Schaub et al., 2012; Maurano et al., 2012; Parker et al., 2013; Pasquali et al., 2014; Gaulton et al., 2010). Here I use an alternative LMM-based variance partitioning methodology proposed by Gusev et al. (2014) and a more extensive catalogue of variation to gain further insight into the genetic architecture of T2D. This approach may help to prioritise potential causal variants.

I use the same datasets here as in the previous chapter for the single-variance component heritability estimates. Thus, most of the information on these datasets is described in Section 3.2. Some further information on the datasets, primarily related to variant annotation, is presented in this chapter in Section 4.2. Similarly, I again use the LMM framework for estimating heritability described in Section 3.3. To partition heritability onto multiple variant classes, extensions to these methods are needed. The necessary methodological extensions are presented in Section 4.3. Primarily, I present a brief exposition of methods proposed by Gusev et al. (2014). Credit for the idea to partition heritability onto different functional classes of variant and development of methods to do so goes to Gusev et al. (2014). However, the suggestion to partition variants by allele-frequency was first made by Lee et al. (2013), albeit with the focus there on obtaining more accurate total heritability estimates, rather than assessing the relative (and absolute) contributions to heritability from variants in different allele-frequency classes. I have made original contributions as described in the introductory paragraphs of Section 3.3.

The analyses here are complicated. I fit many different models on three datasets, and there are many factors that could influence the results obtained. To simplify the narrative for the reader, I present "streamlined" versions of the main results, followed by detailed examination of the robustness of those results. Section 4.4 presents the main results when partitioning into multiple allele frequency classes using Integrated Panel data. Section 4.5 describes the main results when partitioning into multiple functional classes using Integrated Panel data. In Section 4.6, I present results when fitting the same models on the two imputed datasets both for allele-frequency partitioning (Section 4.6.1) and functional-class partitioning (Section 4.6.2). Following the presentation of the main results in those sections, Section 4.7 addresses the robustness of the variance partitioning results in many different ways, with discussion about many possible sources of bias in variance partitioning analyses. Finally, Section 4.8 contains discussion and conclusions.

Learning about the different contributions to variance in T2D risk from variants in different allele-frequency or functional annotation classes could be helpful for our understanding of the disease. Such information would complement everything the field has learned from association approaches, and could help with the prioritisation of associated variants for follow-up study (see Flannick et al., 2015), help with future study design and enable other forms of analysis, and help researchers to gain a better understanding of the disease using a top-down (overall aggregate effects of classes of variation) to complement the usual bottom-up (individual variant associations) approach.

4.2 Variant annotation

I use the same datasets for the variance partitioning analyses in this chapter as for the single-variance component analyses in the previous chapter (described in Section 3.2). To conduct partitioning analyses based on functional classes, however, more information on the variants is needed than was provided in Section 3.2, namely functional annotations. Exactly the same variant annotation approach was taken for the Integrated Panel, Imputed-GoT2D and Imputed-1000G datasets. Genomic and functional information from a number of sources was combined to annotate variants, work carried out by Kyle Gaulton.

The decision was taken to do variant annotation directly using raw data sources instead of using standard automated tools, as discussed in Chaper 2, for two reasons. First, for variance partitioning, we can only use a relatively small number of broad annotation classes, so fine-grained annotation of exonic variants (e.g. stop-gain, stop-loss, non-synonymous and so on) is not required. For these analyses it is sufficient simply to know if a variant falls in any coding or non-coding transcript or UTR. Such information can be obtained easily from GENCODE (Harrow et al., 2012). Thus, for this annotation there was no need to use standard tools and chose to minimise the issues that come from their use. Second, standard tools do not annotate variants in regulatory regions in sufficient detail for this study. Here, I am particularly interested in annotating promoter, insulator, enhancer (for many cell types) and transcription factor binding site (TFBS) variants. These annotation classes are not provided by standard tools, which focus on variants in protein-coding regions.

Gene transcript annotations were obtained from GENCODE version 14 (Harrow et al., 2012). For protein-coding genes, transcripts with a 'protein-coding' tag were filtered for either presence in the conserved coding DNA sequence (CCDS) database (Pruitt et al., 2009b) or with experimentally confirmed start and end, and 5' UTR, exon, and 3' UTR regions were identified from the resulting set of transcripts. Transcripts with a tag of 'lncRNA', 'miRNA', 'snoRNA' or 'snRNA' were identified for non-coding genes.

Regulatory function information was obtained using published resources for epigenetic data. Data were used from nine ENCODE cell types (ENCODE Project Consortium et al., 2012), pancreatic islet cells from a T2D study (Pasquali et al., 2014) and two types of human adipose stromal cells (hASC-t1 and hASC-t4) from a study of adipogenesis (Mikkelsen et al., 2010). Sequence reads were collected from these sources for five chromatin immunoprecipitation (ChIP) assays (H3K4me1, H3K4me3, H3K27ac, H3K36me3 and CTCF), using the following nine ENCODE cell types (ENCODE, 2012):

GM12878 B-lymphocyte, lymphoblastoid, cells with Epstein-Barr Virus exposure from a European Caucasian individual in the International HapMap Project

hESC embryonic stem cells

HepG2 hepatocellular carcinoma (liver cancer)

K562 leukemia cell line established from lung fluid of a 53-year-old female with chronic myelogenous leukemia in terminal blast crisis (the final phase in the evolution of CML)

HSMM normal skeletal muscle myoblasts

HUVEC umbilical vein endothelial (organ-lining) cells

NHEK epidermal keratinocytes (normal skin)

NHLF normal lung fibroblasts

HMEC normal mammary epithelial cells

Reads were mapped to the human reference sequence, hg19 (Rhead et al., 2010b) using `BWA` (Li & Durbin, 2009). `ChromHMM` (Ernst & Kellis, 2012) was applied to the resulting mapped reads for all cell types to call regulatory states, assuming 10 states. The following names were assigned to the resulting 10 state definitions, characterised by the following combinations of chromatin marks:

Active promoter: High H3K4me3 and H3K27ac

Strong enhancer 1: H3K4me3 and H3K27ac and H3K4me1

Strong enhancer 2: H3K27ac and H3k4me1

Weak enhancer: H3K4me1

Poised promoter: H3K27me3 and H3K4me3 and H3K4me1

Repressed: H3K27me3

Low/no signal: the state identified by the algorithm was characterised by having few or no chromatin marks present

Insulator: CTCF

Low/no signal: another state characterised by few or no chromatin marks

Transcription: H3K36me3

Regulatory state maps for pancreatic islets were also collected from Parker et al. (2013). Transcription factor binding ChIP sites were obtained from three sources: 141 proteins from ENCODE, five proteins from Pasquali et al. (2014) and one protein from Mikkelsen et al. (2010).

From these collective annotation data were defined seven functional classes across cell types, and given interest in enhancer elements, twelve cell-type specific enhancer annotations were also created (Table 4.1). Thus, there are also annotation classes for islet-enhancers (enhancer elements identified specifically in pancreatic islet cells), skeletal-muscle-enhancers (enhancer elements identified in HSMM cells), hASC-t1 and hASC-t4 adipose-enhancers (enhancer elements from the hASC-t1 and hASC-t4 cells), and specific classes for the remaining ENCODE cell types.

Using this annotation approach, the following seven broad functional classes were defined, using the generic promoter, insulator and enhancer classes:

Functional Class	Description
Coding	Exons from GENCODE protein coding transcripts
UTR	3′ and 5′ UTR regions from GENCODE protein coding transcripts
Promoter	Active and poised promoter elements from ChromHMM definitions pooled across 12 cell types
Insulator	Insulator elements from ChromHMM definitions pooled across 12 cell types
Enhancer	Strong and weak enhancer elements from ChromHMM definitions pooled across 12 cell types
TFBS	ChIP-seq binding sites pooled across 165 transcription factors
ncRNA	non-coding RNA transcripts from GENCODE
Gm12878 enhancer	Strong and weak enhancer elements identified in Gm12878 cells
hESC enhancer	Strong and weak enhancer elements identified in hESC cells
hASC-t1 enhancer	Strong and weak enhancer elements identified in pre-adipose stem cell (hASC-t1) cells
hASC-t4 enhancer	Strong and weak enhancer elements identified in mature adipose stem cell (hASC-t4) cells
HepG2 enhancer	Strong and weak enhancer elements identified in HepG2 cells
HI (Islet) enhancer	Strong and weak enhancer elements identified in pancreatic islet cells
HMEC enhancer	Strong and weak enhancer elements identified in HMEC cells
HSMM enhancer	Strong and weak enhancer elements identified in HSMM cells
HUVEC enhancer	Strong and weak enhancer elements identified in HUVEC cells
K562 enhancer	Strong and weak enhancer elements identified in K562 cells
NHEK enhancer	Strong and weak enhancer elements identified in NHEK cells
NHLF enhancer	Strong and weak enhancer elements identified in NHLF cells

Table 4.1: Genomic annotation categories used in variance component analysis. All enhancer elements are assigned using the ChromHMM definition of enhancer elements, namely the presence of H3k4me1 marks, where co-occurrence with H3K27ac marks or both H3K27ac and H3K4me3 marks defines a strong enhancer.

Coding protein coding transcripts

ncRNA non-coding RNA transcripts

UTR 3′ and 5′ UTR regions of coding transcripts

Promoters active and poised promoter elements, union across all cell types

Insulators union across all cell types

Enhancers strong and weak enhancer elements, union across all cell types

TFBS transcription factor binding sites, with sites pooled across all factors.

I used these annotation classes (plus an "Other" class for remaining variants) for several of the annotation models discussed below.

Across the analyses here, I only use the generic promoter and insulator categories. However, prompted by general interest in the influence of enhancer and other noncoding regulatory variants in T2D and other complex traits (Harismendy et al., 2011; Parker et al.,

2013; Pasquali et al., 2014; Gusev et al., 2014), I looked at the enhancer variants more closely. In models discussed in this section, I define finer-grained, cell-type specific categories for enhancer variants for each of the twelve cell types. That is, for each cell type (such as islet cells) separately, we identify strong and weak enhancer elements by conducting genome segmentation only using the ChIP-seq data for that cell type (see Section 4.2). This approach allows the definition of Islet-Enhancers, Gm12878-Enhancers, K562-Enhancers and so on, as listed in Table 4.1. The Islet-Enhancers, strong and weak enhancer elements identified from pancreatic islet cells (Pasquali et al., 2014), become a major focus of the analysis, so I also define the "Other-Enhancer" class, which comprises variants that fall in the union of strong and weak enhancer regions combined across the eleven non-islet cell types, but do not overlap Islet-Enhancer regions.

These annotation categories enable the exploration of the contribution to variance in susceptibility to T2D through the lens of genomic function. The aim is to identify certain functional categories that explain a large proportion of phenotypic variance. Further, through enrichment scores one can assess which functional categories appear to explain more genetic variance than expected under the null model that all variants contribute equally to phenotypic variance explained, regardless of the functional class to which they belong. The enrichment scores (see Section 4.3.2) compare the observed proportion of total genetic variance explained by a given variance component to the proportion expected. For the allele-frequency dependent effect-size model presented in this section, the expected proportion of variance explained is simply the proportion of all variants used in the model that appear in the given annotation class.

4.3 Methods

This section presents some extensions to the linear mixed model methods for estimating heritability in a single-variance component model described in Section 3.3. Section 4.3.1 characterises the multiple-variance component LMMs used to partition heritability onto different classes of variant. Section 4.3.2 describes the "enrichment scores" that we use to quantify the extent to which a class of variants contributes more or less to the total heritability than would be expected.

4.3.1 Extending the model to multiple variance components

The first model, presented in Section 4.3.1.1, is a non-overlapping, multiple-component model, as proposed by Gusev et al. (2014). In this model, any given variant is only used for one variance component. This application of this model is obvious when partitioning variants into different components by allele frequency (as a variant cannot have more than one value for its allele frequency), but more subtle when partitioning variants by functional

class. It is not uncommon for a variant to have more than one possible annotation. If this is the case, then a choice must be made with regard to which component the variant contributes. This leads to a "hierarchical" prioritisation of annotation classes to give us disjoint sets of variants to use for each variance component. This approach is closely related to the "MultiBLUP" method from Speed & Balding (2014), although MultiBLUP focuses on prediction of SNP effect sizes.

The second model, in Section 4.3.1.2 offers one possible alternative to enforcing a hierarchy on annotation classes. It is a simple extension to the first model that allows a given variant to contribute to multiple variance components, useful when a variant has multiple possible annotations and to provide another perspective on the contributions to heritability from different variant classes. Hilary Finucane suggested this approach to me (personal communication), but I believe that the details of this particular non-hierarchical partitioning approach are novel and have not been described in the literature.

4.3.1.1 Non-overlapping (hierarchical) variance components

Let us first consider a disjoint categories of variants—these could be functional or MAF classes—each containing a set of variants defined by M_c, $c = 1, \ldots, a$. The categories are non-overlapping, so a variant will appear in only one category. This setting leads to a hierarchy of annotation categories to enforce disjoint variant sets. We extend the model described in preceding sections to express the phenotype as a sum of individual variant effect sizes, but here allow different variance components (and therefore effect-size distributions) for the different categories:

$$\boldsymbol{Y} = Z\boldsymbol{\beta} + \sum_{c=1}^{a} \sum_{k \in M_c} X^{(k)} u_k + \boldsymbol{e} \tag{4.1}$$

where variant effect sizes u_k are drawn from individual normal distributions:

$$u_k \sim \mathrm{N}\left(0, \frac{w_k}{\sum_{k \in M_c} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} \sigma^2_{gc}\right), \tag{4.2}$$

and $\boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, I_n \sigma^2_e)$ as previously. This gives us a vector of variance components $\boldsymbol{\sigma}^2_g = (\sigma^2_{g1}, \ldots, \sigma^2_{ga})$.

Extending the result from Equation 3.11, we then model the variance of the phenotype as:

$$\mathrm{V}(\boldsymbol{Y}) = \mathrm{var}(\boldsymbol{Y}) = \sum_{c=1}^{a} K_c \sigma^2_{gc} + I_n \sigma^2_e \tag{4.3}$$

where each K_c is a GRM computed from the SNPs in category c:

$$K_c = \frac{1}{\sum_{k \in M_c} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} X_c \cdot \mathrm{diag}(\boldsymbol{v}) \cdot X_c^T, \tag{4.4}$$

where X_c represents the genotype matrix for category c, and $\boldsymbol{v}$ is a vector of weights, $(w_1, w_2, \ldots, w_m)$. The corresponding variance components (σ^2_g, σ^2_e) are then jointly inferred

using the REML algorithm in GCTA. This yields the term for the variance explained by each category of variant:

$$VE_{gc} = \frac{\sigma^2_{gc}}{(\sum_{j=1}^{a} \sigma^2_{gj} + \sigma^2_e)}. \tag{4.5}$$

The REML output yields an estimate of the error-covariance matrix of the variance component estimates, whichever fitting algorithm is used. For example, if the Average Information algorithm is used, the inverse of the final AI matrix gives an estimate of the error-covariance matrix.

From the variance-component estimates we can compute the total variance explained by measured genetic factors, $VE_g = \frac{\sum_{j=1}^{a} \sigma^2_{gc}}{(\sum_{j=1}^{a} \sigma^2_{gj} + \sigma^2_e)}$ and an enrichment score for each category of variant, $\%VE_{gc}$. Using the error-covariance matrix and the delta method we can compute standard errors on VE_g and each $\%VE_{gc}$ while accounting for error correlations. I refer to standard errors from the delta method as "analytical" standard errors.

4.3.1.2 Non-hierarchical partitioning of variants

As alluded to above, many variants will have multiple valid functional annotations. For example, a variant could be at a location that is in both an enhancer region and a TFBS region. In this setting, we might want to permit a variant to contribute to more than one annotation class, if appropriate. If we could do this, then imposing a hierarchy on the annotation categories (where we must prioritise, say, the enhancer annotation over the TFBS annotation) would not be required. A "non-hierarchical" partitioning of variants could provide us with an alternative perspective on the relative contibutions to heritability of variants in different annotation classes. The non-hierarchical multiple-component model presented in this section allows us to do this, in one particular way. As explained above, the description of this model is, to the best of my knowledge, a novel contribution.

As above, let us consider a categories, each containing a set of variants defined by M_c, $c = 1, \ldots, a$, which may be overlapping. I use the same model as in Equation 4.1, but now we assume the following effect-size distributions:

$$u_k \sim \mathrm{N}\left(0, \left(\frac{\mathbb{1}_1 \sigma^2_{g1} \cdot w_k}{\sum_{k \in M_1} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} + \frac{\mathbb{1}_2 \sigma^2_{g2} \cdot w_k}{\sum_{k \in M_2} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} + \cdots + \frac{\mathbb{1}_a \sigma^2_{ga} \cdot w_k}{\sum_{k \in M_a} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)}\right)\right), \tag{4.6}$$

where $\mathbb{1}_c$ is an indicator variable, which is equal to 1 if variant $k \in M_c$ and zero otherwise. This model effectively assumes additive effects from the different annotation classes, so that a variant with both an "enhancer" annotation and a "TFBS" annotation contributes both the "enhancer effect" and the "TFBS effect" to variance explained. Put another way,

we effectively use a variant with multiple annotations multiple times in the model. Other approaches to allowing non-hierarchical partitioning of variants are certainly possible; this is a simple approach that fits easily into the previously developed LMM framework.

As above, we then model the variance of the phenotype as:

$$\mathrm{V}(\boldsymbol{Y}) = \mathrm{var}(\boldsymbol{Y}) = E(X \cdot \mathrm{var}(\boldsymbol{u}) \cdot X^T) + I_n \sigma_e^2. \tag{4.7}$$

If we let $A = E(X \cdot \mathrm{var}(\boldsymbol{u}) \cdot X^T)$, then we can decompose this matrix, as:

$$\begin{aligned}
A_{ij} &= \sum_{k=1}^{m} w_k X_{ik} X_{jk} \left(\frac{\mathbb{1}_1 \sigma_{g1}^2}{\sum_{k \in M_1} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} + \frac{\mathbb{1}_2 \sigma_{g2}^2}{\sum_{k \in M_2} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} + \cdots \right. \\
&\qquad \left. + \frac{\mathbb{1}_a \sigma_{ga}^2}{\sum_{k \in M_a} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} \right) \\
&= \frac{\sum_{k=1}^{m} w_k X_{ik} X_{jk} \mathbb{1}_1 \sigma_{g1}^2}{\sum_{k \in M_1} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} + \frac{\sum_{k=1}^{m} w_k X_{ik} X_{jk} \mathbb{1}_2 \sigma_{g2}^2}{\sum_{k \in M_2} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} + \cdots + \frac{\sum_{k=1}^{m} w_k X_{ik} X_{jk} \mathbb{1}_a \sigma_{ga}^2}{\sum_{k \in M_a} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} \\
&= \frac{1}{\sum_{k \in M_1}^{m} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} \sum_{k \in M_1}^{m} w_k X_{ik} X_{jk} \sigma_{g1}^2 + \cdots \\
&\qquad + \frac{1}{\sum_{k \in M_a}^{m} w_k \cdot 2\hat{p}_k(1-\hat{p}_k)} \sum_{k \in M_a}^{m} w_k X_{ik} X_{jk} \sigma_{ga}^2.
\end{aligned}$$

This then gives us the same variance function as for the hierarchical model above (Equation 4.3), but we can allow a variant to contribute to multiple variance components. The GRMs are computed as above, but we use all variants with the appropriate annotation even if they are also used for the computation of other GRMs. The particular assumptions of this model enable us to compute the necessary GRMs in this convenient way.

Just as above, REML estimates are obtained for the variance components and variance explained and enrichment scores can be computed for the various categories in the same way. Larger standard errors are expected for these estimates due to the overlap in variants between classes, but such an "unbiased" partitioning of variants into annotation classes may prove useful in interpreting relative contributions from different variant classes.

4.3.2 Enrichment scores

Following the approach proposed by Gusev et al. (2014), we define an enrichment score $\%VE_{gc}$ for each category c that expresses the proportion of the total phenotypic variance attributable to genetic factors, VE_g, that can be attributable to variants in category c. The enrichment score is defined in the obvious way:

$$\%VE_{gc} = \frac{\sigma_{gc}^2}{\sum_{j=1}^{a} \sigma_{gj}^2}. \tag{4.8}$$

It is trivial to show that $\%VE_{gc}$ has the same form if we define it as a function of $\boldsymbol{VE}_g = (VE_{g1}, \ldots, VE_{ga})$ instead of $\boldsymbol{\sigma}_g$:

$$\%VE_{gc} = \frac{VE_{gc}}{\sum_{j=1}^{a} VE_{gj}}.$$

If we assume the allele-frequency dependent effect size model (in which each variant explains the same proportion of the overall heritability) then the expected value for $\%VE_{gc}$ is simply the proportion of variants in category c relative to the total number of variants m, that is $|M_c|/m$ (Gusev et al., 2014). To assess the significance of an enrichment score we can define a Z-score:

$$Z_c = \frac{\%VE_{gc} - 1/m|M_c|}{\text{se}(\%VE_{gc})}, \tag{4.9}$$

where $\text{se}(\%VE_{gc})$ denotes the standard error of $\%VE_{gc}$. We can use the delta method to obtain these standard errors, as described below.

4.3.2.1 Delta method for enrichment score standard errors

To conduct inference on the enrichment scores obtained, we need to compute standard errors for the estimated enrichment scores. Gusev et al. (2014) mention using the delta method (Dorfman, 1938; Ver Hoef, 2012) to compute standard errors for enrichment scores, but provide no details. Here, for completeness and ease of reading, I summarise the derivation of the standard errors for enrichment scores using the delta method. This is not a particularly novel contribution, but it does not seem to appear elsewhere in the literature.

Let us use the standard notation of ∇ for the vector differential operator, del, so that the vector derivative of a scalar field f is represented as ∇f (the gradient of f, "grad f"). Then, given a parameter vector $\boldsymbol{\beta}$ with a consistent estimator $\hat{\boldsymbol{\beta}}$ and a function h with the property that ∇h exists and is invertible, the delta method implies that:

$$\sqrt{n}(h(\hat{\boldsymbol{\beta}}) - h(\boldsymbol{\beta})) \xrightarrow{D} \text{N}(\mathbf{0}, \nabla h(\boldsymbol{\beta})^T \cdot \Sigma \cdot \nabla h(\boldsymbol{\beta})). \tag{4.10}$$

Given that we have estimators $\boldsymbol{\sigma}_g^2$ and $\boldsymbol{VE}_g$ which can be used to compute enrichment scores using the same functional form, we define our function h to be $h(\boldsymbol{x}) = h(x_1, \ldots, x_a) = (\frac{x_1}{\sum_{j=1}^{a} x_j}, \ldots, \frac{x_a}{\sum_{j=1}^{a} x_j})$ apply h either to estimates $\hat{\boldsymbol{\sigma}}_g^2$ or $\hat{\boldsymbol{VE}}_g$ to obtain estimates of the enrichment score, $\%\hat{\boldsymbol{VE}}_g$.

Consider ∇h:

$$\nabla h(\boldsymbol{x}) = \nabla h(x_1, \ldots, x_a) \tag{4.11}$$

$$= \begin{pmatrix} \frac{\partial h(\boldsymbol{x})_1}{\partial x_1} & \frac{\partial h(\boldsymbol{x})_1}{\partial x_2} & \cdots & \frac{\partial h(\boldsymbol{x})_1}{\partial x_a} \\ \frac{\partial h(\boldsymbol{x})_2}{\partial x_1} & \frac{\partial h(\boldsymbol{x})_2}{\partial x_2} & \cdots & \frac{\partial h(\boldsymbol{x})_2}{\partial x_a} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h(\boldsymbol{x})_a}{\partial x_1} & \frac{\partial h(\boldsymbol{x})_a}{\partial x_2} & \cdots & \frac{\partial h(\boldsymbol{x})_a}{\partial x_a} \end{pmatrix}. \tag{4.12}$$

We see that:

$$\frac{\partial h(\boldsymbol{x})_c}{\partial x_c} = \frac{\partial}{\partial x_c}\left(\frac{x_1}{x_1 + \cdots + x_a}\right) \tag{4.13}$$

$$= \frac{(1)(x_1 + \cdots + x_a) - (x_c)(1)}{[x_1 + \cdots + x_a]^2} \tag{4.14}$$

$$= \frac{\sum_{j=1, j\neq c}^{a} x_j}{\left(\sum_{j=1}^{a} x_j\right)^2} \tag{4.15}$$

and for $k \neq c$:

$$\frac{\partial h(\boldsymbol{x})_c}{\partial x_k} = \frac{\partial}{\partial x_k}\left(\frac{x_1}{x_1 + \cdots + x_a}\right) \tag{4.16}$$

$$= \frac{(0)(x_1 + \cdots + x_a) - (x_c)(1)}{[x_1 + \cdots + x_a]^2} \tag{4.17}$$

$$= \frac{-x_c}{\left(\sum_{j=1}^{a} x_j\right)^2}. \tag{4.18}$$

Combining these results gives us $\nabla h(\boldsymbol{x})$:

$$\nabla h(\boldsymbol{x}) = \frac{1}{\left[\sum_{j=1}^{a} x_j\right]^2} \begin{pmatrix} \sum_{j=2}^{a} x_j & -x_1 & \cdots & -x_1 \\ -x_2 & \sum_{j=1, j\neq 2}^{a} x_j & \cdots & -x_2 \\ \vdots & \vdots & \ddots & \vdots \\ -x_a & -x_a & \cdots & \sum_{j=1}^{(a-1)} x_j \end{pmatrix}. \tag{4.19}$$

Thus, if we let Σ_g denote the error-covariance matrix for the variance components σ_g or the variance explained $\boldsymbol{VE}_g$ as appropriate, then we have from the delta method (say for $h(\hat{\sigma}_g^2)$):

$$\sqrt{n}(h(\hat{\sigma}_g^2) - h(\sigma_g^2)) \xrightarrow{D} \mathrm{N}(\mathbf{0}, \nabla h(\sigma_g^2)^T \cdot \Sigma_g \cdot \nabla h(\sigma_g^2)). \tag{4.20}$$

Say we have an estimate $\hat{\Sigma}_g$ of Σ_g from REML, as well as $\hat{\sigma}_g^2$, and we estimate $\nabla h(\sigma_g^2)$ with $\nabla h(\hat{\sigma}_g^2)$, then we would estimate the covariance matrix for our enrichment scores $h(\hat{\sigma}_g^2)$ as:

$$V_{h(\hat{\sigma}_g^2)} = \frac{1}{n} \nabla h(\hat{\sigma}_g^2)^T \cdot \hat{\Sigma}_g \cdot \nabla h(\hat{\sigma}_g^2). \tag{4.21}$$

The standard error for each $\%VE_{gc}$ is then the square root of the c^{th} diagonal entry of the variance matrix $V_{h(\hat{\sigma}_g^2)}$.

With these methods available to us we can fit multiple-variance component models and assess the relative contributions of different variant classes to variance in risk for T2D.

4.4 Results when partitioning into multiple classes by allele frequency

This section presents a focused version of the variance partitioning results when partitioning by allele frequency. Two different approaches for MAF-stratified models are investigated: one that seeks to determine the contributions to VE from rare, low-frequency and

common variants (Section 4.4.2), and one that uses more MAF bins to try to get greater resolution in the VE estimates (Section 4.4.1). In both settings I will investigate the proportion of phenotypic variance explained by variance components constructed from variants in different allele-frequency classes, and look at enrichment results to find allele-frequency classes that explain more phenotypic variance than expected.

Picking up on the differences between the two effect size models on single-variance component estimates of heritability in Section 3.6.1, I discuss the effects of the effect-size assumptions on heritability estimates from multiple allele-frequency class models throughout this section. However, in an attempt to present a more streamlined thread to follow through the analysis I deliberately postpone discussion of the effects of LD-pruning. The effects of LD-pruning on results when partitioning into multiple allele-frequency classes are covered in Section 4.7.1.1, part of the section of the chapter that addresses the robustness of variance partitioning results, both partitioning by allele frequency and partitioning by functional class. Section 4.7 delves into many different factors that could possibly affect the results to assess the robustness of the partitioning results observed.

4.4.1 Partitioning into eight allele-frequency classes

To try to gain finer resolution information on the contribution to VE from variants across the allele-frequency spectrum, I bin the Integrated Panel variants into the following MAF bins: 0.1–0.5%, 0.5–1%, 1–5%, 5–10%, 10–20%, 20–30%, 30–40% and 40–50% (see Table 4.2 for the number of variants and total genotypic variance for variants in each MAF bin). That is, one bin for rare variants (MAF 0.1–0.5%), two bins for low-frequency variants (MAF 0.5–5%) and five bins for common variants (MAF greater than 5%). A GRM is then computed, K_{Mp}, $p = 1, \ldots, 8$, from the variants in each MAF-bin. As above, each GRM K_{Mp} defines the covariance structure for a polygenic random effect M_p. We then have an eight-variance component model:

$$\boldsymbol{Y} = \mathbf{1}_n \mu + \sum_{p=1}^{8} M_p + \boldsymbol{\epsilon} \tag{4.22}$$

where

$$\boldsymbol{M_p} \sim \text{MVN}_n(\mathbf{0}, K_{Mp}\sigma_p^2).$$

The variance components σ_p^2 are estimated using REML, fitting all 8 variance components jointly. The "default" modeling approach as described in Section 3.4.4 is used.

Lee et al. (2013) proposed this MAF-binning of variants as a way to obtain more accurate heritability estimates from dense genotype data. The idea is that fitting multiple variance components jointly, where each component is computed only using variants in

a relatively narrow allele-frequency range weakens the assumptions on effect-size underpinning the LMM approach. A multiple MAF-bin model fits many more parameters than a model that computes a single GRM across all variants, yielding a more flexible model.

MAF Range	Number of Variants	Total Genotypic Variance
0.1–0.5%	3,123,033	14,933
0.5–1%	1,157,909	16,559
1–5%	2,392,952	114,073
5–10%	1,101,527	148,186
10–20%	1,402,087	348,857
20–30%	1,060,766	394,312
30–40%	928,498	420,350
40–50%	867,663	428,037
Rare (0.1–0.5%)	3,123,033	14,933
Low-frequency (0.5–5%)	3,550,861	130,633
Common (5–50%)	5,360,541	1,739,745

Table 4.2: Number of variants and total sample genotypic variance in different minor allele frequency ranges

When fitting MAF-binned variance components, the estimates from the AFD and CES models are almost identical, with the only noticeable difference appearing in the 0.01%–0.5% MAF bin (Figure 4.1a). The point estimates of VE remain stable as lower-frequency MAF-bin components are omitted, raising the minimum MAF included from 0.1%, to 0.5%, to 1% and finally to 5% (Figure 4.1). The total variance explained from the MAF-binned models (the sum of the contribution from each bin) is slightly higher (0.74, s.e. 0.23 with the AFD model) than from the corresponding single-VC AFD models (recall this was approximately 0.68, s.e. 0.19) and substantially higher than the corresponding single-VC CES models (0.47, s.e. 0.11; compare with Table 4.3). This difference is expected, because fitting a single variance component across the full MAF range with the CES model gives very little weight to low-frequency and rare variants. When fitting the CES model with multiple MAF bins, however, the model estimates a separate variance parameter for each MAF range, allowing the possibility of contributions to variance explained from rare and low-frequency variants. Whereas in the single-VC models there were large differences in estimates from the AFD and CES models, such differences are not apparent for the MAF-bin models. This suggests that the MAF-binning approach removes (or at least weakens) dependency in the results on the distributional assumptions made about effect sizes.

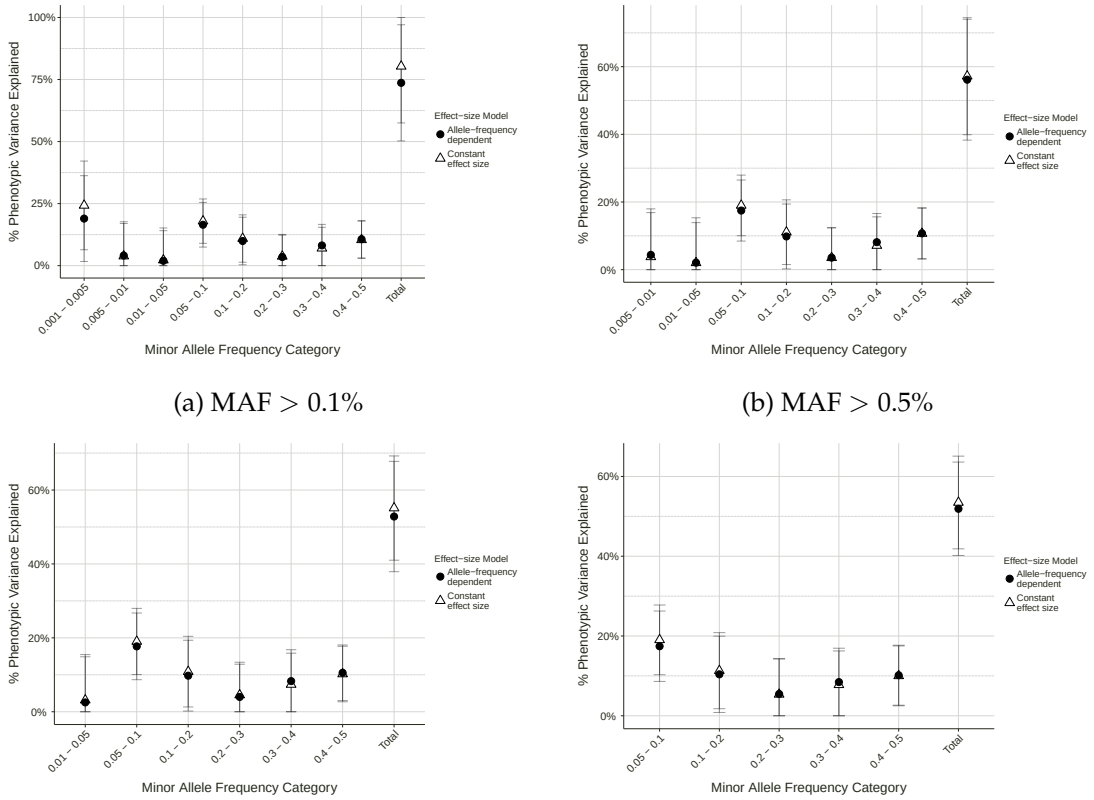
As in the single-variance component results (see Sections 3.5.1 and 3.6.1), the multiple-variance component results here, partitioning by allele frequency, show a large contribution to heritability from common variants (MAF greater than 5%). Splitting the common variants into five variance components in this eight-VC model reveals that the contributions to variance explained come more or less evenly across the common allele-frequency

spectrum, and these estimates are stable whether or not rare and low-frequency components are included in the model (Figure 4.1).

These results suggest only a small contribution to total genetic variance explained from low-frequency variants (MAF 0.5%–5%). The point estimates for the low-frequency variant components are close to zero (and not significantly different from zero), and the sum of total variance explained by additive genetic effects hardly changes when moving from a minimum MAF of 0.5% through 1% to 5% (Figure 4.1 and Table 4.3). These findings hold across the two effect-size models. In contrast to the very substantial contribution to variance explained from common variants, seen across all of the models investigated so far, the contribution from low-frequency variants appears negligible.

It is more difficult to interpret the results obtained here for the contribution from rare variants (MAF 0.1%–0.5%). The estimates are substantial under both the AFD and CES models, even allowing for the large standard errors. In both cases the rare component yields the highest point estimate from any component. However, the estimate for the contribution from rare variants is the only one for which there is any notable discrepancy in estimates under the AFD and CES models. Perhaps surprisingly, the estimate from the CES model is actually larger than that from the AFD model, but given the uncertainty in the estimates I am wary of drawing any conclusions from this. The total sum for variance explained when the rare component is included is higher for both the AFD model (0.74, s.e. 0.23) and the CES model (0.80, s.e. 0.23) than for the single-variance component AFD model (0.68, s.e. 0.19). The standard errors overlap between these estimates, so we cannot claim significant differences in these results, but the fact that the total estimate from the multiple-component CES model is so high, and higher than the estimates from the AFD models, suggests that these estimates should be treated with caution. The total sum for variance explained is higher for the CES model than the AFD model in the eight, seven, six and five-component MAF-binned models (Table 4.3), but the difference is not significant in any of the four models, and since there is no prior expectation that either the AFD or CES model should yield higher totals, seeing all four CES models give a higher total is not significantly unlikely due to chance (binomial p-value 0.06). Section 4.7.1 explores effects that inflate total estimates from MAF-binned models, but further investigation in larger cohorts will be required to clarify the contribution to variance in risk for T2D from rare variants.

The enrichment results also look similar across the AFD model (Figure 4.2) and CES model (Figure 4.3) in terms of the percentage of the genetic variance explained (GVE) by each MAF-bin (top panels of these plots). Recall that the GVE for a given component is the proportion of the total genetic variance explained by the model, where the total GVE for the model is the sum of the variance explained estimates for all of the genetic components in the model. This is expected, since results above showed that the VE estimates (upon which the enrichment scores are based) are very similar. However, close inspection reveals



(a) MAF $> 0.1\%$

(b) MAF $> 0.5\%$

(c) MAF $> 1\%$

(d) MAF $> 5\%$

Figure 4.1: Variance component estimates for allele frequency classes in the eight-variance component partitioning for multiple minimum MAF thresholds using the Integrated Panel data. Results are show for the AFD model (black circles) and CES model (open triangles) when there is no LD-pruning of variants. Variance components for the given MAF-ranges are fit jointly in a multiple-variance component model. Error bars show standard errors (truncated at zero). Table 4.2 gives the number of variants in the corresponding MAF-bin used for computation of variance components. Table 4.3 lists the estimates of the total variance explained estimates for these models more precisely.

that there are differences in interpretation of the enrichment results driven by differences in the proportion of genetic variance *expected* to be explained by each component under the two effect-size models.

Under the AFD model, each variant is expected to contribute the same amount to the overall genetic variance explained, as discussed in the previous chapter. Thus, under the AFD model, the rarer MAF-bins (MAF 0.1%–5%), which contain a sizeable proportion of variants used, are expected to explain a correspondingly substantial percentage of GVE. Exactly this result is observed in Figure 4.2, where the rare variants account for just over 25% of all variants used, and the rare variant component accounts for just over 25% of the total genetic variance explained by the model. The "fold-enrichment" for the rare component is almost exactly one (Figure 4.2, lower panel), re-inforcing the fact that using the AFD model yield exactly the expected contribution from rare variants. In contrast, there is de-

MAF Threshold	Number of genetic components in model	AFD Model	CES Model
MAF $> 0.1\%$	8	0.737 (0.234)	0.795 (0.228)
	1	0.682 (0.193)	0.466 (0.112)
MAF $> 0.5\%$	7	0.562 (0.179)	0.572 (0.173)
	1	0.536 (0.154)	0.461 (0.111)
MAF $> 1\%$	6	0.528 (0.149)	0.544 (0.141)
	1	0.502 (0.140)	0.458 (0.111)
MAF $> 5\%$	5	0.519 (0.117)	0.535 (0.116)
	1	0.490 (0.115)	0.443 (0.106)

Table 4.3: Total estimates of liability-scale variance in risk for T2D explained (standard error in brackets) by multiple allele-frequency component and single variance-component model for the Integrated Panel data. Estimates were obtained using the default model in which variant effect-sizes are allele-frequency dependent (AFD model) and the alternative model in which variant effect-sizes are constant (CES model). The column "Number of genetic components in model" gives the number of genetic variance components used in the model for each minimum MAF threshold. This is 1 for the single-variance component results and the appropriate number based on the minimim MAF threshold for the multiple allele-frequency class models, where MAF bins are 0.1–0.5%, 0.05–1%, 1–5%, 5–10%, 10–20%, 20–30%, 30–40% and 40–50%. Estimates were obtained for minimum MAF thresholds of 0.1%, 0.5%, 1%, 5% and 30%. The AFD 1VC and CES 1VC results are heritability estimates from the single-variance component allele-frequency dependent and constant effect size models, respectively (see Sections 3.5.1 and 3.6.1).

pletion in the contributions from low-frequency components and mild enrichment in four of the five common-variant components (although none of these enrichment or depletion results are significant).

Alternatively, when using the CES model a different picture emerges in terms of enrichment relative to expectation. Under the CES model, each variant is expected to contribute to the overall GVE in proportion to its genotypic variance. Rarer variants (say with MAF 0.1%–5%) have much smaller genotypic variance than more common variants. As such, even if there are many variants in the rare and low-frequency MAF-bins (which there are), the rarer MAF-bin components are expected to contribute only a very small proportion of the total genetic variance explained under the CES model (see Table 4.2). The expectation of smaller contributions from less common variants under this model drastically changes the interpretation of the enrichment results for this model. Despite the point estimates for the variance components and the enrichment results being very similar for the AFD and CES models, the fold-enrichment for rare variants in the CES model is over 32 (although not significant). This large fold-enrichment arises because the total genotypic variance, and thus the expected enrichment, for rare variants is so low. There are differences relative to the AFD model in the fold-enrichment across the MAF spectrum, but these differences are not significant.

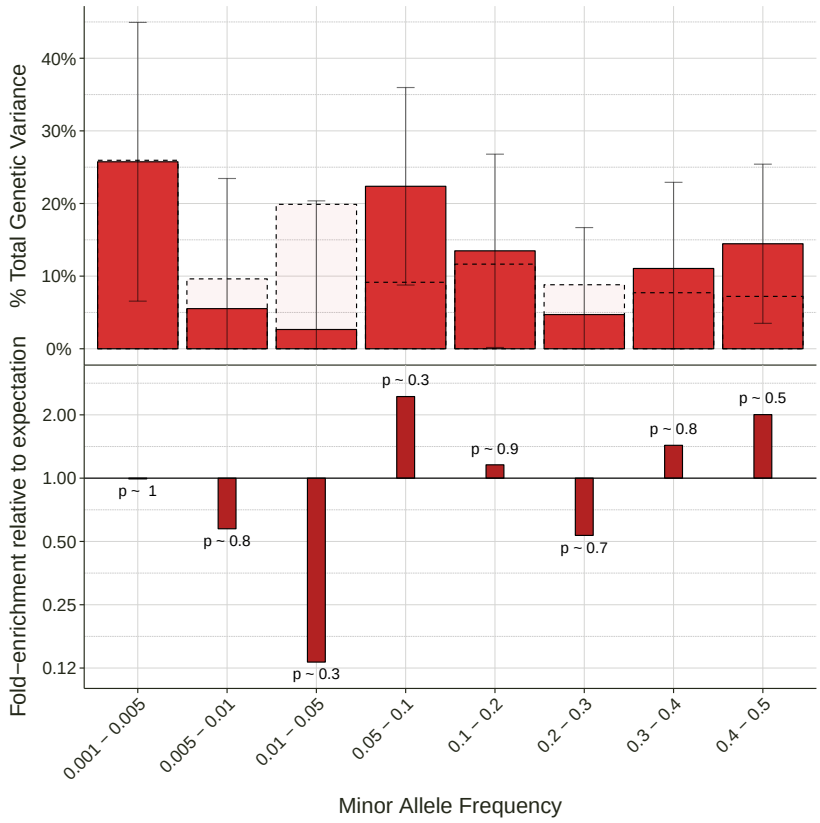


Figure 4.2: Enrichment estimates for allele frequency classes in the 8VC model using the AFD model (i.e. $s = -1$). Results are shown for different MAF thresholds, where variance components for the given MAF-ranges are fitted jointly. P-value and fold-enrichment is shown above the bar for each annotation class. Error bars show standard errors. Dotted bars with transparent fill show the expected proportion of the total genetic variance for each MAF-bin given the proportion of all variants used for the analysis fall into that bin. Table 4.2 gives the number of variants in the corresponding MAF-bin used for computation of variance components.

Ultimately, given the low precision for the enrichment results, I cannot make strong claims about differences in enrichment implied by the two effect-size models. No MAF-bins show significant enrichment under the AFD model, even the MAF 5%–10% component, which exhibits greater than two-fold enrichment. This same component shows three- to four-fold enrichment using the CES model, although not at anything more compelling than the most nominal significance level. Despite the lack of significant enrichment results here, the influence of modeling assumptions on the interpretation of heritability results is again apparent. In Section 3.6.1 results showed the effect that the changing the effect-size model has on single-variance component heritability estimates. When partitioning heritability onto multiple allele-frequency classes as done here, the heritability estimates stabilise and are almost independent of the variant effect-size distribution assumed. However, our effect-size assumptions determine the *expected* relative contributions to heritability from variants in different allele-frequency ranges. Thus, modeling assumptions influ-

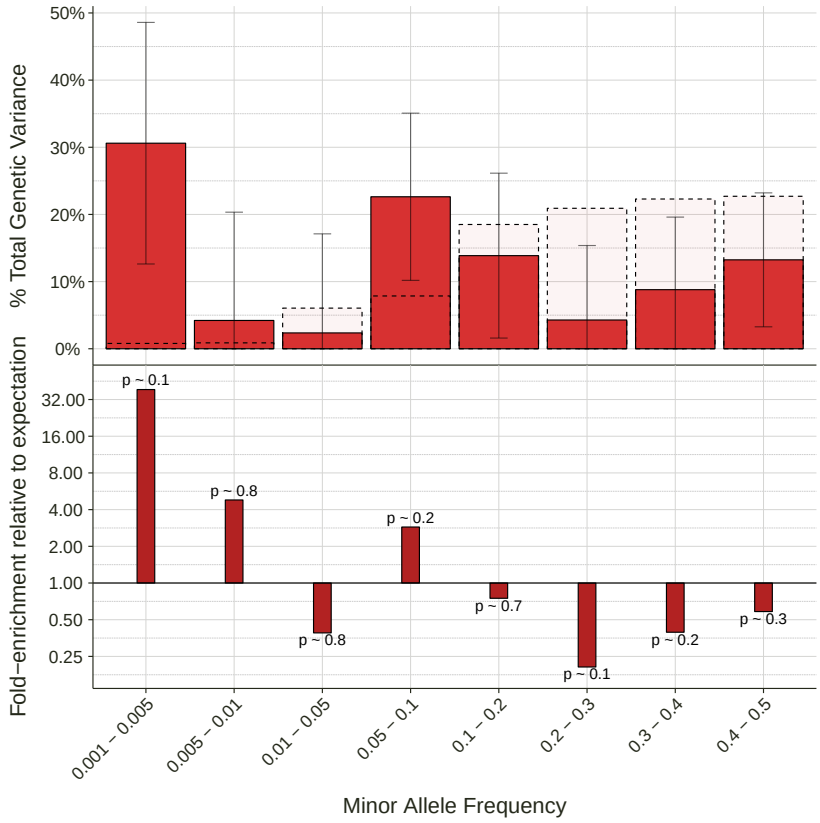


Figure 4.3: Enrichment estimates for allele frequency classes in the 8VC model using the CES model (i.e. $s = 0$). Results are shown for different MAF thresholds, where variance components for the given MAF-ranges are fitted jointly. P-value and fold-enrichment is shown above the bar for each annotation class. Error bars show standard errors (truncated at zero). Dotted bars with transparent fill show the expected proportion of the total genetic variance for each MAF-bin given the proportion of total genotypic variance accounted for by variants in that bin. Table 4.2 gives the total genotypic variance for each MAF-bin.

ence the interpretation of enrichment results for the different MAF classes, even though the variance component estimates from the two models are very similar.

4.4.2 Partitioning into three allele-frequency classes

This section looks at results for a three-component MAF-bin model that aims to estimate variance explained from rare, low-frequency and common variants. I condense the eight variance components explored in the previous section to three here for two reasons. First, it could be informative to look at the effect of including variants in a wider MAF range in components. Second, researchers in the field often distinguish between rare, low-frequency and common variants, but rarely use finer resolution bins, so this three-VC model reflects the way that researchers most often discuss questions of genetic architecture.

Here rare variants are defined to be those with MAF $< 0.5\%$ (although a minimum MAF of 0.1% is applied), low-frequency variants to be those with MAF 0.5–5% and com-

mon variants to be those with MAF greater than 5%. Applying the same variant exclusions as previously retains 3,123,033 rare variants, 3,550,861 low-frequency variants and 5,360,541 common variants (Table 4.2). One can assume an effect size distribution using the AFD model (Equation 3.13) or the CES model (Equation 3.16) and compute a GRM from each set of the rare, low-frequency and common variants, denoted by K_{rare}, K_{low} and K_{com}. Each of these GRMs defines the covariance structure for a polygenic random effect, denoted for an individual i by R_i, L_i and C_i respectively. We thus have a three-variance component model (with bold symbols representing vectors):

$$\boldsymbol{Y} = \mathbf{1}_n\mu + \boldsymbol{R} + \boldsymbol{L} + \boldsymbol{C} + \boldsymbol{\epsilon} \tag{4.23}$$

where

$$\boldsymbol{R} \sim \text{MVN}_n(\mathbf{0}, K_{\text{rare}}\sigma^2_{\text{rare}}),\ \boldsymbol{L} \sim \text{MVN}_n(\mathbf{0}, K_{\text{low}}\sigma^2_{\text{low}}),\ \boldsymbol{C} \sim \text{MVN}_n(\mathbf{0}, K_{\text{com}}\sigma^2_{\text{com}})$$

The variance components σ^2_{rare}, σ^2_{low} and σ^2_{com} are estimated using the standard REML approach as described previously.

The variance explained results, with no LD-pruning, from this three-VC model (Figure 4.4) are broadly very similar to those obtained from the eight-VC model (Figure 4.1). For both the AFD and CES models, the largest contribution to variance explained comes from common variants, the contribution from low-frequency variants is negligible and the point estimates and their standard errors for the rare variant component is almost identical to those obtained with the eight-VC model. There are small, and not significant, differences in the estimates from the AFD and CES models.

The enrichment results for the three-VC model with no LD-pruning (Figure 4.5, bottom row) show broadly similar results to those from the eight-VC models (Figure 4.2 and Figure 4.3). None of the enrichment estimates are significant for either the AFD model or the CES model, even though, as above, the interpretation of fold-enrichment changes substantially depending on which model is used. Unfortunately, with the uncertainty in estimates from the Integrated Panel data we cannot tease out any real differences in the results from the different effect-size models.

The estimates for the total variance explained by additive genetic effects in multiple allele-frequency component models (Table 4.3) (the "Total" components in Figure 4.1) obtained for the MAF $>$ 0.5%, MAF $>$ 1%, and MAF $>$ 5% models are plausible given the standard errors for these estimates and what is known about the heritability of T2D. As mentioned in the previous chapter, heritability estimates for T2D range from 30% to 70% in twin- and family-based studies (Poulsen et al., 1999; Almgren et al., 2011). Estimates at the higher end of that range come from studies that looked at individuals with early-onset T2D, most comparable to the "extreme sampling" design of the GoT2D project.

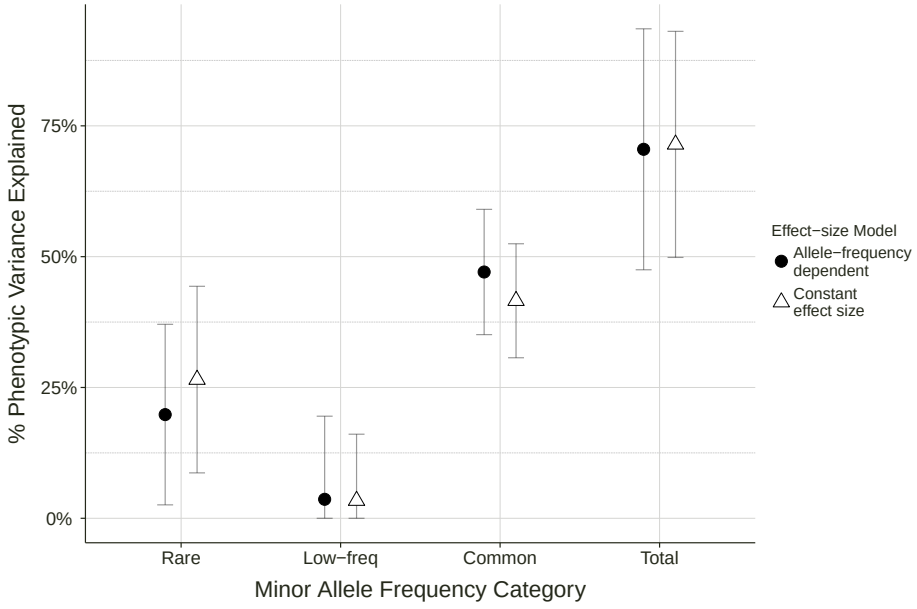
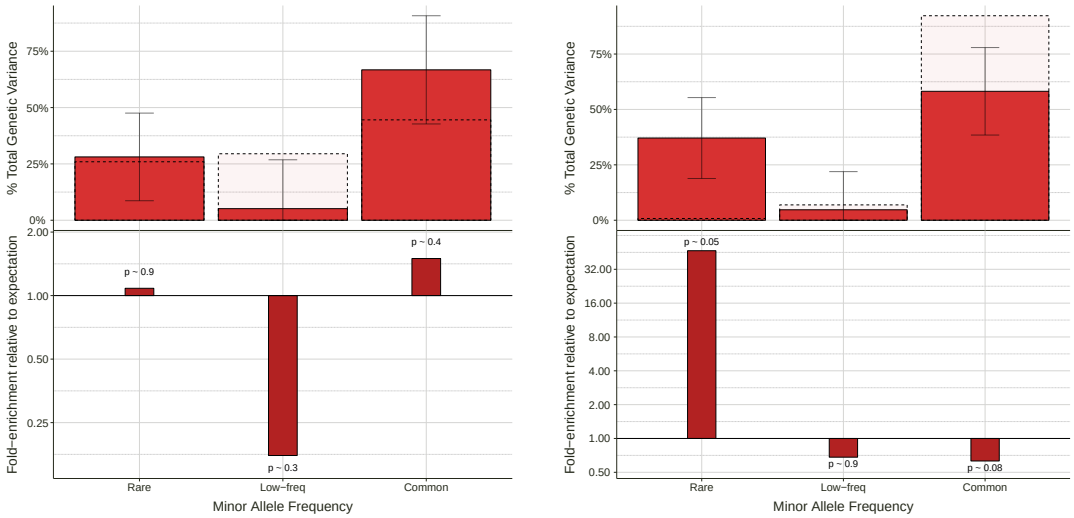


Figure 4.4: Variance component estimates (percentage of phenotypic variance explained, or heritability, on the liability scale) for three allele frequency classes using the Integrated Panel data. Results using allele-frequency dependent and constant effect size models when there is no LD-pruning of variants. Variance components for the given MAF-ranges are fit jointly in either three-variance component model. Error bars show standard errors (truncated at zero). Table 4.2 gives the number of variants in the corresponding MAF-bin used for computation of variance components.

SNP-chip heritability estimates using variance components analysis similary range from roughly 30% to 60% based on variant inclusion and modeling strategies (Gusev et al., 2013). I compare and discuss the total heritability estimates from different models in more detail in Section 4.7.4, but estimates of 70–80%, such as we obtain using the AFD and CES models when fitting three and eight allele-frequency classes in which rare variants are included are not unreasonable.

The jointly fit eight-variance component MAF-bin model can be helpful in gaining insight into the performance of LMMs for estimating heritability. The MAF-bin estimates in the eight-VC model are very similar for the AFD and CES models (Figure 4.1). However, when fitting a single variance component, as in the previous chapter, the AFD and CES model results diverge as the minimum MAF threshold is lowered. Results from the three-VC MAF-bin model fall in between these extremes, but observed differences are not significant. One possibility is that jointly fitting variance components constructed from variants in small MAF ranges weakens the influence of the modeling assumptions on whether or not effect-sizes increase with decreasing MAF (AFD model or CES model), but further investigation would be required to confirm this suggestion.

Our results for variance-partitioning by allele frequency are not conclusive with regard to the contribution to T2D heritability from rare variants. As in the single-variance com-



(a) AFD Model

(b) CES Model

Figure 4.5: Enrichment results estimates for allele frequency classes in the 3VC model. Results are shown for the AFD model ($s = -1$) and CES model ($s = 0$) when there is no LD-pruning. Variance components for the given MAF-ranges are fit jointly. P-value and fold-enrichment is shown in the lower panel of each plot. Error bars show standard errors (truncated at zero). Dotted bars with transparent fill show the expected proportion of the total genetic variance for each MAF-bin given either the proportion of all variants used that are that MAF-bin (AFD model) or the proportion of total genotypic variance accounted for by variants in that bin (CES model). Table 4.2 gives the number of variants for each MAF-bin and the total genotypic variance for each MAF-bin. Fold-enrichment results vary greatly depending on the assumptions about effect-size distributions even though the VE estimates are almost identical.

ponent results, the MAF-binning results suggest a small-to-negligible contribution from low-frequency variants and a substantial contribution from rare variants. These results are strengthened somewhat by the fact that the AFD and CES models agree closely when multiple allele-frequency classes are fitted, but the high uncertainty in the estimates prevents us from calling these observations significant (Figures 4.1 and 4.4). Large standard errors on the estimates for variance components, as with the single-variance component results, prevent us from drawing strong conclusions about the extent to which rare and low-frequency variants contribute to T2D risk. These results suggest a potentially important collective contribution from rare variants, but further analyses in larger cohorts would be required to state this with confidence. If we could obtain variance components estimates with better precision it may be possible to use these models to answer questions about the genetic architecture of T2D definitively.

4.5 Results when partitioning into functional classes

This section presents variance partitioning results when partitioning into multiple functional classes. To try to keep the narrative coherent, I provide a streamlined version of the

results for the Integrated Panel data, only showing results from models assuming allele frequency-dependent effect sizes and using hard genotype calls for all variants that pass QC and have MAF greater than 0.1%. As with previous results, I use data after applying quality control procedures described in Section 3.4.1 and use the default model parameters (unless otherwise specified) as defined in Section 3.4.4.

When partitioning into multiple functional classes we are primarily interested in the relative contributions from different classes, rather than the absolute liability-scale heritability (or variance explained) estimates. As such, I focus on enrichment results in this section, and defer consideration of the underlying heritability estimates to Section 4.7.3. In Section 4.6, I show results attempting to replicate the key findings from the Integrated Panel data using two imputed datasets on a different set of samples. Results for other modelling settings are presented in Section 4.7, where I focus specifically on the robustness of the results.

4.5.1 Partitioning into broad functional classes

First, phenotypic variance will be partitioned using eight broad functional classes previously defined: Coding, noncoding ribonucleic acid (ncRNA), UTR, Promoter, Insulator, Enhancer, TFBS, and Other. Attempting to partition phenotypic variance by annotation class is complicated by the fact that variants commonly have more than one applicable annotation (see Chapter 2). In particular, there is substantial overlap in annotations between the enhancer and TFBS classes for this study, because enhancer and TFBS regions are functionally similar, and these annotations came from separate sources, not designed to give mutually exclusive annotations.

The methods for obtaining the functional annotations are described in Section 4.2, but I recap them briefly here to avoid confusion. Recall that the coding, ncRNA and UTR classes are derived from GENCODE, and so are independent of any particular cell types. A variant can be annotated as TFBS if it falls in the union of all transcription factor binding sites across all cell types and transcription factors used. For the other classes derived from ChIP-seq data (promoters, insulators and enhancers), it is possible to use one or more cell types to define the genomic regions used as the basis for annotation. Specifically, the genome segmentation was conducted for each of the twelve cell types separately (nine ENCODE cell types, plus pancreatic islet cells, pre-adipose and mature adipose stem cells, per Table 4.1). Thus, promoter, insulator and enhancer regions "specific" to each of the twelve cell types are available. I use the union of promoter regions across all cell types for generic "promoter" annotations, the union of insulator regions across all cell types for generic "insulator" annotations and the union of enhancer regions across all cell types for generic "enhancer" annotations. I will ignore cell-type specific promoter and enhancer annotations, but will explore cell-type specific enhancer annotations in later analyses.

When fitting hierarchical partitioning models (see Section 4.3.1.1), annotations need to be prioritised so that each variant is only used for the variance component corresponding to its highest-priority annotation. The priority of annotations in terms of interest is clear for the transcript-based categories (Coding takes precedence over ncRNA over UTR), and promoters take precedence over insulators, which are in turn prioritised over enhancers. The annotation approach yields mutually exclusive annotations for promoters, insulators and enhancers (within cell type), so there is little overlap in promoter, insulator and enhancer annotations and prioritising annotations is unproblematic. However, the enhancer and TFBS annotations come from different sources (not designed to be mutually exclusive) and as functional classes are of similar interest. Further, many variants are annotated as both Enhancer and TFBS. Thus, it is not straight-forward to decide whether the enhancer or TFBS class should be prioritised in a hierarchical partitioning model (Section 4.3.1.1), which affects the interpretation of the results.

Thus, the first model combines enhancer and TFBS annotations in a 7-variance component hierarchical annotation model (see Section 4.3.1.1 for details of the hierarchical multiple variance component model). Variants are prioritised roughly in order of expected functional "importance" (see Table 4.4: classes are listed from highest priority at the top of the list to lowest priority). Mathematically, the model is the same as the model above partitioning variants into eight allele-frequency classes (Equation 4.22, but the (seven, instead of eight) GRMs, K_{Mp}, are computed from variants in a given annotation category, instead of a given MAF range. Table 4.4 provides the number of variants in each annotation category.

Model	Annotation	MAF $> 0.1\%$
Hierarchical	Coding	84,903
Model	UTR	138,792
	Promoter	600,000
	Insulator	324,829
	Enhancer	2,939,356
	TFBS	209,842
	ncRNA	33,281
	Other	7,703,454
Enhancer	Islet-Enhancer	1,677,562
3VC Model	Other-Enhancer	1,853,377
	Other	8,503,498

Table 4.4: Number of variants in annotation categories for the 8-variance component hierarchical model and the islet-enhancer 3-variance component model.

The primary result from partitioning into these seven broad functional classes is that by far the largest contribution to variance in susceptibility to T2D comes from variants in (combined) enhancer and TFBS regions. Enhancers and TFBS variants explain approximately 75% of the total genetic variance, and coding variants, the next largest contributor, explain approximately 10% (Figure 4.6). The contribution to total genetic variance from

enhancer/TFBS is almost 3 times what would be expected given the proportion of all variants that are annotated as enhancer/TFBS, which is substantial, if not significant ($P \approx 0.1$). All other annotated categories, except insulators, show positive, but not significant, enrichment, but the remaining variants ("Other" class) explain significantly less than expected (<0.05-fold enrichment, i.e. 20-fold depletion; $P \approx 0.02$). Coding variants, with 15-fold enrichment, potentially explain a large proportion of genetic variance relative to their number in the genome, but the statistical evidence for the significance of the effect in these data is weak ($P \approx 0.4$).

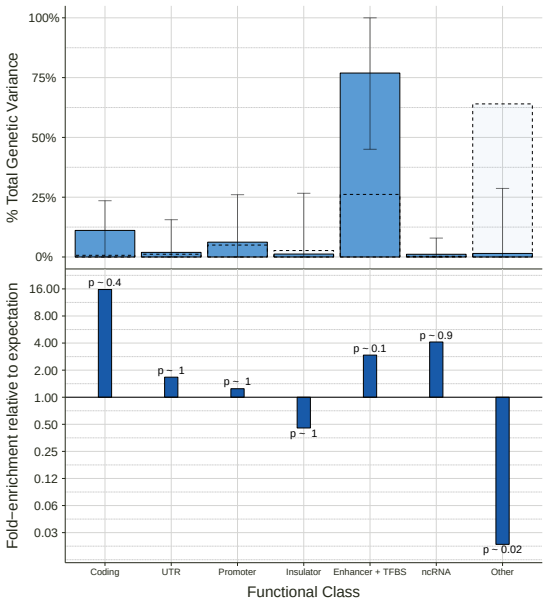


Figure 4.6: *Enrichment when partitioning into 7 broad functional classes using the Integrated Panel data.* Functional class enrichment results for the 7-variance component model assuming allele-frequency dependent effect sizes and using variants with MAF $> 0.1\%$. Variance components for the different functional classes are fit jointly. The upper panel shows the percentage of total genetic variance contributed by each functional class (total given by the sum of the estimate for all of the genetic components in the model). Dotted bars with transparent fill show the expected proportion of the total genetic variance for each functional class given the proportion of all variants used that are in that class. Error bars show ± 1 standard error (truncated at zero). The lower panel shows fold-enrichment for each class, that is the ratio of percentage of total genetic variance explained to the percentage expected, which is the percentage of all variants that are assigned to that class. P-values testing for a difference between observed and expected enrichment are shown at the end of the bar for each annotation class. Table 4.4 gives the number of variants for each functional class.

To tease out the contributions from enhancer and TFBS variants to see which of the two classes might be more important I fit eight-variance component models in which enhancer and TFBS variants define different components. This can be done in two ways:

- A non-hierarchical model (see Section 4.3.1.2 for model details) in which variants with multiple annotations contribute to each relevant functional class, and

- a hierarchical model in which functional classes do not overlap, and for which the prioritisation of classes described above is used (Table 4.4).

As discussed in Section 4.3.1.2, the non-hierarchical approach effectively double-counts the effect of variants with multiple annotations, but is unbiased with regard to the functional classes, as no prioritisation of the classes is imposed.

Results look very similar for the non-hierarchical and hierarchical models (Figure 4.7). The main observable differences are a slightly larger enrichment estimate for TFBS variants and correspondingly smaller estimate for the enhancer variants in the non-hierarchical model (as expected) and much larger standard errors in the non-hierarchical model where overlapping variant classes are allowed.

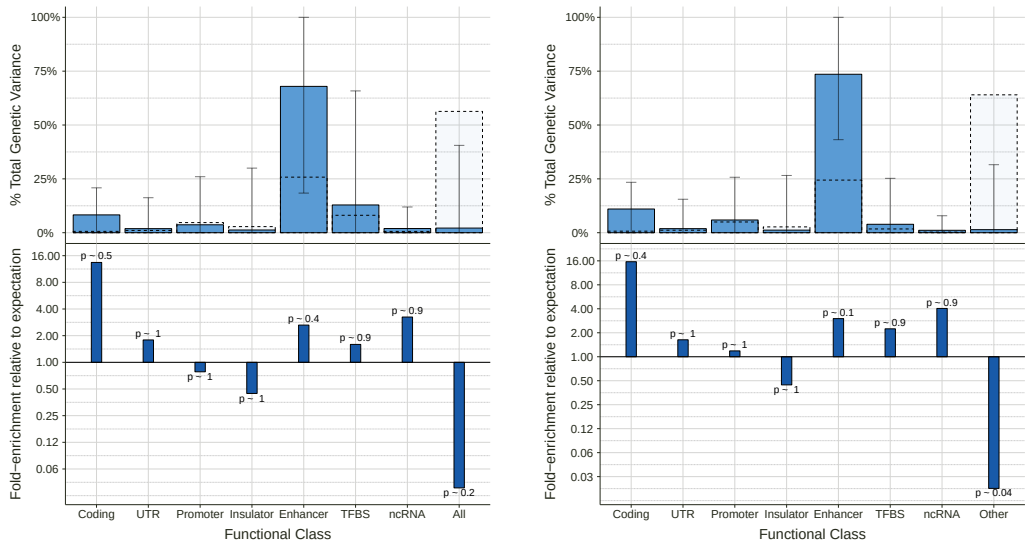
In these models, the contribution and enrichment from enhancer variants looks large (greater than 60% GVE, 3-fold enrichment), but uncertainty in estimates of variance explained is high and statistical evidence for the significance of the enrichment is modest. Again, there are high fold-enrichment for coding variants and very strong depletion for other variants ($P < 0.05$ in the hierarchical model). One concern is the high estimate for total percentage of phenotypic variance explained ($> 90\%$, s.e. 20%), even though it is not statistically different from the eight-variance component MAF-bin estimate of variance explained with MAF greater than 0.1% (74%, s.e. 23%). I discuss this concern in detail in Section 4.7.4, but note here that the results for enrichment (relative contribution from variant classes) remain valid even if there is inflation in the total phenotypic variance explained by these models.

Given that the enhancer class seems to be more important in enrichment and fold-enrichment terms than the TFBS class and that there is substantial biological interest in enhancer variants with regard to T2D (Harismendy et al., 2011; Parker et al., 2013; Pasquali et al., 2014; Gusev et al., 2014), I now focus our analysis on approaches to identifying a cell type-specific source for this large enhancer enrichment.

4.5.2 Partitioning into enhancer classes

Following up on the strong enrichment seen in enhancer variants in the preceding analysis I explore further ways to examine whether or not the effect observed from enhancers is real and whether or not it stems from a particular type of enhancer variants. Partitioning more specifically into enhancer classes is done using the following models:

1. 14VC non-hierarchical model, partitioning into 12 cell-type specific enhancer classes plus "other-functional" (any functional class but not enhancer) and "other-nonfunctional" (all remaining variants) classes,



(a) Non-Hierarchical Model

(b) Hierarchical Model

Figure 4.7: *Enrichment when partitioning into 8 broad functional classes using Integrated Panel data.* Functional class enrichment results for the non-hierarchical (a) and hierarchical (b) 8-variance component models assuming allele-frequency dependent effect sizes and using variants with MAF $> 0.1\%$. Variance components for the different functional classes are fit jointly. The upper panel shows the percentage of total genetic variance contributed by each functional class (total given by the sum of the estimate for all of the genetic components in the model). Dotted bars with transparent fill show the expected proportion of the total genetic variance for each functional class given the proportion of all variants used that are in that class. Error bars show ± 1 standard error (truncated at zero). The lower panel shows fold-enrichment for each class, that is the ratio of percentage of total genetic variance explained to the percentage expected, which is the percentage of all variants that are assigned to that class. P-values testing for a difference between observed and expected enrichment are shown at the end of the bar for each annotation class. Table 4.4 gives the number of variants for each functional class.

2. 4VC hierarchical model partitioning into islet-enhancer, other-enhancer, other-functional and other-nonfunctional classes, and

3. 3VC hierarchical model partitioning into islet-enhancer, other-enhancer and other (all remaining variants) classes.

Why these partitionings were pursued should become clear as I present and discuss the results.

4.5.2.1 Partitioning into cell type-specific enhancer classes

I would like to identify any specific cell-types that seem to be driving the observed signal from enhancer variants. This analysis begins by fitting a 14-variance component model that partitions variance into 12 cell-type enhancer classes plus other functional variants and other non-functional variants. When combined across all cell types, there is a large

number of enhancer variants. However, it may be possible to identify a small number of cell types as being disproportionately important.

Seeking a relatively unbiased approach to determining which cell-type enhancers might be more important in explaining variance in liability to T2D, I fit non-hierarchical cell-type enhancer classes. This means that variants that fall in enhancer regions identified in multiple cell types are used for all relevant variance components. Allowing variance components to overlap in terms of variants used increases the uncertainty of the estimates, but means that one does not need to decide *a priori* how to prioritise different possible annotations of variants. Finding a satisfactory prioritisation of cell-type enhancer classes here would be all but impossible, as there is a large degree of overlap between enhancer regions identified from different cell types. Using a non-hierarchical partitioning side-steps this problem, at the cost of increasing the uncertainty in the variance component estimates.

This partitioning approach identifies pancreatic islet enhancers as by far the most important class of enhancer variants (Figure 4.8). The islet-enhancer class accounts for 50% of the total genetic variance, with over 4-fold enrichment ($p \approx 0.2$). The only other class that looks of potential importance is Hmec-enhancers (mammary epithelial cell line). The Hmec-enhancers show greater than 3-fold enrichment (not significant), but explain less than 15% of the total genetic variance. In this model, there is substantial depletion of variance explained by the other functional (i.e. functional but not enhancer) variants. The previous enrichment observed for coding and ncRNA gets "washed out" when these variants are included in a single component with UTR, promoter, insulator and TFBS variants, for which either minimimal enrichment or depletion was observed. As in previous models, there is very strong depletion for other non-functional variants in this model (greater than 20-fold depletion; $P < 0.1$). Given the aetiology of T2D, pancreatic islet cells are highly relevant, so it is tantalising to see such strong enrichment for variants located in enhancer regions identified in pancreatic islet cells.

4.5.2.2 Partitioning into islet and non-islet enhancers

The analysis above highlights islet-enhancer variants as explaining the vast majority of the enhancer signal. None of the other cell-types look remotely as important, so I focus now on the islet-enhancer variants. I do this by partitioning into three- and four-variance component hierarchical class models. This partitioning, of course, gives islet-enhancers the best possible chance of explaining the variance in susceptibilty to T2D because there are now non-overlapping annotation classes and any variant that falls in an islet-enhancer region is allocated to that class. However, this seems appropriate given how much stronger the enrichment (both percentage of total genetic variance and fold-enrichment) for islet-enhancers is compared with other cell type-enhancers.

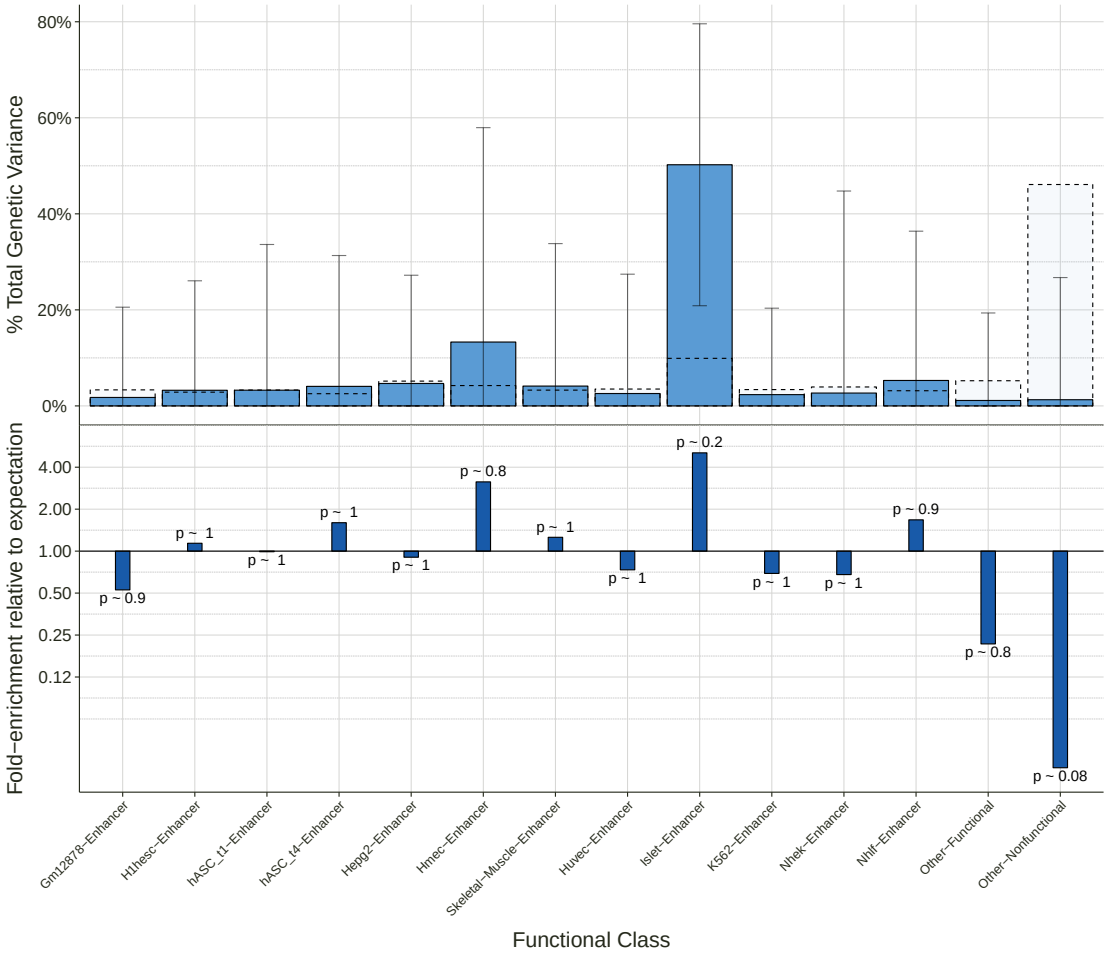
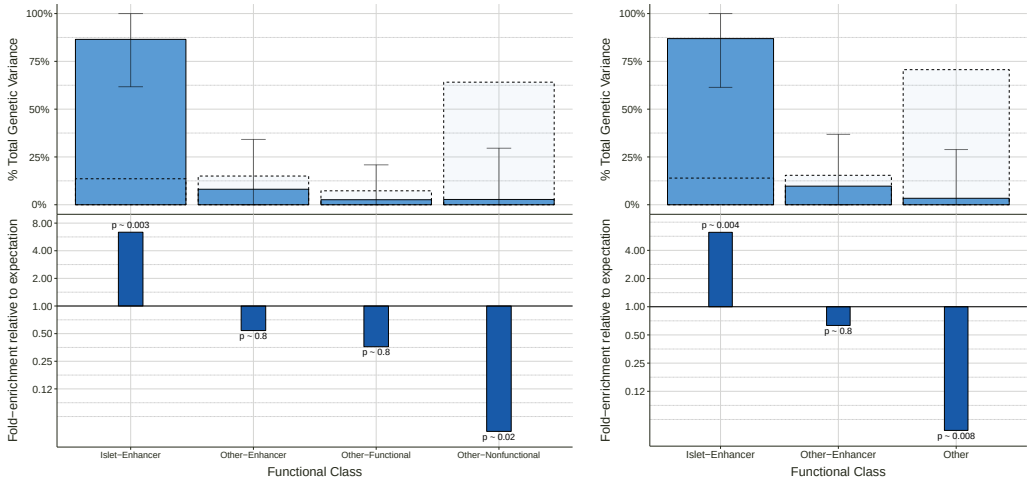


Figure 4.8: *Enrichment when partitioning into non-hierarchical cell-type enhancer classes using Integrated Panel data*. Functional class enrichment results for a model with 14 variance components: 12 non-hierarchical cell-type enhancer classes, plus a component containing all variants with other (non-enhancer) functional variants and a component using all remaining non-functional variants. The non-hierarchical enhancer classes mean that variants that are identified as falling in an enhancer region in multiple cell types are used in all relevant cell-type enhancer components. Results shown here assume allele-frequency dependent effect sizes and use variants with MAF $> 0.1\%$. Variance components for the different functional classes are fit jointly. The upper panel shows the percentage of total genetic variance contributed by each functional class (total given by the sum of the estimates for all of the genetic components in the model). Dotted bars with transparent fill show the expected proportion of the total genetic variance for each functional class given the proportion of all variants used that are in that class. Error bars show ± 1 standard error (truncated at zero). The lower panel shows fold-enrichment for each class, that is, the ratio of percentage of total genetic variance explained to the percentage expected, which is the percentage of all variants that are assigned to that class. P-values testing for a difference between observed and expected enrichment are shown at the end of the bar for each class.

Partitioning variance in susceptibility to T2D onto islet-enhancer, other-enhancer and other (functional and non-functional, separate or combined) variants shows that islet enhancers explain almost all of the total genetic variance (Figure 4.9) in such models. This is the case whether or not other-functional and other-nonfunctional are fitted variants in separate components (Figure 4.9a) or all non-enhancer variants are used in one component (Figure 4.9b). The islet-enhancer class shows significant, greater than 6-fold enrichment relative to the expected amount ($P < 0.005$), while the other class in the 3VC model and the other-nonfunctional class in the 4VC model explain significantly less than expected ($P < 0.01$, 3VC; $P < 0.05$, 4VC), with greater than 10-fold depletion. The "other-enhancer" class in these models explains less than the expected proportion of the total genetic variance (not significant).



(a) 4VC islet-enhancer model

(b) 3VC islet-enhancer model

Figure 4.9: *Enrichment when partitioning into islet- and non-islet enhancers*. Functional class enrichment results for the 4-variance component (a) and 3-variance component models (b) partitioning into islet-enhancer, non-islet enhancer and other classes. In the 4VC model there are other-functional (i.e. non-enhancer functional variants) and other-nonfunctional variants as the two non-enhancer components, whereas in the 3VC model there is a single other component for all non-enhancer variants. For these models we assume allele-frequency dependent effect sizes and use variants with MAF $> 0.1\%$. Variance components for the different functional classes are fit jointly. The upper panel shows the percentage of total genetic variance contributed by each functional class (total given by the sum of the estimate for all of the genetic components in the model). Dotted bars with transparent fill show the expected proportion of the total genetic variance for each functional class given the proportion of all variants used that are in that class. Error bars show ± 1 standard error (truncated at zero). The lower panel shows fold-enrichment for each class, that is the ratio of percentage of total genetic variance explained to the percentage expected, which is the percentage of all variants that are assigned to that class. P-values testing for a difference between observed and expected enrichment are shown at the end of the bar for each class.

As a final attempt to specify the set of variants driving the enhancer enrichment results I fitted a 4-variance component model partitioning into "islet-only enhancers" (variants identified as located in enhancers in islet cells but not other cell types), "islet-shared en-

hancers" (variants identified as located in enhancer regions in islet cells and at least one of the other cell types), "other enhancers" (remaining non-islet enhancer variants) and all other variants. This model revealed significant eight-fold enrichment in islet-shared enhancers ($P < 0.01$) and significant 20-fold depletion for the "other" category ($P < 0.01$). Islet-only enhancers showed non-significant three-fold enrichment and other enhancers showed non-significant depletion.

Taken together, the results from cell type-specific partioning of enhancer variants show that islet enhancers are very strongly enriched, and significantly so, and explain a large proportion of variance in susceptibility to T2D. Enrichment in enhancers seen in previous models seems primarily to be driven by effects from islet enhancer variants, in particular variants that are located in enhancer regions identified in islet cells and at least one other cell type. After effects from islet enhancers have been accounted for, other enhancer variants, other functional variants and other non-functional variants appear to have little capacity to explain variance in susceptibility to T2D. However, the dampening effect of including so many different types of variant in the other category (for example a relatively small number of coding variants with a very large number of unannotated variants) should be borne in mind. In such cases we fit one parameter in the model (a variance component) that tries to account for the effects of all of those variants and so signal from a small number of variants with larger effects can get "washed out" when included in the same component as a large number of variants that make little contribution to variance explained.

Coding and ncRNA variants look substantially enriched relative to expectation, but these enrichment results are not significant in these models and compared to islet enhancer variants coding and ncRNA variants explain a very small proportion of the total genetic variance. Partitioning into multiple functional classes, with potentially enriched functional classes such as coding and ncRNA variants, islet-enhancer, other-enhancer, other-functional and other-nonfunctional classes could be worthwhile, but only when larger sample sizes become available in which the precision of estimates could be substantially better than in the sample analysed here. Given the uncertainty in the partitioning estimates obtained here, I do not expect a different six-variance component partitioning to yield conclusive results on this dataset.

These results from the Integrated Panel data indicate strong enrichment from islet-enhancer variants, but there remains substantial uncertainty in the results and there are many factors that could affect enrichment estimates. To address these considerations I replicate, in the next section, the analysis in a second, larger cohort for which imputed data are available (Section 4.6), and then take a detailed look at the robustness of variance partitioning results in Section 4.7.

4.6 Results using imputed data in a larger UK cohort

The variance partitioning analyses to this point have provided further insight into the partitioning of phenotypic variance by allele-frequency and shown strong enrichment for enhancer variants, particularly those identified from pancreatic islet cells. However the uncertainty in heritability and enrichment estimates was high for most analyses. Thus, I sought to replicate the findings from the Integrated Panel data in a second cohort of individuals.

Through the GoT2D project I have access to a set of over 4,500 UK individuals who have been genotyped using a SNP array (see Section 3.2.5 for details). For these individuals further genotypes imputed using the 1000 Genomes reference panel (Imputed-1000G data) and, separately, using the GoT2D Integrated Panel as a T2D-specific reference panel (Imputed-GoT2D data) are available. Imputation was done by Kyle Gaulton for the Imputed-1000G data, and by Loukas Moutsianas for the Imputed-GoT2D data (see Section 3.2.5 for details). After removing individuals who were sequenced (and thus appear in the Integrated Panel dataset) and applying quality control filtering to variants and individuals, there are 4,525 individuals (1,587 cases and 2,938 controls) in the UK imputed dataset. Imputation errors are inevitable in the imputed datasets, but imputation errors are expected to introduce noise into variance component estimates rather than produce specific biases. I undertake the same variance partitioning analyses that were conducted on the Integrated Panel data (described above) on these imputed datasets to see if the heritability and enrichment patterns replicate.

4.6.1 Imputed data: partitioning by allele frequency

I partition the variants for both the Imputed-1000G and Imputed-GoT2D datasets into three allele-frequency classes—rare, low-frequency and common variants—and, in a finer-grained model, into eight allele-frequency classes. In both cases, multiple variance components are fitted jointly in the LMM. Again, as for the Integrated Panel data, I compare results for the default, allele-frequency dependent effect-size model and the alternative, constant effect size model. In the Integrated Panel data a large contribution to phenotypic variance explained was observed from common variants, a very small contribution from low-frequency variants and a substantial point estimate (though with high uncertainty) for rare variants (see Section 4.4.2).

Results for variance explained from the imputed data (Figure 4.10) are entirely consistent with the results from the Integrated Panel data (Figure 4.4). Again, there is a large contribution from common variants, a small contribution from low-frequency variants, a

substantial estimate for rare variants, and in each model the estimate for common variants is larger than for low-frequency or rare variants. Observed differences between the estimates from the imputed datasets and the Integrated Panel data are not significant.

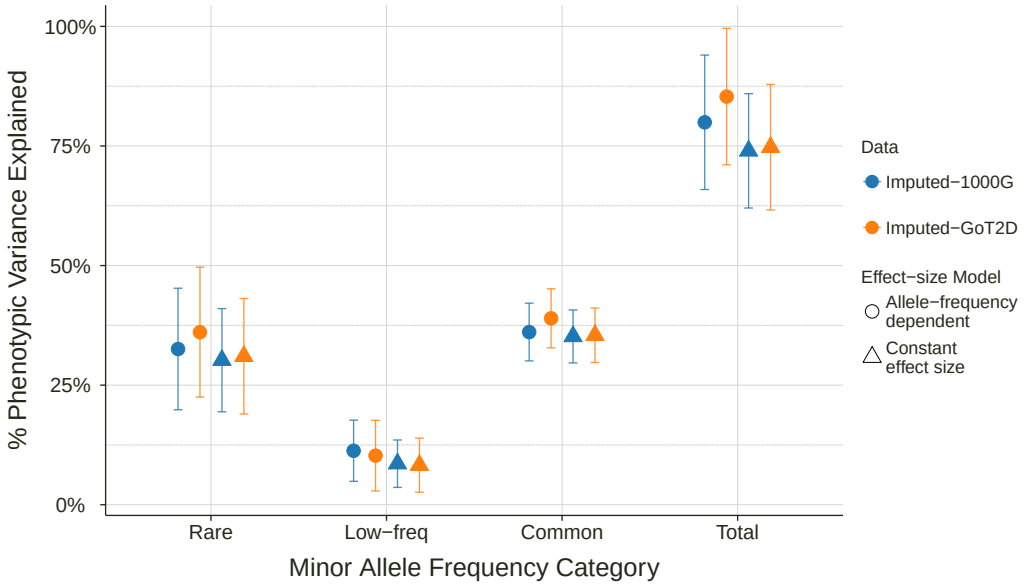


Figure 4.10: *Phenotypic variance explained using the Imputed-1000G and Imputed-GoT2D datasets when partitioning into three allele frequency classes: rare, low-frequency and common*. Percentage of phenotypic variance explained by each MAF class in the 3-variance component and 8-variance component models assuming allele-frequency dependent effect sizes (circles) or constant effect sizes (triangles). The total shows the sum of the contributions from each allele frequency class, with standard error computed using the delta method. Variance components for the different allele frequency classes are fitted jointly and error bars show ± 1 standard errors (truncated at zero and 100%).

Estimates for the total phenotypic variance explained for the imputed data from the three allele-frequency class model are 85.3% (s.e. 14%; AFD model) and 74.7% (s.e. 13%; CES model) for the Imputed-GoT2D data, and 80.0% (s.e. 14%) and 74.0% (s.e. 12%) for the Imputed-1000G data. These point estimates are higher than, but not significantly different from, the corresponding estimates from the Integrated Panel data (70.5%, s.e. 23% for the AFD model; 71.5%, s.e. 22% for the CES model).

Across the two imputed datasets and two effect-size models, the rare variants explain 30–36% of the liability-scale variance. The standard errors of the estimates are smaller than for the Integrated Panel data, but are large enough that one cannot claim to observe any significant difference between the variance component estimates or total heritability estimates between the two imputed datasets themselves or between the imputed data results and the Integrated Panel results. However, the imputed data results do provide further modest evidence of a non-zero contribution from rare variants. It is problematic to assign significance to individual variance component estimates in this type of model, but all of

the imputed data estimates for the heritability from rare variants are at least two standard errors away from zero. Thus, I would be cautious about claiming exactly what the collective contribution from rare variants is, but one can be reasonably confident that the contribution is truly non-zero.

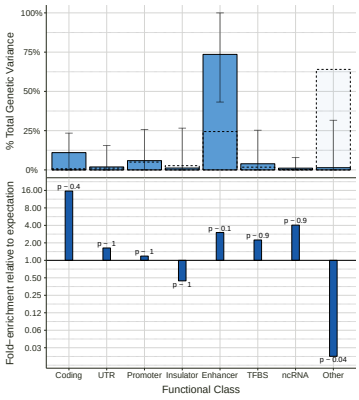
Partitioning variants into eight allele frequency classes using both imputed datasets yielded qualitatively very similar results to those obtained with the Integrated Panel data (data not shown). In both imputed datasets, the estimated contribution from rare variants in the eight-variance component model was almost identical to that obtained in the three-variance component model. Enrichment results for the two imputed datasets agree with each other very closely and tell the same story as the Integrated Panel data (data not shown). Again, interpretation of the enrichment results when partitioning by allele frequency depends greatly on the modeling assumptions made about the variant effect-size distributions.

4.6.2 Imputed data: partitioning by functional class

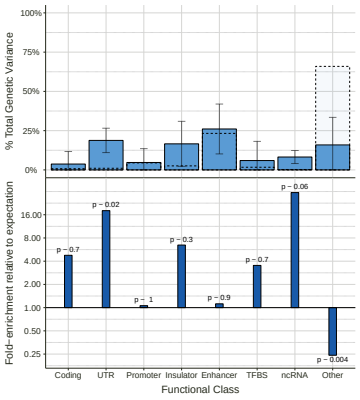
Partitioning into functional classes using the Integrated Panel data revealed strong enrichment for islet-enhancer variants and significant depletion for variants without a functional annotation. The same models are now fitted to the imputed data. Exactly the same segmentation of the genome into functional regions was used to assign variants to functional classes as was used for the integrated panel data. I show results using all variants in the Imputed-1000G dataset (after quality control and applying a minimum MAF threshold of 0.1%). No significant differences between the Imputed-1000G and Imputed-GoT2D results were observed for any of the models discussed below (data not shown). I also obtained results using the set of variants that appear both in the integrated panel dataset and the Imputed-1000G dataset and found that these gave qualitatively very similar results with only small, non-significant changes in variance component estimates (data not shown).

The Imputed-1000G results for the 8-variance component hierarchical model yield significant depletion for the "other" class, as variants without a functional annotation exhibit 4-fold depletion ($P < 0.005$), which is less extreme in terms of effect size but more significant than the depletion in this category observed in the Integrated Panel data (Figure 4.11a). There is no significant enrichment after accounting for multiple testing for any of the other classes. (Figure 4.11b). There are no significant differences between the enrichment or fold-enrichment estimates for any functional classes between the imputed and Integrated Panel data.

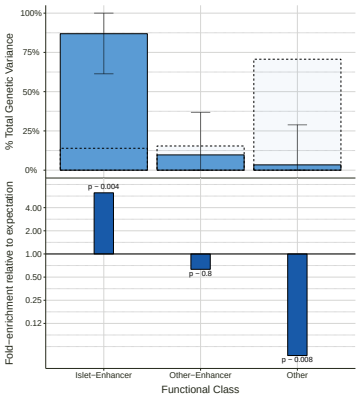
Fitting the 14VC non-hierarchical cell type-enhancer model (see Section 4.5.2.1 yields no significant enrichments for either the Imputed-1000G or Imputed-GoT2D data, and no significant differences in enrichment compared to the Integrated Panel results (data not shown).



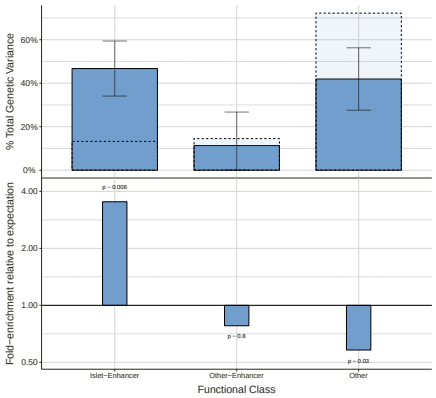
(a) Integrated Panel, All Samples



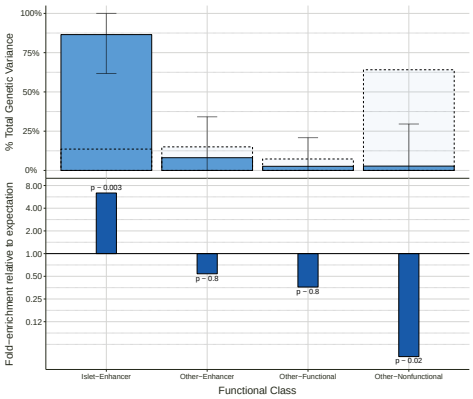
(b) Imputed-1000G Data, UK samples



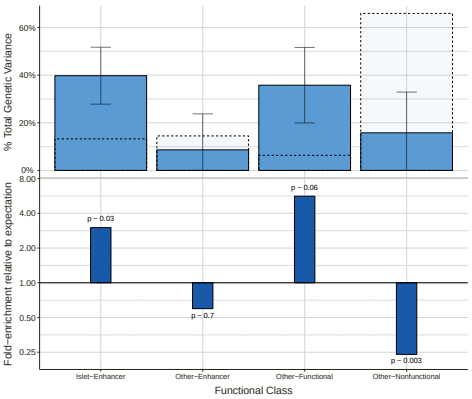
(c) Integrated Panel, All Samples



(d) Imputed-1000G Data, UK samples



(e) Integrated Panel, All Samples



(f) Imputed-1000G Data, UK samples

Figure 4.11: *Enrichment results using imputed data when partitioning into functional classes*. Results are shown for: the 8-variance component hierarchical model using Integrated Panel data (a) and Imputed-1000G data (b), the 3-variance component islet-enhancer hierarchical model using Integrated Panel data (c) and Imputed-1000G data (d), and the 4-variance component islet-enhancer hierarchical model using Integrated Panel data (e) and Imputed-1000G data (f). For each plot, the percentage of genetic variance explained by each functional class in the model assuming allele-frequency dependent effect sizes is shown in the top panel, and the fold-enrichment is shown in the bottom panel. Variance components for the different functional classes are fitted jointly. Results shown here are when using all variants passing QC from each dataset (both integrated panel and imputed data) using variants with MAF $> 0.1\%$. Enrichment results when using only variants that are shared between the imputed and integrated panel datasets are very similar (data not shown). Error bars show ± 1 standard errors (truncated at zero).

Fitting a three-variance component hierarchical islet-enhancer model (see Section 4.5.2.2 for details) leads to the same conclusions as for the Integrated Panel results. There is significant, 3-fold enrichment for islet-enhancer variants ($P < 0.01$), nominally significant depletion for the "other" class ($P < 0.05$) and no significant enrichment for the "other-enhancer" class (Figure 4.11d).

Fitting the four-variance component hierarchical model with a component for islet-enhancer variants, one for other enhancer variants, one for other functional variants and a final component for all other, non-functional variants to the imputed data produces very similar results (Figure 4.11d). We observe nominally significant, 3-fold enrichment for islet-enhancer variants ($P < 0.05$) and no significant enrichment for other-enhancer variants. Splitting the "other" class from the 3-VC model into an "other-functional" and "other-nonfunctional" class gives significant, 4-fold depletion for the "other-nonfunctional" variants ($P < 0.005$) and a substantial enrichment estimate that is not significant for the "other-functional" variants. Further investigation in larger cohorts would be required to clarify if there is any true signal for the other functional variants.

Smaller estimates are observed for total variance explained from these models for the imputed datasets than for the Integrated Panel data (data not shown), but the differences are not significant.

Overall, the imputed datasets provide a useful set of results against which to compare the Integrated Panel results. The key findings from the Integrated Panel data replicate in the imputed data.

4.7 Robustness and exploration of factors affecting variance partitioning results

The previous three sections presented streamlined results in which I laid out the narrative for the variance partitioning analyses. Recall that I undertook analyses partitioning variance by allele frequency (Section 4.4) and functional class (Section 4.5) using the GoT2D Integrated Panel data, and replicated the key findings in a larger UK cohort using imputed data (Section 4.6). In the preceding sections we did not, generally, discuss changes to parameter settings that could potentially affect the partitioning results. In this section I examine in detail the robustness of the variance partitioning results.

For the analyses partitioning variance by allele frequency I discussed the effects of changing model assumptions in Section 4.4. Assuming either the allele-frequency dependent effect-size model or the constant effect-size model had no substantial impact on heritability estimates, but strongly influenced the interpretation of enrichment results. I did not

discuss, in Section 4.4, the effect that LD-pruning could have on the allele-frequency partitioning results. I explore the effects of LD-pruning on partitioning heritability by allele-frequency class in Section 4.7.1.

In Section 4.5, above, I presented one strand of the enrichment analysis when partitioning variance by functional class. Only results obtained using hard genotype calls for all variants passing QC with MAF greater than 0.1% were discussed. There was evidence of enrichment in enhancer variants, particularly those identified in pancreatic islet cells, and depletion for variants without a functional annotation. However, I noted in passing that raw estimates of variance explained were higher than might be expected, and that results from these models can be sensitive to assumptions and model parameters.

I discuss many aspects of the robustness of the enrichment results in Section 4.7.2 and the robustness of the liability-scale variance explained, or heritability, estimates in Section 4.7.3. I undertook many parallel analyses on partitioning into multiple functional classes to assess the effects on the results of varying modeling parameters. These included varying assumptions about effect-size distributions, setting different minimum allele frequency thresholds, excluding regions around known T2D GWAS loci, LD-pruning variants, and using genotype dosages instead of hard genotype calls. I also undertook "shifted-enhancer" and "pseudo-enhancer" analyses to probe further the enhancer signal, which I will explain and discuss in Sections 4.7.2.2 and 4.7.3.2, respectively.

To get another perspective on how likely or unlikely it is to obtain certain partitioning results due to chance alone, I conducted analyses in which I permuted the phenoype (T2D case-control status) and then fitted variance partitioning models. I discuss permutation results in Sections 4.7.1.2 (allele-frequency partitioning), 4.7.2.3 and 4.7.3.3 (functional-class partitioning).

When partitioning by functional class, due to space considerations, I will generally only present robustness results for the three-variance component hierarchical islet-enhancer model (see Section 4.5.2.2). We carried out the same robustness analyses for the eight-variance component broad functional classes model (cf. Section 4.5.1) and other cell-type specific enhancer partitionings (data not shown), and found that the general conclusions drawn about robustness hold across different functional-class partitioning models. Due to the consistency of conclusions about robustness across partitioning models I do not repeat all of the robustness analyses for the imputed data in Section 4.7.2.4, focusing instead solely on the robustness of the partitioning results for the imputed datasets to changing the minimum MAF threshold, changing the effect-size model and using only variants shared between the Imputed-1000G and Integrated Panel datasets.

Across many of the partitioning analyses there were high total heritability estimates. I discuss this issue in Section 4.7.4. The Integrated Panel cohort consists of individuals from five distinct European populations. This structure could potentially influence partitioning

results, so in Section 4.7.5 I discuss results obtained from fitting the partitioning results on sub-populations within the Integrated Panel cohort.

4.7.1 Robustness of allele-frequency partitioning results

In Section 4.4, which presented the main allele-frequency partitioning results, I discussed the effect of assumptions about the variant effect-size distribution on the interpretation of results. In this section, I check the robustness of multiple allele-frequency class heritability estimates to LD pruning (Section 4.7.1.1) and assess permutation results for the allele-frequency partitioning models.

4.7.1.1 Effects of LD-pruning variants

Section 3.6.3.1 showed that LD-pruning variants had a noticeable effect on single-variance component heritability estimates. Given the considerations about LD-effects on estimates of heritability discussed in Section 3.6.3, I repeated the 3VC MAF-bin analysis described above, using variants after LD-pruning with the `PLINK` software, applying maximum R^2 thresholds of 0.8, 0.5 and 0.3 (see Section 3.6.3.1). LD-pruning drastically reduces the number of variants used for analysis (Table 4.5). Patterns in variance partitioning results when applying LD-pruning appear slightly different from those obtained when all variants are used, but differences are not statistically significant (Figure 4.12).

MAF Range	No LD pruning	$R^2 < 0.8$	$R^2 < 0.5$	$R^2 < 0.3$
0.1–0.5%	3,123,033	2,110,658	1,576,114	1,221,178
0.5–1%	1,157,909	595,468	405,021	282,961
1–5%	2,392,952	867,434	559,887	367,140
5–10%	1,101,527	236,602	137,316	81,837
10–20%	1,402,087	239,020	120,741	66,392
20–30%	1,060,766	170,743	78,524	35,901
30–40%	928,498	145,963	61,194	23,666
40–50%	867,663	144,469	78,690	56,925
Rare (0.1–0.5%)	3,123,033	2,110,658	1,576,114	1,221,178
Low-frequency (0.5–5%)	3,550,861	1,462,902	964,908	650,101
Common (5–50%)	5,360,541	936,797	476,465	264,721

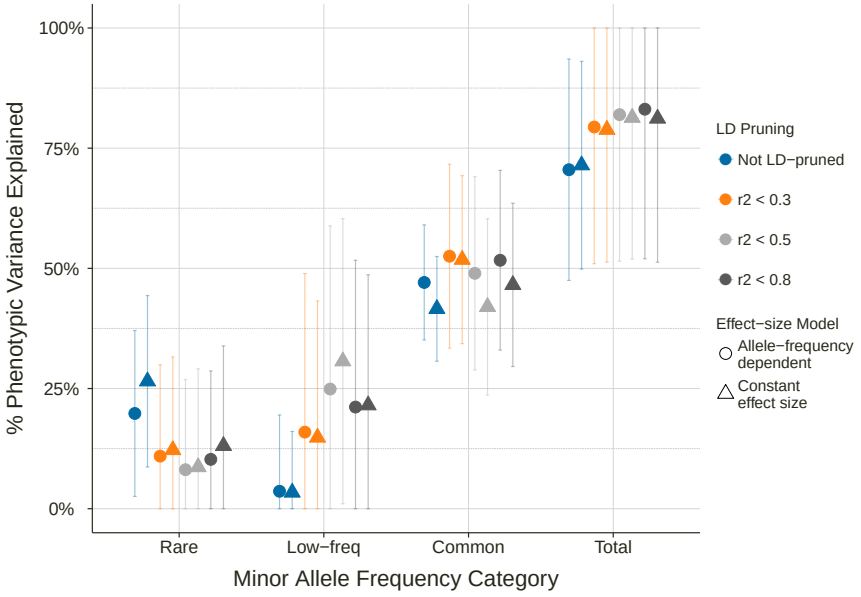
Table 4.5: Number of variants and total sample genotypic variance in different minor allele frequency ranges when LD pruning has been applied to variants at maximum R^2 thresholds of 0.8, 0.5 and 0.3. Numbers of variants when no LD pruning is done are shown for comparative purposes.

The estimates of total variance explained are higher when using LD-pruned variants than non-pruned variants for both the three-variance component model (Figure 4.12a) and the eight-variance component model (Figure 4.12b). Given the size of the standard errors for these estimates, one cannot confidently claim significant differences in the total estimates, but the pattern is consistent across both the three- and eight-variance component

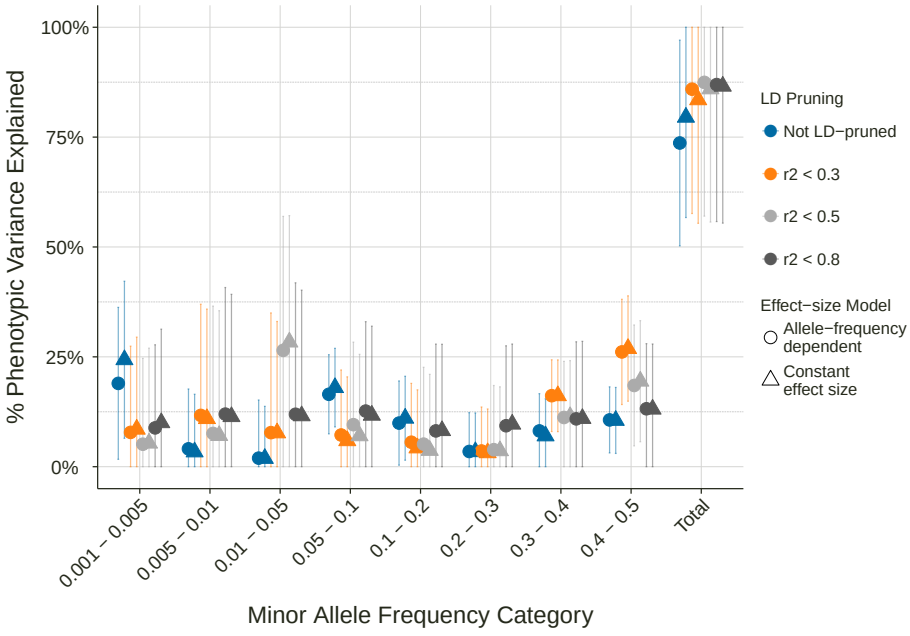
models here and are reasonably consistent with the single-variance component results (Section 3.6.3.1). The total liability-scale heritability estimates are high using LD-pruned variants. For the three-variance component model they range from 79% (s.e. 28; $R^2 < 0.3$, CES model) to 83% (s.e. 31%; $R^2 < 0.8$, AFD model). For the eight-variance component model total variance explained estimates range from 83% (s.e. 28%; $R^2 < 0.3$, CES model) to 87% (s.e. 30%; $R^2 < 0.5$, AFD model). These estimates are a little larger than the corresponding estimates obtained without LD-pruning, but are not significantly different. Again, the total estimates from the eight-variance component models are slightly higher than the total estimates from the three-variance component models, suggesting that fitting more variance components slightly inflates the total variance explained by the model. The increases observed in the total variance explained from the models with LD-pruning could be driven by higher correlation between GRMs induced by LD-pruning. This idea is explored in further detail in Section 4.7.4.4.

The total variance explained estimates are relatively consistent across LD-pruning thresholds, but there is much more variability when one looks at individual variance component estimates (Figure 4.12). Across the individual components in the two models there does not appear to be a consistent pattern as to which LD-pruning settings lead to higher or lower heritability estimates. Higher estimates are observed for the rare variant component when there is no LD-pruning than when there is LD-pruning at any threshold, higher estimates for the low-frequency component for all of the LD-pruning approaches than when there is no LD-pruning, but differences are not statistically significant. Enrichment results across varying LD-pruning do not show significant enrichment for any frequency classes, or any significant differences from results obtained without LD-pruning. Thus, bearing in mind the large standard errors for the estimates, the general picture is one of consistency, especially for estimates from common variants.

LD-pruning produced larger standard errors on obtained variance-component estimates, so as previously discussed, it appears that LD-pruning variants may introduce unwanted effects into the variance partitioning results. LD-pruning is expected to have limited effects on heritability estimates because genomic segments shared from a recent common ancestor will usually extend well beyond the usual range of LD, and so LD-pruning should have little consequence for estimating the relatedness between individuals that define the GRMs in the LMMs (Speed et al., 2012). However, if one looks closely at the set of variants retained for analysis after LD-pruning (Table 4.5), one sees that at an R^2 threshold of 0.3, over 60% of rare variants, 80% of low-frequency variants and over 95% of common variants are filtered out. The filtering is less extreme for higher R^2 thresholds, but still substantial: even at an R^2 of 0.8 over 80% of common variants are filtered out. Thus, LD-pruning dramatically changes the allele-frequency distribution of the variants used for the analyses. This may affect the partitioning results obtained, and if applied naively could



(a) Three allele-frequency classes



(b) Eight allele-frequency classes

Figure 4.12: Variance component estimates for multiple allele frequency classes with LD-pruning of variants. Results using AFD model (circles) and CES model (triangles) when there is no LD-pruning (blue points) and LD-pruning using PLINK with a maximum R^2 threshold of 0.8 (dark grey points), 0.5 (light grey points) and 0.3 (orange points). Here, variance components for the given MAF-ranges are fitted jointly in a three-variance component model (a) or an eight-variance component model (b). Error bars show standard errors (truncated at zero). The total shows the sum of the contributions from each allele-frequency class, with standard error computed using the delta method. Table 4.2 gives the number of variants in the corresponding MAF-bin used for computation of variance components.

affect inferences about the relative importance of rare and low-frequency variants to explaining variance in susceptibilty to T2D.

With very large numbers of rare and low-frequency variants in the dataset, it seems that the default `PLINK` algorithm for LD-pruning preferentially filters out common variants. Different settings for the LD-pruning could possibly yield better results, but I conclude from this analysis, as from the analysis of single-variance component estimates, that LD-pruning (at least using default settings) is not a good approach to apply to these sorts of variance partitioning analyses.

4.7.1.2 Permutation results for partitioning by allele frequency

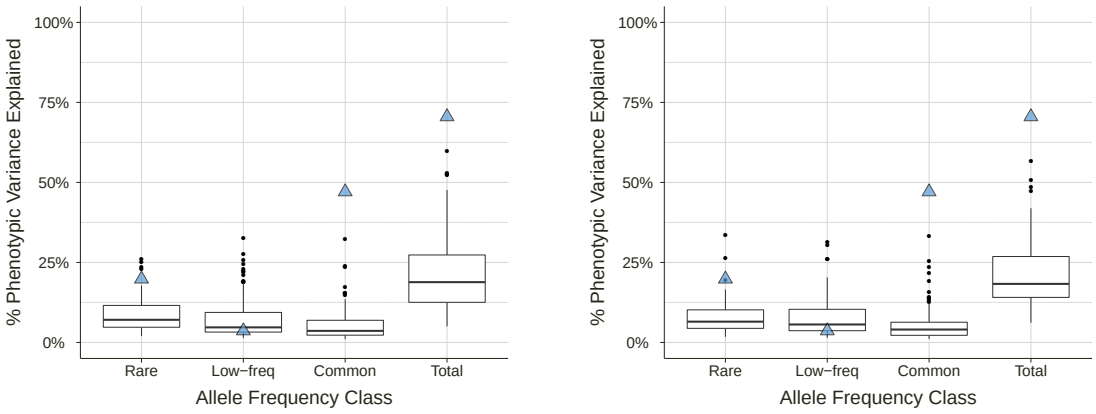
I conduct permutation studies to test empirically to see if the estimates of variance explained and enrichment from variants in different allele frequency classes are larger than expected by chance. To study the model partitioning into three variance components by allele frequency, I use the same GRMs as for the "real" results, but permute case-control status. I permute in two ways:

1. permuting case-control status across all samples, and
2. permuting case-control status between individuals within each population (country of origin).

For each permutation approach, REML estimates of the variance components and corresponding enrichment results are obtained in the usual way for the rare, low-frequency and common variants for $n = 100$ permutations of the phenotype (Figure 4.13).

When case-control status is permuted there should be no relationship between genetic effects and phenotype, so one can obtain a null distribution for the variance component estimates. I show results only for the allele-frequency dependent effect size model. Section 4.4 showed that the variance component estimates for the three-variance component model are almost identical for the AFD and CES models, so it seems reasonable to expect that using either of these models for the permutation study would yield qualitatively identical results. These results show that there is little difference in permutation results whether permuting phenotype across all samples or permuting phenotype only among individuals within the same country of origin (Figure 4.13).

These permutation results tell broadly the same story as the analyses of the real phenotypes. That is, it looks like there is a real effect from common variants, negligible effect for low-frequency variants and the contribution from rare variants is substantial, but could be obtained by chance. The contribution from common variants in real data is much larger than the largest estimate using permuted phenotypes, and so is much larger than you would expect to obtain if there was no true contribution from common variants. Thus, one



(a) Permuting across all samples

(b) Permuting within country

Figure 4.13: *Results permuting phenotype for a model with three allele frequency classes using the Integrated Panel data.* Results for the percentage of phenotypic variance explained by each allele frequency class when permuting case-control status across all individuals in the sample (a) and permuting case-control status across individuals separately within sub-populations (b). Boxplots show results from $n = 100$ permutations. Blue triangles show the results obtained when the true case-control status is used.

would conclude that there is a true non-zero heritability estimate for common variants. The contribution observed for low-frequency variants is very close to the median from the permuted phenotype results, consistent with a negligible contribution.

The heritability estimate for the rare component is higher than all but 2–4 of the estimates for that component when using permuted phenotypes. Thus, it is unlikely to observe an estimate greater than that observed if there is no contribution from rare variants, but it is possible to obtain such a large estimate, even from only 100 permutations. This accords with the conclusions drawn from the standard errors for the rare component estimates: the contribution from rare variants to variance in susceptibilty to T2D could be substantial, but one cannot claim this with high confidence.

The total variance explained when using the real phenotypes is substantially larger than any of the totals when using permuted phenotypes. One concludes, then, that it would be very unlikely to have obtained such a large total for variance explained solely from chance effects. I note, however, that the total variance explained, even with permuted phenoytypes, can be large. The upper quartile total heritability estimate from the permutations is over 25% and we see several totals of over 50% for the liability-scale heritability. These results establish that there are substantial upward biases in the heritability estimates. The REML fitting approach used here gives non-zero variance component estimates, even in "null" cases such as with the permuted phenotypes here. The median permutation liability-scale heritability estimate across the rare, low-frequency and common components can be as high as 10%. When taking the sum of several variance com-

ponents, the total variance explained by the model can appear much higher than might be expected. These permutation results go some way to explain the higher totals seen in the eight-variance component models compared with the three-variance component models. When more upwardly-biased variance estimates are obtained, then the total is correspondingly inflated. Our enrichment estimates, however, are relative measures of contribution to heritability, so enrichment results should not be adversely affected by a small amount of inflation, or upward bias, in the estimate for each individual variance component. I check this notion via permutation analyses in Section 4.7.2.3.

4.7.2 Robustness of functional class enrichment results

Section 3.6 demonstrated that the estimates of heritability from single-variance component models can change to a large extent based on the prevalence assumed for the disease, the MAF threshold applied and the effect-size model used. In this section, I explore the effects of many different modeling parameters on enrichment results when partitioning by functional class. Due to space constraints, I show results only for the three-variance component hierarchical islet-enhancer model. Results shown should be taken as representative across all different models unless otherwise stated. We conducted the same investigations for the eight-variance component hierarchical model and observed qualitatively identical robustness results (data not shown). We observe that over many different models enrichment results are robust to many different parameter settings.

4.7.2.1 Varying modeling parameters

To investigate the effects of varying modeling parameter settings on enrichment results I present results for the 3-variance component hierarchical islet-enhancer model. I inspect the impact of:

- Varying the minimum MAF threshold;
- Changing the effect-size model;
- Excluding variants in known T2D-associated loci;
- Using only variants shared between the Integrated Panel and the Imputed-1000G datasets;
- Using genotype dosages instead of hard genotype calls;
- LD-pruning variants.

I display enrichment results (Figure 4.14) and fold-enrichment results (Figure 4.15) for the different modeling parameters side-by-side and discuss findings below.

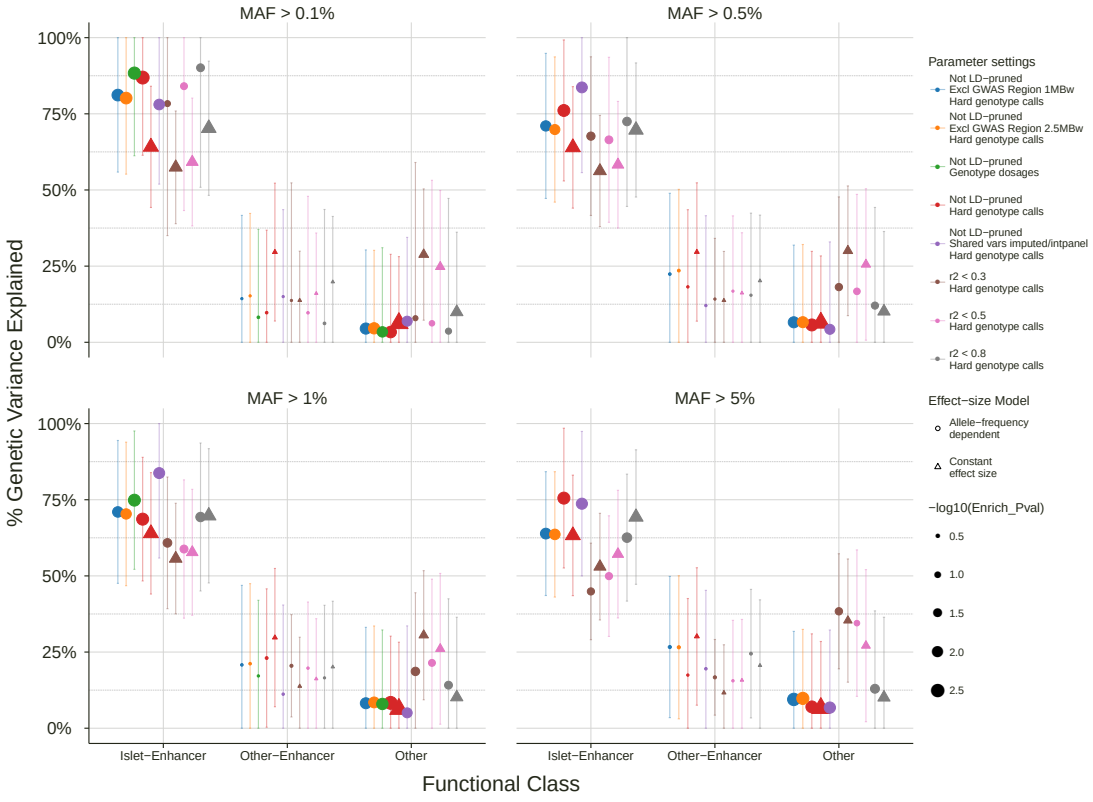


Figure 4.14: *Robustness of enrichment results for many parameter settings in the 3-variance component islet-enhancer model using Integrated Panel data.* Variance components for the given classes are fit jointly. Error bars show standard errors (truncated at zero and 100%). Results are shown for many different parameter settings, including: different minimum MAF threshold, when excluding variants in GWAS regions (a window of either 1Mb or 2.5Mb around the GWAS lead SNP), using genotype dosages instead of hard calls, using variants shared between the integrated panel and the Imputed-1000G datasets, and for various LD-pruning settings. I also show results using the allele-frequency dependent and constant effect size models for variant effect size. The size of the points on the plot reflect the p-value for enrichment for the class.

Varying MAF threshold Enrichment results are very consistent across different minor allele frequency thresholds (Figure 4.14). This consistency holds even though (as shown below) the raw variance explained estimates differ greatly across differing minimum MAF thresholds. There is a slight tendency for enrichment estimates for the islet-enhancer class to decrease as minimum MAF threshold increases. For example, enrichment estimates when using the allele-frequency dependent effect-size model and all variants passing QC are somewhat higher if a minimum MAF threshold of 0.1% is applied than if a minimum MAF threshold of 5% is applied. When using the constant effect size model, which allows for very little contribution from rare and low-frequency variants, there is practically no change in enrichment when all variants (that is, without any LD-pruning) are used.

Fold-enrichment results are even more consistent across minimum MAF thresholds than the enrichment results (Figure 4.15). There is virtually no change in fold-enrichment

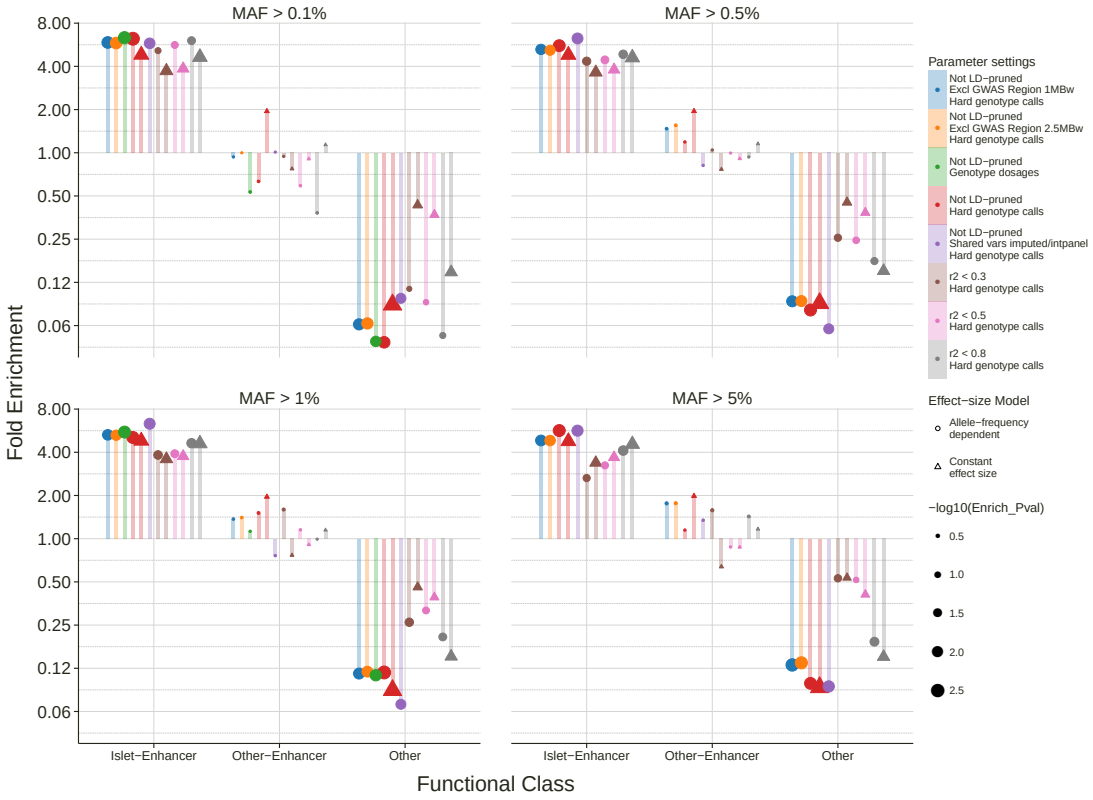


Figure 4.15: *Robustness of fold-enrichment results for many parameter settings in the 3-variance component islet-enhancer model.* Variance components for the given classes are fit jointly. Error bars show standard errors (truncated at zero and 100%). Results are shown for many different parameter settings, including: different minimum MAF threshold, when excluding variants in GWAS regions (a window of either 1Mb or 2.5Mb around the GWAS lead SNP), using genotype dosages instead of hard calls, using variants shared between the integrated panel and the Imputed-1000G datasets, and for various LD-pruning settings. I also show results using the allele-frequency dependent and constant effect size models for variant effect size. The size of the points on the plot reflect the p-value for enrichment for the class.

for islet-enhancer variants and fold-depletion for other variants depending on whether rare and low-frequency variants are included or only common variants are used.

Overall, the pattern of results, both for total genetic variance explained and fold-enrichment across the functional classes, is consistent as with changes to the minimum MAF threshold. Our conclusions do not change depending on whether or not we include rare and low-frequency variants in the analysis.

Varying effect size model Using multiple MAF-binned variance components (Section 4.4) there were negligible differences in the heritability estimates between the allele-frequency dependent effect-size model and the constant effect size model. However, there were substantial differences between the two effect size models for single variance-component heritability estimates (Section 3.6.1). When partitioning by functional class, there is only one

MAF bin for each class, so one might expect to see differences in enrichment results for the functional class depending on the effect-size model.

In fact, the enrichment results are reasonably robust to changing the effect size model assumed (Figure 4.14). We see slightly lower enrichment using the CES model with a minimum MAF of 0.1%, but for higher minimum MAF thresholds the enrichment results are very similar whether using the allele-frequency dependent effect-size model (the default) or the constant effect-size model. The fold-enrichment results are almost identical across the two effect-size models (Figure 4.15). Differences between the equivalent models are not significant.

The CES model can result in lower total heritability estimates than the AFD model. Reassuringly, however, the enrichment results, giving the relative contribution from different functional classes of variant are highly robust to changes in the total variance explained from a model. As seen here, the enrichment results are also robust to the effect-size model assumed. It is interesting, and perhaps surprising, that major differences in modeling assumptions do not translate into major differences in which functional classes of variant appear to be the major contributors to explaining variance in risk for T2D.

Excluding variants in known T2D GWAS loci With over 80 loci associated with T2D, one might wonder if the partitioning results seen are driven by variants in these T2D GWAS loci. To test this, I fitted partitioning using models that excluded variants in known T2D GWAS loci. The 99% credible sets produced by the GoT2D project for the known loci were used to generate "known GWAS regions" as the region between the two variants in the set most distant from each other. These regions were then extended either by 0.5Mb on each side (to get 1Mb exclusion regions) or by 1.25Mb on each side (2.5Mb exclusion regions).

Enrichment results are robust to excluding variants in 1Mb and 2.5Mb windows around known GWAS loci (Figure 4.14). There is slight reduction (not significant) in the enrichment from islet-enhancer variants, hinting at a possible attenuation of signal when associated regions are removed. However, fold-enrichment results are almost identical whether variants around known GWAS loci are excluded or not (Figure 4.15).

Overall, one concludes that the enrichment results observed are not driven by variants in T2D GWAS regions. This suggests a couple of possibilities:

1. that variants with explanatory power may not be confined to previously associated loci, and

2. variants outside known T2D loci may effectively tag variation in the T2D loci, even with exclusion regions of 1Mb or 2.5Mb at these loci.

Both of these possibilities could simultaneously be true. Using only variants that passed QC in the first Wellcome Trust Case-Control Consortium project gives similar enrichment results (data not shown).

Using genotype dosages instead of hard genotype calls I noted above (Section 3.2.4) that measured genotypes could be expressed either as hard genotype calls (the default used) or as genotype dosages. The models presented thus far can be fitted to either type of data. Thus, one may wonder if enrichment results differ when using genotype dosages instead of hard calls.

Reassuringly, enrichment and fold-enrichment results are robust to using genotype dosages instead of hard genotype calls (Figure 4.14 and 4.15). In fact, results using genotype dosages are all but indistinguishable from results using hard calls, so it does not matter for these models if the observed genotypes are expressed in one form or the other.

LD-pruning variants I discussed earlier the possible effects of linkage disequilibrium (LD) on variance partitioning results using LMMs (Sections 3.6.3.1 and 4.7.1.1). From those analyses I concluded that LD-pruning of variants did not appear to be a good approach for the heritability analyses we are interested in. However, the concern here is with relative contributions of classes of variants to VE instead of absolute estimates of heritability or VE. In the main results section above (Section 4.5), I fitted models ignoring any LD effects, so here apply a sanity check on the effects of LD-pruning on enrichment and fold-enrichment results. As previously, I look at results using LD-pruned variants with maximum R^2 thresholds of 0.8, 0.5 and 0.3.

LD-pruning variants does not materially affect the enrichment and fold-enrichment results (Figure 4.14 and 4.15, respectively). The enrichment and fold-enrichment results for the islet-enhancer variants are are very similar with and without LD-pruning. However, the depletion of the other (non-enhancer) class decreases when using LD-pruned variants, especially when using the CES model, and this decrease becomes more pronounced as the the minimum MAF threshold increases.

The overall robustness of functional-class enrichment results to LD-pruning is perhaps surprising given LD-pruning had a noticeable effect on the allele-frequency partitioning models. On reflection, however, this robustness to LD-pruning is perfectly consistent with the robustness to MAF thresholds for the enrichment results shown above. If LD structure is more or less equivalent across functional classes, then one would not expect LD pruning to change the enrichment results. This appears to be the case for the data here.

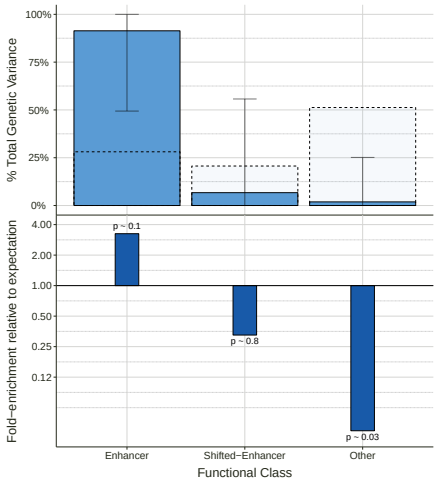
4.7.2.2 Shifted-enhancer models

In the results presented above, enhancer variants explained a large amount of the genetic variance in susceptibility to T2D. However, it might be possible that the signal seen for enhancer variants is due not to the variants in enhancer regions themselves, but rather to variants in LD with (i.e. correlated with) variants in enhancer regions. To exclude this explanation for the signal attributed to enhancer variants, "shifted-enhancer" regions were defined by Kyle Gaulton. The "shifted-enhancer" regions corresponding to each enhancer region are regions flanking the enhancer region on either side, of equal size in base-pairs to the enhancer region, and not overlapping any other enhancer regions. The idea is that these shifted-enhancer regions will contain many of the variants in LD with variants in the enhancer regions. Thus, we can fit enhancer and shifted-enhancer components in the model and compare their contributions to variance in T2D risk. If the shifted-enhancer component explains a large amount of phenotypic variance, then one would doubt that the enhancer variants themselves are as important for the genetic architecture of T2D. This approach is similar to the very recently published "GoShifter" method (Trynka et al., 2015), which develops such ideas to a greater extent than was attempted here.

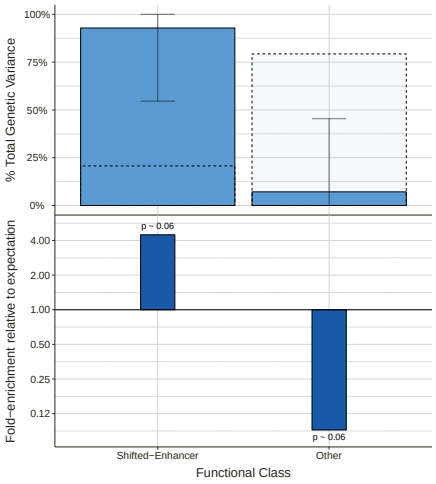
When fitted jointly in a three-variance component model applying minimum MAF thresholds of 0.1% and 0.5%, the enhancer component shows three-fold enrichment, whereas the shifted-enhancer component exhibits roughly 1.4–3-fold depletion and explains less than 20% of total genetic variance (Figures 4.16a and 4.16a). The large standard errors on the estimates preclude one from making strong statements about the significance of this result, but the implication is nevertheless clear. When enhancer and shifted-enhancer components are fit jointly, the model shows a very strong tendency to assign variance in T2D risk to the enhancer variants rather than variants in the flanking shifted-enhancer regions.

I also fit the shifted-enhancer component together with the other component in a two-variance component model, from which enhancer variants are excluded completely (Figures 4.16b and 4.16d). Across minimum MAF thresholds, the shifted-enhancer class exhibits enrichment of roughly 90%, four-fold enrichment relative to expectation. This very strong enrichment from shifted-enhancers shows that these variants successfully tag the enhancer variants. When enhancer variants are excluded from the model, their signal is captured by the shifted-enhancer variants. However, when enhancer and shifted-enhancer components are fit jointly in an LMM, they are made to "compete" to explain the phenotypic variance, and this case, the real enhancers explain almost all of the total genetic variance, with only a small contribution from shifted-enhancer variants.

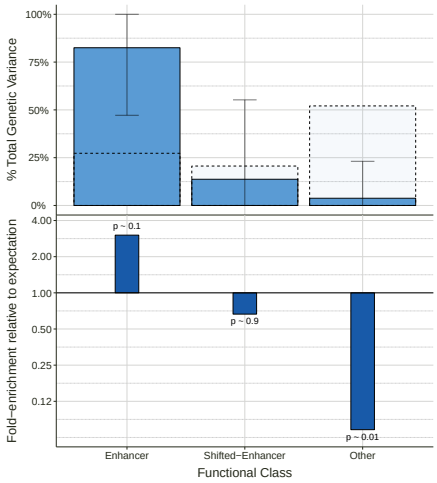
Results from shifted-enhancer analyses indicate that while shifted-enhancer variants effectively tag enhancer variants, the true signal comes from the enhancer variants themselves. These results support the claim made by Gusev et al. (2014) that jointly fitting in an



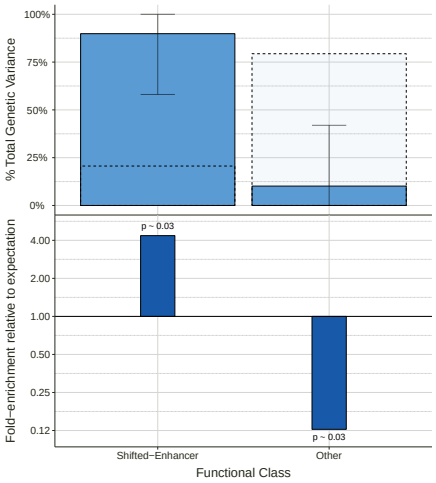
(a) 3VC Shifted-Enhancer Model, MAF $> 0.1\%$



(b) 2VC Shifted-Enhancer Model, MAF $> 0.1\%$



(c) 3VC Shifted-Enhancer Model, MAF $> 5\%$



(d) 2VC Shifted-Enhancer Model, MAF 5%

Figure 4.16: *Enrichment when partitioning into enhancers and shifted-enhancers using Integrated Panel data*. Functional class enrichment results for the 3-variance component model partitioning into enhancer, shifted-enhancer and other classes (a,c) and the 2-variance component model that fits just the shifted-enhancer and other classes (b,d). Results presented here assume allele-frequency dependent effect sizes and use variants with MAF $> 0.1\%$ (a,b) or MAF $> 5\%$ (c,d). The "shifted-enhancer" class consists of variants in regions flanking enhancer regions (of the same size as the enhancer regions, and excluding variants that overlap with other enhancer regions). Variance components for the different functional classes are fit jointly. The upper panel for each plot shows the enrichment (percentage of total genetic variance contributed) for each functional class (total given by the sum of the estimates for all of the genetic components in the model). Dotted bars with transparent fill show the expected proportion of the total genetic variance for each functional class given the proportion of all variants used that are in that class. Error bars show ± 1 standard error (truncated at zero). The lower panel shows fold-enrichment for each class, that is the ratio of percentage of total genetic variance explained to the percentage expected, which is the percentage of all variants that are assigned to that class. P-values testing for a difference between observed and expected enrichment are shown at the end of the bar for each class.

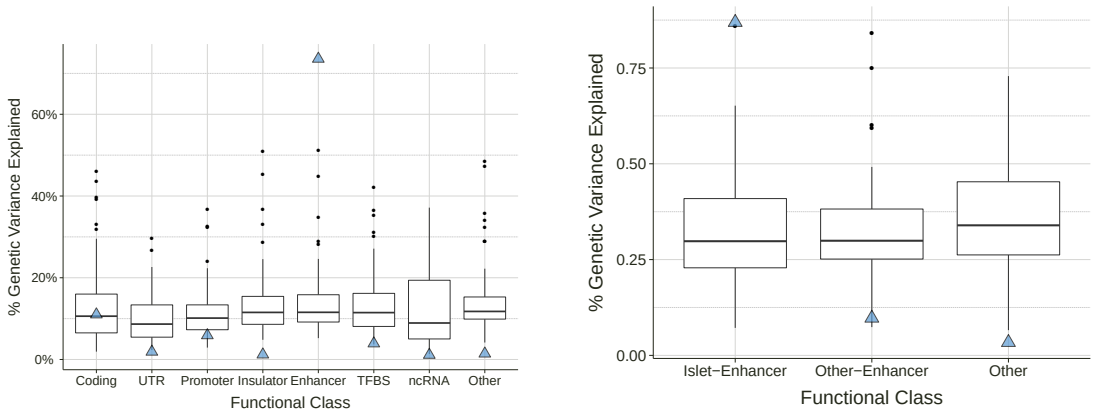
LMM variance components constructed from variants in various functional classes forces the variance components to "compete" to explain phenotypic variance, and that yields enrichment estimates that capture true signal even if there is extensive LD between variants in the various functional classes.

4.7.2.3 Permutation results for enrichment

Permuted case-control results for percentage of genetic variance explained support the notion that the observed enrichment in the enhancer class is unlikely to occur by chance (Figure 4.17). I conducted analyses permuting case-control status across individuals, as outlined previously. Here, I present results when permuting case-control status across all individuals in the Integrated Panel cohort, but observe qualitatively identical results if I permute phenotype within country of origin (data not shown). Similarly, I restrict myself to the allele-frequency dependent model using a minimum MAF threshold of 0.1%, but note that changing the effect-size model or the minimum MAF threshold yield equivalent results (data not shown). I inspect permutation results for both the eight-variance component hierarchical model and the three-variance component hierarchical islet-enhancer model.

In 100 permutations, no estimate of percentage of genetic variance (%GVE) explained for the enhancer class is obtained that is as large as that observed when the true case-control labels are used (Figure 4.17a). Furthermore, across functional classes, the distribution of %GVE estimates is centred roughly on the expected value of 12.5% for the eight-variance component model (expected if the total genetic variance were evenly spread across the eight classes), suggesting that calibration here is reasonable.

Considering the percentage of total genetic variance explained by the islet-enhancer, other-enhancer and other classes in the three-variance component model confirms that the enrichment in islet-enhancers is unlikely due to chance (Figure 4.17b). The distributions of %GVE from the three components are centred on roughly 33% (as expected when there are three classes), again, suggesting resonable calibration of the permutation estimates for enrichment. Although there is substantial variability in the %GVE values obtained from the permutations, there are no permutation enrichment estimates greater than the value obtained for the true case-control status (87%). There are not any enrichment values for the other class that are lower than that observed for the true case-control status. The caveats about inflation in the estimates as previously discussed apply here too, but the overall picture supports the proposition that the enrichment in islet-enhancer variants and the depletion in non-enhancer variants are real effects. Similar permutation results for other models discussed above yield the same conclusions as those presented here (data not shown).



(a) 8-VC hierarchical model

(b) 3VC hierarchical islet-enhancer model

Figure 4.17: *Permutation results for enrichment when permuting phenotype across all samples with minimum MAF of 0.1% and using Integrated Panel data.* Results for the percentage of genetic variance explained (enrichment) by each functional class in (a) an hierarchical 8-variance component model with broad functional categories (b) and a 3-variance component hierarchical islet-enhancer model. These results were obtained using the allele-frequency dependent effect-size model when permuting case-control status across all individuals in the sample (boxplots; $n = 100$). Blue triangles show the results obtained when the true case-control status is used.

4.7.2.4 Robustness of results from imputed data

The previous sections assessing the robustness of enrichment results have focused on results obtained using the Integrated Panel data. I fitted the same models to the two imputed datasets, which generally produced results comparable with those from the Integrated Panel data (see Section 4.6). Thus, we do not repeat all of the above robustness checks on the imputed datasets, assuming that results on imputed data will be similarly robust to effects of excluding variants in known T2D loci and LD-pruning variants. Using the CES model across minimum MAF thresholds yields enrichment results very similar to those obtained with the AFD model and a minimum MAF threshold of 5% (data not shown). In this section I briefly discuss the effects on the imputed data results of varying the minimum MAF threshold and using only variants shared between the Integrated Panel and the Imputed-1000G datasets.

In Section 4.6.2, I only showed results for the Imputed-1000G data. Here, I observe that the enrichment results are very similar for the Imputed-1000G and Imputed-GoT2D data, especially if the minimum MAF is increased from 0.1% to 5% (Figures 4.18a & 4.18b). Perhaps as a side-effect of poorer imputation for rare and low-frequency variants, we see close to perfect agreement in enrichment results when only using common variants in the imputed datasets. Therefore, the imputed data enrichment results are not as robust to varying the minimum MAF threshold as the Integrated Panel results. When only using variants shared between the Integrated Panel and Imputed-1000G datasets (Figures 4.18c

& 4.18d), the results agree closely with those obtained using all variants passing QC in the Imputed-1000G dataset. These results suggest, therefore, that the enrichment results obtained with the imputed datasets are robust to changing the minimum MAF used for variants and also to changing the set of variants used for the analysis.

Taken together, the results presented in this section show that the observed enrichment in enhancer and islet-enhancer variants is robust to many different parameter settings and other factors that could affect enrichment results.

4.7.3 Robustness of functional class variance-explained results

This section explores the robustness of the variance explained, or heritability, estimates underlying the enrichment results. Again, the effects of varying modeling parameters are investigated then "pseudo-enhancer" and permutation results are assessed to gain further insight into the enhancer and islet-enhancer signals observed.

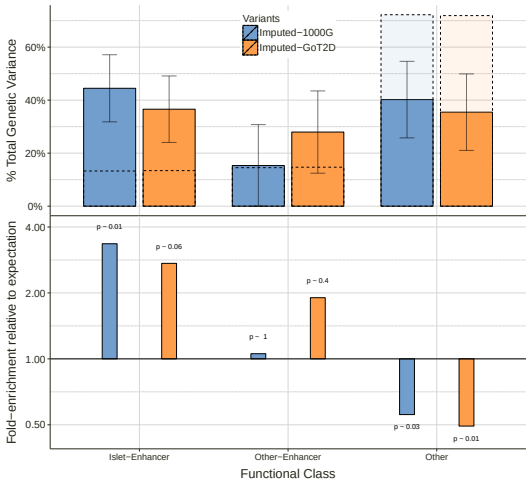
4.7.3.1 Varying modeling parameters

Following on from the investigation of the robustness of enrichment results in Section 4.7.2.1, I assess the impact of the same factors on the underlying variance explained or heritability estimates:

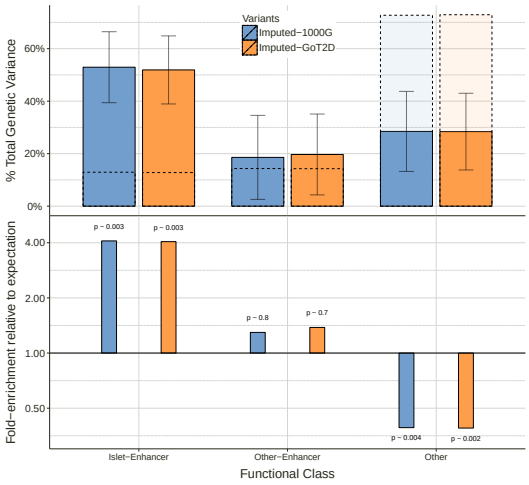
- Varying the minimum MAF threshold;
- Changing the effect-size model;
- Excluding variants in known T2D-associated loci;
- Using only variants shared between the Integrated Panel and the Imputed-1000G datasets;
- Using genotype dosages instead of hard genotype calls;
- LD-pruning variants.

Here, I show results for the three-variance component hierarchical islet-enhancer model using Integrated Panel data (Figure 4.19).

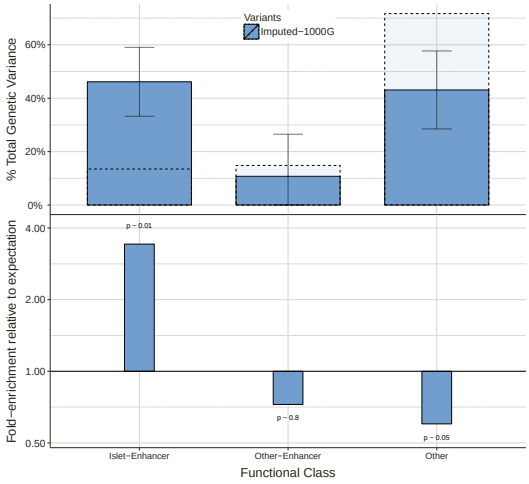
Broadly speaking, there are differences in heritability estimates (at the level of individual components and in the total from the model) when changing these parameter settings. The overall pattern of the contributions from the different functional classes stays very consistent. Across all modeling settings that we explore here, the highest contribution to heritability always comes from the islet-enhancer component. For some of the modeling parameters, differences are observed at lower minimum MAF thresholds (0.1% or 0.5%) that diminish and even disappear for higher minimum MAF thresholds (1% or 5%). Given



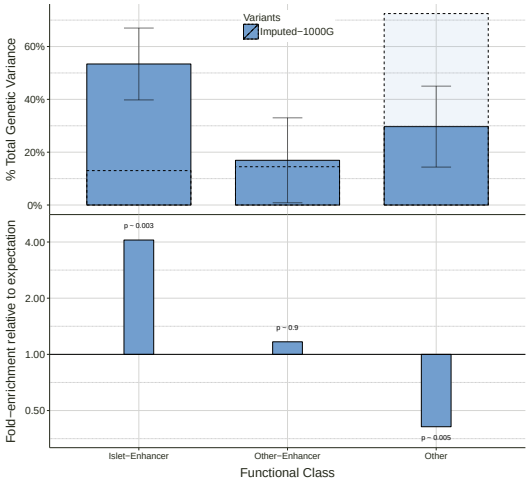
(a) All imputed variants, MAF $> 0.1\%$



(b) All imputed variants, MAF $> 5\%$



(c) Variants also in Integrated Panel, MAF $> 0.1\%$



(d) Variants also in Integrated Panel, MAF $> 5\%$

Figure 4.18: *Enrichment results using imputed data when partitioning into three hierarchical islet-enhancer classes*. Results are shown here for a minimum MAF of 0.1% (a,c) and 5% (b,d), when using either all imputed variants (a,b) or only variants shared between the Integrated Panel and Imputed-1000G datasets. Percentage of genetic variance explained by each functional class in the 3-variance component islet-enhancer hierarchical model assuming allele-frequency dependent effect sizes is shown in the top panel for each plot. The bottom panel shows fold-enrichment for each functional class. Variance components for the different functional classes are fit jointly. Both imputed datasets use the same set of individuals, but the "Imputed-1000G" dataset consists of 1000 Genomes variants imputed into those individuals and the "Imputed-GoT2D" uses GoT2D Integrated Panel variants imputed into those individuals. Results shown here are when using all variants passing QC from each dataset (both imputed datasets) or a subset of variants from the Imputed-1000G dataset that also pass QC in the Integrated Panel dataset. Error bars show ± 1 standard errors (truncated at zero).

the uncertainty in the estimates obtained, very few observed differences could be claimed to be significant.

In Section 4.7.2.1, I discussed the effects of the various factors listed above in detail. Many of the observations made regarding effects on enrichment results hold for variance explained results, so I will not discuss all of these factors again in great detail. Instead, I will focus on some of the larger effects and simply note that, as for the enrichment results, heritability or variance explained estimates differ little when excluding regions around known T2D GWAS loci and using genotype dosages instead of hard genotype calls.

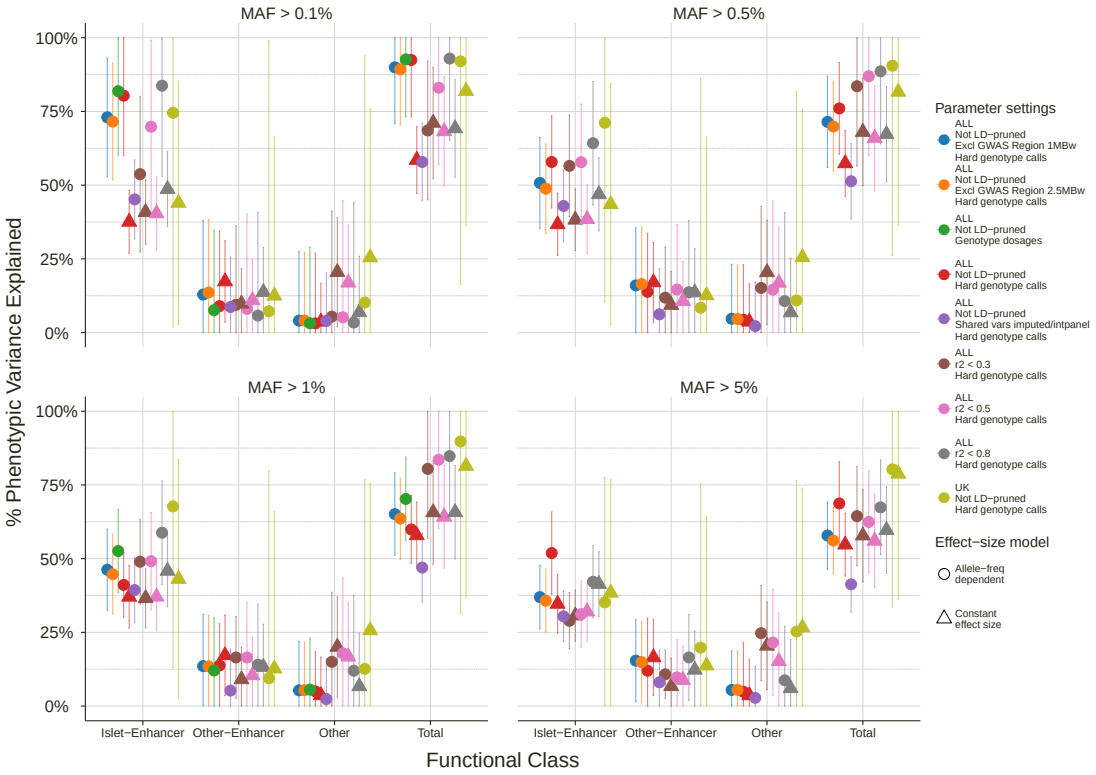


Figure 4.19: *Robustness of phenotypic variance explained results for the islet-enhancer partitioning*. Percentage of phenotypic variance explained by each functional class in the 3-variance component model islet-enhancer model under many different parameter settings, using variants with MAF >0.1% (top left), MAF >0.5% (top right), MAF >1% (bottom left) and MAF >5% (bottom right). Variance components for the different annotation classes are fit jointly. Error bars show ± 1 standard errors (truncated at zero). The total shows the sum of the contributions from each functional class, with standard error computed using the delta method.

The minimum minor allele frequency threshold is the factor with the largest effect on heritability estimates. The total percentage of phenotypic variance explained by the model increases as the minimum minor allele frequency decreases from 5% to 0.1%, universally across other parameter settings (Figure 4.19). In making these comparisons, it becomes clear that the total variance explained by multiple-variance component models partitioning by functional class can become very, almost implausibly, large. An AFD model with MAF

>0.1% gives a liabilty-scale heritability estimate of over 90%, indicating issues with using such models to obtain an estimate of the total heritability for a binary trait. In Section 4.7.4 I examine effects that could inflate the total heritability estimates from models partitioning into multiple functional classes and discuss this issue in more detail.

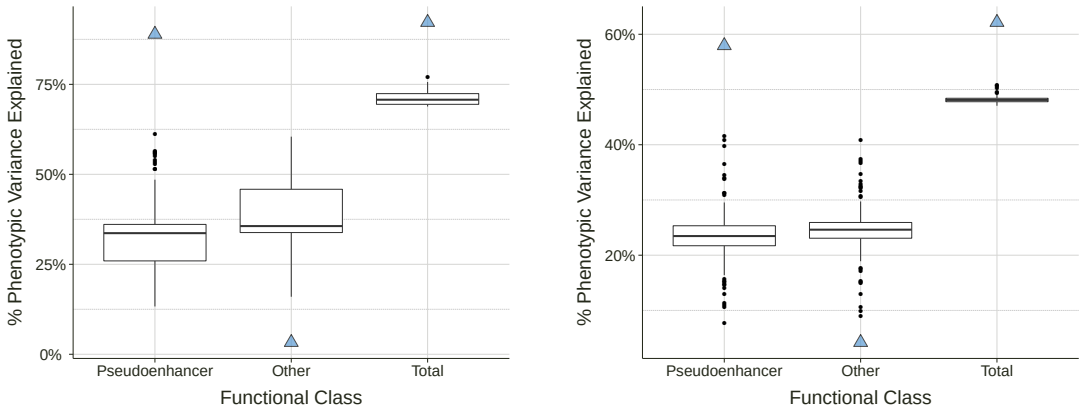
Compared with the effects of changing the minimum MAF threshold, the differences in VE estimates introduced by changing other parameters are subtle. There are some substantial differences in estimates from the AFD model compared with the CES model, with differences generally more pronounced for lower minimum MAF thresholds (as we have come to expect). LD-pruning variants yields many larger estimates for heritability, but these are not significantly different from results from corresponding models different levels of LD-pruning (including no LD-pruning). When the analysis is restricted to use the much smaller set of variants that are shared between the Integrated Panel and the Imputed-1000G datasets, the heritability estimates are lower (as to be expected when using fewer variants), but the overall pattern for the partitioning results is concordant with that observed with other parameter settings.

The heritability estimates, for each component and the total, vary in functional-class partitioning models when changing modeling assumptions and parameters. This variability means that making statements about "the" heritability of a trait is difficult, especially from these types of partitioning models. However, the differences in raw variance explained estimates observed here do not greatly affect the relative contributions from different classes of variant, as demonstrated in the robustness of the enrichment and fold-enrichment results discussed in Section 4.7.2.

4.7.3.2 Pseudo-enhancer results

Another approach to determining if the observed enrichment in enhancers reflects a real effect is to check results obtained when fitting two classes in the model, one "pseudo-enhancer" class and one "other" class. The pseudo-enhancer class consists of the same number of variants as the true enhancer class, with very similar allele-frequency distribution. The aim is to test whether any randomly chosen selection of the same number of variants with similar characteristics could give heritability estimates as large as the actual enhancer variants. No pseudo-enhancer estimates are observed to be as large as the value obtained for the true enhancer variants across 150 pseudo-enhancer sets (Figure 4.20). There are also no estimates for the "other" class in the pseudo-enhancer models that are as small as the estimate for other variants when the true enhancer variants define the second component of the model.

Across many replications for four different MAF thresholds (only MAF 0.1% and 5% shown) the distributions for the pseudo-enhancer estimates are very similar to those from the other variants. Only for MAF greater than 0.1% is there a tendency for larger estimates



(a) MAF >0.1%

(b) MAF >5%

Figure 4.20: *Variance explained results for pseudoenhancer models using Integrated Panel data.* Results are shown for the allele-frequency dependent effect-size model for (a) MAF >0.1% and (b) MAF >5%. Variance components for the "Pseudoenhancer" and "Other" classes are fit jointly. The boxplots show results for 150 replications of fitting this two-variance component model with different sets of pseudoenhancer variants. The blue triangles show the results obtained when the real enhancer variants were used (so are actually "Enhancer" rather than "Pseudoenhancer" variants, but are included for ease of comparison).

from the other variants. The distributions for the totals from the pseudo-enhancer models are very tight, indicating that there is some balancing between the pseudo-enhancer and other estimates in a given model. That is, if the pseudo-enhancer estimate in a given model is higher, then the estimate from the other class will be correspondingly lower such that the total of the two components remains consistent across replications.

	Single-VC Model	Two-VC True Enhancer Model	Median Total from Pseudo-enhancer Models (IQR)
MAF $>$ 0.1%	68.15 (19.25)	92.22 (19.44)	70.73 (69.45–72.43)
MAF $>$ 0.5%	53.61 (15.44)	77.81 (15.58)	56.87 (56.18–57.31)
MAF $>$ 1%	50.15 (14.02)	70.43 (14.14)	52.91 (52.31–53.32)
MAF $>$ 5%	49.03 (11.46)	62.15 (11.54)	48.12 (47.79–48.41)

Table 4.6: Estimates of total liability-scale percentage of variance in risk for T2D explained (standard error in brackets) by single variance-components, two-variance component models with true enhancers and other variance, compared with the median total from two-variance component pseudo-enhancer models. The median is obtained from $n = 150$ replications and the interquartile range (IQR) is shown in brackets. Estimates shown for the single-VC model were obtained using the default model in which variant effect-sizes are allele-frequency dependent (AFD model). Estimates were obtained for MAF thresholds of 0.1%, 0.5%, 1% and 5%.

The total variance explained in the pseudo-enhancer models is very similar to the single-variance component results (Table 4.6). The median total percentage phenotypic variance explained in the pseudo-enhancer models tends to be slightly larger than the estimate from a single-VC model. This makes sense, as the same variants are used, they are just (mostly) randomly assigned to two variance components in the pseudo-enhancer models

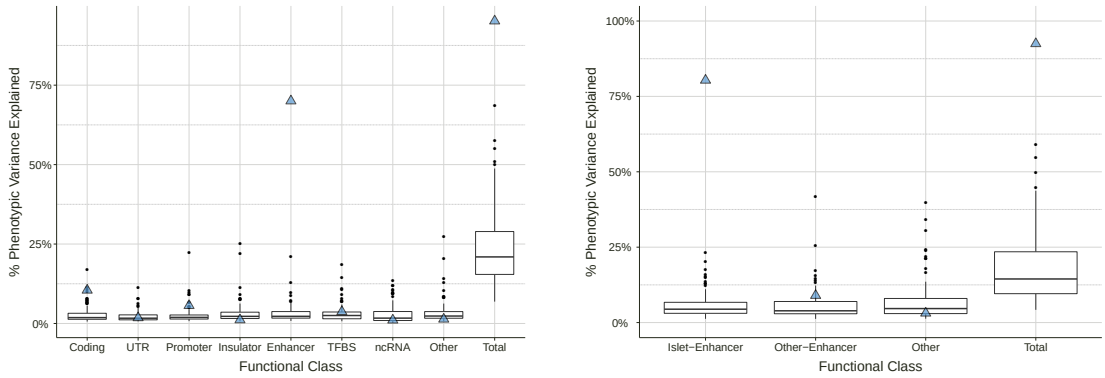
instead of all being used in one variance component in the single-VC model. Partitioning variants into two large sets without separating variants by functional class and estimating a variance component for each set leads to, on average, roughly equal partitioning of the phenotypic variance between these two components.

These results present a contrast with the results from the two-VC model with "real" enhancers and other variants, in which the VE estimates for the enhancer class and the total are substantially higher. These pseudo-enhancer results suggest that partitioning of variants into classes that are more specific and relevant to the disease allows these models to explain more of the variance in risk for T2D. Variability in the pseudo-enhancer estimates is high, but for all replications the estimates are substantially lower that the estimates observed for true enhancer variants. These results provide further support for the idea that enhancer variants explain a large proportion of variance in susceptibility to T2D, and the observed estimates are likely to represent a true effect.

4.7.3.3 Permutation results for variance explained results when partitioning by functional class

To test empirically if the observed contribution from enhancers is likely occur by chance I fit the eight-variance component hierarchical model and the three-variance component hierarchical islet-enhancer model, but permute case-control status to obtain a set of estimates under the null of genetic effects having no capacity to explain phenotypic variance. In Section 4.7.2.3 I looked at enrichment results from these permutation analyses. This section assesses the estimates of variance explained from the permutation analyses and compare them against the results obtained with the true phenotypes. As above, I only present results for the allele-frequency dependent model, applying a minimum MAF threshold of 0.1% and permuting case-control status across all individuals in the Integrated Panel cohort. Permuting case-control status within country of origin instead of across all samples yields qualitatively identical results (data not shown).

In 100 permutations of case-control status across all samples there were no estimates of the percentage of phenotypic variance explained by enhancer or islet-enhancer variants close to the value obtained with the true case-control status (Figure 4.21). Similarly, the total phenotypic variance explained with permutated phenotypes never reached the value observed with true case-control status, in either model. However, even when permuting case-control status, the total phenotypic variance explained could be substantial, with a median of approximately 15% and 75th percentile of 22% in the three-VC model and median of approximately 22% and 75th percentile of roughly 29% in the 8-VC model. Analyses described in Section 3.6.2 suggested that at this MAF threshold, the inflation effect of population structure is at least 2% (percentage points) and could be as high as 10% if fitting a single-variance component model. The permutation results here are consistent



(a) % 8VC hierarchical model

(b) % 3VC hierarchical islet-enhancer model

Figure 4.21: *Permutation results for variance explained when permuting phenotype across all samples using Integrated Panel data.* Results for the percentage of phenotypic variance explained (VE) by each functional class in (a) an hierarchical 8-variance component model and (b) a 3-variance component hierarchical islet-enhancer model. These results were obtained when permuting case-control status across all individuals in the sample (boxplots; $n = 100$). Blue triangles show the results obtained when the true case-control status is used.

with this, as the 75th percentile of variance explained estimates across functional classes is 3.4%, with median 2.1% in the 8VC model and 75th percentile of 7.1% and median 4.5% in the 3-VC model, although individual estimates can be substantially larger.

Thus the per-component inflation in the 3-VC model is higher than in the 8-VC model, but the total inflation is higher in the 8-VC model by virtue of having many more components. Taking the sum of estimates over eight or three functional classes (as appropriate) can therefore lead to a substantial total, even if each individual estimate is inflated only a little by structure, bias and other effects.

4.7.4 High estimates for total phenotypic variance explained

High total heritability estimates have been observed in models with multiple variance components both when partitioning by allele frequency (recall Figure 4.22) and when partitioning into multiple functional classes, especially if enhancer or islet-enhancer variants are used in a distinct component (recall Figure 4.19). Here, I investigate the high totals in greater depth and discuss possible causes for inflated totals for variance explained. I examine increased totals when fitting more variance components in allele-frequency partitioning models, effects of changing prevalence values, effects of fitting highly-correlated variance components, and inflation from population structure and other biases.

4.7.4.1 Higher totals when fitting more variance components

Across all of the analyses partitioning into multiple variance components (whether by allele frequency or functional class), higher totals for variance explained are obtained than

when fitting a single-variance component. One might ask to what extent the increase in the total is due solely to fitting more variance components in the model. This can be investigated in the allele-frequency partitioning setting by comparing the total heritability estimates, with both effect-size models and LD-pruning of variants, from single-component, three-component and eight-component models (Figure 4.22).

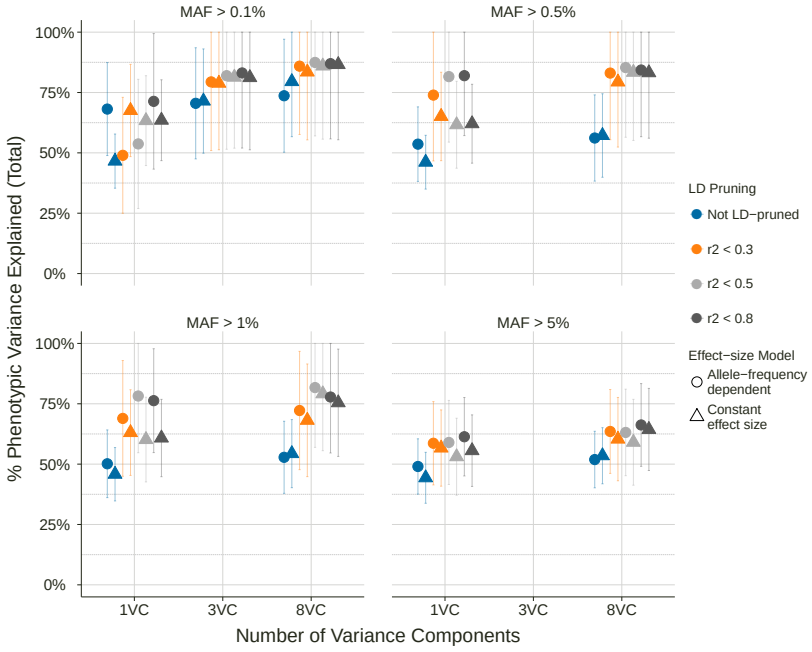


Figure 4.22: *Comparing total phenotypic variance explained on the liability scale for different parameter settings for models with different allele-frequency partitioning of variants using Integrated Panel data.* Here we can directly compare total variance explained Totals are the sum of all variance component estimates for a given model. For the MAF $>$ 0.5%, $>$ 1% and $>$ 5% results, the "8VC" model actually has 7, 6 and 5 components respectively, as we drop variance components for lower MAF ranges, raising the minimum MAF threshold. Error bars represent ± 1 standard error, truncated at 100%.

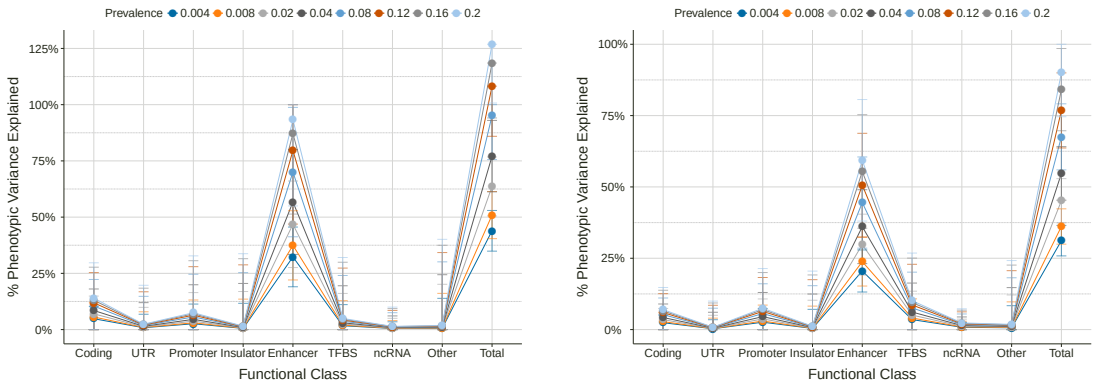
With a minimum MAF threshold of 0.1% there is a steady increase in total from the 1-VC to the 3-VC to the 8-VC model. There is a similar increase in total from the 1-VC model to the 8-VC model as the minimum MAF threshold is increased to 0.5%, 1% and 5%. (Note that for the minimum MAF of 0.5%, 1% and 5% results, the "8-VC" model actually has 7, 6 and 5 components respectively, as variance components are dropped for lower MAF ranges, raising the minimum MAF threshold.) The increase in totals as we use more variance components appears across all LD-pruning approaches. As discussed previously, substantially higher totals are obtained when applying LD-pruning, for reasons that are not entirely clear.

Overall, the inflationary effect of fitting more variance components is relatively small, accounting for an increase in total heritability estimates of 5–10%. This size of increase

does not account for all of the increase in total seen in models partitioning into multiple functional classes, but appears very likely to contribute to some of the inflation observed.

4.7.4.2 Effects of changing disease prevalence value

As previously discussed, it is desirable to convert variance-component estimates from the observed scale to the more interpretable liability scale (see Section 3.3.8). One only needs to apply a linear transformation to the observed-scale estimates, and the transformation depends only on the proportion of cases in the sample and the proportion of cases in the population (the prevalence). Results above showed that heritability results are generally robust to changes in the subset of variants used. However, the assumed value for the prevalence of the disease has a large effect on the estimates of heritability on the liabilty scale, as discussed in the case of single-variance component estimates in Section 4.7.4.2. This is expected to be the case in multiple-variance component models too, so in this section I show the effects of changing the prevalence value from the default of 8% in the eight-variance component hierarchical model. I present results for the allele-frequency dependent effect-size model and minimum MAF thresholds of 0.1% and 5%.



(a) MAF >0.1%

(b) MAF >5%

Figure 4.23: *Robustness of phenotypic variance explained results to changing the prevalence value for the eight-variance component hierarchical model using Integrated Panel data.* Percentage of phenotypic variance explained by each functional class in the eight-variance component hierarchical model with many different prevalence values. Results are shown using the allele-frequency dependent effect-size model and variants with (a) MAF >0.1% and (b) MAF >5%. Variance components for the different annotation classes are fit jointly. Error bars show ±1 standard errors (truncated at zero). The total shows the sum of the contributions from each functional class, with standard error computed using the delta method.

I investigated the effect on VE estimates of changing the assumed prevalence for T2D (Figure 4.23). When partitioning the variance into eight variance components using the hierarchical model with MAF greater than 5% and applying prevalence values ranging from 0.4% to 20%, the point estimate for the total percentage of phenotypic variance explained

by the genetic effects ranges from approximately 30% for a prevalence of 0.4% to about 90% for a prevalence of 20% (Figure 4.23b). Even a small change in the prevalence value can lead to a substantial change in liability-scale estimates of variance explained. For example, the default prevalence value of 8% gives a total variance explained estimate of approximately 68%, but if one uses a prevalence value of 4% the VE estimate reduces to roughly 55%. If one uses a prevalence value of 2%, then the VE estimate reduces further to 45%. If the appropriate prevalence were actually 10% of that which has been taken as the default (actually 0.8% instead of 8%), then the VE estimate would be 37%, only just over half the estimate obtained with the default prevalence.

Conversely, if higher prevalence values are used, the liability-scale estimate of the VE increases. If one uses an inappropriate prevalence value, then VE estimates can become nonsensical. This effect is obvious when using variants with MAF greater than 0.1% (Figure 4.23a). An estimate of approximately 95% for the total %VE is obtained when using the default prevalence of 8%. If the prevalence value is increased, to 12% or more, then the VE estimates increase to over 100%. It is, of course, impossible for the VE or heritability to be truly greater than 100%, so applying the wrong prevalence value could lead to problems with inference. The effect of reducing the prevalence value is emphatic with a minimum MAF threshold of 0.1% than MAF 5%. The absolute differences in the VE estimates when using different prevalence values are larger when the underlying observed-scale estimates from the REML fit are larger, which is the case when including more variants with lower MAF.

Overall, the assumed value for the disease prevalence has a large effect on the interpretation of the model results. Liability-scale estimates of VE depend strongly on the prevalence value assumed for the disease. Larger prevalence values lead to larger VE estimates, and smaller prevalence values reduce VE estimates. Applying an inappropriate prevalence value can lead to nonsensical VE estimates or otherwise misleading inference. On one hand, this strong effect of the prevalence value is known. Prevalence appears in the equation for converting observed-scale heritability estimates to the liability scale in the original paper by Lee et al. (2011), and is necessarily quoted for any liability-scale heritability estimates for binary traits obtained using the LMM approach. On the other hand, I am not aware of any discussion in the literature quantifying the effect on liability-scale heritability estimates of varying the prevalence value, as shown here. In the analyses of variance partitioning for the GoT2D data presented above I noted that there appeared to be inflation in the VE estimates in the multiple-component models, and that the total VE estimates from these models were implausibly high (for example, total %VE of 95% when partitioning into eight annotation categories using variants with MAF greater than 0.1%).

One possible explanation for why the total VE for these models is so high is that the prevalence value used is too high. Given that an opportunistic "extreme phenotype" sam-

pling approach was taken to ascertain the GoT2D samples, it seems plausible that a prevalence of 4% or 2% (or even lower) could be more appropriate than the standard prevalence value for T2D of 8%. If one were to adopt a lower prevalence value, then the total VE would be substantially lower, and one would have less concern about the total VE or heritability estimates being too high. If a lower prevalence value were used then all liability-scale heritability estimates would be reduced. Thus, relative differences in the totals between the single-variance component models, allele-frequency partitioning models and functional-class partitioning models would still exist. In this case, one would not be as concerned about high total heritability estimates overall, but would still seek an explanation for why functional-class partitioning models tend to yield higher total heritability estimates than allele-frequency models with the same number of variance components.

4.7.4.3 Inflation from population structure and other biases

Population structure and permutation results have been discussed in greater detail previously, but I briefly recap results here and discuss their consequences in relation to the consideration of high total heritability estimates observed from multiple-component variance partitioning models.

I previously noted upward bias in variance component estimates when looking at permutation results (Sections 4.7.1.2 and 4.7.3.3). Null estimates, obtained when permuting case-control status, are always non-zero with a median heritability estimate of 3–5%. The total heritability estimates for permuted phenotypes from models with multiple variance components, correspondingly, commonly fall in the range of 15–30%. The permutation estimates are likely inflated by both upward bias in the raw REML variance component estimates and population structure effects. It is problematic to relate the inflation observed in permutation results directly to results obtained using the true phenotypes, because inflation at the boundary of the parameter space (namely zero, for variance component estimates) may well be different from inflation for estimates away from the boundary (and heritability estimates for many components in many models are well away from the boundary). Nevertheless, the most reasonable conclusion from the permutation results and estimates of population structure effects is that population structure and upwardly-biased REML estimates are likely to contribute to inflation in total heritability estimates from variance partitioning models.

When fitting variance components constructed by partitioning variants by functional class, several variance components are defined that use tens of thousands to millions of variants across the full MAF spectrum. These variance components each contain very many variants that are slightly informative for population structure, leading to the cumulative effect that each variance component captures population structure effectively.

Thus, a possible reason for greater inflation of total heritability estimates in the functional-class partitioning models is that there are several components in the model which are all inflated by population structure effects. Compared with the allele-frequency partitioning, functional-class variance components tag population structure to a greater extent. If functional-class partitioning is affected by population structure in this way, then it would also lead to higher correlation between variance components, which as the next section shows, appears related to higher total heritability estimates.

Assessing the overall inflationary effect of population structure over all modeling contexts is challenging, but it seems reasonable to assume that population structure effects may inflate total heritability estimates in the partitioning models by 5–10%.

4.7.4.4 Effect of correlated variance components

Another posited explanation for inflated total heritability estimates in multiple-variance component models is LD "leakage" between components. We might imagine that computing GRMs from sets of variants that tend to be correlated to a reasonable extent would lead to correlated relatedness values in the GRMs. These GRMs are used to define the covariance structure of the random effects in the LMM. It may be the case that fitting more highly correlated variance components, in which LD or correlation between variants "leaks" between variance components, inflates heritability estimates (per component and in total) from the LMM.

To test the effect of correlated variance components on total heritability estimates, I compare total variance explained in a selection of representative examples of three-variance component models with different levels of correlation between variance components. For this comparison I use the allele-frequency dependent effect size model and minimum MAF of 0.1%. I define correlation between components as the Pearson correlation between relatedness values from the GRMs that define the variance components. Correlations were computed from a random sample of 10,000 off-diagonal relatedness values from the GRMs.

Fitting more highly correlated variance components relates to higher total variance explained by the model (Table 4.7). As a basis for comparison I start with the 3-VC MAF-bin model without LD-pruning using Integrated Panel data. This model has a moderate degree of correlation between variance components and yields an estimate for total variance explained of 70.5%. Fitting this same model, but applying LD-pruning of variants with a maximum R^2 threshold of 0.3, leads to the degree of correlation between components increasing markedly, as does the total variance explained, to 79.4%. When one moves from the 3-VC MAF-bin model to a 3-VC hierarchical islet-enhancer model (no LD-pruning) even higher correlation between variance components and the highest total seen for any three-variance component model (total heritability estimate of 92.5%) is observed.

Model	Data	Correlation between components	Total VE
3VC MAF-bin model, no LD-pruning	Integrated Panel	0.63, 0.56, 0.79	70.5%
3VC MAF-bin model, LD pruned $R^2 < 0.3$	Integrated Panel	0.78, 0.71, 0.90	79.4%
3VC islet-enhancer model, no LD-pruning	Integrated Panel	0.89, 0.86, 0.91	92.5%
3VC islet-enhancer model, LD pruned $R^2 < 0.3$	Integrated Panel	0.57, 0.72, 0.67	68.6%
3VC islet-enhancer model, no LD-pruning	Imputed-1000G	0.54, 0.46, 0.59	73.5%

Table 4.7: Comparing total variance explained (VE) in 3-variance component models with different levels of correlation between variance components. I defined correlation between components as the Pearson correlation between relatedness values from the GRMs that define the variance components. Correlations were computed from a random sample of 10,000 off-diagonal relatedness values from the GRMs. Correlations are given in order to represent correlation between: first and second components, first and third components, second and third component. For the MAF-bin model, variance components are "Rare", "Low-frequency" and "Common". For the islet-enhancer model, variance components are "Islet-Enhancer", "Other-Enhancer" and "Other". Results presented here were obtained using allele-frequency dependent effect sizes and used variants with MAF $> 0.1\%$.

Next, to change the direction of these comparisons, I look at a 3-VC islet-enhancer model with lower correlation between components by LD-pruning variants with maximum R^2 of 0.3. For this model there is slightly less correlation between components than for the first model (3-VC MAF-bin) and also a lower total (68.6%). So even when fitting the islet-enhancer partitioning model which initially yields a very high total, if the correlation between variance components is reduced then there is a drop in the total variance explained. Finally, I compare these results to results from the 3-VC islet-enhancer model when using the Imputed-1000G datasets without LD-pruning. Here, there is slightly lower correlation between components, and a total variance explained estimate of 73.5%, which is comparable if slightly higher than the totals for the 3-VC MAF-bin model without LD-pruning and the 3-VC islet-enhancer model with strong LD-pruning.

Overall, the relationship between increasing correlation between variance components and increasing total variance explained by the model appears strong. This effect could go a long way to explaining why such high total heritability estimates are observed from some of the multiple-component partitioning models. One perhaps counter-intuitive result arising from this investigation is that there is higher correlation between variance components for Integrated Panel data when LD-pruning is applied to variants. This increase in correlation between variance components provides the best explanation so far for why higher total heritability estimates are obtained when LD-pruning variants than when variants are

not LD-pruned, even though variants are removed from the model that could potentially tag genetic signal.

4.7.5 Results in sub-populations

The GoT2D cohort that I have been using in this study consists of individuals from five distinct sub-populations: Botnia, Finland, Germany, Sweden and the UK. In the analyses described above, I have controlled for population structure effects by fitting principal components as fixed effects in the models (see Sections 3.4.4). In Section 3.6.2.1 I determined that fitting more than 10 principal components in the model made no further difference to single-variance component heritability estimates, and I attempted to calculate the remaining effect of population structure in Section 3.6.2.2. Nevertheless, one might worry that population structure has further effects on the enrichment results observed beyond what could be accounted for with principal components or captured by computing inflation from population struture. To test this idea, I fit the eight-variance component hierarchical functional class model separately on individuals in the distinct sub-populations and look for major discrepancies in the partitioning results.

I first split the GoT2D cohort roughly in half, looking at the German, Swedish and UK individuals as a group (1,290 individuals) and the Finnish and Botnian individuals as a group (1,264 individuals). This is a natural division of the cohort, as analysis of GRMs and principal components cleanly separated Finnish and Botnian individuals from German, Swedish and UK individuals (data not shown). I then conduct the same analysis separately on the Finnish cohort (968 individuals), the UK cohort (643), the Swedish cohort (442), the Botnian cohort (296) and the German cohort (205). For each sub-population I recompute principal components using only the individuals in that sub-population and fit the first ten population-specific principal components along with a Sex effect and (where appropriate) a Batch effect as fixed effects in the LMM. It is expected that the uncertainty in the estimates increases substantially as the sample size decreases, but one hopes to see similar patterns in the partitioning of variance in the sub-populations as seen when studying the whole GoT2D cohort combined.

4.7.5.1 Enrichment results

I investigate enrichment results for variants with MAF greater than 0.1% (Figure 4.24). As the robustness results above suggested, the enrichment results are qualitatively the same for different minimum MAF thresholds (data not shown). For the smaller cohorts, there is less stability in the estimates as the minimum MAF changes, but high standard errors on estimates mean that differences are not significant.

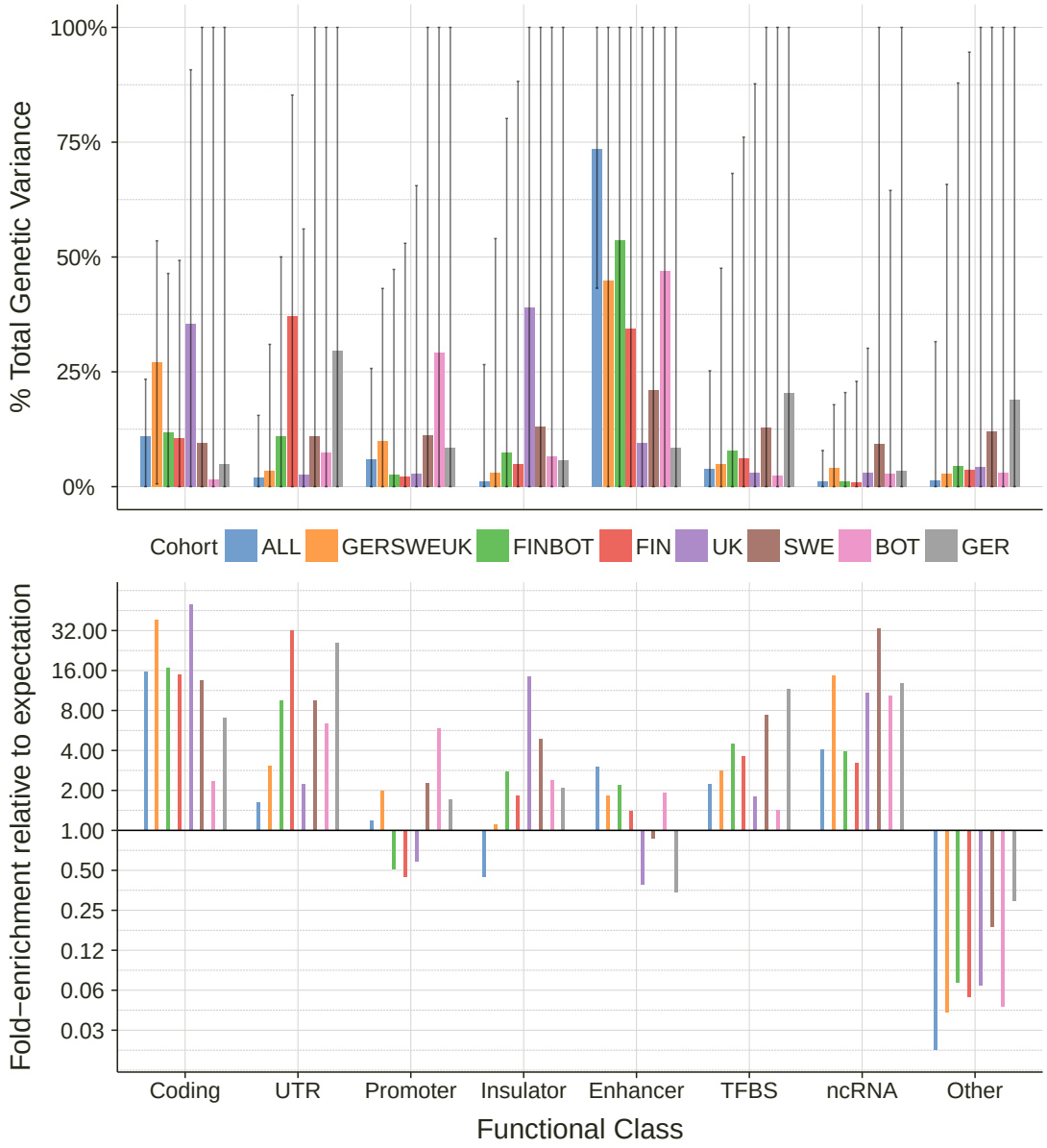


Figure 4.24: *Enrichment results for sub-populations when partitioning into eight functional classes: MAF >0.1%.* Results using Integrated Panel data for the percentage of genetic variance explained and fold-enrichment for each functional class for the true phenotypes comparing results obtained in sub-populaitons: "ALL" (all individuals), "GERSWEUK", (German, Swedish and UK individuals), "FINBOT" (Finnish and Botnian individuals), "FIN" (Finnish individuals), "UK" (UK individuals), "SWE" (Swedish individuals), "BOT" (Botnian individuals) and "GER" (German individuals). The cohorts as written here appear in decreasing order of sample size. Results shown here are for variants with MAF $> 0.1\%$. Enrichment p-values are not shown here, as with the exception of the "ALL" cohort, uncertainty in estimates are so large that all p-values are greater than 0.2. Error bars show ± 1 standard error, truncated at 0 and 100%.

The German, Swedish and UK individuals make up slightly more than one half of the GoT2D sample after filtering (1,290 individuals with relatedness less than 0.05). The variance partitioning results for these individuals are very comparable to those obtained from the whole sample combined (Figures 4.24). With this sample size, the standard errors of the variance component errors are large so no significant enrichment results are obtained, but there is very similar apportioning of the genetic variance across functional classes as seen for the full GoT2D cohort. The enhancer class here explains around half of the total genetic variance (more than 50% for variants with MAF greater than 5%, slightly less than 50% for MAF greater than 0.1%). Such enrichment corresponds to approximately two-fold enrichment. Greater than 20-fold enrichment is observed for the coding variants and strong depletion for the "Other" class of variants.

There are 1,264 individuals (after filtering) in the Finnish and Botnian cohorts combined, just under half of the whole GoT2D cohort. Just as for the German, Swedish and UK samples, above, the variance partitioning results for these individuals are very comparable to those obtained from the whole sample combined (Figures 4.24). Again, there are large standard errors and correspondingly high p-values (not shown) for enrichment. Nevertheless, the enhancer variants still explain 50% of the total genetic variance, exhibiting two-fold enrichment. Just as above, the coding variants show approximately 20-fold enrichment and there is strong depletion in the "Other" class.

Variance component and enrichment estimates become very noisy for the individual-country cohorts, and standard errors for enrichment results cover the whole parameter space (0–100% of total genetic variance). Thus, the enrichment results for the Finnish, UK, Swedish, Botnian and German cohorts are consistent with results in larger, combined cohorts as there are no significant differences in results. Reassuringly, the patterns of enrichment looks generally similar to those observed in larger samples and strong depletion is consistently observed (not significant) in the "other" class along with high estimates for fold-enrichment for coding variants.

4.7.5.2 Permutation results

Permutation results in sub-populations provide modest support for the idea that the observed enrichment in the enhancer class is unlikely to have been observed by chance (Figure 4.25). In at most one out of 100 permutations is there an estimate for the enhancer variants greater than that observed for the true phenotypes for the German, Swedish and UK individuals combined, Finnish and Botnian individuals combined, or Finnish individuals. In the smaller cohorts, as discussed above, estimates are less stable. This is reflected in the permutation results, where for the UK, Swedish, Botnian and German individuals, there are numerous permuted results producing an estimate larger than that obtained with

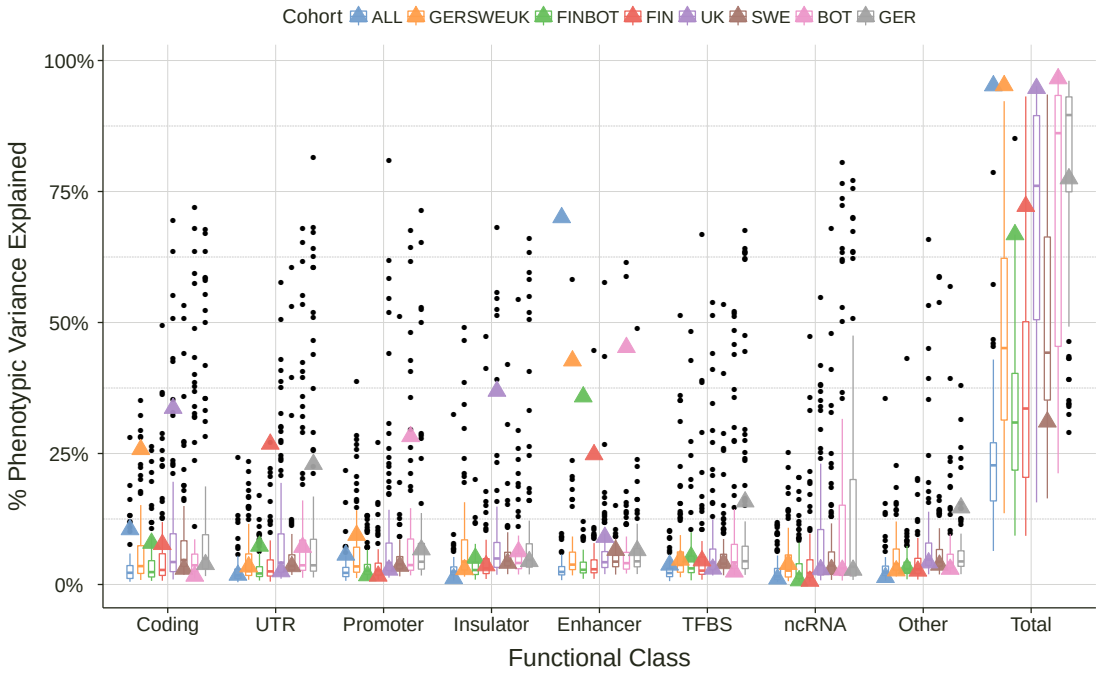


Figure 4.25: *Permutation results for sub-populations when partitioning into eight hierarchical functional classes.* Results for the percentage of the phenotypic variance obtained when permuting case-control status across individuals within sub-populations. comparing results obtained in sub-populations: "ALL" (all individuals), "GERSWEUK", (German, Swedish and UK individuals), "FINBOT" (Finnish and Botnian individuals), "FIN" (Finnish individuals), "UK" (UK individuals), "SWE" (Swedish individuals), "BOT" (Botnian individuals) and "GER" (German individuals). The cohorts as written here appear in decreasing order of sample size. Results shown here are for variants with MAF $> 0.1\%$. Boxplots show results from $n = 100$ permutations. Coloured triangles show the results obtained when the true case-control status is used.

the true phenotypes. Compared with the previous permutation results on the whole cohort, there is much higher variability in the permutation results across all functional classes for the smaller cohorts, and the variability increases as the sample size decreases. Seeing higher variability in the estimates from a smaller sample is, of course, what one would expect.

Overall, patterns of enrichment are consistent as the GoT2D sample is broken down into smaller sub-populations. It does not look like the enrichment results we observe in the full sample are being driven by enrichment signals arising in just one or two of the distinct country cohorts. The possiblities for interpretation of the results from smaller cohorts is limited by high uncertainty in estimates, but we conclude that population structure effects are not a major factor in the observed enrichment results.

4.7.5.3 Higher totals for smaller sample sizes

Throughout these variance partitioning analyses, particularly when partitioning by functional class, we have seen high estimates for the total percentage of phenotypic variance

explained by these models. In Section 4.7.4.1 I noted slightly higher total heritability estimates when more variance components are fitted in the model and in Section 4.7.4.2 I demonstrated that changing the assumed prevalence value for the disease has a large effect on the variance component estimates and the total variance explained by the model.

In the sub-population permutation analyses above, another factor is observed that seems to influence the total variance explained by model: sample size. In an eight-variance component model partitioning by hierarchical functional class on our "full" cohort (2,554 individuals), the median total for the percentage of variance explained was approximately 20% (Figure 4.21a). The median totals were lower (approximately 15%) for models in which fewer variance components were included (see Figure 4.21b). When the sample size is halved when permuting phenotype (by taking either the 1,290 German, Swedish and UK individuals as a group, or the 1,264 Finnish and Botnian individuals) then the median total percentage of phenotypic variance explained by the model increases to 30-45%. When the sample size is reduced further by looking at permutation results for the individual sub-populations, the median total variance explained steadily increases, from approximately 30% for the Finnish cohort (964 individuals) to approximately 75% for the UK cohort (643 individuals), 45% for the Swedish cohort (442 individuals), 87% for the Botnian cohort (296 individuals) and finally over 90% for the German cohort (205 individuals). Note that the median total variances can vary considerably for different minimum MAF thresholds (data not shown), but in a similar sort of range, and generally with higher totals for smaller sample sizes.

Total estimates of phenotypic variance explained from LMMs are sensitive to sample size. The individual variance components estimates become more variable as the sample size decreases and the overall phenotypic variance explained by the model appears to increase (albeit with large variability in these totals). This means that, especially for small sample sizes, models that should not actually explain the phenotypic variance (as when phenotypes are permuted) can give sizeable totals for the phenotypic variance coexplained. Thus, it is possible that the high totals observed in the partitioning analysis on the real data are influenced by this sample-size effect. In the context of association studies and LMM analyses, 2,554 individuals is a relatively small sample.

In a paper proposing an alternative approach to REML for estimating heritability from genotype data, Golan et al. (2014) argue with analyses simulating case-control status that REML estimates are downwardly biased as the sample size increases. Here, the same effect is apparent: higher estimates for smaller samples. Estimates of the total variance explained by these models tends to increase as the sample size decreases and lower estimates for total variance explained are obtained in imputed data sets with more samples (although there are other reasons why one could see these results, sample variation being one). It is

difficult to tease out the differences that are due to increased sample size from those due to a different set of variants and samples used for the analysis.

The results from Golan et al. (2014) and the results here indicate that the LMM approach when partitioning into multiple functional classes has weaknesses for estimating the absolute value of the heritability. The particular models fitted here, partitioning variance into many variant classes, do not give reliable estimates for the absolute heritability of T2D. Happily, this is not the aim. The enrichment results, which are based on the relative contributions to variance explained from different classes of variants, do not depend on the absolute estimates of variance explained and are robust.

4.8 Discussion and conclusions

In this and the previous chapter I have presented many analyses investigating the genetic architecture of type 2 diabetes by estimating and partitioning heritability using whole-genome sequence data for a cohort of 2,554 individuals with and without T2D. I estimated heritability using linear mixed model methods and used multiple variance components to partition variance into multiple allele-frequency classes and multiple functional classes. To the best of my knowledge, this is the first application of these variance-partitioning models to whole-genome sequence data. I replicated key findings using imputed genotype data from a larger UK cohort.

Using variants with minor allele frequency greater than 0.1% in the Integrated Panel dataset I obtained a heritability estimate of 68% (s.e. 19%) with the most standard approach currently used in the field. When I repeated this analysis in a second, larger cohort using two sets of imputed variants I obtained heritability estimates of 69% (s.e. 10%; Imputed-1000G data) and 74% (s.e. 10%; Imputed-GoT2D data). These estimates are higher than previous esimates from chip genotype data that do not study rare variation (Gusev et al., 2013), but comparable to family-based heritability estimates from T2D patients with age at onset 35–60 years (Almgren et al., 2011).

Single-variance component heritability estimates confirmed that common variants explain a large proportion of variance in risk for T2D. However, single-variance component results did not clarify the collective contribution that low-frequency and rare variants make to T2D heritability. To probe further the contribution from low-frequency and rare variants I analysed multiple variance component models partitioning variance into multiple allele-frequency classes. This approach was suggested by Lee et al. (2013) as a better way to obtain heritability estimates using dense genotype data, where there is extensive linkage disequilibrium between assayed variants.

When partitioning variance into multiple allele-frequency classes, results confirmed that common variants explain a large amount of variance in liability for T2D. The contri-

bution from rare variants remains unclear, as point estimates were high, but so was the uncertainty in the estimates. The contribution from low-frequency variants appeared negligible. I observed very similar results in the imputed datasets, and these estimates did not depend greatly on assumptions about the effect-size model for variants. The smaller standard errors on the imputed data estimates provided modest evidence that the collective contribution of rare variants to T2D heritability is indeed non-zero. Thus, the results suggest a potentially important role for the collective contribution of rare variants to T2D liability, but larger sample sizes with greater precision in the estimates will be required to place meaningful bounds on the extent of the contribution to variance explained from rare variants.

These results suggest a future approach to seeking a parsimonious model that remains flexible enough to capture differences in the effects of variants across the allele-frequency spectrum. It appears a four- or five-variance component model should perform well, with variance components from rare (MAF 0.1–0.5%), low-frequency (MAF 0.5–5%) and MAF 5–10% variants, and then either one MAF 10–50% component or a MAF 10–30% and a MAF 30–50% component. Such a MAF-partitioning model could provide the right trade-off between fitting a small number of variance components and having sufficient flexibility to characterise the aggregate contribution of variants across the allele-frequency spectrum to variance in T2D risk. With more data, such a model could yield more definitive results.

Changing tack, from allele-frequency to function, I then analysed multiple-variance component models partitioning variance into different functional classes. Results showed that different functional classes explain substantially different proportions of the variance in susceptibility to T2D. I observed significant three- to four-fold enrichment for variants in enhancer regions identified from pancreatic islet cells and significant depletion for variants that did not receive any functional annotation. These findings replicate in a second, larger cohort using imputed data. In the Integrated Panel data, shared islet enhancers (variants in enhancer regions identified in islet cells and at least one other cell type) showed significant eight-fold enrichment, potentially further specifying the set of variants driving enhancer enrichment results.

Permutation and empirical testing results confirm that the large estimates for the contributions from enhancer, particularly islet-enhancer, variants are estimating real contributions. Permutation results also show that the raw estimates of variance explained are inflated, possibly due to population structure and upward bias in estimates. However, high totals for variance explained from some models do not invalidate the enrichment results. As enrichment is computed relative to the total variance explained by the model, enrichment results hold even if the totals are higher than they should be. Lower total estimates from a larger sample using imputed data are consistent with observations about the effect of sample size on estimates of total variance explained. The permutation and other

robustness results suggest that partitioning by functional class would be a poor approach to obtaining an estimate of the narrow-sense heritability, if that is the goal of an analysis. However, the primary interest is in the relative contributions from different classes of variation and for this purpose the permutation results confirm that the observed enrichment estimates for islet-enhancer variants and depletion for non-functional variants are unlikely to occur by chance.

I conducted many parallel analyses partitioning into multiple functional classes to assess the robustness of the results. I investigated the effects of varying assumptions about effect-size distributions, setting different minimum allele-frequency thresholds, excluding regions around known T2D GWAS loci, LD-pruning variants and using genotype dosages instead of hard genotype calls. Enrichment results are highly concordant across these different settings even though the underlying raw variance-explained estimates change. Thus, I conclude that the islet-enhancer enrichment and non-functional depletion results are robust to varying experimental assumptions and settings.

Throughout my analyses various technical effects were encountered from the LMM methods used that can affect estimates of heritability. I observed that liability-scale heritability estimates (preferred because they are more interpretable and can be compared across different datasets and traits) are strongly affected by the prevalence assumed for the binary trait. Misspecifying the prevalence by even a small amount can lead to substantial changes in inference on heritability. This is a general and important point for interpreting any results from variance component analyses on binary traits. LD-pruning does not seem to be a necessary or useful approach when estimating heritability from whole-genome sequence data, and it appears that high correlation between relatedness values (kinship matrices) can inflate totals for variance explained by variance-partitioning models. As with all association analyses in genetics, population structure must be carefully considered and accounted for when estimating and partitioning heritability. Colocalisation of unobserved functional annotations as recently described by Trynka et al. (2015) may be one confounding factor not addressed in this analysis that could be worth exploring in detail in future work.

In summary, these results show that variation in islet enhancers explains a large proportion of variance in susceptibility to T2D, with significant enrichment for variants in enhancer regions active in pancreatic islet cells, and significant depletion for variants with no functional annotation. Future work could include applying two or three MAF classes to some of the analyses partitioning into functional classes, for example having rare, low-frequency and common components for each class among islet-enhancers, other-enhancers and other variants. However, we would require a substantially larger sample to obtain estimates with sufficient precision to characterise definitively enrichment in T2D heritability from islet-enhancer and other functional variants across the MAF spectrum. Conducting

these or similar analyses on a larger dataset could potentially clarify the importance (or otherwise) of other functional classes, some of which had fold-enrichment that was high but non-significant in these analyses.

Overall, the variance partitioning methods used here appear to translate well from the chip-genotype data context for which they were originally conceived, and have been successfully used, to the whole-genome sequence data from the GoT2D study. The sorts of analyses undertaken here should be useful for other complex traits as well. As larger cohorts with whole-genome sequence data become available, we should be able to obtain more precise estimates for enrichment, and will be able to compare MAF and functional class partitioning results across traits. Enrichment of functional classes could be used to inform priors for variant prioritisation in association studies. Developing such approaches was beyond the scope of the work here, but represents a potentially useful application of this type of analysis. Our variance partitioning analyses demonstrate a role for genomic variation in pancreatic islet enhancer regions in explaining variance in risk for T2D. These results also offer further tantalising, if not entirely conclusive, possibilities for the genetic architecture of T2D. Further application of these approaches should contribute to advancing our understanding of the genetic architecture of type 2 diabetes and other complex traits.

Chapter 5

Introducing SCATER: Software tools for the pre-processing, quality control and visualisation of single-cell RNA-sequencing data

5.1 Introduction and background

In the preceding chapters I have discussed the identification and annotation of genomic variation and some of the relationships between genomic variation and human disease. Studying variation in genomic DNA is powerful, especially when utilising knowledge of genetics to guide analyses. A person's germline genome, encoded in their DNA, is generally stable and largely consistent across individual cells and over time, although somatic variation is implicated in ageing, neurodegeneration and, of course, cancer (Kennedy et al., 2012). In addition, assaying genomic variation is becoming ever cheaper with SNP chips and, increasingly, whole-exome and whole-genome sequencing. The combination of affordable assays and the idea that a person's genome need only be assayed once (and conveniently so from a blood or saliva sample) has driven population-scale genome sequencing studies and the push to establish whole-genome and whole-exome sequencing as routine aspects of clinical medicine.

To understand how and why certain genetic variants are associated with disease we must go beyond changes to DNA sequence and explore function. Outside the disease context, better understanding genomic function would greatly benefit basic science as well. For example, it would aid research on cellular differentiation and developmental biology. There are many critically important functional mechanisms that operate in the genome, including gene and protein expression, epigenetic processes like DNA methylation and histone modification, non-coding RNA and many more. Here the focus will be on the transcriptome and, specifically, the measurement of cellular mRNA levels. Even though large-scale investigations have been undertaken to characterise genomic function through

epigenomic, transcriptomic and proteomic variation, until recently such studies have been limited to interrogating variation between individuals (or cell lines) in bulk samples of cells (for example ENCODE Project Consortium et al., 2012; Lonsdale et al., 2013; Lappalainen et al., 2013; Maarten Altelaar et al., 2012). The next frontier for the study of genomic variation in human health and disease is to investigate intra-individual variation by applying genome-scale assays at the level of individual cells.

In a sense, single-cell analysis is nothing new in biology. Since the identification of the cell as the fundamental unit in biology, biologists have been studying the intricacies of individual cells. A technological breakthrough, the invention of the microscope, allowed Hooke and van Leeuwenhoek to identify cells (Gest, 2004), and the microscope has become a fundamental instrument for cellular and molecular biology in the centuries since. The close study of the behaviour of cells has underpinned major advances in molecular biology and medical science. Further technological advancements are now driving changes in the world of single-cell biology that could prove as transformative to the field as the microscope. Astonishing technological breakthroughs in recent years in nucleic acid chemistries and cell handling (for example, microfluidics) are enabling high-throughput genomic assays on individual cells. Today, we can perform whole-genome, whole-transcriptome, epigenomic and proteomic studies on hundreds to thousands of individual cells, with measurements becoming more precise at the same time as assays are scaled up to process tens of thousands of cells, or more, in parallel.

Single-cell genomics enables the exploration of new data dimensions. For example, bulk-tissue RNA sequencing (RNA-seq) experiments (Mortazavi et al., 2008; Marioni et al., 2008) have been highly successful for investigating the transcriptome, but with single-cell RNA-seq we can go further and begin to understand the bimodal, or even multimodal, distribution of gene expression levels across cells (Hebenstreit et al., 2011; Shalek et al., 2013). With single-cell genomics we can begin to understand heterogeneity as never before (Figure 5.1A). Rare cell types can be interrogated, known and unknown cell sub-populations uncovered, effects of the cell cycle studied, high-resolution data on catalytic processes obtained, insights gained into the dynamics of developmental processes such as stem cell differentiation and much more (Figure 5.1B). Within the last five years, single-cell "omics" approaches have blossomed, enabling unbiased identification of cell "type", broadly conceived.

Most of the major bulk-tissue "omics" technologies now have their single-cell counterpart. We can sequence DNA in individual cells, and this has already been used for preimplantation genetic diagnosis (Van der Aa et al., 2013). Many epigenomic assays can now be conducted at single-cell level, such as bisulfite sequencing to measure methylation (Guo et al., 2013; Smallwood et al., 2014; Farlik et al., 2015) and methods for sequencing regions of accessible chromatin (ATAC-seq; Cusanovich et al., 2015; Buenrostro et al.,

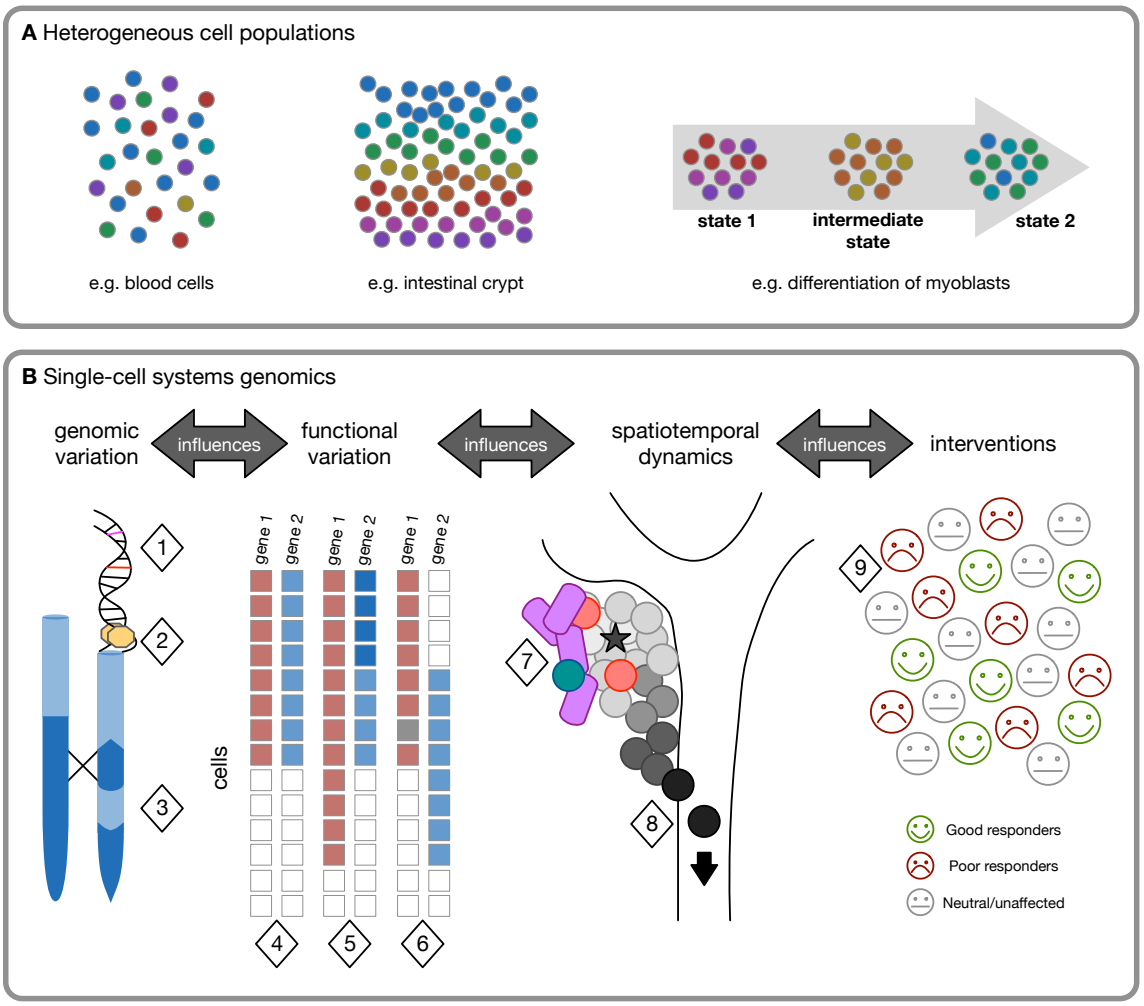


Figure 5.1: *Single-cell genomics: a tool for the post-GWAS era*. Single-cell genomics can be used to study intra-individual variability through inter-cellular heterogeneity. (A) Heterogeneous cell populations can be explored at the resolution of individual cells with single-cell genomics assays. Single-cell RNA-seq, in particular, can be used to interrogate inter-cellular transcriptomic heterogeneity in mixed cell populations. (B) Modelling intra- and inter-individual heterogeneity requires four levels of information, the first being high-resolution estimates of (1) genetic, (2) epigenetic and (3) structural variation both in germline and, where relevant, cancer cells. This is complemented by integration with high-resolution estimates of functional variation, such as the example gene-expression heatmaps in samples (4–6). Cells in sample 4 form two clusters, based on low-level gene expression (shown as red and blue squares) and undetectable or no expression (shown as white squares). The genes in sample 5 show different patterns of altered expression. While there is an increase in the proportion of cells expressing gene 1 at a low level, gene 2 suggests a new sub-population of cells in which it is highly expressed (shown as dark blue squares). Cells in sample 6 cluster into the same four groups as the cells in sample 5. However, this is due to differential co-expression rather than altered expression level or expression prevalence. Bulk sequencing would not be able to differentiate sample 4 from sample 6. The grey square in sample 6 indicates a "dropout event" where a gene with substantial true expression in a cell appears not to be expressed because the gene's transcripts were "missed" for technical reasons. Spatiotemporal information during treatment is required to understand the influence of genomic variation, intervention and cell population dynamics on emergent behaviours such as differentiation processes or drug resistance. Cell microenvironment (such as cells in colour in 7) is thought to play a major role in most cancers, as well as in healthy tissues. Cell microenvironment is also thought to affect the plasticity of cell phenotype over time to allow distant metastases (8) and processes of cell differentiation. Translating models of intra-individual heterogeneous processes into models of heterogeneous individual response, as shown in (9) by the cartoons of good and poor responders to treatment amongst a population of neutral or unaffected individuals, is a goal of so-called "precision" and/or "stratified" medicine. Adapted from Figure 2 of Wills & Mead (2015).

2015a,b). Single-cell Hi-C (Nagano et al., 2013) and single-cell ChIP-seq are also available, although utility of the latter is currently limited by low sensitivity of antibody-based work on single cells. Single-cell proteomics is also developing rapidly with technologies such as inductively coupled plasma mass spectrometry (Miyashita et al., 2014), mass cytometry (Bendall et al., 2011, 2014) and proteomic chips (Shi et al., 2012) enabling experiments on thousands of individual cells. Multiple single-cell genomic assays can even be applied in parallel to the same cell, such as with "G&T-seq" (Macaulay et al., 2015) and other methods to sequence the genome and transcriptome of the same cell (Klein et al., 2002; Dey et al., 2015; Li et al., 2015). Recently, Wilson et al. (2015) combined various single-cell functional assays with single-cell gene expression data to design a method to sort haematopoietic stem cells from other blood cells. Live cell imaging (Hoppe et al., 2014) will enable the study of individual cell behaviour over time before the application of genomic assays to those individual cells. Approaches that collect multiple data modalities from a single cell look set to gain momentum as single-cell genomics technologies develop further.

Single-cell transcriptomics has progressed a long way already. Single-cell RNA sequencing (scRNA-seq), discussed in detail below, can disentangle cell progressions and be used for marker discovery, from which cell-specific markers can be used to purify cell populations. Studying transcriptomic heterogeneity in space and time is possible with scRNA-seq, and spatial transcriptomics is developing quickly. Building on single-molecule fluoresence in-situ hybridisation (FISH; Taniguchi et al., 2010), methods such as MERFISH (multiplexed error-robust FISH; Chen et al., 2015b), FISSEQ (Lee et al., 2015), Tomo-seq (Junker et al., 2014), Seurat (Satija et al., 2015) and others (Achim et al., 2015) are making high-throughput spatial mapping possible, assaying expression levels of thousands of genes in parallel (Crosetto et al., 2015).

With the marriage of single-cell biology and high-throughput genomic assays, we stand on the threshold of the single-cell genomics era. Single-cell genomics enables the exploration of both genetic and non-genetic heterogeneity in individual cells (Spudich & Koshland, 1976; Kærn et al., 2005). Like the microscope centuries before it, the potential for single-cell genomics is immense, and it could revolutionise whole-organism science (Shapiro et al., 2013). Of the new single-cell genomics methods, scRNA-seq is arguably the most mature and is the focus of the work described here. However, new analytical tools are needed to take full advantage of this new data type. Pre-processing, quality control and data normalisation are crucial to account for its novel characteristics and biases. Numerous statistical methods and several software tools have been published for scRNA-seq data (reviewed in detail below), but there is currently a gap in the scRNA-seq workflow between raw read data and clean, tidy gene- or transcript-level expression data ready for downstream analysis. In this chapter I present a new R software package, SCATER, which provides tools for the pre-processing, quality control, normalisation and visualisation of

single-cell RNA-seq data. Such fundamental tools are not necessarily sexy, but they are vitally important to the successful application of single-cell RNA-sequencing to myriad questions about inter-cellular, intra-individual transcriptomic variation.

5.1.1 Chapter outline

This chapter introduces a new R software package called SCATER: "single-cell analysis tools for expression in R". The chapter describes, demonstrates and discusses the functionality of SCATER and how it interacts with other tools for scRNA-seq analysis. Development of `scater` was supervised by Quin Wills, who made many suggestions for functionality to include. I made all decisions about software design and implementation, and wrote all of the code, with the exception of two functions contributed by Kieran Campbell. I conducted all of the analyses presented in the chapter, using data provided by collaborators as described in Section 5.2.

The chapter comprises three main sections, which correspond to the major themes of SCATER:

1. Data pre-processing and quality control (Section 5.3);
2. Data visualisation (Section 5.4);
3. Software and data integration (Section 5.5).

Describing an R package is not straightforward, so I demonstrate the use of SCATER on real data before discussing further features and aspects of the package in greater detail. Datasets used for the demonstration of SCATER are described in Section 5.2.

Section 5.3 walks through the use of the SCATER package for pre-processing, quality control and normalisation in the context of real data analysis. This section acts like a "vignette" of the software, demonstrating the capabilities of SCATER with live code and output. I introduce the tools used in SCATER for the quantification of transcript abundance from raw RNA-seq reads. After expression quantities are obtained I discuss strategies for: (1) accessing annotation information on genomic features; (2) computing quality-control metrics in an (almost) automated way; (3) undertaking quality control of problematic and lowly-expressed genes or features; (4) conducting quality control of cells; (4) applying simple normalisation methods and (5) identifying covariates that likely need to be accounted for in further normalisation steps or downstream statistical modeling. Throughout, I will demonstrate how SCATER's plotting capabilities can be used to explore the dataset and guide decisions on pre-processing and quality control.

Section 5.4 provides more detail on SCATER's data visualisation capabilities. Specifically, I introduce cumulative expression plots as an approach to gaining an overview of

expression across all cells in a dataset. Cumulative expression plots are superior to boxplots for this purpose in the scRNA-seq context. I also discuss in greater detail the use of reduced-dimension representations of cells to explore cell-type structure, including subpopulations of cells. Further, I demonstrate the use of gene sets defined using *a priori* knowledge with projection plots to investigate certain effects in the data, such as those due to the cell cycle. At the end of the section, useful gene-wise expression plots and density plots of expression values before and after normalisation are shown.

Section 5.5 discusses "software and data integration", broadly conceived. Specifically, I outline SCATER's integration with other software, providing further details on the implementation of SCATER and how it builds on the R and Bioconductor ecosystems of statistical and bioinformatic tools. The SCESet class and its advantages are described in detail. Various options for tools that can integrate with SCATER for data normalisation, differential expression analysis, heterogeneous gene expression analyses, and cell and gene clustering are suggested. I also provide in-depth discussion of rapid transcript quantification with KALLISTO and SCATER, and the automated quality control output possible with SCATER. The section ends with a discussion of integration of other data modalities with SCATER, an important aspect of scRNA-seq analysis particularly and, looking ahead, single-cell genomics generally.

To set the scene for the SCATER package, the remainder of this section comprises two main subsections. First, Section 5.1.2 introduces the opportunities and challenges of scRNA-seq. I review existing technologies and experimental protocols, characteristics of scRNA-seq data, approaches to normalisation and methods for exploring expression heterogeneity. I then discuss the importance of quality control, particularly for scRNA-seq data, and make the case for producing a new software package for this task. Second, Section 5.1.3 provides an overview of the SCATER package, introducing its key features, summarising the SCATER workflow and discussing general recommendations for pre-processing and quality control.

The chapter concludes with some general discussion and plans for future work on the SCATER package.

5.1.2 Single-cell RNA-seq: data, methods, opportunities and challenges

5.1.2.1 Single-cell RNA-seq technologies and data generation

To sequence mRNA from individual cells, one must overcome two challenges not present when sequencing from populations of cells in bulk RNA-seq:

1. Capturing single cells, and

2. Amplifying minute amounts of mRNA from a cell.

These two steps form the basic strategy of any scRNA-seq experiment (Kolodziejczyk et al., 2015). A single cell is captured and lysed, then reverse transcription is performed to select for mRNA (with poly[T] priming) and to obtain complementary DNA (cDNA). The very small amounts of cDNA are amplified by either polymerase chain reaction (PCR) or in vitro transcription (IVT), and then used for sequencing library preparation. After sequencing, expression levels of genomic features (typically transcripts or genes, or both) are quantified, and then analysis of the data can proceed (Figure 5.2B).

Manual and automatic methods are available for capturing single cells (Figure 5.2B). Manual methods such as micromanipulation or micropipetting (Tang et al., 2009; Grindberg et al., 2013; Xue et al., 2013; Yan et al., 2013) or laser capture microdissection (Keays et al., 2005; Frumkin et al., 2008) are time consuming and low throughput, but enable cells to be derived directly from tissues and ensure that a single cell is captured at each isolation attempt. Automatic methods for cell capture are much higher throughput. Fluorescence-activated cell sorting (FACS), pysical capture microfluidic and reverse emulsion microfluidic (microdroplet) methods require cells to be dissociated and suspended in a buffer, a step that can be challenging as enzymatic treatment may have an effect on cell viability and thus a cell's transcriptional profile (Kolodziejczyk et al., 2015). FACS enables very fast sorting of hundreds of cells into microtitre plates and can be used to enrich for particular cells of interest (Hayashi et al., 2010; Wilson et al., 2015). Using the Fluidigm C1 microfluidic robot platform (Streets et al., 2014; Fluidigm Corporation, 2015), up to 96 cells can be analysed per chip after capture with integrated fluidic circuits. Captured cells can be inspected under a microscope and the nanolitre reaction volumes save on reagent costs (Shalek et al., 2014; Trapnell et al., 2014; Treutlein et al., 2014). However, the method requires at least 1000 cells as input, only works for cells relatively homogeneous in size and can have low capture efficiency for sticky or non-spherical cells (Kolodziejczyk et al., 2015). Microdroplet-based methods are the highest-throughput technology currently available with the ability to capture thousands to tens of thousands of cells in a single experiment (Brouzes et al., 2009; Mazutis et al., 2013; Fan et al., 2015; Macosko et al., 2015; Klein et al., 2015). Their highly parallel nanolitre and picolitre reaction volumes further reduce costs for large-scale scRNA-seq experiments.

After cells have been captured, clever chemistry is required to produce sequencing libraries from the very small amounts of mRNA present in individual cells. The first scRNA-seq protocol was developed by Tang et al. in 2009 (see also Tang et al., 2010). The STRT-seq protocol introduced the approach of sequencing the 5-prime end of mRNA transcripts and increased throughput (Islam et al., 2011, 2012). In 2012, Smart-seq enabled full-length transcript pre-amplification (Ramsköld et al., 2012), and CEL-seq developed multiplexed linear amplification with IVT to avoid PCR bias (Hashimshony et al., 2012). The following year, Smart-seq2 reduced cost and increased sensitivity for full-length transcriptome profiling

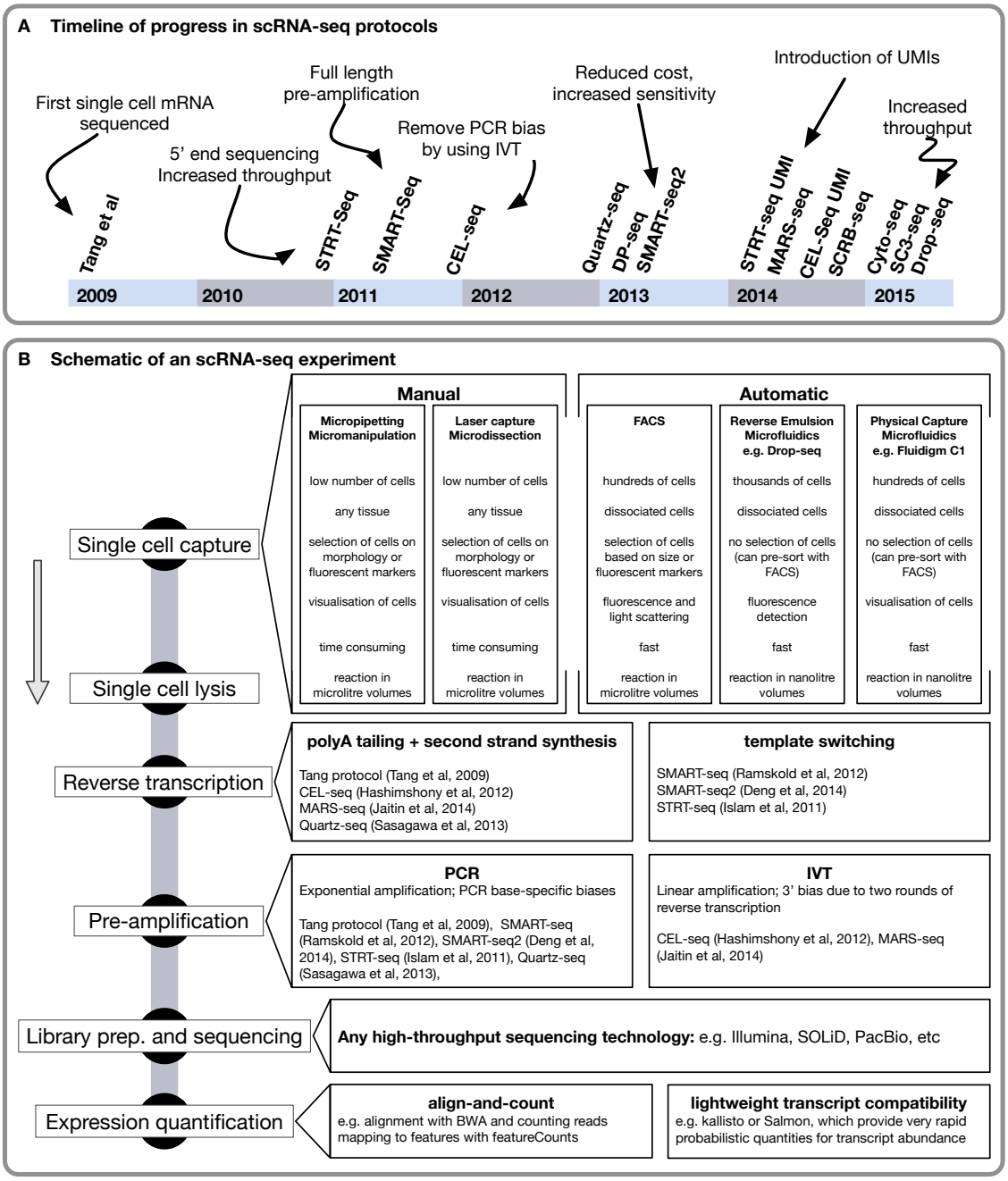


Figure 5.2: Single-cell RNA-seq technologies. (A) Timeline of the development of protocols for scRNA-seq. (B) Schematic diagram of the basic steps of an scRNA-seq experiment from single cell capture to expression quantification. Reverse emulsion microfluidics are also known as "microdroplet" methods. The most common physical capture microfluidics platform is Fluidigm's C1 chip, which uses programmable valves to capture single cells. Abbreviations: fluorescence activated cell sorting (FACS), polymerase chain reaction (PCR), in vitro transcription (IVT), unique molecular identifier (UMI). This figure was developed from a similar figure by Kolodziejczyk et al. (2015).

(Picelli et al., 2013, 2014), and further protocols Quartz-seq (Sasagawa et al., 2013) and DP-seq (Bhargava et al., 2013) appeared. The MARS-seq (Jaitin et al., 2014) and SCRB-seq (Soumillon et al., 2014) protocols arrived in 2014. Further innovation was achieved in the form of unique molecule identifiers (UMIs), which barcode individual mRNA molecules within a cell during reverse transcription (Hug & Schuler, 2003; Fu et al., 2011; Shiroguchi et al., 2012). Use of UMIs can improve the quantification of expression levels in single cells by counting RNA molecules directly. The UMIs technique has been adapted for the STRT-seq (Kivioja et al., 2012; Islam et al., 2014), CEL-seq (Grün et al., 2014) and MARS-seq (Jaitin et al., 2014) methods. Recently, the new protocols Cyto-seq (Fan et al., 2015), SC3-seq (Nakamura et al., 2015) and Drop-seq (Macosko et al., 2015) have increased throughput for scRNA-seq experiments even further (Figure 5.2A). The sequencing libraries produced with any of these protocols are sequenced (usually with multiplexing) on any high-throughput sequencing platform to produce raw read data for each cell.

Once raw read data, either single-end or paired-end reads, have been obtained, the expression level of genomic features must be quantified. Essentially, the problem is to identify from which genomic feature each read originated. The term "read" refers to both single-end or paired-end reads, as a single single-end read or a pair of paired-end reads each represents an mRNA fragment that was sequenced. Expression features can be any genomic regions containing a sequence that can normally appear in an RNA-seq experiment, so are typically transcripts, genes, transcript isoforms or exons. Approaches to expression quantification from raw reads are exactly the same for scRNA-seq as they are for bulk RNA-seq. Here, I focus on using reference genomes/transcriptomes. De novo assembly and analysis of RNA-seq data is possible, but I will not discuss de novo approaches here as, to this point, the application of scRNA-seq has been limited to humans, cell lines and model organisms.

Let us use the random variable Y_i to denote the observed expression from a feature of interest i. There are two general approaches to obtaining Y_i, and two natural kinds of raw expression values:

1. **"align-and-count"**: in this approach, sequenced reads are aligned against a reference genome or transcriptome. There are many alignment tools available that align reads in ways that account for transcript isoforms and (possibly novel) exon skipping in transcripts (for example Trapnell et al., 2009; Dobin et al., 2013; Liao et al., 2013). A set of genomic features of interest is defined (typically transcripts, genes or exons), reads are assigned to features and a "count" of the number of reads assigned to each feature is returned as the expression value, Y_i. Assignment can either be done in a "naive" way (simple counting, as with Anders et al., 2015; Liao et al., 2014) or a "probabilistic" way in which $E(Y_i)$ is used, estimated using the EM algorithm with a

method like eXPRESS (Roberts & Pachter, 2013), RSEM (Li & Dewey, 2011), SAILFISH (Patro et al., 2014) or CUFFLINKS (Trapnell et al., 2010), among other tools. With the align-and-count approach, the raw expression data are counts (or expected counts) for each feature.

2. **"transcript compatibility"**: this approach builds on the idea of "lightweight alignment" introduced by Patro et al. (2014) with SAILFISH. Read alignment expends unnecessary computational effort finding the precise location in a transcript from which a read originates, which often is not possible to do unambiguously. Instead, the "transcript compatibility" approach taken by very recent methods such as KALLISTO (Bray et al., 2015) and SALMON (Patro et al., 2015) aims to determine the set of transcripts with which each sequenced read is *compatible*. An expectation maximisation approach is then used to obtain estimates of transcript expression levels. This approach does not requre traditional read alignment and can be done extremely quickly, with apparently little or no decrease in accuracy of expression quantification (Bray et al., 2015; Patro et al., 2015). The transcript compatibility approach naturally gives transcripts-per-million (TPM) values as raw expression values, although expected counts are typically also returned.

The raw expression values then form the basis of all downstream analyses. Until mid-2015, almost all RNA-seq expression quantification was done with some version of the align-and-count approach, but I predict that the sheer speed of the KALLISTO and SALMON tools will lead to their rapid adoption. Removing the need for read alignments (although lightweight or pseudo-alignments can be produced) alters standard RNA-seq read processing pipelines. New workflows with these tools will be much faster and need significantly fewer computational resources, but adjustment will be required for established quality control procedures. I discuss the issues of quality control in detail below.

5.1.2.2 Characteristics and novelties of single-cell RNA-seq data

Single-cell RNA-seq data has characteristics that distinguish it from bulk-tissue RNA-seq data. This fact has implications for the analysis of scRNA-seq data, as the well-developed statistical methods for bulk RNA-seq will not necessarily be appropriate for scRNA-seq.

RNA-seq has become the dominant platform for so-called "bulk-tissue" gene expression studies. In bulk experiments, we combine RNA across a pool of thousands to millions of cells. Pooling cells increases the amount of mRNA available and thus reduces the extent of amplification of the cDNA reverse-transcribed from RNA that is required to generate sufficient material for sequencing. One effect of pooling across cells, however, is that we are only able to measure the "average" expression levels of genes. The information on the

gene expression behaviour of individual cells is lost. Even so, studying the average expression of genes across cells is of much interest, a fact to which the near-ubiquitous use of bulk-tissue RNA-seq attests. Single-cell RNA-seq offers a higher resolution look at expression heterogeneity, not limited to studying mean expression across cells (Figure 5.1B).

Three prominent characteristics of scRNA-seq data make it substantially different to analyse compared with bulk RNA-seq data:

1. Zero-inflation;
2. Overdispersion;
3. Highly prone to certain biases.

Zero-inflation and overdispersion in scRNA-seq data are closely linked and arise for both biological and technical reasons, also in connection to certain experimental biases. Depending on the units used for expression in scRNA-seq data (discussed below), the data used for analysis may be treated either as count data, or something closer to log-normal. Whatever units are used, scRNA-seq data are both "overdispersed", that is, more variable than allowed for by many standard statistical models, and "zero-inflated", meaning that there are more zero values (or values at a minimum threshold, if transformed) than accounted for by standard models for count or log-normal data.

For example, a Poisson model would be a natural starting point when analysing count data, but, as has been very clearly established for bulk RNA-seq data, there is much more variability in RNA-seq data (especially when biological replicates are present) than can be accounted for by the Poisson model. This observation drove the widespread adoption of Negative Binomial models for count data in bulk RNA-seq analysis (Robinson & Smyth, 2007, 2008; Robinson et al., 2010; Anders & Huber, 2010; Hardcastle & Kelly, 2010; McCarthy et al., 2012; Anders et al., 2013, among others). Single-cell RNA-seq data is even more overdispersed, due to the extra sources of technical and biological variability discussed below. The same sources of variability drive an excess of zero-values in the data.

As discussed above in relation to scRNA-seq experimental protocols, there are substantial technical challenges when studying gene expression in individual cells. Consideration of these challenges informs approaches to analysing RNA-seq data from single-cell studies. Islam et al. (2014) identify the two main challenges in single-cell RNA-seq as:

1. The efficiency of cDNA synthesis, which sets the limits of detection, and
2. Amplification bias, which reduces quantitative accuracy.

With tiny amounts of RNA available in individual mammalian cells ($\sim$10 picograms), currently-available protocols for scRNA-seq have limits of detection of between five and

ten mRNA molecules (Islam et al., 2011; Ramsköld et al., 2012; Hashimshony et al., 2012). This corresponds to a capture efficiency of approximately 10%.

All current methods use amplification, either by PCR or by *in vitro* transcription, before sequencing reverse-transcribed cDNA (Figure 5.2). Amplification biases, discussed above, and their effects on obtained transcript abundance measurements vary depending on the specific protocol used. Allelic dropout is also a key technical effect in scRNA-seq data. Dropout refers to the false quantification of a gene or transcript as "unexpressed" due to the corresponding transcript being "missed" during the reverse-transcription step (Stegle et al., 2015). If a transcript is not reverse-transcribed then it cannot be detected during sequencing. Furthermore, cell lysis does not always work perfectly, and it is also possible to sequence "background RNA" from a chip or plate, which can be wrongly assigned to individual cells. Inevitable variability introduced in library preparation and sequencing adds further to the overall technical variability. Protocols for scRNA-seq are still relatively new. As such, substantial variability and unexpected problems often arise even when they are "correctly" applied.

There are three dominant sources of biological variation in scRNA-seq data: (1) heterogeneity arising from sub-populations, (2) heterogeneity arising from transcription kinematics and (3) heterogeneity arising from biological process like cell cycle (Kolodziejczyk et al., 2015). If a sample of cells contains multiple sub-populations of cells (as, for example, a sample of blood cells does), then substantially different transcriptomic profiles for the different cell types (large sets of genes with high expression in one sub-population and little or no expression in other sub-populations) will lead to zero-inflation and overdispersion in measured expression levels across the sample as a whole.

Transcription kinematics, reflecting dynamic transcript expression and its regulation, also generates zero-inflation and overdispersion in scRNA-seq data. For example, "bursty" expression is a feature of single-cell transcriptomic behaviour, thought to occur due to promoters flipping between "on" and "off" states (Dar et al., 2012). Gene expression is assayed with scRNA-seq as a "snapshot" at a specific point in time, so the stochastic nature of gene expression in single-cells (Elowitz et al., 2002; Raj & van Oudenaarden, 2008, 2009) can lead to measured expression values of zero (or near-zero) for a gene in a set of cells, even if the gene is otherwise "highly expressed" in cells of that type.

Biological processes, often unrelated to the biological sources of variability of interest, add further levels of heterogeneity between cells, and produce zero-inflation and overdispersion in expression measurements. For example, a cell's RNA content and transcriptional profile changes dramatically throughout the cell cycle, regardless of what else might be happening inside or outside the cell. The cell cycle and other such processes can thus have large effects on assayed expression values and will often need to be accounted for in statistical analyses (see Buettner et al., 2015, for a detailed discussion).

The richness of single-cell data allows us to ask, and possibly answer, questions of biological interest that were not possible with bulk-tissue RNA-seq data. However, the specific characteristics of scRNA-seq data need to be taken into account in statistical models and analysis.

5.1.2.3 Normalisation of single-cell RNA-seq data

In almost all scRNA-seq experiments, many sources of unwanted variation should be accounted for before proceeding with more sophisticated analyses (Bhargava et al., 2014). Data normalisation is therefore necessary to make meaningful comparisons between cells of different known and unknown types. Library size and composition will inevitably vary between cells, and this needs to be taken into account. Often, correct choice of expression units for scRNA-seq data will suffice for library size differences and in some settings size-factor normalisation can ameliorate apparent expression differences driven by differing library composition. Feature controls (such as spike-in gene controls) can help correct technical artifacts appearing in scRNA-seq data, but there are known issues with their use. Important experimental variables and latent factors (if used) can be regressed out, so that normalised data have these effects removed. Standard quantile normalisation, as commonly performed for microarray analysis, for example, is not appropriate for scRNA-seq data (discussed further below). However, single cell-specific quantile normalisation and rank-based normalisation methods hold potential (Scialdone et al., 2015).

Appropriate units for single-cell RNA-seq expression data. There are several distinct units for RNA-seq expression: raw counts, normalised counts, counts-per-million (CPM), TPM, and fragments per kilobase per million mapped (FPKM). The units CPM, TPM, and FPKM all seek to normalise the data (in a simple way) to account for differences in sequencing depth between libraries, an unavoidable feature of RNA-seq data generally.

For bulk RNA-seq data, TPM are the most appropriate units, although FPKM has been widely used, and raw counts have been successfully used for differential expression testing (Robinson & Smyth, 2007, 2008; Robinson et al., 2010; Anders & Huber, 2010; Hardcastle & Kelly, 2010; McCarthy et al., 2012; Anders et al., 2013, among others). In packages for differential expression using count data, CPM are often used for reporting (mean) expression levels (introduced in Robinson et al., 2010). However, there are issues for certain types of inference when using counts as the unit of expression that can be avoided by using TPM instead (Trapnell et al., 2013).

The most appropriate units for scRNA-seq data depend on the experimental protocols used. For example, TPM is most appropriate for protocols that sequence full-length transcripts, although care may be required for highly biased protocols, such as those with pronounced 3-prime bias (Ramsköld et al., 2012). For other protocols, for example when

using unique molecular identifiers (UMIs) (Kivioja et al., 2012; Islam et al., 2014; Grün et al., 2014; Jaitin et al., 2014) or protocols that only sequence the 5-prime end of transcripts (Islam et al., 2011), counts and CPM could be appropriate units, although TPM should still be appropriate in most settings. The SCATER package endeavours to support the widest range possible of single-cell RNA-seq data, so it enables computation of TPM, CPM and FPKM values. The package is designed so that scRNA-seq data expressed in any of the units discussed here can easily be stored, accessed and used with appropriate methods.

Sequencing depth and library composition. The RNA-seq expression units TPM, CPM, and FPKM account for differences in sequencing depth between libraries. There are, however, further factors to consider in the normalisation of scRNA-seq data. In particular, differences in library composition or "use of sequencing real estate" can create problems for comparisons between cells, just as they can be problematic for between-sample comparisons in bulk RNA-seq. Such problems can be ameliorated in bulk RNA-seq analysis by using "size-factor normalisation". Size-factor normalisation refers to a suite of methods that essentially compare libraries based on some robust metric and compute normalisation factors by which expression quantities are multiplied or divided so that the expression quantities are more comparable between libraries. The canonical example, described by Robinson & Oshlack (2010), involves comparing bulk RNA-seq expression levels between liver and kidney tissue. The total RNA output differs between samples from the two tissues, as there is a prominent set of genes with high expression in liver, which skews apparent differences in expression levels between kidney and liver for other genes. The different methods for size-factor normalisation aim to account for differences between libraries due to such effects.

There are three main methods for computing size-factors for RNA-seq normalisation:

- TMM: "trimmed mean of M-values" from Robinson & Oshlack (2010);
- RLE: "relative log expression" from Anders & Huber (2010);
- upperquartile: "upper quartile of expression values" from Bullard et al. (2010).

They generally produce similar size-factors, so results tend not to vary greatly depending on which method is used (data not shown). I prefer to use the TMM method due to its explicit expectation of, on average, no bias in fold changes between libraries.

For scRNA-seq data, normalisation should generally use feature controls, if they are available (discussed in the next section). If feature controls are not available, then library size and size-factor normalisation can be done using all genes (as is usually the case for bulk RNA-seq) or, if desired, a set of housekeeping genes (Treutlein et al., 2014). The difficulty with using housekeeping genes for normalisation is a problem that has dogged such approaches since the era of microarrays, namely the challenge of defining a large enough,

but specific, set of genes that should display constant expression across cells and conditions. This is rarely achievable in practice.

Using extrinsic spike-in feature controls. A strongly recommended approach to aid quality control and normalisation for all scRNA-seq experiments is to use extrinsic "spike-in" molecules as feature controls. Either a whole-transcriptome spike taken from a different species from the cells of interest or a specially designed set of artificial spike-in moleclues is added to each cell's lysate (Stegle et al., 2015). The most widely used artificial spike-in mix is that from the External RNA Control Consortium (ERCC; Jiang et al., 2011), which comprises a set of 92 synthetic spikes based on bacterial sequences. The spike-in mix should be added in constant volumes to each cell extract. Some care needs to be taken, as variation in cell size or cell cycle could lead to substantial differences in RNA content, in which case cell extracts may need to be normalised themselves before spike-in molecules are added, or the spike-in volume may need to be varied. An equal number of molecules of each spike-in RNA species should be present in each single-cell library, a fact that can be exploited to normalise gene expression levels and to estimate technical sources of variation (Stegle et al., 2015).

With extrinsic spike-in controls, it is possible accurately to estimate relative differences in the total RNA content between cells. The assumption of a constant amount of spike-in material across cells makes it easy to compute the ratio of assayed transcript quantities for the transcriptome of interest to assayed "transcript" quantitites for the spike-ins, allowing differences in the amount of RNA between cells to be inferred (Stegle et al., 2015).

Extrinsic spike-ins are unequivocally useful for scRNA-seq experiments, but there are some caveats regarding their use. The 92-spike ERCC set of spike-ins, the most common set, are 500 to 2,000 nucleotides in length, which is shorter than the average human mRNA (approximately 2,100 nucleotides including UTRs; Krebs et al., 2014). The issue is that many scRNA-seq protocols have an inherent 5-prime-to-3-prime length bias, where the 3-prime bias is more pronounced for longer transcripts (Ramsköld et al., 2012). Thus, a conversion based on the shorter ERCC spike-ins may be problematic, and, additionally, the ERCC spike-ins have comparatively short poly(A) tails and lack 5-prime caps, which may cause different reverse transcription efficiency for ERCC spike-ins relative to endogenous RNA molecules (Stegle et al., 2015).

With the many different scRNA-seq protocols available, with different characteristics and biases, in addition to complications with using ERCC spike-ins, it is challenging to devise a universally applicable normalisation strategy. For the time being, normalisation strategies will need to be developed or adapted carefully for each experiment to properly acount for variability in sequencing depth, cellular RNA content and further technical effects. Stegle et al. (2015) propose calculating two alternative size factors when spike-ins are

available: one for the spike-ins and one for the endogenous mRNA molecules. The size factor from spike-ins accounts for sequencing depth (as the spike-in molecules are present at the same quantity in all cells). The size factors from the endogenous features, in contrast, enable normalisation for relative differences in RNA content. The normalised spike-ins can be used to estimate the degree of technical variability across the whole dynamic range of expression, because they are adjusted for library size (Brennecke et al., 2013; Grün et al., 2014). The ratio of the two normalisation factors could be used to estimate the total mRNA content of each cell (Stegle et al., 2015).

Very recently, more sophisticated normalisation methods using spike-in controls have been proposed using integrated Bayesian hierarchical models (BASiCS; Vallejos et al., 2015) and a gamma regression model (GRM; Ding et al., 2015). Both are implemented in R, so they could easily be incorporated into a SCATER workflow, as normalised expression data can be manually added to and easily accessed from SCATER data objects.

Removing effects of known covariates. Sometimes in genomics it feels like with the arrival every new technology all of the previous knowledge about the importance of experimental design is lost or ignored. It took some years of bulk RNA-seq experiments before experimental design for bulk RNA-seq studies was directly addressed (Auer & Doerge, 2010). Similarly, it seems that in the general excitement at the possibilities for scRNA-seq discussion of experimental design has been largely neglected, except insofar as regards requirements for the number of cells and depth of sequencing per cell (Stegle et al., 2015). However, scRNA-seq studies, like any in genomics (or science broadly), require careful experimental design to avoid confounding and batch effects (Leek et al., 2010). Even assuming a carefully designed experiment, there will be a large number of covariates (recorded by the conscientious researcher) affecting measured expression levels.

With judicious exploratory data analysis as part of the pre-processing, quality control and normalisation of scRNA-seq data, important known covariates can be identified. Once identified they can be flagged for inclusion in downstream statistical models, or their effects regressed out of normalised expression values and the residuals used as expression values for downstream analyses. The latter approach will be preferable in many cases, as many of the most recent statistical methods for scRNA-seq data (discussed below) are not able to handle arbitrarily complex experimental designs, in contrast to many bulk RNA-seq methods, such as those introduced by McCarthy et al. (2012) in EDGER.

Removing effects of unknown latent factors. It is one thing to remove the effects of known covariates on expression values, but another to identify and possibly remove effects of *unknown* factors. Latent variable approaches have proven useful in bulk RNA-seq analysis for removing unwanted variation from expression data, with tools such as

SVASEQ (Leek & Storey, 2007; Leek, 2014), RUV (Risso et al., 2014), PEER (Stegle et al., 2012) and CELLCODE (Chikina et al., 2015) proving popular. Details differ between the various methods, but all aim to learn a low-dimensional set of hidden factors from the dataset that capture unwanted variation in the data. Latent variables can be estimated using a subset of features. Sometimes this is an explicit feature of the method, for example RUV, which uses control features to characterise and remove unwanted variation. In bulk RNA-seq data, latent factors commonly identify batch or population effects, such as the effect of sequencing time or location.

Given the many extra possible sources of variability in scRNA-seq data discussed above, extending successful latent variable methods to scRNA-seq seems a sensible course of action. Indeed it has been, with Buettner et al. (2015) demonstrating the utility of a single-cell latent variable model (SCLVM) method for identifying hidden sub-populations of cells. In principle, such a method can identify whatever latent variables are present in an scRNA-seq dataset, but identification of latent variables can be guided by prior knowledge used to define specific feature sets. In particular, sets of cell cycle genes can be used for latent factor analysis, which greatly improves the ability of such methods for ascertaining the cell cycle phase of assayed cells (Buettner et al., 2015; Scialdone et al., 2015). The identification of cell cycle phase is relevant to most scRNA-seq studies (typically as a nuisance factor to be conditioned out), so I expect such methods to prove popular and be developed further in coming years. Just as with bulk RNA-seq data, latent variables identified from scRNA-seq data can be regressed out or included in downstream modeling.

Non-parametric methods. The final class of normalisation methods to explore is the as-yet under-explored area of non-parametric normalisation approaches. Quantile normalisation proved highly successful for microarray data, but standard microarray-style quantile normalisation is not appropriate for scRNA-seq data. I applied the QUANTRO method (Hicks & Irizarry, 2015) to the datasets analysed later in this chapter, which uses a data-driven approach to determine whether or not multi-sample global quantile normalisation is appropriate for a dataset, based on comparing expression distributions between samples. The QUANTRO results (not shown) strongly confirm the expectation that global expression distributions between individual cells are too variable to apply standard quantile normalisation on scRNA-seq data, especially as large differences in expression distributions will often reflect true biology of interest. Nevertheless, non-parametric normalisation techniques do hold promise. Scialdone et al. (2015) used rank-based normalisation as part of their best method for identifying the cell cycle phase of cells. Work in Oxford also suggests that rank-based and quantile normalisation methods adapted specifically for single-cell data could prove effective (Quin Wills, personal communication). Although far

from mature at this point, non-parametric normalisation methods warrant further, serious attention.

5.1.2.4 Methods for exploring expression heterogeneity

Characterising heterogeneity between single cells is of substantial scientific interest in many areas of biology and medicine (see Saadatpour et al., 2014, for example). This applies to finding genes (transcripts) with high variability and/or differential variability between groups of cells. It also applies to the identification of novel cell types and sub-populations in samples of cells (see Zeisel et al., 2015, for one recent example among many). With suitably normalised expression data, there are many methods available for exploring expression heterogeneity. These methods include both approaches developed for the analysis of bulk RNA-seq data that may be applied to scRNA-seq data and approaches specifically designed for single-cell data.

Methods fall into two broad categories, either studying known heterogeneity in designed experiments (e.g. differential expression) or discovering unknown heterogeneity (e.g. clustering, uncovering heterogeneous expression). Methods can also be categorised as either cell-centric (focusing on making inference about cells) or gene-centric (for inference on genes or features).

Naturally, given the new avenues opened for exploration by scRNA-seq data, cell-centric methods have prodominated in the single-cell milieu thus far. Cell-centric methods can be used for:

- Identifying known and novel cell types;
- Cell clustering;
- Tracing differentiation or cell progression along some sort of pathway;

among many other possible applications. To enable these goals, substantial effort has already been placed in developing appropriate visualisation techniques for scRNA-seq data.

Principal components analysis (PCA) is a standard element of the statistician's toolkit, which produces a set of orthogonal components consisting of linear combinations of a set of variables with maximum variance subject to the orthogonality constraint (Pearson, 1901; Hotelling, 1933). PCA can be useful for visualising scRNA-seq data, but as a linear method is unlikely to be optimal for exploring the often highly non-linear relationships within scRNA-seq data. Alternatives such as probabilistic PCA (Buettner et al., 2014), multiresolution correlation analysis (MCA; Feigelman et al., 2014), diffusion maps (Haghverdi et al., 2015) and t-distributed stochastic neighbour embedding (t-SNE) from Van der Maaten & Hinton (2008) have been used for better visualisation of cell-type structure from scRNA-seq data, with t-SNE proving particularly popular (Amir et al., 2013; Bendall et al., 2014; Macosko et al., 2015). Closely related and often overlapping with visualisation techniques are

methods for clustering single cells from their expression data, such as SNN-Cliq, a shared nearest-neighbour approach (Xu & Su, 2015). Latent variable models, such as SCLVM (Buettner et al., 2015) and others discussed above in the context of normalisation, can also be useful for exploring cell population structure.

One of the largest areas of interest for RNA-seq is exploring expression relationships in cells through their progression along some sort of development or differentiation pathway. The MONOCLE package (Trapnell et al., 2014) is one of the most popular for this purpose, using dimension reduction and minimum spanning trees to produce a "pseudo-temporal ordering" of cells. The SINCELL package provides metrics to evaluate cell-to-cell similarities and a graph-building algorithm to assess cell-state hierarchies (Juliá et al., 2015). The SEURAT package (Satija et al., 2015) is primarily for spatial reconstruction of single-cell gene expression data, but includes visualisation methods applying t-SNE and can identify rare cell populations and cell sub-types (Macosko et al., 2015). Taking a slightly different tack, Efroni et al. (2015) attempt to identify cell types using a method for the quantification of cell identity that compares single-cell expression profiles to repositories of cell type-specific transcriptomes. Chen et al. (2015a) combine cell lineage trees with scRNA-seq data to uncover regulatory circuits driving cell fate decisions. The methods developed by Bendall et al. (2014) for uncovering progression trajectories in human B cells using mass cytometry should also be applicable to scRNA-seq data.

Gene-centric analyses seek to identify features with particular characteristics, typically those that differ in some way between cell groups of interest. An obvious application in the single-cell setting is to find differentially genes to use as biomarkers to help with the identification of cell types in a sample. The "classical" differential expression problem from bulk RNA-seq, finding genes that differ in average expression level across conditions, is also of interest in the single-cell world. In addition, scRNA-seq data can be used to find genes that are statistically more or less variable in a group of cells and to find differentially variable genes between conditions.

As in bulk RNA-seq studies, "known heterogeneity"' can be studied using scRNA-seq data, for example differential gene expression between groups of cells defined in the experimental design. In principle, count-based methods for differential expression (DE) testing, such as EDGER's exact test (Robinson & Smyth, 2007, 2008) and EDGER's likelihood ratio test (McCarthy et al., 2012), and approaches derived from these, such as EDGERUN (Dimont et al., 2015), DESEQ (Anders & Huber, 2010), and DESEQ2 (Love et al., 2014)), handle zeroes in the data. However, with zero-inflated data, these models can only account for large numbers of zeroes by estimating a large value for the dispersion parameter (variance). The effect of this is that statistical significance is reduced, possibly unnecessarily. The quasi-likelihood F-test in EDGER (Lund et al., 2012), should be the best of the bulk-tissue methods (Gordon Smyth, personal communication) because the variance function can account

for "bursty" expression characteristic of scRNA-seq through the quasi-likelihood variance function. I found the QLF-test to be very conservative in comparison to methods like MONOCLE and SINGLECELLASSAY (those DE methods discussed below; data not shown). Further investigation would be required to determine if this conservatism is an appropriate response to highly variable data, or if this approach is poorly powered compared with other methods.

DE methods based on transformed data have also been proposed for bulk RNA-seq data, the most successful of which is VOOM (Law et al., 2014) in the LIMMA package (Ritchie et al., 2015). Unfortunately, VOOM does not work well, because the observation-weighting approach developed for bulk RNA-seq does not work for scRNA-seq data exhibiting high levels of zero-inflation (data not shown). CELLCODE (Chikina et al., 2015) implements a latent variable model explicitly for DE analysis with heterogeneous cell mixtures. The description of the method suggests that it could be useful for scRNA-seq analyses, but it was designed for bulk transcriptomic data and it is not yet established how well it will perform with scRNA-seq data.

Methods designed with scRNA-seq data in mind account for the particular characteristics of scRNA-seq data in various ways. Most focus, sensibly enough, on accounting for zero-inflation, but typically not on other issues like biases in the data. Such unwanted sources of variability are implicitly assumed to have been dealt with in pre-processing, quality control and normalisation of the data, before these methods are applied. The MONOCLE package (Trapnell et al., 2014) uses a thresholded "Tobit" model (Tobin, 1958) for differential expression testing. SINGLECELLASSAY (Finak et al., 2014; McDavid et al., 2014) explicitly models the bimodality of gene expression and can test for differential expression frequency ("binary"), for differential mean expression ("continuous") or a combined version ("hurdle"). This style of hurdle model was proposed earlier for single-cell quantitative PCR data by Wills et al. (2013). SEURAT provides a likelihood ratio test for zero-inflated data as its default DE method, but also includes methods for a receiver operating characteristic test, a t-test and a likelihood ratio test based on Tobit-censoring models.

The SCDE package (Kharchenko et al., 2014) takes a more complicated approach using hierarchical Bayesian models that aim to characterise individual-cell noise models that can account for allelic dropout and incorporate them into a DE testing framework. With explicit modelling of error models and many characteristics of scRNA-seq data, SCDE provides an intuitively pleasing model. However, I found that the software implementation occasionally returned internally inconsistent results (significance levels for a gene differing depending on whether it was tested individually or with all other genes), and the method appeared to perform poorly when very heterogeneous libraries were present in the sample of cells (data not shown). Careful quality control should limit the impact of the latter issue, while the former, one presumes, simply reflects a bug in the software implementation.

Several of the clustering/sub-type methods described above aim to identify DE genes between cell groups inferred from the dataset, but so far these have not been done with particularly sophisticated significance-testing frameworks. Correct assessment of statistical significance can be challenging in settings where differential expression testing is carried out between groups learned from the dataset as if it were a designed experiment. In particular, the false discovery rate may not be correctly controlled. To this point, this issue seems not to have been discussed in the scRNA-seq literature.

There is insufficient space in this chapter to discuss all of these methods in detail, but to provide an impression of how some of the challenges of DE are tackled I briefly discuss the models from MONOCLE and SINGLECELLASSAY. In MONOCLE a Tobit model (Tobin, 1958) is used to model each gene's expression levels across cells. Under this model, each gene's observable (log-transformed) expression level, Y, depends on an underlying latent variable Y^*:

$$Y = \begin{cases} Y^* & \text{if } Y^* > \lambda \\ \lambda & \text{if } Y^* \leq \lambda \end{cases} \tag{5.1}$$

where λ is a detection threshold. The latent variable Y^* may depend on covariates x_i, such as experimental group, day cells were collected, processing batch and so on. The parameter λ is a user-specified value, taken to be FPKM $= 0.1$ by default in MONOCLE. Of course, one could allow $Y = 0$ for genes not expressed, but the censoring approach allows for "background" expression to be taken into account. If there is a threshold below which expression values are very noisy and not regarded as reliable, then the Tobit model allows these values to be censored at the threshold value. If conducting an analysis using pseudo-temporal ordering, MONOCLE uses a generalized additive model (GAM):

$$E(Y) = s(\Psi_t(b_x, s_i)) + \epsilon, \tag{5.2}$$

where $\Psi_t(b_x, s_i)$ denotes the assigned pseudotime of a cell and s is a cubic smoothing function with (by default) three effective degrees of freedom (Trapnell et al., 2014). If testing for differential expression between cells in different experimental groups, then the GAM simply uses the categorical labels as predictor variables, with no smoothing. In both cases, testing for differential expression is performed with an approximate chi-square likelihood ratio test.

Bimodality of gene expression is an established feature of single-cell gene expression data (Hebenstreit et al., 2011; McDavid et al., 2013; Shalek et al., 2013). Both technical effects and true biology are likely to drive bimodal expression. Despite its utility, the MONOCLE approach avoids the bimodality of single-cell expression data (through censoring) rather than modeling it explicitly. Wills et al. (2013) introduced the use of the hurdle model for single-cell quantitative PCR, and in SINGLECELLASSAY McDavid et al. (2014) adopt this semi-continuous modeling framework that explicitly accounts for expression bimodality.

Their hurdle model (Cragg, 1971; Jones, 1989) uses a probabilitistic mixture-model-based approach to separate positive expression values from background noise with gene-specific thresholds. Then, their framework models separately the frequency of expression and the continuous, positive expression values. Using the general linear model, they are able to test arbitrary contrasts and, if desired, estimate variance components or allow for random effects with mixed models.

Non-parametric methods for differential expression hold promise for scRNA-seq. As discussed above, Scialdone et al. (2015) reported that pre-processing scRNA-seq data using rank-based normalisation improved the robust capture of transcriptional cell-cycle signatures. Given the levels of inherent technical noise in currently-available scRNA-seq protocols, non-parametric, rank-based methods may prove to be appropriate for many more types of analysis in scRNA-seq data. The scRNA-seq analysis methods described so far are all parametric and have been designed to handle characteristic features of single-cell RNA-seq data. However, their current utility may be limited by the degree of variability in present scRNA-seq datasets.

Although not the focus of this chapter, I have conducted some investigations into the use of non-parametric methods for single-cell DE analysis. In particular, I explored an approach called "rank product", which was originally proposed for DE analysis for microarray data (Breitling et al., 2004; Breitling & Herzyk, 2005; Hong et al., 2006). Adapted to the scRNA-seq context, the rank product approach assumes appropriately normalised expression values Y_{gi} for gene g in cell i. We assume that cells in two experimental conditions are to be compared, and define the set $\mathcal{A}$ to contain the cells under the first condition and set $\mathcal{B}$ to contain the cells under the second condition. Fold-change is computed for all pairs of cells under the two conditions, that is, all pairs i, j such that $i \in \mathcal{A}$ and $j \in \mathcal{B}$. For each pair i, j we thus obtain a fold-change value FC_{gij} for each gene g, where $\text{FC}_{gij} = Y_{gi}/Y_{gj}$ if expression values are on the natural scale, and $\text{FC}_{gij} = Y_{gi} - Y_{gj}$ if expression values have been log2-transformed.

For each pair of cells, a fold-change rank r_{gij} is computed across all genes as $r_{gij} = \text{rank}(\text{FC}_{gij})$. The rank product for gene g is then simply the product of the ranks across all pairs, that is, $\text{RP}_g = \prod_{i \in \mathcal{A}} \prod_{j \in \mathcal{B}} r_{gij}$. The rank-product statistics are then used to rank genes in terms of differential expression. As written, these fold-change values treat higher expression in condition A to be "up-regulation". To compute rank-products treating higher expression in condition B as up-regulation, i and j simply need to be switched in the computation of the fold changes FC_{gij}.

The attractive features of rank-product approaches are that they are simply understood, preserve the correlation structure of genes in the dataset, are robust to departures from expected (parametric) expression distributions and perform well (on microarray data) when

samples are inhomogeneous. However, apart from the limitation of only being able to compare two conditions at a time, there are two major challenges for applying rank-product to scRNA-seq data:

1. Rank-product assumes equal relative measurement variance for all genes, and
2. Assigning significance to rank-product statistics is problematic for large sample sizes.

The first challenge can be tackled with judicious variance-stabilising data normalisation. The second challenge is more difficult to resolve even with methods available for calculating exact *P*-values (Eisinga et al., 2013) and a fast algorithm for approximate *P*-values (Heskes et al., 2014). The crux of the problem is that raw rank-products become extremely large in the setting of thousands of genes and hundreds of cells, and the established methods for assigning significance to rank-product statistics either require the raw rank-product statistics (for computing exact or approximate *P*-values) or time-consuming permutation methods. Rank-product methods were developed in the context of small, designed microarray experiments or gene expression meta-analyses, but scRNA-seq experiments have sample sizes (i.e. number of cells) at least an order of magnitude larger.

In my experience, rank products provide good ordering of genes for differential expression (data not shown), but alternatives to direct computation of exact or approximate *P*-values for rank-product statistics are required. One alternative is FCROS, which uses "fold-change rank ordering statistics" (which rank genes very similarly to rank products) and an approximate normal distribution to determine significance of the FCROS values (Dembélé & Kastner, 2014). Treating pairwise fold-change rankings as empirical *P*-values and applying methods for combining correlated *P*-values could be an effective approach similar to rank-products for which assigning significance would be feasible even in very large samples. This approach would require further work to develop. The SCATER package includes a fast implementation of the original rank-product approach, but conducts rank-product computations on the log scale to avoid numeric overflow that arises if using raw rank-products even on "small" scRNA-seq datasets. Though further work is required to assign significance for DE genes in a dataset, rank-product can be used effectively for "meta-analysis", combining results from different datasets or methods. The `aggregateResults` function in SCATER enables such meta-analyses, with exact and approximate rank-product *P*-values, which can be used to aggregate results to find a consensus set of DE genes based on DE rankings across different datasets or from a set of different DE methods.

Discovering genes with differential mean expression levels is often of interest, but the higher resolution of scRNA-seq data compared with bulk RNA-seq data means that new questions can be asked. One novel possibility of scRNA-seq data is the identification of differentially variable genes. Exploring heterogeneity of expression variability remains a

nascent area of scRNA-seq analysis. Brennecke et al. (2013) made an important contribution with an approach to distinguish biological variability from the high levels of technical variability in scRNA-seq data. Their approach quantifies the statistical significance of observed cell-to-cell variability in expression on a feature-by-feature basis, based on characterising the relationship between the squared coefficient of variation and average expression for spike-in features, and defining a chi-square test statistic for high expression variability for each gene.

Following the Brennecke et al. (2013) paper, Kumar et al. (2014) used a permutation approach to find genes with significantly high or low coefficiant of variation. Dueck et al. (2015) use an F-statistic as a measure of variation across single-cell samples, scaling observed variation by variation measured in spike-in control features at matched expression levels. Through its hierarchical Bayesian model, the BASiCS method (Vallejos et al., 2015) quantifies technical variabiliity using spike-in features and decomposees the total variability of expression counts into technical and biological components. BASiCS also includes a method to identify highly or lowly variable features.

One might also be interested in how feature variability is affected under different conditions in a designed experiment. Phipson & Oshlack (2014) introduced the DiffVar method for testing for differentially variable features in data from bulk transcriptomic and epigenomic assays. DiffVar is based on a robust extension of Levene's test (Levene, 1961) and uses the empirical Bayes linear model and testing framework from limma (Ritchie et al., 2015), proven to be efective in genomic contexts. Further work is required to determine how well this approach works for scRNA-seq data, but at the very least it suggests a direction worth pursuing for identifying differentially variable genes between designed conditions.

The summary of scRNA-seq methods provided here shows that a substantial amount of work has already been done to improve the analysis of scRNA-seq data. The bulk of the effort has been spent, quite reasonably, on cell-centric methods. There is still large scope for improvement in methods to detect differentially expressed genes and, particularly, differentially variable genes. The best bulk RNA-seq methods for DE analysis model mean-variance relationships carefully and share information across genes, but they do not take into account the specific characteristics of scRNA-seq data. Newer methods developed for scRNA-seq data account for the characteristics of scRNA-seq data in various ways, but, with the possible exception of SCDE, do not model gene-specific mean-variance relationships or share information across genes as effectively as the best bulk-tissue methods. Further necessary avenues of development appear to be combining approaches from the best bulk-tissue methods with scRNA-seq methods. Non-parametric approaches look promising. On this point, it may be worth putting more effort into an appropriate framework for determining the significance of rank-product statistics in the scRNA-seq context. Deep

investigation is beyond the scope of this chapter, but more research is required to resolve how best to perform differential expression analyses for scRNA-seq data.

5.1.2.5 The importance of quality control and dedicated software tools

This overview of the data characteristics, experimental protocols and analysis methods for scRNA-seq demonstrates how complicated the data can be and how many known and unknown factors can affect assayed expression values. Everything discussed establishes the crucial importance of proper pre-processing, quality control and normalisation of scRNA-seq data ahead of downstream statistical modeling and inference on scientific questions of interest. Exploratory data analysis is necessary for quality control, so versatile visualisation methods are needed in the preparation of scRNA-seq datasets for further analysis. All of this was true for single-cell quantitative PCR data (McDavid et al., 2013), and remains true for the newer, higher-throughput scRNA-seq platform. There are many analytical methods already available for scRNA-seq data to visualise cells in reduced-dimension spaces and to explore both gene-centric and cell-centric heterogeneity in a dataset. To this point, two normalisation methods designed for scRNA-seq data, BASiCS and GRM, have been proposed. However, all existing methods presuppose a clean, tidy matrix of expression values. Clean, tidy data do not come directly off the sequencer.

There is currently a large, and important, gap in the scRNA-seq workflow between raw, sequenced reads and clean, tidy expression data ready for statistical analysis. Experience analysing scRNA-seq thus far has convinced me that this gap would be best filled by a self-contained set of dedicated software tools for the pre-processing, quality control (QC), normalisation and visualisation of single-cell RNA-seq data. For successful analysis of scRNA-seq data, many different data types must be integrated. This data integration requires an appropriate data structure to organise the expression and accompanying data, but the options currently available, even through the outstanding Bioconductor project (Huber et al., 2015), do not have all of the required functionality for scRNA-seq data. As a new type of data, scRNA-seq analysis currently needs to be exploratory and hands-on. Thus, built on top of a sensible data structure, a software solution must support interactivity for the user. It thus requires many user-friendly, flexible methods to provide a manageable, but rigorous, workflow to go from raw scRNA-seq data to a clean, tidy dataset ready for downstream analysis. An R package is the best option, as R is the most common environment for interactive data analysis in genomics and the majority of existing tools for scRNA-seq analysis have been written in R.

5.1.3 The SCATER package

Having recognised the need for a self-contained R package for the pre-processing, quality control, normalisation and visualisation of scRNA-seq data, I have produced the SCATER

package to fill the gap in the analysis workflow from raw reads to data ready for downstream analysis. This chapter presents SCATER. I provide an overview of the package in this section, then Section 5.2 describes the datasets used for the analyses presented. Section 5.3 demonstrates the use of the package with a case study on real data, Section 5.4 discusses SCATER's visualisation methods in more depth and Section 5.5 covers broad issues of software and data integration. Key features of SCATER include: (1) the "single-cell expression set" (SCESet) class that I developed in the package as a sensible data structure for scRNA-seq data; (2) automated calculation of quality control metrics; (3) extensive visualisation capabilities and many more capabilities.

5.1.3.1 Workflow for data pre-processing and quality control

The diagram in Figure 5.3 outlines the key steps to the SCATER workflow to go from raw RNA-seq reads to a tidy SCESet object ready for further analysis. The broad workflow is:

1. Obtaining feature abundance values.
2. QC and filtering of features.
3. QC and filtering of cells.
4. Simple normalisation.
5. QC of explanatory variables, such as experimental covariates like batch, cell source and other recorded experimental information.

Optionally, one might do another round of normalisation to remove effects of particular explanatory variables from the data. Automated computation of many quality control metrics and extensive plotting functionality support the workflow.

After this procedure the user has a clean, tidy scRNA-seq dataset ready for downstream statistical modelling and analysis.

5.1.3.2 Architecture of the package

The SCATER package is built around the SCESet class (Figure 5.4), which provides a sophisticated container for scRNA-seq data. The many methods available in the package build on the foundation of the SCESet class, which, briefly, is an R S4 class that inherits the ExpressionSet class from Bioconductor's BIOBASE package (Huber et al., 2015). Further details on the class, including its motivation and execution, are available in Section 5.5.1.2. Discussion of the implementation of SCATER and how the package uses and integrates other software tools is provided in Section 5.5.

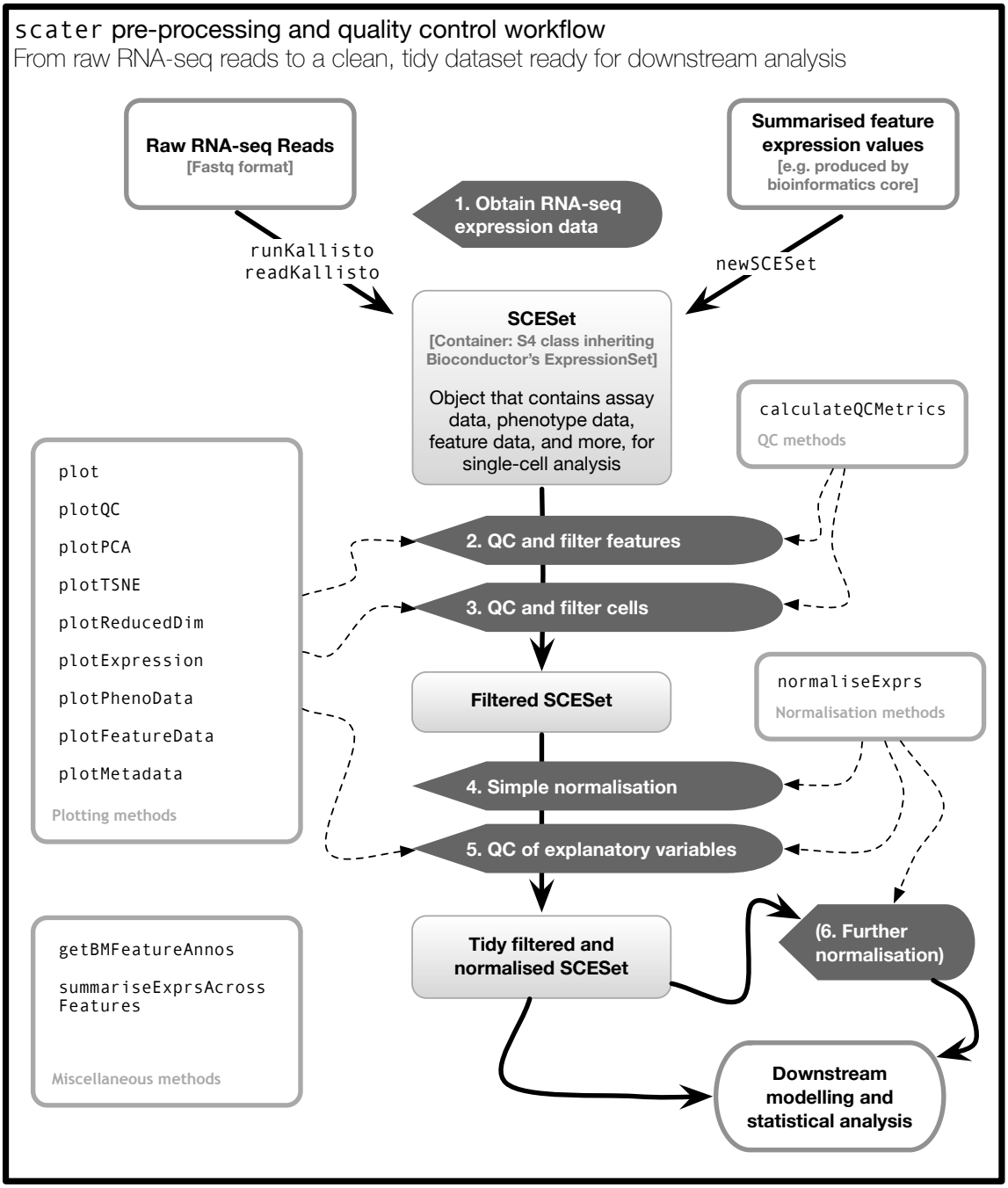


Figure 5.3: Diagram of the SCATER pre-processing and quality control workflow.

5.1.3.3 Recommendations for quality control

In general, it is difficult to make precise prescriptions for how to conduct quality control for scRNA-seq data, because each experiment will have specific requirements or quirks to be taken into account. At this time, scRNA-seq is still new, and so QC requirements are not yet as well established as they are for other types of studies such as GWAS (see Anderson et al., 2010, for example). Nevertheless, I attempt to make some recommendations for quality control here that I will show in action on real data in the next section. The SCATER

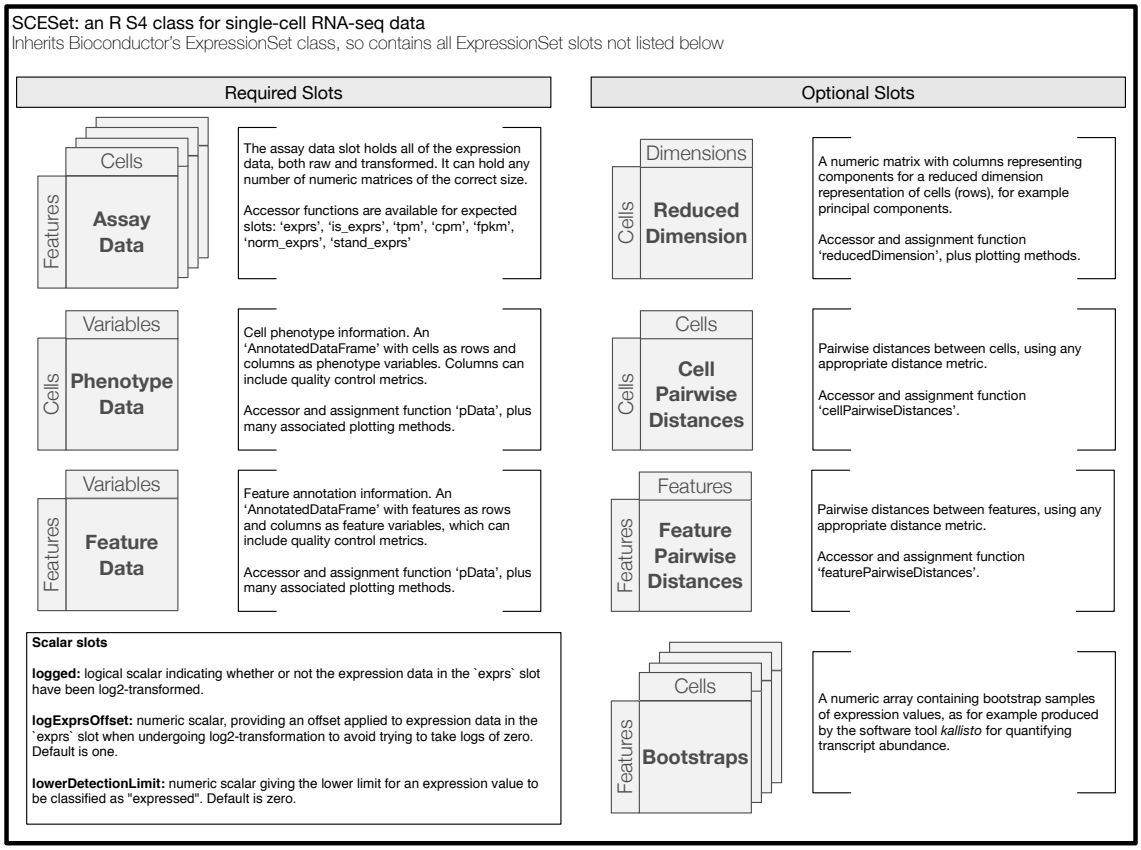


Figure 5.4: Diagram of the most important features of the SCESet class.

package is able to compute a large number of QC metrics automatically, which can be used to guide QC, especially feature and cell filtering. QC and pre-processing for scRNA-seq is necessarily exploratory and hands-on at present, so all of the general recommendations below should be adapted depending on the context and questions of interest for a given study.

Feature-level QC:

- Filter features with very little expression across the dataset, as they contain little to no useful information. Having defined a "detection limit" threshold below which an observation is deemed not to be expressed, only keep a feature if it has measurable expression in at least some given number of cells. A conservative approach is to demand expression in at least 10% of cells to retain the feature. Such filtering can increase power for certain types of inference (Bourgon et al., 2010). In some cases, such filtering will not be appropriate, but in all cases features with no expression at all in any cells should be filtered out.
- Check that the most-expressed features look "sensible" (this is unavoidably subjective), and that the top 20, 50 or 100 most-expressed features do not account for most

of the expression across the whole dataset. It is expected to see many spike-in controls and mitochondrial genes/transcripts among the 50 most-expressed features, but some endogenous features should also be present.
- Check a plot of frequency of expression (number of cells in which a feature is expressed) against mean expression. This can give a broad view of technical noise in the data.
- Check a plot of frequency of expression by feature biotype (protein-coding, pseudogene, etc) that the distribution of expression frequency for protein-coding features and other important categories are reasonable (that is, not overly skewed towards zero).
- Filter out any feature types not of interest.

Cell-level QC:

- Filter out cells with a high percentage of expression from feature controls (e.g. ERCC spike-ins). By default, I filter out cells with greater than 80% expression from feature controls.
- Filter out cells with low sequencing depth. By default, I filter out cells with log10(total counts) more than five median absolute deviations away from the median log10(total counts). Libraries with fewer than 10,000 counts should almost always be discarded, and in certain settings this threshold could be raised to 100,000 (for example, if library size for most cells is in the millions).
- Filter out cells with low "total features" (that is, number of features with detectable expression). By default, I filter out cells with total features more than five median absolute deviations away from the median total features. Some caution should be applied if using very heterogeneous cell populations as, for example, primary tissue cells can have much lower total features than cell line cells for biological rather than technical reasons.
- Inspect plots of reduced-dimension representations of the cells and filter out cells clustering with libraries sequenced from "blank" wells or other cells/libraries known to be problematic. Generally, sequencing a number of blank libraries as negative controls is a useful strategy to aid QC.
- Explore relationships between cell phenotype variables (including experimental variables and QC metrics), including plotting with SCATER functions, and filter cells that appear suspicious.

Normalisation: In SCATER, the following size-factor normalisation methods are available in the function `normaliseExprs`: TMM, RLE, "upperquartile" (in fact any percentile can be used to compute normalisation factors) and "none" (i.e. all size factors are set to

1). See Section 5.1.2.3 for details of these approaches. Internally, `normaliseExprs` calls `calcNormFactors` from the EDGER package (Robinson et al., 2010). I recommend normalising to ERCC spike-in features, if available, and applying TMM size factors.

A strength of SCATER, discussed in Section 5.5.1, is its modularity and ease of interoperability with other packages and normalisation methods. Thus the simple normalisation methods currently included in SCATER are complementary to more specific normalisation methods, which can easily "plug in" to SCATER and its pre-processing workflow.

Experimental variable QC: Experimental design for scRNA-seq studies has been given short shrift in the literature to this point, an oversight reflected in the lack of tools available for exploring the importance of different experimental variables in the dataset. Thus, recommendations for QC of experimental variables include:

- Make full use of the capabilities of the `plotQC` function in SCATER to identify "important" experimental variables. Here, too, "important" is a subjective term, but a good heuristic is to take serious consideration of variables that appear more "important" in QC plots than variables that by design should explain a large amount of variability in the data (for example, tissue of origin or batch).
- Either flag important variables for inclusion in downstream statistical modeling or regress out the effects of those variables and use residuals as expression values in downstream analyses.

As mentioned above, quality control will almost certainly need to be tailored for each individual scRNA-seq dataset. The SCATER package offers a straight-forward pre-processing, QC and normalisation workflow with a great deal of flexibility for the user and an array of visualisation methods for exploratory data analysis vital to quality control.

5.2 Single-cell datasets

This chapter demonstrates the use of the `scater` package on two real data sets. I use a dataset produced to explore heterogeneity in colon and ileum cells as an example of a well-behaved dataset, for which pre-processing and QC are relatively straight-forward. As a contrast, I also use a dataset from a study of the cell cycle in haematopoietic stem cells, which due to various aspects of the study design and execution is substantially more challenging. Both datasets are described below.

For both datasets, sequencing libraries for the individual cells were sequenced in the Genomics Core at the Wellcome Trust Centre for Human Genetics (WTCHG), Oxford using an Illumina HiSeq 2500 machine (Illumina, Inc, 2013b). Sequence reads were generated using the standard Illumina pipeline to produce Fastq files containing raw read data.

These Fastq files were the input files for the transcript abundance quantification carried out with SCATER as described in the next section. Independently, Eshita Sharma and John Broxholme at the WTCHG Genomics Core produced summarised gene-level count data by aligning reads with TOPHAT version 2.0.12 (Trapnell et al., 2009) and assigning reads to genomic features and counting totals per feature using HTSEQ version 0.6.1p1 (Anders et al., 2015).

5.2.1 Dataset: Simmons Data

The first dataset comes from Alison Simmons's lab at the Weatherall Institute of Molecular Medicine, Oxford (hence I call it the "Simmons Data"). The dataset was designed for exploring cellular heterogeneity in primary colon and ileum cells. The data ("Simmons Data") were collected in three phases, a "pilot" consisting of 59 libraries from colon cells, and larger samples of 96 libraries from colon cells and 96 libraries from ileum cells. Libraries from the colon cells (pilot and first experiment) were generated by James Kinchen and libraries from ileum cells were generated by Ana Catuneanu. Cells came from one individual patient in the pilot phase, and from three individual patients for the latter phases. Three separate sequencing runs were done for the Simmons Data, corresponding to the pilot and the two subsequent experiments. The pilot is denoted as "P140471" (59 libraries), the second experiment on colon samples as "P150057" (96 libraries) and the third experiment on ileum samples as "P150058" (96 libraries). In total 229 libraries were generated from individual single cells, 12 libraries were generated from bulk tissue samples and 12 libraries were generated from blank wells as negative controls.

5.2.2 Dataset: Cell Cycle Data

The second dataset that I will use to demonstrate the capabilities of the SCATER package, the "Cell Cycle Data", consists of 348 haematopoietic stem cells (HSCs) from primary mouse tissue, human and mouse cell lines. The study aims to investigate quiescence (cell inactivity or dormancy) and the regulation of the cell cycle in HSCs and the relationship between quiescence and drug resistance in the context of cancer. Synchronising cell cycle across cells is difficult, so conducting bulk-tissue genome-scale assays (such as RNA-seq) to study cell cycle has been challenging. Using scRNA-seq one can study whole-transcriptome gene expression in individual cells. At the time that expression measurements are made, cells naturally occur at different phases in the cell cycle. Thus, with single-cell assays there is higher resolution to relate gene expression to cell cycle and quiescence.

The study sampled HSCs from mouse bone marrow ("primary tissue") at two weeks and six weeks after birth. To compare with the primary mouse cells, data was also obtained

from human cell lines (K562 and HL60) and mouse cell lines (BAF3 and EML). Sequencing libraries were generated from individual cells by Ben Povinelli in Adam Mead's lab at the Weatherall Institute of Molecular Medicine, Oxford. For the purposes of this chapter, I will only use the first dataset to be produced from the single-cell study. This dataset comprises 348 sequenced libraries with scRNA-seq data for 187 primary mouse HSCs at the two-week (114 cells) and six-week (73 cells) timepoints, 144 cells from the four cell lines and 17 "blank" libraries (no cell present; used as negative controls). Three Fluidigm C1 chips were used for the preparation of the cell-line libraries, while FACS was used for quantifying DNA and RNA content (with Hoechst and pyronin-Y staining, respectively) in primary mouse cells and sorting those cells onto three separate plates for the preparation of sequencing libraries (El-Naggar, 2004). Ben Povinelli used the Hoechst and pyronin-Y fluorescence values to assign cell phase status to each of the primary mouse cells.

For the plates containing the primary mouse cells, their provenance is clear, namely that they are primary mouse HSCs, and we know at which timepoint they are from. However, as a result of an experimental design aimed at reducing batch effects (Leek et al., 2010), this is not the case for the cell line cells processed with the C1 system. To avoid introducing confounding of cell-line type (human or mouse, as well as specific cell lines) and batch (here, a chip corresponds to a "batch"), human and mouse cells were mixed prior to isolation on a chip. Mixing cells from different cell lines prevents the situation in which technical (chip or batch) variability cannot be distinguished from biological variability (differences between cells of different cell lines). However, the tradeoff for avoiding batch effects in this way is that we do not know, at the point that cells are sequenced, the source of the individual cells. It should be a straight-forward bioinformatic problem to identify human and mouse cells from the RNA-seq data alone, but analyses presented later in this chapter reveal that for various reasons this is not the case. This fact reinforces the great importance of careful quality control and pre-processing of single-cell RNA-seq at the current time.

5.3 Data pre-processing and quality control

The aim of this section is to demonstrate the functionality and utility of SCATER using real data. I will walk through transcript and gene expression quantification, pre-processing, QC and normalisation procedures for the Simmons Data. A major goal of SCATER is to improve the reproducibility of analyses by providing easy-to-use analytical tools, so all of the analyses following the transcript abundance quantification in this section were produced as "live" code in this document. That is, I wrote this chapter as a KNITR document (Xie, 2013) with the R code embedded and compiled the document to produce the code output and figures shown. Thus, it is a case of "what you see is what you get" with the SCATER

code here: the code shown directly produces the results presented. I used SCATER' version 0.1.6 for the analyses in this chapter.

The analyses presented here are necessarily highly "streamlined". QC and pre-processing analysis for a new dataset typically requires extensive exploratory analyses and visualisations, for which SCATER enables flexible solutions. Unfortunately, there is simply insufficient space in a thesis chapter to do justice to all of the functionality of SCATER. The suite of plotting functions is listed in Figure 5.3 describing the SCATER workflow and key functions, and visualisations are discussed further in Section 5.4. Nevertheless, I cannot cover all the ways in which these functions can be used to explore relationships between experimental metadata, QC metrics computed by SCATER and the expression data. Supplementary Tables A.3, A.2 & A.1 list SCATER's accessor and assignment functions, cell-centric QC metrics and feature-level QC metrics, respectively.

5.3.1 Transcript abundance quantification using wrappers for KALLISTO

Transcript abundances can be computed with KALLISTO from within R using SCATER's wrapper functions. I computed transcript abundances using version 0.42.2 of KALLISTO with the most recent human transcriptomes from Ensembl release 80 from May 2015 (Cunningham et al., 2015) and version 75 from February 2014 (Flicek et al., 2014), which was the version used to generate results previously produced by the "Genomics Core" at the Wellcome Trust Centre for Human Genetics, Oxford.

Before running KALLISTO one needs to generate a "kallisto index", which is a hash table of *k*-mers in the transcriptome. Generating the index beforehand allows subsequent computation by KALLISTO to be very efficient. In this case, I create indexes that incorporate both the human transcripts and the ERCC spike-in controls (Jiang et al., 2011) so that the "expression" of feature controls can be quantified as well as human transcripts (commands not shown here).

The Simmons Data consists of paired-end reads, so KALLISTO estimates the average fragment length (a required parameter for the KALLISTO quantification model) from the reads. Version 0.42.2 of KALLISTO introduced a method for adjusting the estimated transcript abundance values to account for sequence-specific biases. Single-cell RNA-seq protocols are known to have biases, particularly a 3-prime bias to reads, as discussed above, so I obtained transcript abundances using KALLISTO with this bias correction (some comparison of the overall differences in results produced is presented in Section 5.5.1.3).

To run KALLISTO using SCATER, one needs access to the Fastq-format files containing the sequenced reads, one or more KALLISTO indexes (as described above) and a "targets" file, a tab-delimited text file that lists the cell ID names and corresponding Fastq files. The "targets" file is used by the SCATER function `runKallisto` to generate the KALLISTO transcript quantification results. As described earlier, three separate sequencing runs were

done for the Simmons Data, a pilot and two subsequent experiments. The pilot is denoted as "P140471" (59 cells), the second experiment on colon samples as "P150057" (96 cells) and the third experiment on ileum samples as "P150058" (96 cells). I ran KALLISTO separately for each set. Subsequently, I combined the results to obtain a single `SCESet` object that contains all of the transcript abundances for all of the cells.

Here, I show the SCATER commands for quantifying transcript abundance for the 96 "P150057" cells, after having generated an appropriate targets file. I show the code for using the Ensembl release 80 version of the human transcriptome and use KALLISTO's bias correction method. Generating transcript quantities using a different transcriptome build, or without bias correction, simply requires changing arguments to `runKallisto`. I took exactly the same approach for the "P140471" and "P150058" cells (code not shown).

```
kallisto_P150057_rel80_wi_bias_corr <- runKallisto("/data/P150057/targets.txt",
        "/data/annotations/Homo_sapiens.GRCh38.rel80.cdna.all.ERCC.idx",
        output_prefix = "/data/P150057/kallisto_output_rel80_with_bias_corr",
        n_cores = 18, single_end = FALSE, correct_bias = TRUE, verbose = TRUE)
```

```
[1] "Analysis started:  2015-06-29 16:55:29"
[1] "Analysis completed:  2015-06-29 17:26:43"
[1] "Processed 96 samples"
```

The output above shows the extraordinary speed of transcript quantification with KALLISTO. Using 18 threads on a Linux server with moderate specifications, I obtained transcript abundances for 96 cells for over 170,000 transcripts from over 120 million reads in roughly half an hour of clock time. To put this in perspective, the align-and-count approach taken by the WTCHG Genomics Core using TOPHAT version 2.0.12 and counting reads using HTSEQ version 0.6.1p1 would require over 200 times the computing resources (Bray et al., 2015), so processing time would be measured in days rather than minutes. This dramatic increase in speed of processing raw RNA-seq read data has large implications for how we can conduct and analyse scRNA-seq studies (further discussion in Section 5.5.1.3). The `readKallisto` function makes it easy to read KALLISTO results into an SCESet object in an R session.

```
sce_P150057_rel80_wi_bias_corr <- readKallistoResults(
    kallisto_P150057_rel80_wi_bias_corr, read_h5 = TRUE)
```

This SCESet object contains matrices of raw counts, transcripts-per-million and log2-transformed transcripts-per-million data, as well as a logical matrix indicating which observations are "expressed" (here defined as any non-zero expression value). The phenotype data and feature data slots of the SCESet object contain information from the KALLISTO

runs about cells and transcripts, respectively (see Figure 5.4 for more details). Thus, in two lines of R code, the SCATER tools produce an object ready for transcript-level analysis.

For the Simmons Data, I combined the quantification results from the three sequencing runs to produce a single SCESet object (`sce_simmons`) containing all of the transcript-level data, which is analysed below.

5.3.2 Adding feature information and collapsing expression to the gene level

Before proceeding with analysis, it is typically useful to add further feature annotation information. This is done with SCATER's `getBMFeatureAnnos` function. Building on the R package `biomaRt` (Durinck et al., 2005), I define a set of transcript attributes from the Ensembl database that I want to add to the SCESet object. The function `getBMFeatureAnnos` obtains them using `biomaRt` (by default using the current version of Ensembl, here version 80) and adds them to the object. The flexibility of the arguments means that any information accessible with `biomaRt` can be added to an SCESet object. Thus, virtually any annotation information available from Ensembl or other major genomic databases can be added to an SCESet object.

```
attr_wanted <- c("ensembl_transcript_id", "ensembl_gene_id", "hgnc_symbol",
                 "chromosome_name", "transcript_biotype", "transcript_start",
                 "transcript_end", "transcript_count")
sce_simmons <- getBMFeatureAnnos(
    sce_simmons, attributes = attr_wanted, biomart = "ensembl",
    dataset = "hsapiens_gene_ensembl", feature_symbol = "hgnc_symbol",
    feature_id = "ensembl_gene_id")
```

As well as being useful for quality control and analysis, adding transcript annotation information makes it possible to summarise expression data at the gene level, which is still the most common approach to analysing single-cell and bulk RNA-seq data. There is a SCATER function, `summariseExprsAcrossFeatures`, to summarise abundance data at the level of any set of features of interest. I demonstrate the most likely use-case here, which is collapsing transcript-level data to the gene level.

I summarise by `feature_id` in the featureData slot of the SCESet objects. The `feature_id` is the Ensembl gene ID if `biomaRt` was able to locate it and the transcript ID defined by KALLISTO otherwise. The default behaviour of SCATER is to use TPM values for expression where possible.

```
sce_simmons <- summariseExprsAcrossFeatures(
    sce_simmons)
```

```
Collapsing expression to 38102 features.
```

As the output above shows, collapsing transcript-level expression values to Ensembl gene IDs yields expression for 38,102 features. This number of "gene" is much higher than the actual number of human genes thought to exist (likely fewer than 20,000). Nevertheless, there are at least that many unique Ensembl gene IDs, so I collapse expression to these features and use them as "genes" for gene-level expression analyses.

It is straightforward to summarise TPM values at the gene level. As TPM values take into account the length of transcripts, they can simply be summed across all transcripts to obtain accurate expression values for each gene. It is not as straight-forward to summarise read counts at the gene level. The "naive" approach of summing counts across transcripts to get gene-level counts can be problematic for RNA-seq protocols that sequence fragments from full-length transcripts (Trapnell et al., 2013). However, the SCATER function retains all of the data in the transcript-summarised SCESet object, following a general philosophy of respecting and retaining raw data. Thus, the object produced by `summariseExprsAcrossFeatures` sums counts (and CPM and FPKM values too, if present) across all transcripts for each gene. The TPM values are recommended for downstream analyses, but ultimate discretion is left to the user.

To obtain the dataset used in the subsequent "live" parts of the analysis, I add gene annotation information using `getBMFeatureAnnos`, as shown above for transcripts, and use gene symbols (where available) combined with Ensembl gene IDs as feature IDs for the analysis that follows (code not shown).

5.3.3 Adding cell metadata

Henceforth, I go through the pre-processing and QC methods in SCATER to prepare the Simmons Data for downstream analysis, using the dataset produced as described in the preceding sections. From this point until the end of the section, the R code used has been included in the document and compiled using `knitr` (Xie, 2013). Thus the code shown (on a peach-coloured background) directly produces the output provided, including the various different plots.

The first step is to read the SCESet containing the data into R.

```
load("../../022_Simmons/cache/sce_simmons_rel80_wi_bias_corr_gene.RData")
```

This loads the object `sce_simmons`, which contains the gene-level expression values and associated metadata computed above. The file `combined_cell_annotation.csv` contains cell annotation information that I will use for the analysis:

```
cell_annos <- read.csv("../../022_Simmons/metadata/combined_cell_annotation.csv")
```

There are many more rows in the cell annotation file than in the SCESet object, indicating that not all cells annotated were subsequently sequenced. The first task is thus to match the annotations to the cells in the SCESet object. This can be done using the "Readgroup" column of the annotation data-frame, which corresponds to the "sample names" (i.e., cell IDs) in the SCESet.

There is annotation information for all cells for which I have abundance information. Thus, I can match up the cell IDs with the Readgroup IDs and pull out the relevant rows of the cell annotation data-frame.

```
mm <- match(sampleNames(sce_simmons), cell_annos$Readgroup)
cell_annos_filt <- cell_annos[mm,]
```

The information missing from the annotations data-frame is about the tissue used. For this study, the pilot samples and second run were conducted by James Kinchen using colon tissue. The third run was generated by Ana Catuneanu and used ileum tissue. I add this information to the annotations data-frame.

```
cell_annos_filt$tissue_type <- ifelse(cell_annos_filt$study == "ana",
                                      "ileum", "colon")
cell_annos_filt$batch <- reorder(plyr::revalue(
    cell_annos_filt$study, c(pilot = "Batch1", james = "Batch2", ana = "Batch3")))
rownames(cell_annos_filt) <- sampleNames(sce_simmons) <- cell_annos_filt$X
```

Above, I also assign the "X" column of the cell annotations data-frame as the rownames for that data-frame and the sample names for the SCESet, because those names are more informative than the Readgroup IDs (and shorter). The names have to match so that this information can be added to the SCESet, or a (protective) error will be thrown. To add the cell annotation information to the SCESet I use the `pData` function.

```
pdata_new <- new("AnnotatedDataFrame", cbind(cell_annos_filt, pData(sce_simmons)))
pData(sce_simmons) <- pdata_new
```

This approach retains all of the previous phenotype (cell) data generated with the `kallisto` quantification.

The final consideration in organising the dataset at this point is dealing with genes that are not expressed in any cells. Specifically, I remove genes with no expression at all across any cells. Since there is no information available for these genes, they can be of no interest to the study.

```
all_zero <- rowSums(tpm(sce_simmons)) == 0
sce_simmons <- sce_simmons[!all_zero,]
```

This filtering removes from the analysis 5823 genes for which we have no information. The code above also demonstrates the ease with which we can subset SCESet objects. With the expression, feature and cell annotation information all added to the SCESet object I can proceed with QC.

5.3.4 Calculation of QC metrics

To help with the QC of the dataset, SCATER enables straight-forward calculation of many QC metrics with the `calculateQCMetrics` function. To help produce useful metrics, it is good to identify feature and cell controls, if present. In this case, I use ERCC spike-ins and mitochondrial genes as gene controls (called "feature_controls" in SCATER) and blank and bulk "cells" (more correctly "sequenced libraries") as cell controls.

```
ercc_genes <- grepl("^ERCC", featureNames(sce_simmons))
mt_genes <- grepl("^MT-", featureNames(sce_simmons))
blank_cells <- sce_simmons$sample_information == "Blank"
bulk_cells <- sce_simmons$sample_information == "Bulk"
sce_simmons <- calculateQCMetrics(sce_simmons,
    feature_controls = list(ERCC = ercc_genes, MT = mt_genes),
    cell_controls = list(Blank = blank_cells, Bulk = bulk_cells))
```

The `calculateQCMetrics` function calculates many feature-level and cell-level QC metrics directly from the expression data, as discussed in more detail later in the chapter and listed in full in Supplementary Tables A.1 & A.2. Thus, after running `calculateQCMetrics` on an SCESet object, a large amount of useful information is produced, at the cell level (added to the phenoData slot of the object) and the feature level (added to the featureData), that can be used to aid QC.

5.3.5 QC and filtering of features

The first step in the QC process is filtering out unwanted features. It is often desirable, depending on the experimental context, to filter out features with very low overall expression, and any others that plots or other metrics indicate may be problematic, as outlined in Section 5.1.3.3 above, or of no interest to the study at hand.

First let us look at a plot that shows the top 50 (by default) most-expressed features. By default, "expression" for this plot is defined using the feature counts (if available), TPM, CPM or FPKM values. The values in the exprs slot of the object can be used instead, if desired. This plot can be produced for subsets of the cells by simply subsetting the SCESet used.

It can be particularly useful to inspect the most-expressed features in just the cell controls (for example blanks or bulk samples). Here, I show the most-expressed genes in blanks and non-blank cells (Figure 5.5). In the previous section, I defined two sets of cell

controls in the call to `calculateQCMetrics`. That function added the "is_cell_control_Blank" column to the phenotype data of the SCESet object, which simply indicates if a cell is defined as a blank cell control. The '$' operator makes it easy to access the "is_cell_control_Blank" column and use it to subset the SCESet, as below. One can thus compare the most-expressed features in the cell controls (Figure 5.5B) and in the cells of biological interest and the bulk libraries (Figure 5.5A).

The `multiplot` function included in SCATER allows a very simple way to plot multiple GGPLOT2 plots (Wickham, 2009), output by most of SCATER's plotting functions, on the same page. In Figure 5.5, however, we use `plot_grid` from the excellent COWPLOT (Wilke, 2015) package, available on CRAN (R Core Team, 2015a) which provides more sophisticated possibilities for arranging multiple GGPLOT2 plots on a single page.

For blank libraries (Figure 5.5B), feature controls, especially ERCC spike-in controls, account for almost all of the library (>99% of all counts). ERCC spike-ins and mitochondrial genes still contribute a large amount to the libraries for non-blank libraries but their contribution to the total is reduced. Endogenous, but biologically uninteresting genes, such as *GAPDH*, typically appear in the top 50 for non-blank libraries (Figure 5.5A).

Another way to obtain an idea of the level of technical noise in the dataset is to plot the mean expression level against frequency of expression (that is, number of cells with expression for the gene above the defined threshold (default is zero)). A set of specific features to plot can be defined, but need not be. By default, the function will look for defined feature controls (as supplied to `calculateQCMetrics`). If feature controls are found, then these will be plotted, in a different colour from the endogenous features. Figure 5.6 thus shows both ERCC and MT feature controls. The mitochondrial genes are all very highly expressed. The ERCC spike-in genes cover a range of expression levels, and a range of frequency of expression (number of cells in which the gene has non-zero expression). The spike-in genes should, of course, be present in all cells, so looking at their frequency of expression across the mean expression range provides some insight into "dropout" of signal due to technical effects.

Beyond these QC plots, the function `plotFeatureData` exists as a general and flexible function for plotting sets of feature metadata variables. The featureData slot of the SCESet object behaves largely like a data frame, so we can always pass it to GGPLOT directly if we wish. Thus if we want to make more complicated plots we can do this instead of using the convenience function `plotFeatureData`. Taking this approach, Supplementary Figure A.1 shows a sensible distribution for number of cells expressing genes with different biotypes. The median number of cells expressing each protein-coding gene is just under 100 (out of 251), whereas it is much lower for the various kinds of pseudogene. A distribution of expression frequency for protein-coding genes centred closer to zero could indicate large-scale problems with the experiment, as it would be concerning if a substantial proportion

```
p1 <- plotQC(sce_simmons[, !sce_simmons$is_cell_control_Blank],
             type = "highest-expression")
p2 <- plotQC(sce_simmons[, sce_simmons$is_cell_control_Blank],
       type = "highest-expression")
plot_grid(p1, p2, labels = c("A", "B"))
```

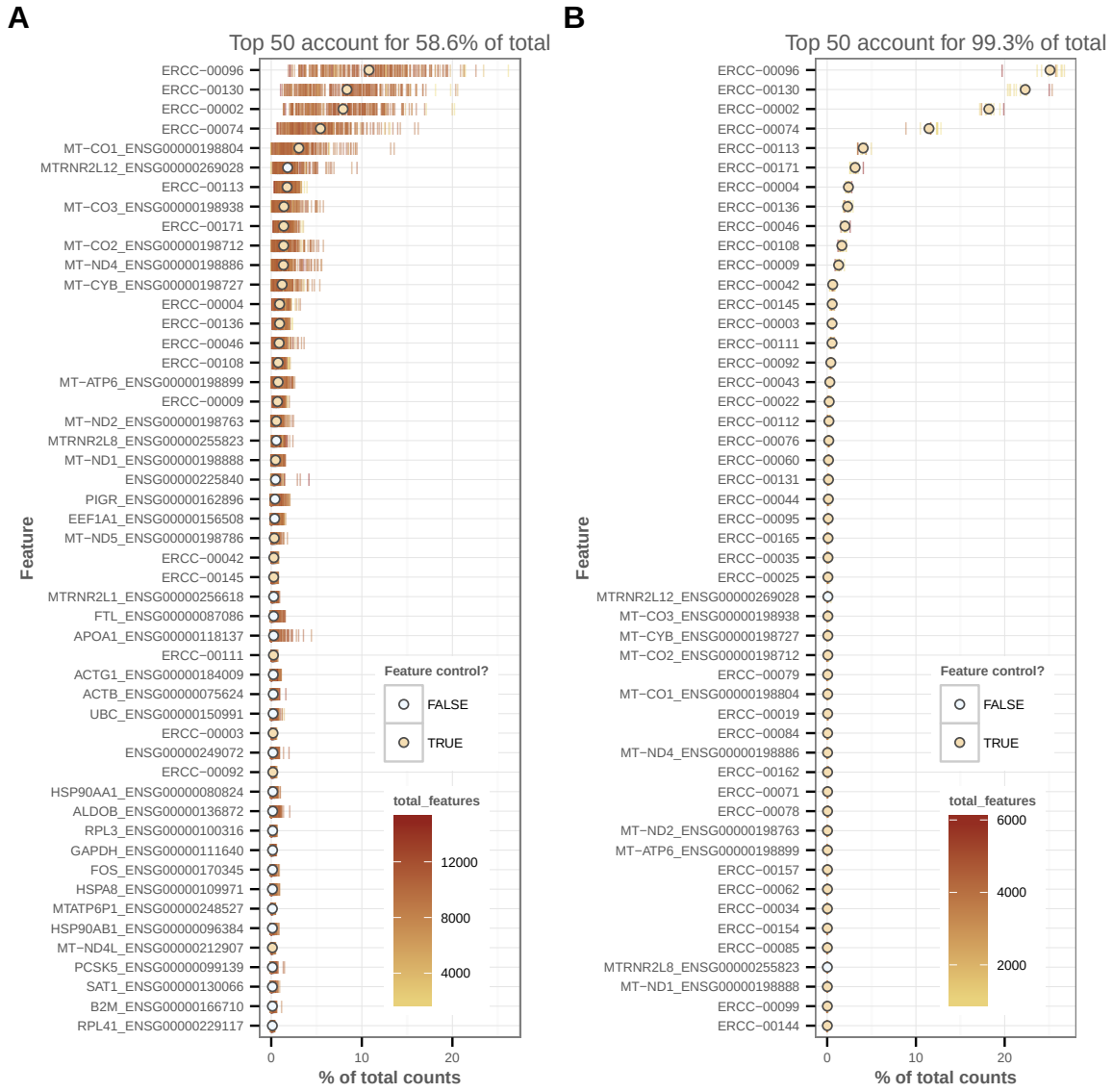


Figure 5.5: The "highest-expression" method for the `plotQC` function on SCESet objects plots the expression levels for each cell of the (by default) 50 most-expressed genes across the whole dataset. The subsetting capabilities of SCESet objects make it easy to look at the most-expressed genes in subsets of the cells. Here, the most-expressed genes are plotted separately for (A) non-blank cells and (B) blank wells (no cell should have been present in the well). If present, feature controls are shown in a different colour from endogenous features. The colour for cell-level values can be defined by cell phenotype data, by default the total features (number of genes with non-zero expression) for the cell.

```
plotQC(sce_simmons, type = "exprs-freq-vs-mean" )
```

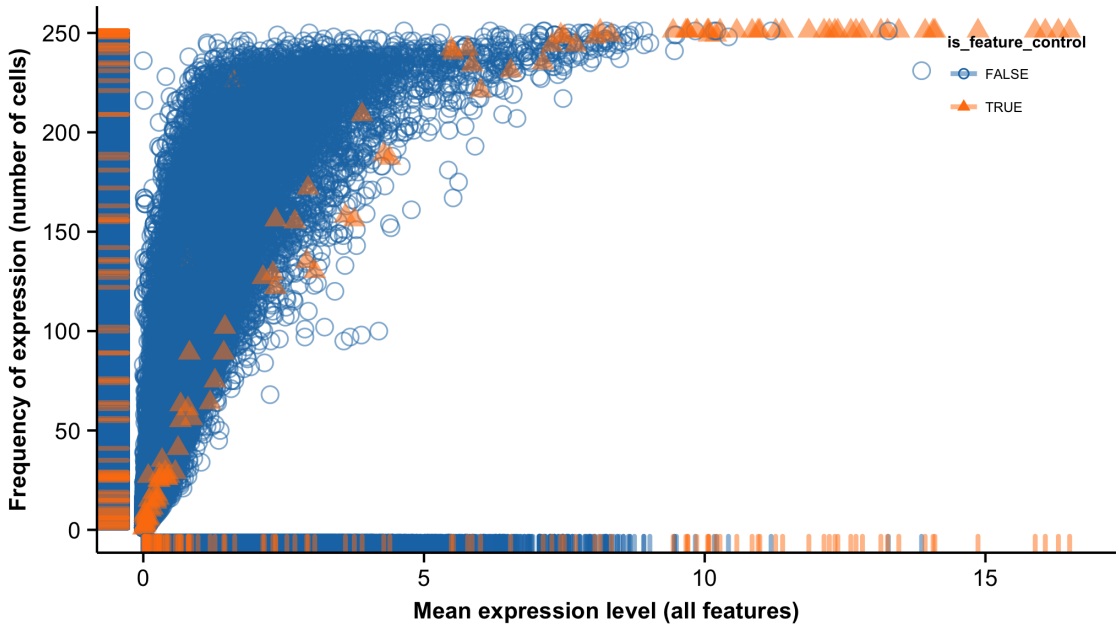


Figure 5.6: The "exprs-freq-vs-mean" method for the `plotQC` function on SCESet objects plots the expression frequency for features against the expression mean. By default, all features are plotted with the feature controls plotted in a contrasting colour. However, any specific subset of the features can be plotted by specifying the appropriate arguments in the `plotQC` function. Here, the default is used with the Simmons Data, so ERCC and mitochondrial genes are plotted in orange and the other features are plotting in blue.

of protein-coding genes were detected in only a small number of cells. The figure also shows that there is a moderate number of *IG* and *TR* genes that are generally expressed in only a handful of cells and are not of biological interest here. I will drop these from the analysis.

```
keep_genes <- !(grepl("^IG", fData(sce_simmons)$gene_biotype)
                | grepl("^TR", fData(sce_simmons)$gene_biotype))
sce_simmons <- sce_simmons[keep_genes,]
```

The subsetting of rows of SCESet objects makes it easy to drop unwanted features. This is generally a well-behaved dataset, so in this case little filtering of genes is required. In other cases, however, a more thorough investigation of feature-level QC metrics (utilising `plotQC` and `plotFeatureData`) may be required. The final step shown here is simply to filter out very lowly-expressed genes as in many cases such features are not of interest because they contain so little information and some types of inference can benefit from removal of very lowly-expressed features (Bourgon et al., 2010). In general, the pattern of expression and non-expression could be interesting, and some information will be lost by discarding such genes. The particular requirements for a given study will dictate whether

lowly expressed features ought to be retained or not. For illustrative purposes here, I show how it can be done. There are 251 cells (including blanks and bulks) in this dataset, so to retain a gene for analysis I demand that it is expressed (above the threshold of zero here) in at least 10% of libraries.

```
drop_genes <- fData(sce_simmons)$n_cells_exprs < ceiling(0.1 * ncol(sce_simmons))
sce_simmons <- sce_simmons[!drop_genes,]
```

This filtering step drops 15388 genes, so the subsequent analysis is restricted to just 16576 genes (or, strictly, features with unique Ensembl gene IDs, as discussed above). Different thresholds for defining an observation as expressed or not, or for proportion of cells expressing a feature may be appropriate in different settings.

I recalculate the QC metrics on the new, smaller SCESet.

```
ercc_genes <- grepl("^ERCC-", featureNames(sce_simmons))
mt_genes <- grepl("^MT-", featureNames(sce_simmons))
blank_cells <- sce_simmons$sample_information == "Blank"
bulk_cells <- sce_simmons$sample_information == "Bulk"
sce_simmons <- calculateQCMetrics(sce_simmons,
    feature_controls = list(ERCC = ercc_genes, MT = mt_genes),
    cell_controls = list(Blank = blank_cells, Bulk = bulk_cells))
sce_simmons$filter_on_total_counts_or_features <- (
    sce_simmons$filter_on_total_counts | sce_simmons$filter_on_total_features
)
```

With genes filtered, analysis can proceed to QC and filtering of cells.

5.3.6 QC and filtering of cells

When conducting quality control of cells, the aim is to identify potentially problematic cells (and there are many ways in which cells could be problematic) and filter them out of the dataset. Plotting functions help the exploration of computed QC metrics and other cell phenotype information, as well as looking at reduced-dimension representations of cells. The subsetting of columns (which correspond to cells) of SCESet objects makes it easy to drop unwanted cells.

The function `plotPhenoData` can be used to plot cell metadata variables. This function is useful for exploring the relationships between the many QC metrics computed by `calculateQCMetrics` and other experimental metadata. Often, problematic cells can be identified from such plots. A particularly useful plot for cell QC is plotting the percentage of expression accounted for by feature controls against total features (Figure 5.7). There is a modest tendency for the percentage of expression from feature controls to decrease as total features increases. One expects to see well-behaved cells with relatively high total features (number of features with detectable expression) and low percentage of expression from

feature controls. High percentage expression from feature controls and low total features are indicative of blank and failed cells.

To decide which cells to filter out, I look at reduced-dimension representations of the cells. The SCATER package provides the functions `plotPCA` and `plotTSNE` to produce principal components analysis plots and t-SNE plots simply from SCESet objects. Briefly, t-SNE is a probabilistic dimension-reduction technique that uses conditional probabilities to represent similarities between high-dimensional datapoints. It is discussed in more detail in Section 5.4.2. Here I use t-SNE (Van der Maaten & Hinton, 2008), which tends to do a better job than PCA of placing similar cells close together in low-dimensional space. Both PCA and t-SNE visualisations are discussed in more detail in Section 5.4. A first step for cell filtering cells is to produce a two-dimensional t-SNE plot colouring cells in the plot according to whether or not filtering the cell is suggested based on the computed QC metrics suggest filtering (Figure 5.8).

By default, the t-SNE plot, like the PCA plot, is produced using the 500 features with the most variable expression across all cells. The number of most-variable features used can be changed with the 'ntop' argument, and alternatively, a specific set of features to use for the t-SNE plot can be defined with the 'feature_set' argument. This option is discussed in detail in Section 5.4.3. One option would be to identify a set of highly variable genes using a method such as that proposed by Brennecke et al. (2013) to produce a t-SNE or PCA plot. The perplexity parameter (analogous to defining the number of neighbours that each cell is expected to have) can be adjusted to explore better visualisations for a given dataset.

Following the general QC recommendations discussed in Section 5.1.3.3, there are good groups for suspicion of cells with greater than 80% of expression from ERCC spike-in genes, so I will remove genes that SCATER's automated QC metrics suggest to filter on percentage of expression from ERCC controls. I will also remove cells clustering with the blanks on the t-SNE plot, so will filter out cells with a value between -1.5 and 2 on the first dimension and less than zero in the second dimension (read off the plot), and also cells with a t-SNE value less than -9 in the second dimension. I also filter out libraries annotated as either "Blank" (ten libraries) or "Bulk" (twelve libraries) to focus on the real, better quality cells. Further, I also filter cells with low total counts and low total features, as identified with the computed QC metrics. The two-dimensional coordinates for the embedding can be saved by `plotTSNE` to the `reducedDimension` slot of the SCESet object, which helps for filtering cells based on t-SNE analysis.

```
sce_simmons <- plotTSNE(sce_simmons, rand_seed = 201506, return_SCESet = TRUE,
                        draw_plot = FALSE, perplexity = 20)
keep_cells <- !(sce_simmons$filter_on_pct_exprs_feature_controls_ERCC |
                    sce_simmons$filter_on_total_counts |
```

```
p1 <- plotPhenoData(sce_simmons, theme_size = 8, aesth = aes(x = total_features,
        y = pct_exprs_feature_controls, colour = sample_information)) +
    theme(legend.position = "bottom")
p2 <- plotPhenoData(sce_simmons, theme_size = 8, aesth = aes(x = total_features,
        y = pct_exprs_feature_controls, colour = filter_on_total_counts)) +
    theme(legend.position = "bottom")
p3 <- plotPhenoData(sce_simmons, theme_size = 8, aesth = aes(x = total_features,
        y = pct_exprs_feature_controls, colour = filter_on_total_features)) +
    theme(legend.position = "bottom")
p4 <- plotPhenoData(sce_simmons, theme_size = 8, aesth = aes(x = total_features,
        y = pct_exprs_feature_controls,
        colour = filter_on_total_counts_or_features)) +
    theme(legend.position = "bottom")
plot_grid(p1, p2, p3, p4, labels = c("A", "B", "C", "D"))
```

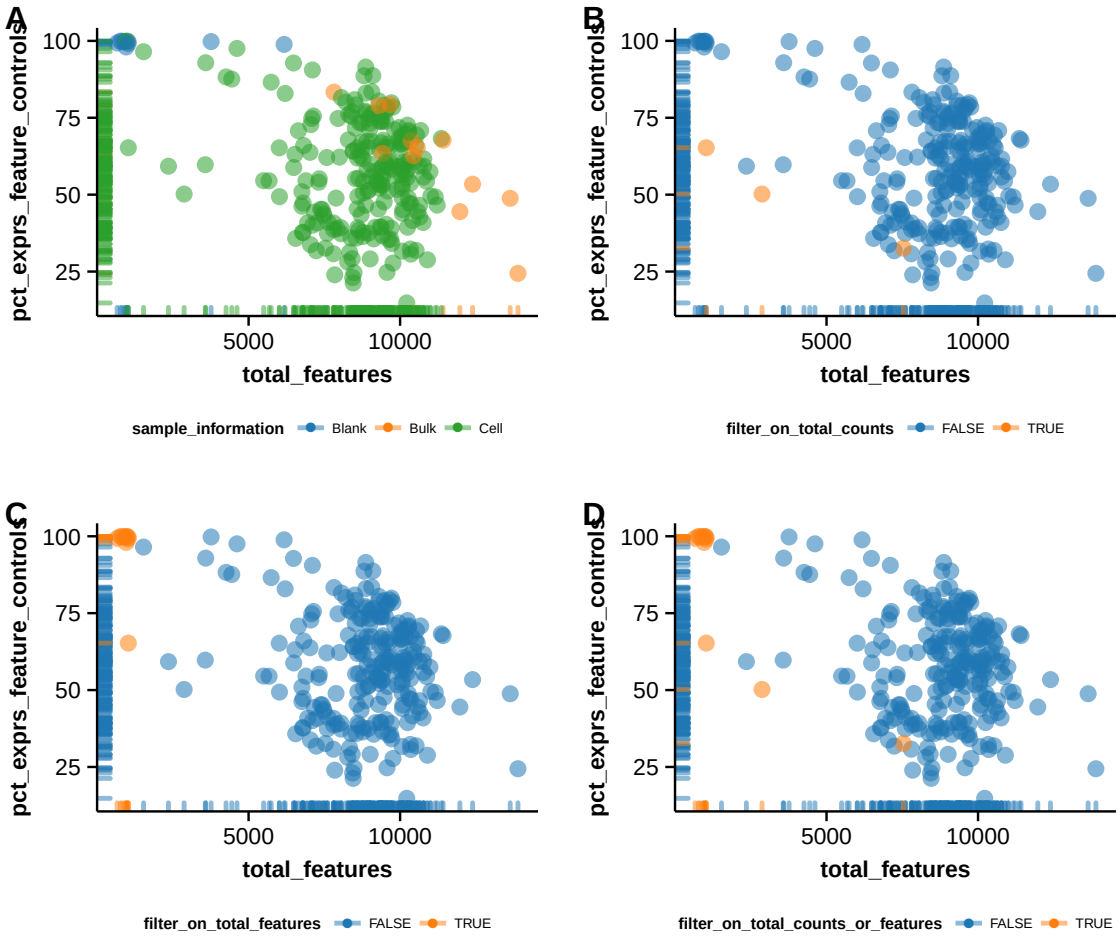


Figure 5.7: The `plotPhenoData` function provides a convenient way to plot cell phenotype, metadata, and QC metrics. Here, for the Simmons Data, the total features (number of genes expressed in each cell) is plotted against the total counts (total size of the library; total counts). Libraries with low total counts or total features are likely to be problematic and will be filtered out of the analysis in later steps. Points are coloured by (A) 'sample_information', and showing that there is a tendency for the bulk libraries to have high total features, which is expected; (B) whether or not to filter on library depth, (C) whether or not to filter based on total features, and (D) whether or not to filter on either total counts or total features.

```
p1 <- plotTSNE(sce_simmons, colour_by = "batch", rand_seed = 201506,
        perplexity = 20) + geom_vline(xintercept = c(-1.5, 2), linetype = 2) +
    theme(legend.position = "bottom")
p2 <- plotTSNE(sce_simmons, colour_by = "sample_information", rand_seed = 201506,
                perplexity = 20) + theme(legend.position = "bottom") +
    geom_vline(xintercept = c(-1.5, 2), linetype = 2)
p3 <- plotTSNE(sce_simmons, rand_seed = 201506, perplexity = 20,
        colour_by = "filter_on_pct_exprs_feature_controls_ERCC") +
    geom_vline(xintercept = c(-1.5, 2), linetype = 2) +
    theme(legend.position = "bottom")
p4 <- plotTSNE(sce_simmons, colour_by = "filter_on_total_features",
                rand_seed = 201506, perplexity = 20) +
    theme(legend.position = "bottom") +
    geom_vline(xintercept = c(-1.5, 2), linetype = 2)
plot_grid(p1, p2, p3, p4, labels = c("A", "B", "C", "D"), ncol = 2)
```

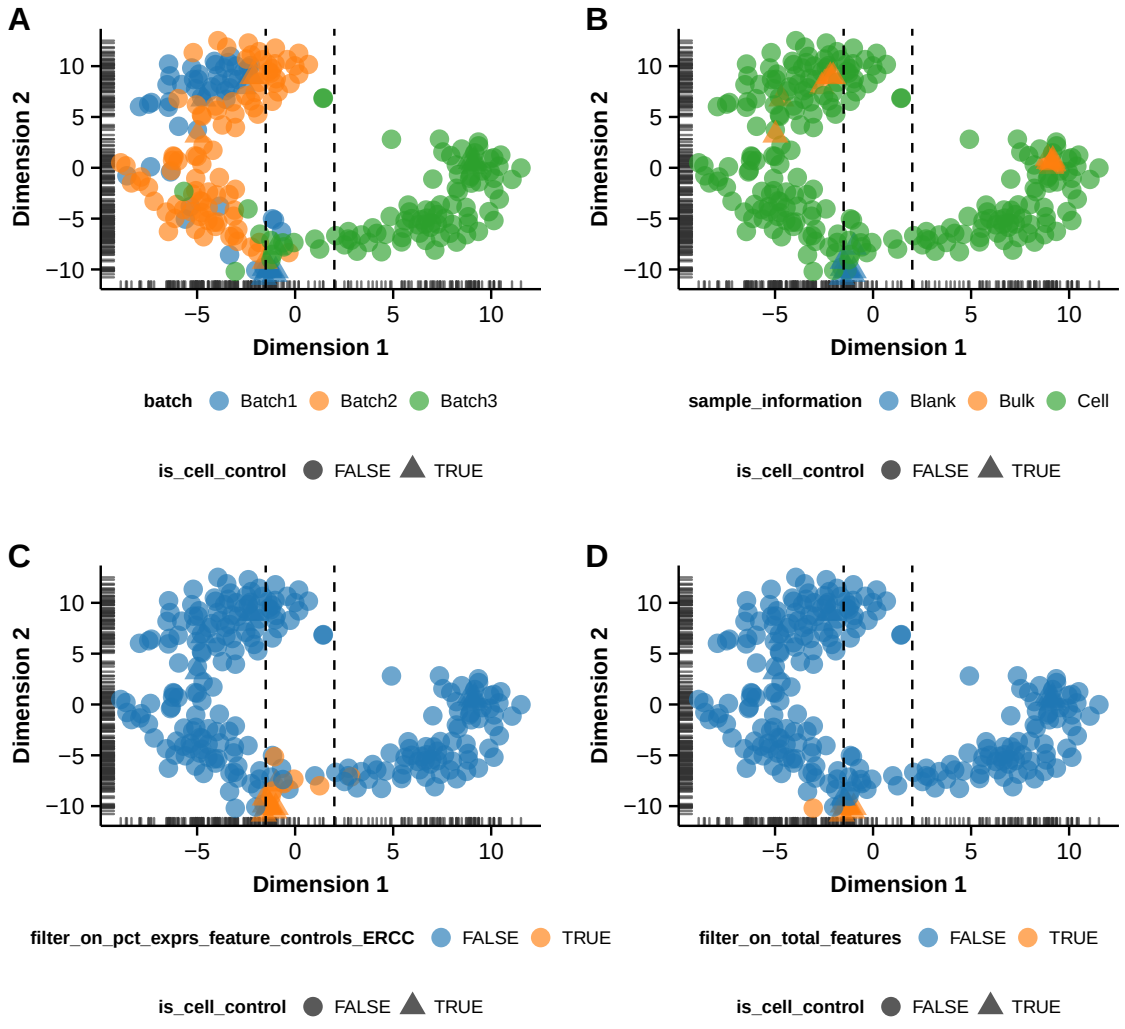


Figure 5.8: Scatter plots of the first two components from a t-distributed stochastic neighbour embedding of the cells of the Simmons Data are plotted with points coloured by (A) percentage of expression from ERCC spike-in control genes, (B) whether or not to filter cells based on percentage of expression from ERCC genes, (C) whether or not to filter on library depth, and (D) whether or not to filter on total features. Vertical lines at x-intercepts of -1.5 and 2 indicate a region of cells that might be filtered out on the basis of being similar to the blank wells.

```
                    sce_simmons$filter_on_total_features |
                    (redDim(sce_simmons)[, 1] > -1.5 &
                         redDim(sce_simmons)[, 1] < 2 &
                         redDim(sce_simmons)[, 2] < 0) |
                    (redDim(sce_simmons)[, 2] < -9 ) |
                    sce_simmons$sample_information == "Bulk" |
                    sce_simmons$sample_information == "Blank")
sce_simmons_filt <- sce_simmons[, keep_cells]
```

In this analysis, 210 cells are retained for analysis. A t-SNE plot can be produced using only endogenous genes (that is, not control genes) after filtering (Figure 5.9). The first dimension of the t-SNE plot strongly separates the ileum cells (batch 3) from the colon cells (batches 1 and 2), although a handful of ileum cells cluster with the colon cells. The second dimension appears generally to order cells by the percentage of the cell's expression accounted for by feature controls.

5.3.7 Simple data normalisation

Normalisation of scRNA-seq is discussed in Section 5.1.2.3 and options currently available in SCATER in Section 5.1.3.3. Here, I undertake normalisation using the ERCC spike-in genes as the feature set to which to normalise expression values. The default behaviour in the `normaliseExprs` function (if the "use_as_exprs" argument is not specified) is to normalise the count data, computing size factors from the raw counts. I use the TMM method to compute normalisation factors. Similar results (not shown) are obtained with RLE or "upperquartile" size factors.

```
ercc_genes <- grep("^ERCC-", featureNames(sce_simmons_filt))
sce_simmons_filt <- normaliseExprs(sce_simmons_filt, method = "TMM",
                                   feature_set = ercc_genes)
```

Since I have normalised to ERCC genes, the normalised counts-per-million values computed are strictly "counts-per-million-ERCC-counts", and expression becomes relative to the TMM-scaled ERCC counts. The effect on the overall expression densities for ERCC genes and all genes is subtle (Supplementary Figure A.3). The normalisation tightens the distributions of highly expressed genes across cells, but there is slightly more variability for moderately expressed genes. The whole distribution is shifted to the right with normalisation since CPM values are relative to just the ERCC genes instead of all genes.

One can also normalise the TPM values instead of the counts by specifying the "use_as_exprs" argument.

```
ercc_genes <- grep("^ERCC", featureNames(sce_simmons))
sce_simmons_filt <- normaliseExprs(sce_simmons_filt, method = "TMM",
            feature_set = ercc_genes, use_as_exprs = "tpm", logratioTrim = 0.1)
```

```
endog_genes <- !fData(sce_simmons_filt)$is_feature_control_ERCC
p1 <- plotTSNE(sce_simmons_filt[endog_genes,], colour_by = "batch",
               rand_seed = 201506, perplexity = 20) +
    theme(legend.position = "bottom")
p2 <- plotTSNE(sce_simmons_filt[endog_genes,], rand_seed = 201506,
              colour_by = "pct_exprs_feature_controls", perplexity = 20) +
    theme(legend.position = "bottom")
plot_grid(p1, p2, labels = c("A", "B"), ncol = 2)
```

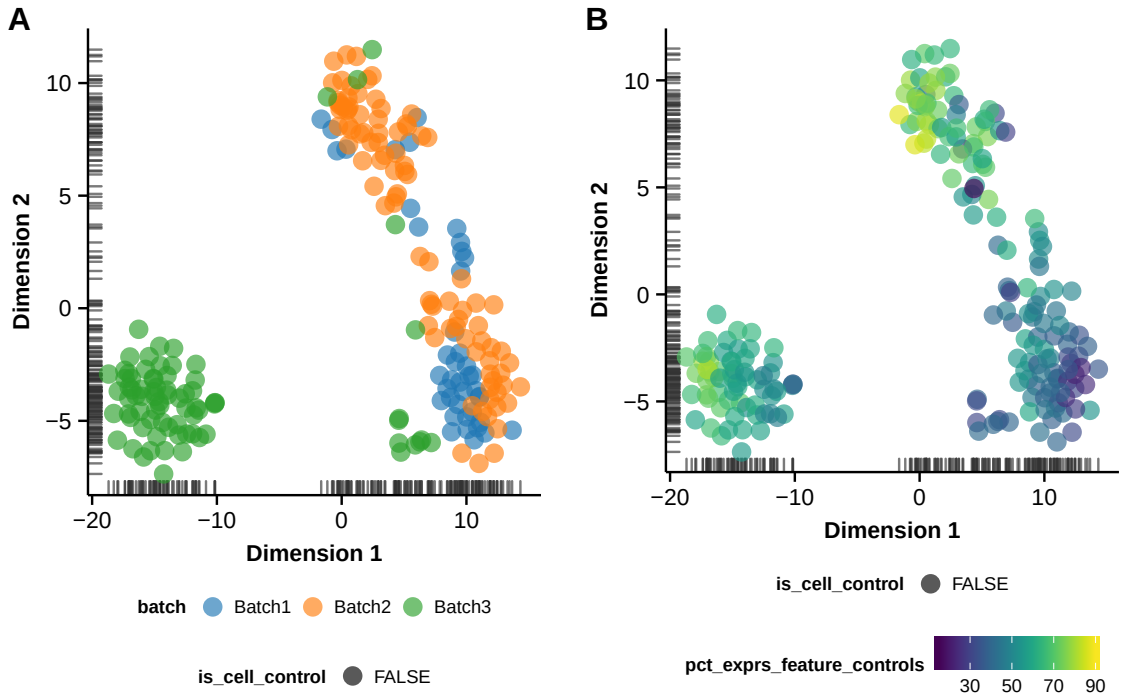


Figure 5.9: Scatter plots of the first two components from a t-distributed stochastic neighbour embedding using only endogenous genes of the cells of the Simmons Data after filtering problematic cells. Points are coloured by (A) batch (batches 1 and 2 are colon samples, batch 3 ileum), and (B) the percentage of expression in the cell accounted for by feature controls.

Using ERCC spike-ins as the features to use for normalisation, means that the unit for expression becomes "transcripts-per-million-ERCC-transcripts". Looking at densities of ERCC genes, and all genes, after normalisation (Supplementary Figure A.4) yields similar conclusions to the ERCC-count normalisation (Supplementary Figure A.3), but the normalisation based on TPM values seems to do a better job of increasing similarity in expression distributions across all genes. For this well-behaved dataset with a relatively high degree of consistency between cell libraries, these simple normalisation approaches do not substantially change the expression distributions across cells. Repeating the t-SNE plot from earlier yields a slightly cleaner separation of colon and ileum cells, but overall a very similar picture (not shown).

More sophisticated, or situation-specific, normalisation methods are possible (for ex-

ample, BASiCS and GRM). Existing normalisation methods can easily be "plugged in" to SCATER, because the accessor and assignment functions (see Supplementary Table A.3) make it simple to extract or replace an SCESet's expression and normalised expression values. Further developments of normalisation methods in SCATER, particularly exploring quantile and rank normalisation, are planned.

5.3.8 QC of experimental variables

Experimental design is a critical, but neglected, aspect of scRNA-seq studies, as discussed in Section 5.1.3.3. To the best of my knowledge, methods like those described in this section for exploring experimental and QC variables and the experimental design, do not feature in any scRNA-seq software packages apart from SCATER. As discussed in Section 5.1.2, there are a very large number of potential confounders, artifacts and biases in scRNA-seq studies. Exploring the effects of such explanatory variables (both those recorded during the experiment and computed QC metrics) is crucial for appropriate modeling of the data. The SCATER package provides a set of methods specifically for quality control of experimental and explanatory variables, which will be demonstrated briefly here.

The relative importance of different explanatory variables can be explored with some of the `plotQC` function options. Supplying the `type = ''expl''` argument to `plotQC` computes the marginal R^2 for each variable in the SCESet when fitting a linear model regressing expression values for each gene against just that variable, and displays a density plot of the gene-wise marginal R^2 values for the variables. The default approach looks at all variables in the phenoData slot of the object and plots the top "nvars_to_plot" variables (default is 10). Alternatively, one can choose a subset of variables to plot in this manner, which I do here (Figure 5.10). The density curves for marginal R^2 show the relative importance of different variables for explaining variance in expression between cells. For this dataset, the marginal R^2 density curves are very similar with and without the ERCC-TPM normalisation carried out in the previous section, reinforcing that this is a nicely behaved dataset.

With the same calls as above, but with `method = "pairs"`, `plotQC` can produce a pairs plot to visualise the relationships between explanatory variables, ranked by their median marginal R^2 (Figure 5.11). From this pairs plot the confounding of batch and tissue type is obvious (of course, we already knew this), demonstrating that this type of plot is useful for finding correlations between experimental and QC variables with substantial explanatory power.

This analysis indicates that batch and tissue type (which are confounded), total features and percentage of expression from feature controls have substantial explanatory power for many genes, so these variables are good candidates for conditioning out in a normalisation step, or including in downstream statistical models. Sequencing depth (on the log10-scale) does not appear to be an important explanatory variable.

```
p1 <- plotQC(sce_simmons_filt, type = "expl", use_as_exprs = "exprs",
       variables = c("total_features", "log10_total_counts",
                     "pct_exprs_feature_controls", "tissue_type", "batch"))
p2 <- plotQC(sce_simmons_filt, type = "expl", use_as_exprs = "norm_exprs",
       variables = c("total_features", "log10_total_counts",
                     "pct_exprs_feature_controls", "tissue_type", "batch"))
plot_grid(p1, p2, labels = c("A", "B"), nrow = 2)
```

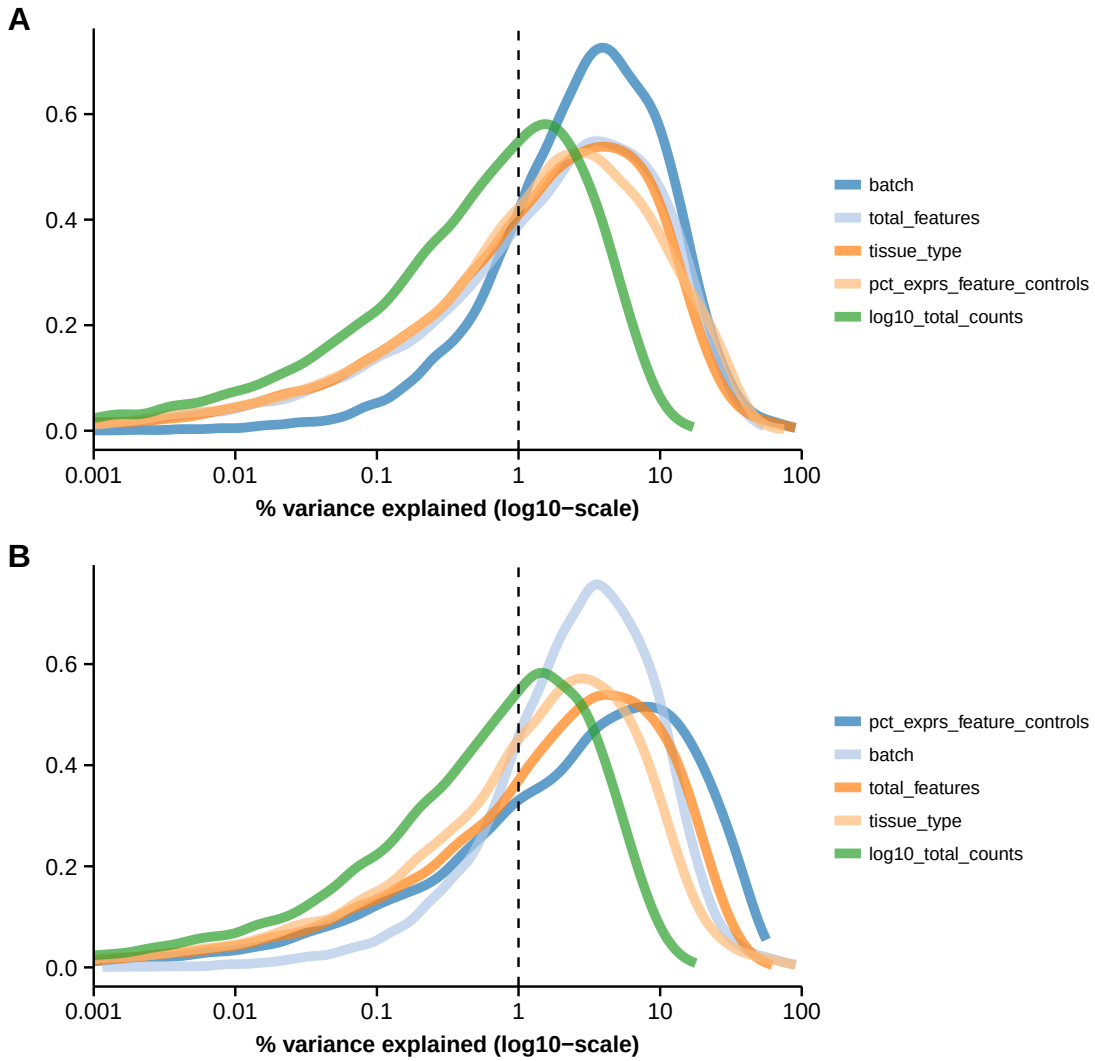


Figure 5.10: Density plots of marginal R^2 values for a set of explanatory variables for the Simmons Data. A linear model is fitted to each feature with one explanatory variable as the only covariate (a "marginal" model) and the R^2 value from the model is recorded. For each explanatory variable, the density of the marginal R^2 values is plotted. The explanatory variables "batch", "total features" (number of features with non-zero expression), "tissue type", percentage expression from feature controls and "total counts" (number of counts for the cell, on the log-10 scale) are shown. The plot (A) uses the expression values in the `exprs` slot of the SCESet object, and plot (B) uses normalised expression values obtained using the TPM-ERCC normalisation approach described in the text.

```
plotQC(sce_simmons_filt, type = "expl", method = "pairs",
       variables = c("total_features", "log10_total_counts",
                     "pct_exprs_feature_controls", "tissue_type", "batch"))
```

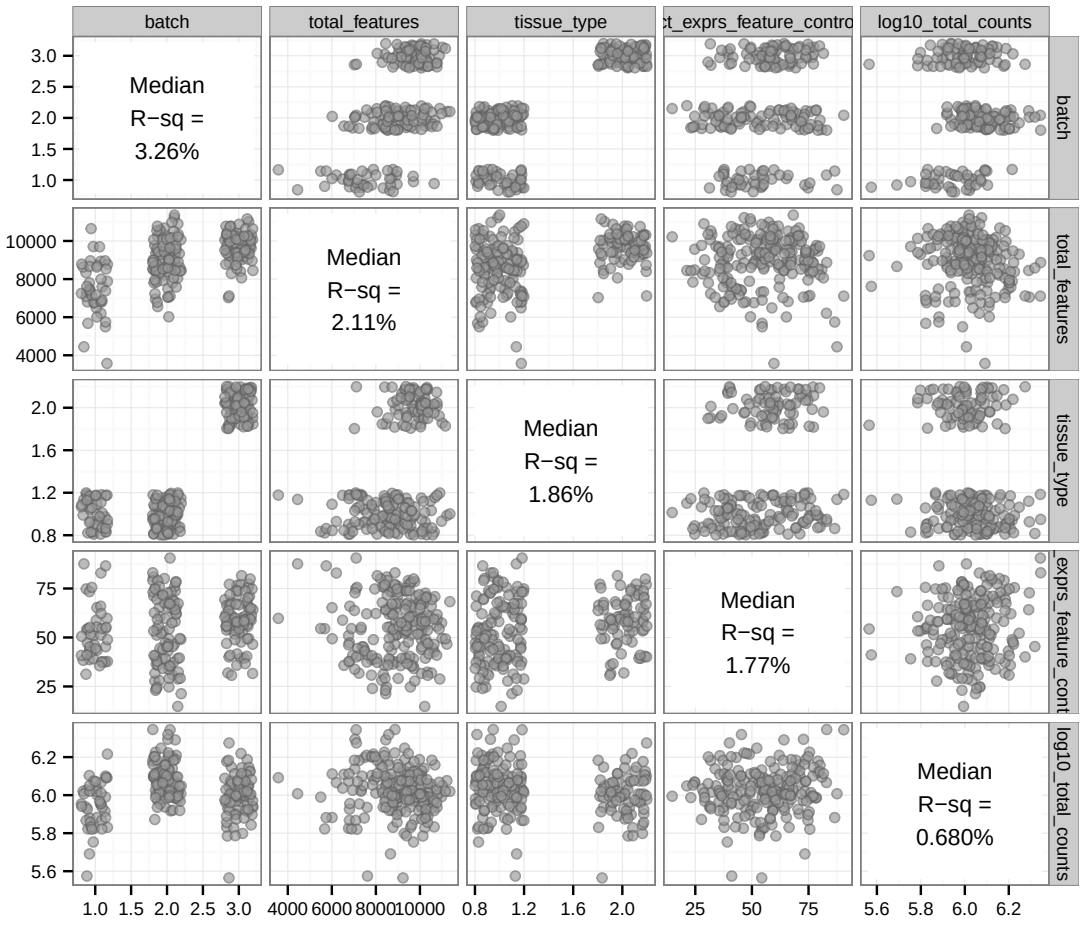


Figure 5.11: Scatter plots (pairs plot) of five explanatory variables ranked by their median marginal R^2. Jittering is used for categorical variables. Each point represents a cell. The variables shown here, from top-left to bottom-right are "batch" (experimental batch), "total features" (number of genes with non-zero expression), "tissue_type" (either colon or ileum), "pct_exprs_from_feature_controls" (percentage of total expression of the cell from feature controls, here ERCC spike-ins and mitochondrial genes) and "log10_total_counts" (sequencing depth, or total counts, on the log10-scale).

One can also easily produce plots to identify principal components that correlate with experimental and QC variables of interest. The function `plotQC` with the option `type = "find-pcs"` ranks the principal components in decreasing order of R^2 from a linear model regressing PC value against the variable of interest. The default behaviour is to show the relationships between the variable of interest and the six principal components with the strongest relationship to the variable (as measured by R^2). This works both for continous and categorical variables (Figure 5.12).

After important explanatory variables have been identified with the tools shown above, their effects can be accounted for in subsequent statistical models, or they can be conditioned out using `normaliseExprs`, if so desired. If a design matrix incorporating a selection of explanatory variables is supplied as an argument to `normaliseExprs`, then normalised expression values returned for each feature will be the residuals from a linear model fitted with the design matrix, after any size-factor normalisation has been applied to the expression data. This functionality is not shown here due to space constraints.

Thus, after convenient pre-processing, QC and normalisation with SCATER, the data are well organised (with feature and cell metadata and many data transformations), clean and tidy, and are ready for further statistical modeling and analysis.

5.4 Data visualisation

Visualising the data and metadata in various ways is crucial for exploring and understanding a dataset, as shown in Section 5.3. In that section, I demonstrated the `plotQC` function to produce various specific types of plot useful for QC. I also introduced `plotPhenoData` for convenient plotting of cell phenotype information (including QC metrics), and the function `plotFeatureData` for plotting feature information. Finally, I used the `plotTSNE` function to produce reduced-dimension t-SNE plots to identify potentially problematic cells to filter out of the dataset. The SCATER package has further plotting capabilities for exploring and analysing aspects of a single-cell RNA-seq dataset.

This section demonstrates more of the SCATER package's suite of plotting functions to produce informative and attractive exploratory, summary and diagnostic plots. I use the Cell Cycle Data as well as the Simmons Data to demonstrate further SCATER's capabilities. As described in detail in Section 5.2.2, the Cell Cycle Data consists of 348 HSCs from primary mouse tissue and cell lines (human and mouse). There are libraries from 144 cell-line cells, 187 primary mouse cells (114 harvested from mice at 2 weeks of age, 73 at 6 weeks) and 17 blank wells. Human and mouse cell-line cells were mixed together before capture on C1 chips, so the specific organism of origin is not known for these cells. Thus, I produced two versions of the Cell Cycle Data, taking the raw RNA-seq reads and quantifying transcript abundance with KALLISTO using the human transcriptome and, also, the mouse

```
p1 <- plotQC(sce_simmons_filt, type = "find-pcs",
             variable = "pct_exprs_feature_controls")
p2 <- plotQC(sce_simmons_filt, type = "find-pcs",
             variable = "tissue_type")
plot_grid(p1, p2, labels = c("A", "B"), nrow = 2)
```

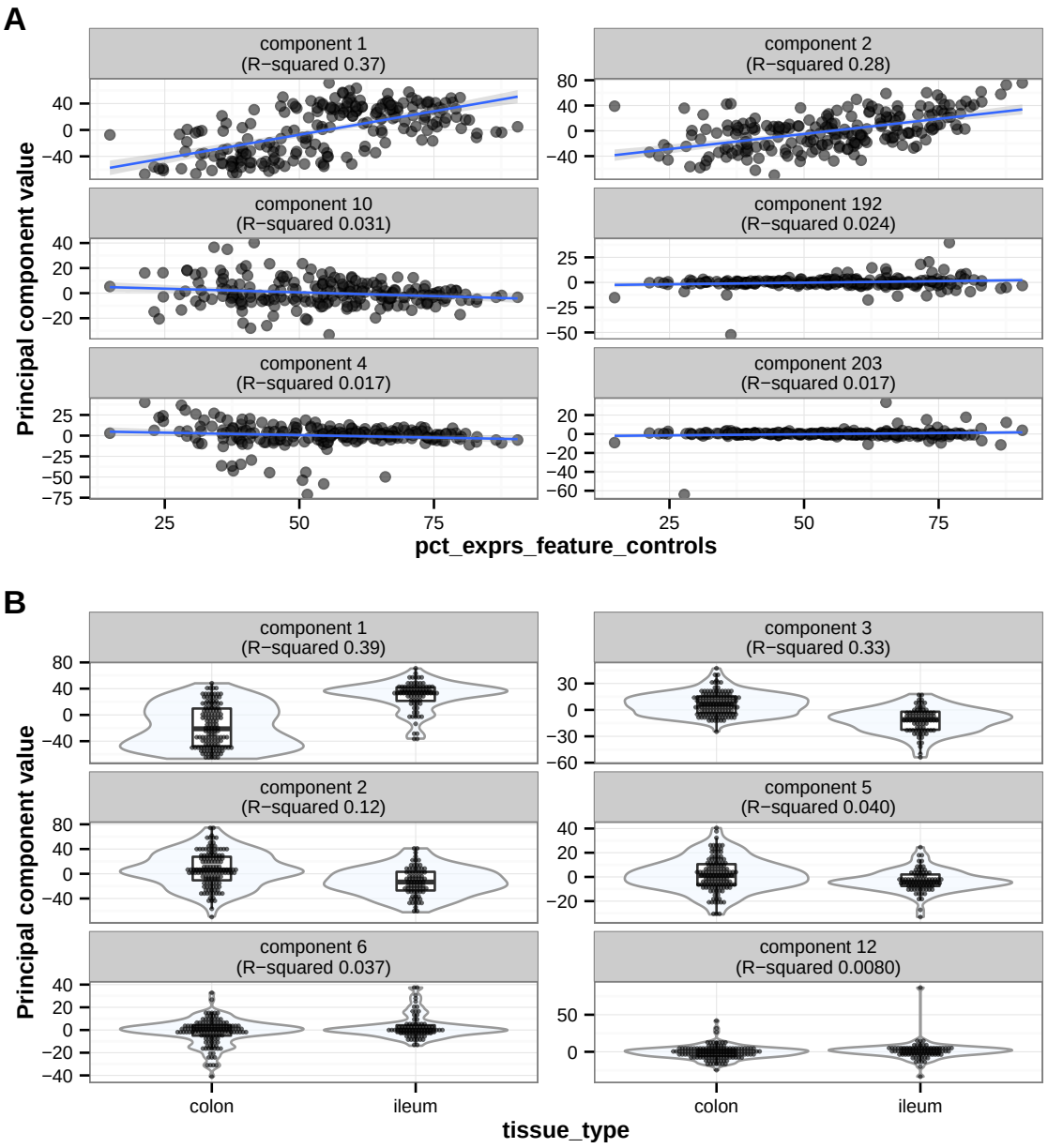


Figure 5.12: Plots of values for the top six most-associated principal components for (A) percentage of expression from feature controls, and (B) tissue type, against the values for the variable of interest. Principal components are ranked by their R^2 when the PC values are regressed against the variable of interest in a linear model. A scatter plot is used for continuous variables, with a fitted line from a linear model (with 95% confidence interval) shown. For categorical variables, a violin plot overlaid with a boxplot and a dotplot showing the actual PC values is plotted.

transcriptome (Ensembl release 80, in both cases), including ERCC spike-in sequences. Thus, I have a "human" version and a "mouse" version of the whole dataset.

I prepared the Cell Cycle Data for the results presented in this section by following the same procedure for pre-processing and QC as for the Simmons Data in Section 5.3, in parallel, for the mouse and human versions of the data. For the primary mouse cells, cell-cycle phase was identified by Ben Povinelli using fluorescence values from Hoechst 33342 staining (which stains cellular DNA) and pyronin-y staining (which stains RNA). Thus, we can explore differences between mouse and human cell line cells and primary mouse cells in different phases of the cell cycle with visualisation tools available in SCATER.

In particular, I will look at:

1. Cumulative expression plots;
2. Exploring cell-type structure with reduced-dimension representations
3. Using gene sets from *a priori* knowledge with projection plots

A further important function not previously mentioned, is `plotExpression` for plotting expression levels (using any available transformation of the data) against any of the cell phenotype variables. The function has options to use cell phenotype variables to define the colour, size and shape of points plotted. Often, in the course of an analysis it is necessary to inspect the expression levels of a feature or set of features in full detail, instead of relying only on summary information and plots. The `plotExpression` function in SCATER enables the user to do this conveniently, with great flexibility (see Figure 5.19).

5.4.1 Cumulative expression plots

Boxplots are a standard method used to gain an overall impression of expression distributions across samples in microarray and bulk RNA-seq datasets. However, boxplots do not work well for single-cell data. For many cells, the median expression level for features, especially before any feature-level filtering is done, is zero. It is not uncommon for the 75th percentile of expression values also to be zero for a substantial number of cells. Thus, boxplots are not particularly informative when trying to obtain an overview of expression across the whole dataset. Instead, plotting the cumulative expression of the most-expressed features for each cell provides a better overview of expression distributions across all cells. In scRNA-seq data there is a very long tail of low-expression and zero-expression observations, so it is more meaningful to focus on the expression of the most-highly expressed features in a dataset.

In SCATER, the default plot method for an SCESet object produces a cumulative expression plot, so simply applying the function `plot` to an SCESet object produces this useful overview of the dataset. By default, the cumulative proportion of each library accounted

```
plot(sce_simmons, block1 = "batch", colour_by = "sample_information") +
    theme(legend.position = "top")
```

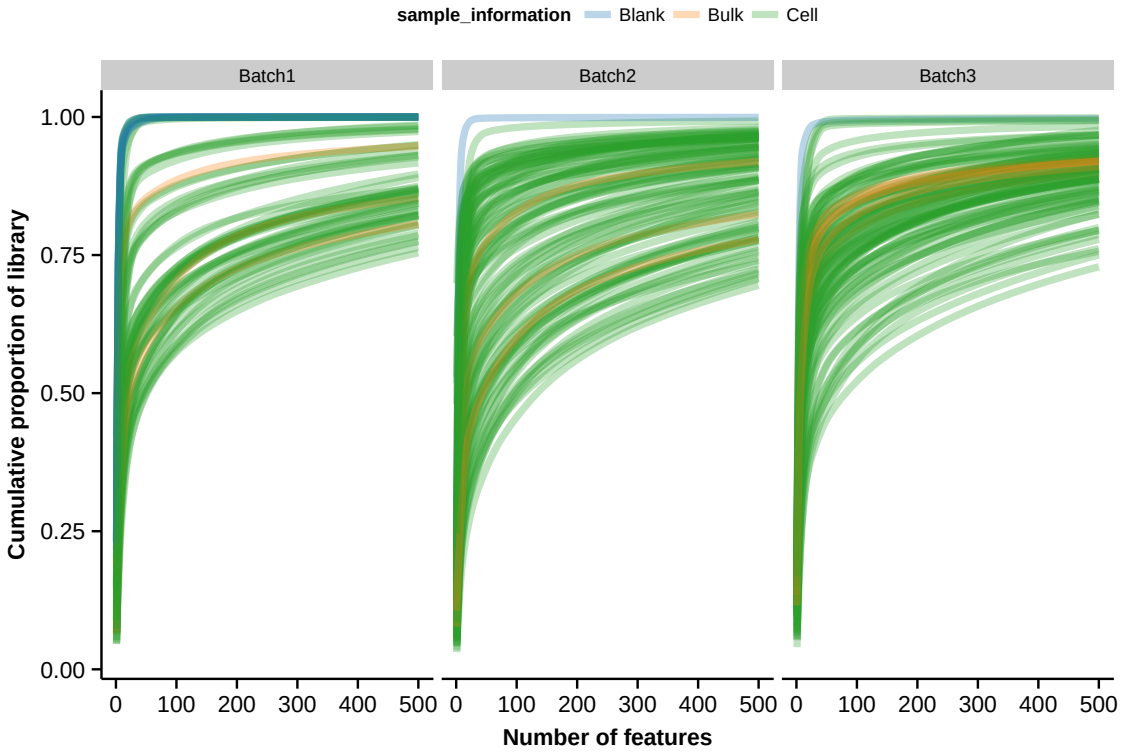


Figure 5.13: The `plot` method for SCESet objects plots the cumulative proportion of each library's total expression accounted for by the 500 highest-expressed genes. The "block1" and "block2" arguments can split the cells to be plotted onto multiple panels, while the "colour_by" argument can be used to define different colours for different cells. The plot method accesses variables defined in the phenotype data slot of the SCESet object.

for by the top 500 most-expressed genes is plotted. For the Simmons Data (Figure 5.13), there is substantial variability in use of "sequencing real estate" (or transcriptional complexity) between cells across the three experimental batches. The 500 most-expressed features (here, genes) account for between 70% and 100% of a cell's total expression. For the blank "cells", fewer than 50 genes account for practically all of the expression; ERCC spike-ins and mitochondrial genes as shown by Figure 5.5. There are some cells with profiles very similar to the blanks, suggesting the failure of these cells at some point in the experiment.

Cell phenotype variables can be used as "blocking factors" to split the cumulative expression plot into distinct "facets", showing different sets of cells (here, cells are faceted by experimental batch), as shown in Figure 5.13. Cell phenotype variables can also be used to colour cumulative expression curves (here, curves are coloured by the type of the sample, "Blank", "Bulk" or "Cell"), helping to highlight differences in expression distributions for different types of cells.

Cumulative expression plots are a useful innovation for scRNA-seq data and should take over from boxplots as a default method for gaining an overview of expression distributions across all cells in a dataset.

5.4.2 Exploring cell-type structure with reduced-dimension representations

To understand relationships, similarities and differences between cells it can be helpful to visualise the cells in a reduced-dimension space. With expression values typically available for thousands to tens of thousands of features, it is not possible to visualise cells in "expression space" with thousands of dimensions. Reducing the dimensionality of the cell information enables visualisation of cell relationships in a small number of dimensions, which can be plotted. Often just plotting the first two dimensions of a reduced-dimension representation of cells can highlight important structure in the data. Dimension reduction is not only useful for visulaisation: some downstream analyses, for example clustering and pseudotemporal ordering methods (see Section 5.1.2), are better in a reduced space. Reduced-dimension representations can be used for cell filtering, for example by identifying cells that cluster with blank wells and are likely to be problematic. Once outlier or otherwise problematic cells have been identified they can be filtered out of the analysis, as in Section 5.3.6.

The SCATER package makes it easy to conduct principal components analysis (PCA; see Jolliffe, 2014, for example) and t-distributed stochastic neighbour embedding (t-SNE; Van der Maaten & Hinton, 2008). PCA is arguably the most widely-used method for dimension reduction across many fields, and t-SNE is a newer dimension-reduction method that has been proven to be an excellent tool for visualising high dimensional data (Van der Maaten, 2009; Van der Maaten & Hinton, 2012). It has recently been successfully applied to scRNA-seq data (Amir et al., 2013; Bendall et al., 2014; Macosko et al., 2015). In Section 5.3.6, above, I demonstrated the utility of visualising cells in reduced-dimension spaces (specifically, using t-SNE) for identifying cells to filter out of the dataset. More generally, reduced representations of cells enable the exploration of cell population structure.

Again, SCATER provides simple, but flexible, ways to visualise cells in reduced dimensions directly from an SCESet object, with the functions `plotPCA` (for PCA plots), `plotTSNE` (for t-SNE plots) and `plotReducedDim`, which will plot whichever values are stored in the `reducedDimension` slot of an SCESet object. This final approach means that any reduced-dimension representation of cells (for example, an independent component analysis produced by MONOCLE or similar) can be stored in an SCESet object and plotted conveniently. I include PCA methods in SCATER because PCA is a very widely used and well-understood method that many users will be comfortable using.

While less familiar than PCA, t-SNE tends to produce better visualisations. However, t-SNE is a stochastic method, which means that plots can change, typically in a minor

fashion but occasionally in a major way, with repeated runs. In practice, I have found it valuable to set a "random seed" for t-SNE plots (done for the t-SNE plots in this chapter) to make them reproducible. Unlike PCA, which is deterministic and will only produce one visualisation from a given dataset, t-SNE has several parameters that can be adjusted. Most important is the perplexity parameter. The perplexity parameter can be loosely thought of as the expected number of neighbours for each datapoint and its value can be experimented with to improve visualisations (Van der Maaten & Hinton, 2008). The default plot for the above methods shows the first two components. If any cell controls have been defined, then those cells are plotted in a different colour. For all of SCATER's reduced-dimension plotting functions, cell phenotype variables can be used to define the colour, shape and size of points on the plot, which is convenient for exploratory analyses.

Using t-SNE is particularly good for cell filtering and exploring cell populations, as it tends to do a better job at placing similar cell types close together in the reduced-dimension space. The idea forming the basis of stochastic neighbour embedding is converting high-dimensional Euclidean distances between datapoints (in our context, cells in expression space) into conditional probabilities that represent similarities. The idea of measuring similarity between datapoints using conditional probabilities is extended to the "map points", the low-dimensional representation of the cells, and t-SNE aims to find a low-dimensional data representation that minimises the mismatch between the conditional probabilities in the high-dimensional space and the conditional probabilities in the low-dimensional space (Van der Maaten & Hinton, 2008). Producing two-dimensional representations of high-dimensional data is inherently challenging because pairwise distances in a two-dimensional map cannot faithfully model distances between points in the high-dimensional space due to the "crowding problem". The crowding problem is that the area of the two-dimensional map available to accommodate moderately distant datapoints will not be nearly large enough compared with the area available to accommodate nearby datapoints. Very briefly, t-SNE generally produces better two-dimensional visualisations of high-dimensional data than PCA or other approaches because it better solves the "crowding problem" (see Van der Maaten & Hinton, 2008, for detailed discussion).

For the Simmons Data, the first dimension of the t-SNE plot clearly distinguishes between cells from ileum and colon samples (as seen in the previous section), whereas PCA needs a combination of the first two principal components to distinguish between the tissue types (Supplementary Figure A.5). Both PCA and t-SNE do a good job of clustering blank wells (and with them, presumably failed cells).

Observing separation of colon and ileum cells is good, but a more subtle problem is distinguishing between human and mouse cell-line cells in the Cell Cycle Data. One basic attempt to determine the species of origin of the cell-line cells (conducted by Eshita Sharma and John Broxholme) was to take 100 random sequenced reads for each cell and

align them against sequences in the BLAST "nr" (non-redundant) database (Altschul et al., 1990; Johnson et al., 2008). The best species match was recorded for each read, and the species for the cell was assigned based on which species had the most "votes" from the 100 reads. In some cases, this approach identified with confidence which species the cell was most likely to have come from, but in many cases it did not, with 30–40% of reads mapping best to human and a similar proportion maping best to mouse. Thus, this approach was not particularly effective, and it may be possible to do better at determining the species of origin for the cells and uncovering other structure in the sample of cells using visualisations. Just looking at the cells from Chip 12 with t-SNE shows four distinct "populations" (Figure 5.14). Two clusters (when using either the mouse or human transcriptome data) can clearly be identified as either human or mouse cells, while two other clusters appear "mixed", possibly due to lower quality sequencing libraries being produced for these cells due to biological or technical reasons (or both). PCA plots show a similar picture, but with noisier clusters (Supplementary Figure A.6).

When all cells in the Cell Cycle Dataset are included in a t-SNE plot, the primary mouse cells separate clearly from the cell-line cells, using human or mouse data (Figure 5.15). Several distinct clusters of cell-line cells can be observed. Using PCA does a reasonable job of separating cell types but, as noted above, gives a noisier visualisation (not shown).

The `plotPCA` and `plotReducedDim` functions allow more than just the first two components to be plotted by specifying the `ncomponents` argument (Supplementary Figure A.7). When more than two components are plotted, the boxes on the diagonal in the scatter plot matrix show the density for each component. If so desired, the top principal components can be added to the reducedDimension slot, as so:

```
sce_simmons <- plotPCA(sce_simmons, ncomponents = 30, return_SCESet = TRUE,
                       draw_plot = FALSE)
```

The function `reducedDimension` (as well as the shorthand `redDim`) can be used to access and assign the reduced dimension coordinates. This means that any other dimension reduction method can be applied and the coordinates stored. For example, we might wish to use diffusion maps (Haghverdi et al., 2015) or Gaussian process latent variable models (Buettner et al., 2015) or any other dimensionality reduction method. One can store these in the SCESet object and produce plots just as with PCA and t-SNE, using the `plotReducedDim` function. By default, for both t-SNE and PCA, feature expression is scaled to have mean of zero and unit variance, but both can alternatively be done without variance scaling by changing the "scale_features" argument to `plotTSNE` and `plotPCA`. For the Simmons Data, not scaling the expression data makes only a slight difference (rotation) to the visualisation produced (not shown), but with other data the effects could be greater.

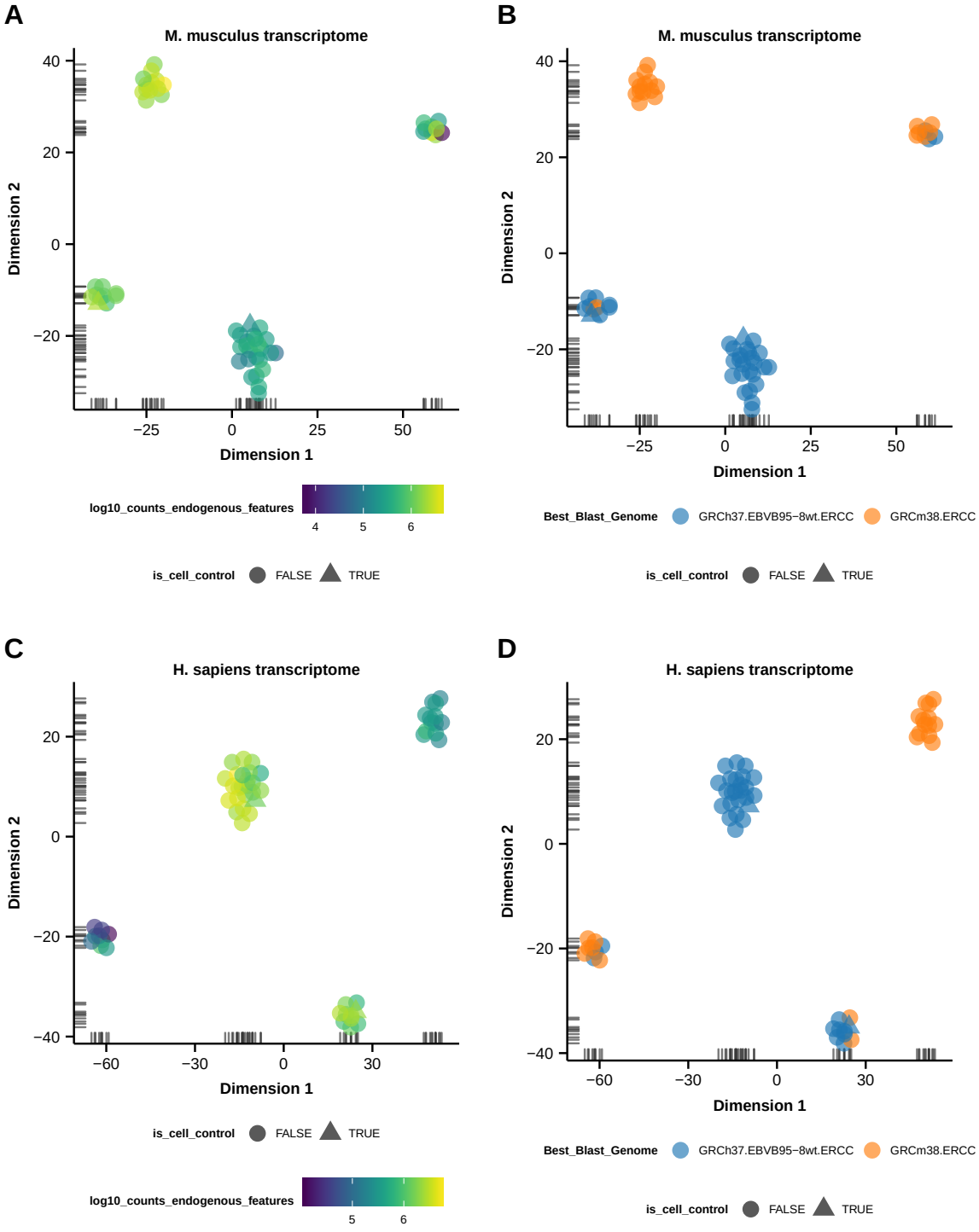


Figure 5.14: Scatter plots of the first two components from a t-distributed stochastic neighbour embedding of the cells from Chip 12 of the Cell Cycle Data are plotted with points coloured by (A,C) log-10 counts from endogenous genes (i.e. non control genes) tissue type and (B,D) best genome hit from blasting 100 random reads from the library against the non-redundant BLAST database, when using expression quantities obtained with the mouse transcriptome (A,B) and the human transcriptome (C,D).

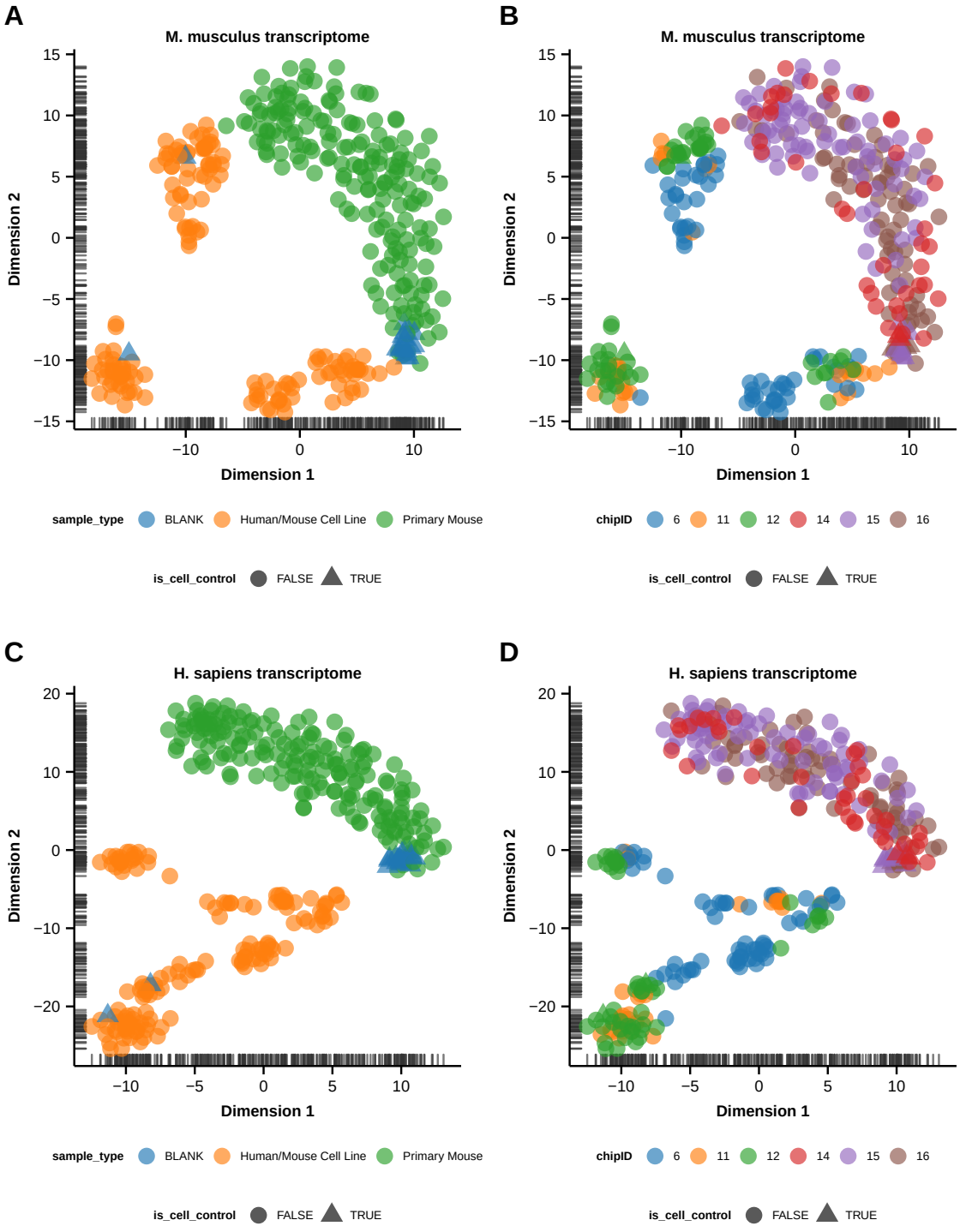


Figure 5.15: Scatter plots of the first two components from a t-distributed stochastic neighbour embedding of all cells from the Cell Cycle Data are plotted with points coloured by (A,C) sample type and (B,D) chip ID using the mouse transcriptome (A,B) and the human transcriptome (C,D).

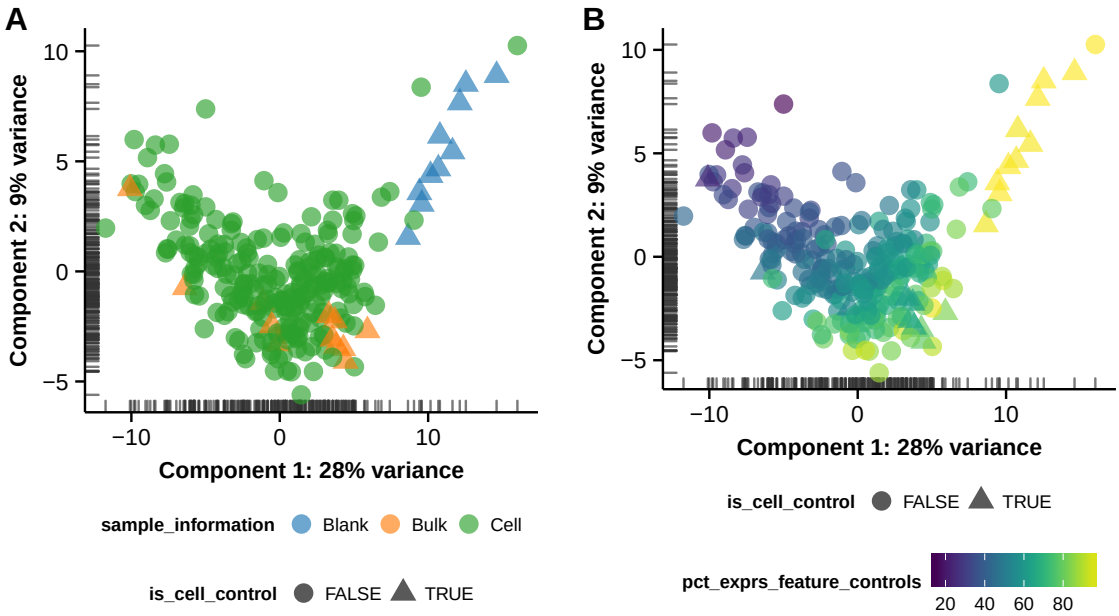


Figure 5.16: Scatter plots of the first two principal components of the Simmons Data computed only using feature controls with points coloured by (A) tissue type, and (B) percentage of expression accounted for by feature controls.

As the plots shown in this section demonstrate, reduced-dimension representations of the cells can be very informative. Providing simple, flexible functions for producing PCA and t-SNE visualisations in SCATER is thus of great utility for exploring cell populations in an scRNA-seq dataset. Furthermore, SCATER provides the means to store and plot any reduced-dimension representation of cells.

5.4.3 Using feature sets from *a priori* knowledge with reduced-dimension plots

Thus far, the plots of reduced-dimension representations of cells (or projection plots) have used all available features in the SCESet object. By default, the PCA and t-SNE plots are produced using the 500 features with the most variable expression across all cells, which can be changed with the `ntop` argument. This default approach produces general reduced-dimension visualisations of the cells that are agnostic towards any particular sources of variation. This approach is useful for exploring large-scale similarity and difference between cells. However, in many settings we will want to investigate specific processes affecting cell-to-cell similarity. Applying *a priori* knowledge to define feature sets of interest for particular processes can be highly informative. For example, Scialdone et al. (2015) recently found that using prior knowledge to define feature sets is vital for exploring processes like cell cycle, which can have substantial effects on single-cell expression measurements (Buettner et al., 2015).

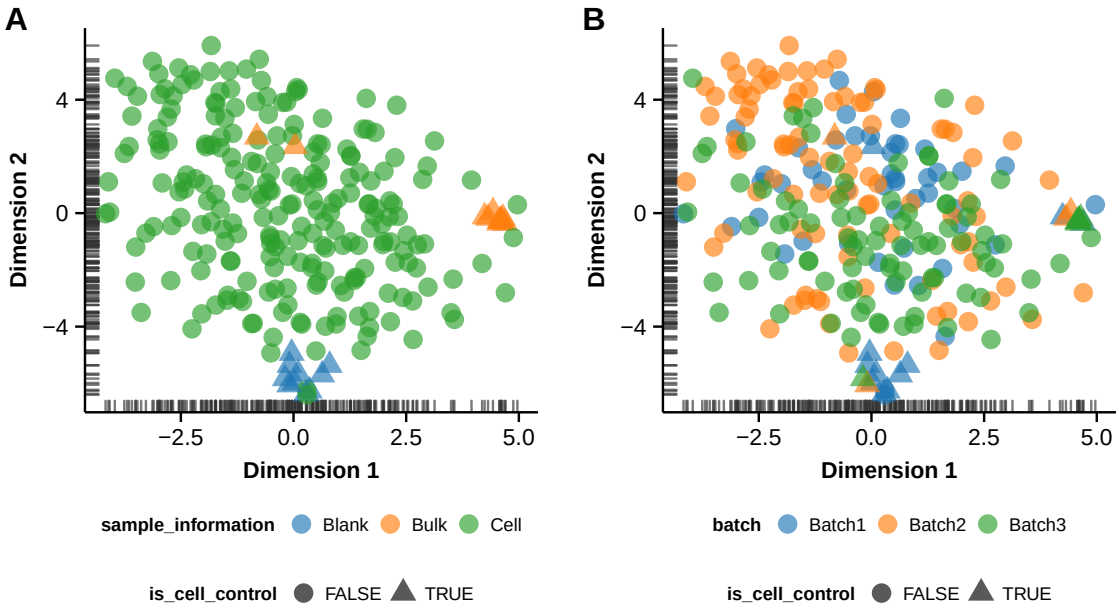


Figure 5.17: Scatter plots of the first two t-SNE dimensions for the Simmons Data computed only using feature controls with points coloured by (A) sample information, and (B) batch.

With SCATER, any specific set of features based on prior knowledge can be used for PCA or t-SNE. A feature set to use can be defined by supplying the `feature_set` argument to `plotPCA` or `plotTSNE`. This allows, for example, using only housekeeping features or control features or cell cycle genes to produce reduced-dimension plots. In Figure 5.16, only the features previously defined as "feature controls" for the Simmons Data, namely ERCC spike-ins and mitochondrial genes, are used for the PCA. When restricted to feature controls, the tissue types are no longer separated by the first two principal components, but there is still a strong relationship visible between the principal components and the percentage of expression from feature controls. Similarly, in t-SNE plots for the Simmons Data using only the features defined as "feature controls" the tissue types and three batches are no longer separated in the first two dimensions (Figure 5.17).

In the Cell Cycle Data there are cells in different phases of the cell cycle, as described briefly above. A result of this is that the RNA content of cells differs greatly between cells. This is one feature of single-cell data that often ought to be accounted for in a normalisation procedure, and one that can be assessed by looking at how gene expression distributions for sets of genes change after normalisation. I use a set of known cell-cycle genes taken from Cyclebase 3 (Santos et al., 2015) and augmented with a set of quiescence genes and lineage differentiation genes from MSigDB (Liberzon et al., 2011) with additional curation from the literature by Ben Povinelli. I load these gene sets, which I previously saved into an `RData` object, and define an SCESet with only the primary mouse cells. I now explore the cell population structure shown when using only the cell-cycle genes compared with

using all genes. I also show results before and after normalising TPM expression values to ERCC spike-ins as done for the Simmons Data in Section 5.3.7.

Reduced-dimension plots produced using a specific gene set can provide a clearer view of processes operating in a data set. Here, using only cell-cycle genes to produce t-SNE and PCA plots produces much clearer separation of cells in different phases of the cell cycle than the corresponding plots using all genes (Figure 5.18). In both the t-SNE and PCA plots when using cell cycle genes, most of the G2/M and S phase cells cluster together and are distinguishable from the G0 and G1 cells. However, even when using only cell cycle genes it is difficult to distinguish between G0 and G1 cells, a fact that reflects the underlying biology. When using all genes, distinct clusters for the different cell phases are not obtained.

The simple normalisation approach does have a useful effect on adjusting expression levels. Before normalisation there is a noticeable downward trend in average expression levels when looking at cells in the G0, G1, G2/M and S phases (Figure 5.19A). This effect is likely driven by the fact that cells in the G0 and G1 phases will have lower RNA content than cells in the G2/M and S phases. Of course, the ERCC spike-ins should show, on average, the same expression levels across all cells. After normalisation this is closer to being the case, with similar distributions achieved for ERCC spike-ins across genes in the four defined cell phases (Figure 5.19B). Thus, normalising to expression levels of ERCCC spike-ins does seem to account for this RNA content effect.

Looking at expression densities for cells before and after normalisation shows more consistency across cells for the expression distributions of the ERCC genes, but large differences in expression distributions including all genes for a number of cells (Figure 5.20). After normalisation, expression distributions across cells are much more similar than before normalisation, where there are numerous cells with expression distributions very different from the majority of the cells. For the well-behaved Simmons Data, with a relatively high degree of consistency between cell libraries, the normalisation to ERCC controls did not drastically change the relationships between cells. For the Cell Cycle Data's primary mouse HSCs, where we observe much larger differences in expression profiles between cells (likely driven, it appears, by cell cycle effects), normalisation to ERCC controls has a large effect on expression distributions, stabilising expression distributions to a large extent between cells (Figure 5.20).

After data normalisation, the t-SNE and PCA plots generated using only the cell cycle genes look qualitatively very similar to those produced with data before normalisation (Figure 5.21A,C). However, the t-SNE and PCA plots generated using all genes after normalisation (Figure 5.21B,D) show a little more separation of cells in different phases of the cell cycle than in the corresponding plots before normalisation (Figure 5.18B,D). These results suggest that even applying this simple normalisation method may usefully remove

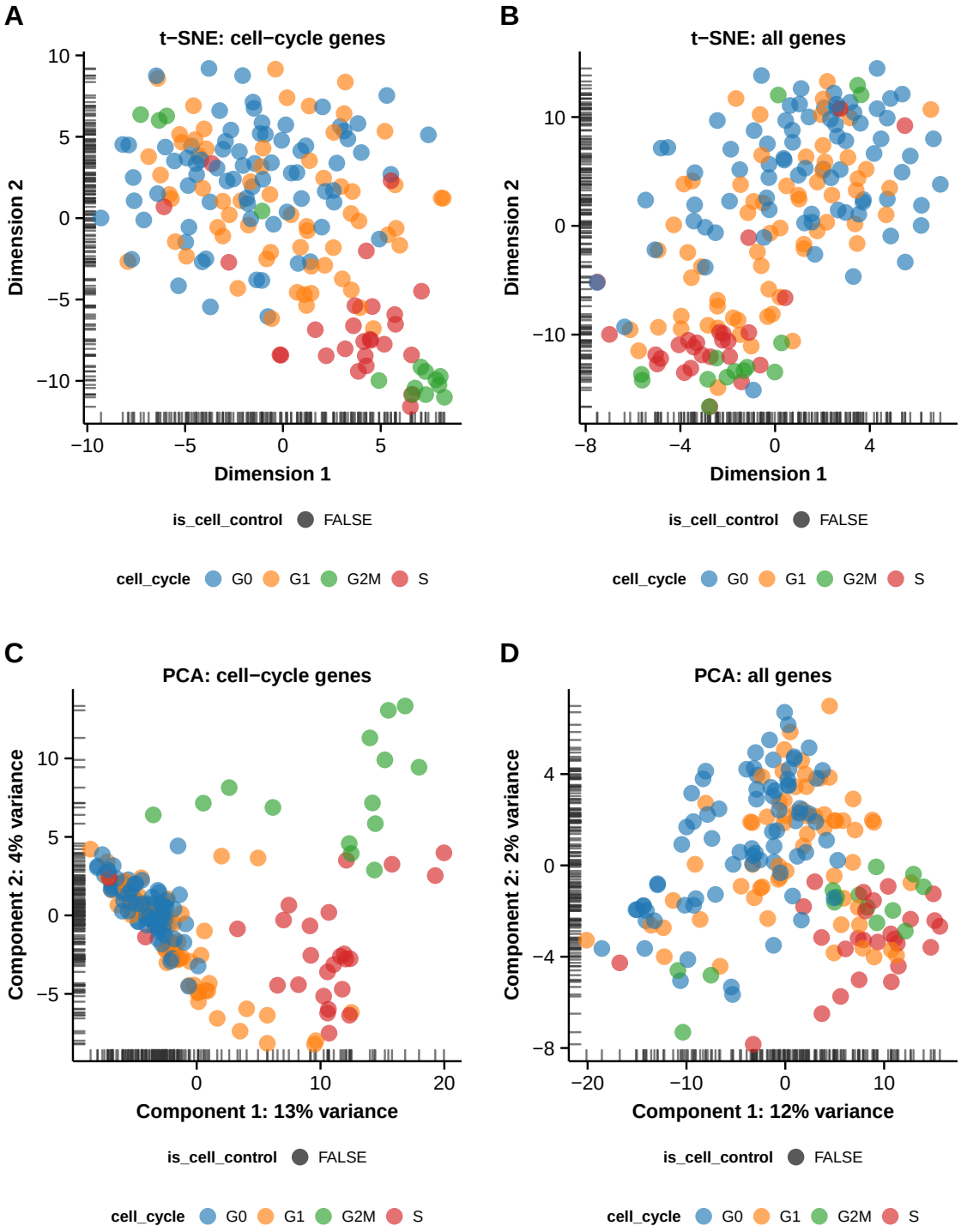


Figure 5.18: For the primary mouse cells in the Cell Cycle Data, t-SNE plots using log2-transformed TPM values before normalisation using (A) only cell cycle genes and (B) all genes. PCA plots after ERCC TPM normalisation using (C) only ERCC genes and (D) all genes. Cells are coloured by observed cell cycle phase, determined by Hoechst and pyronin-Y staining.

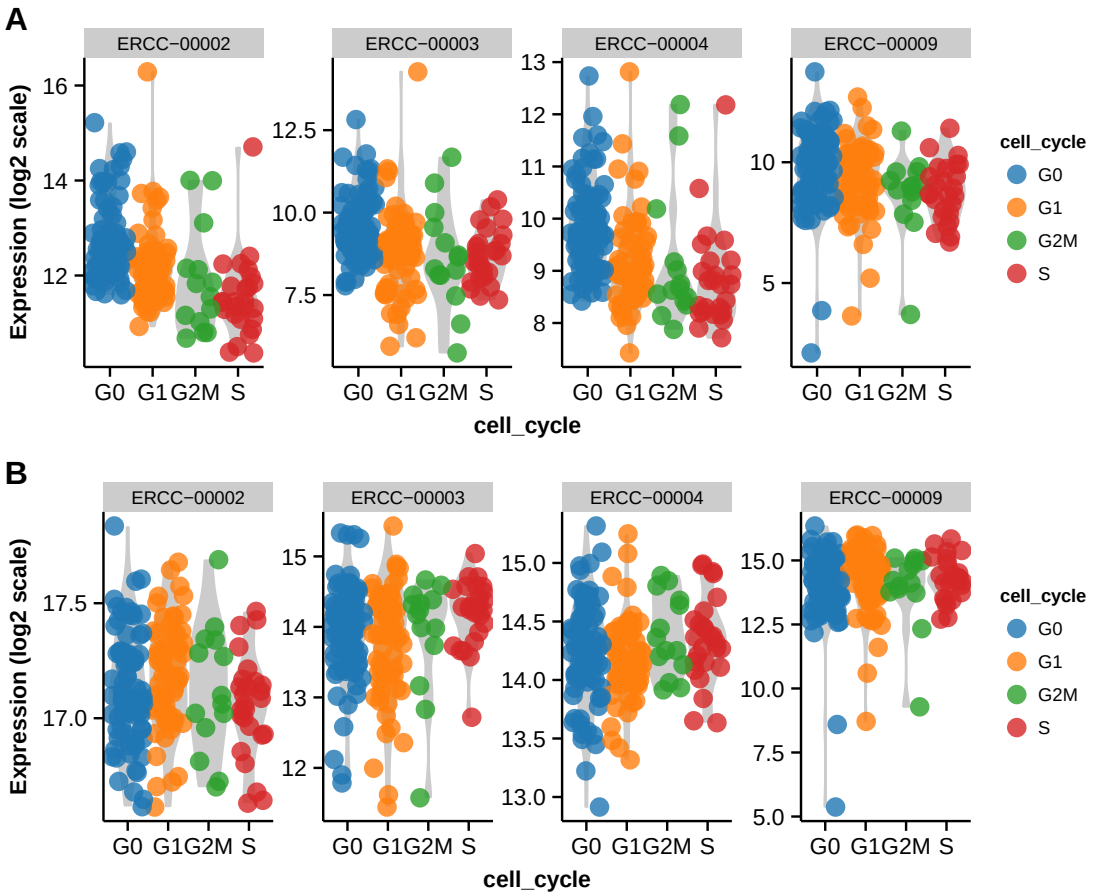


Figure 5.19: Expression plots for four ERCC spike-in genes for the primary mouse cells in the Cell Cycle Data (A) before ERCC TPM normalisation, and (B) after normalisation.

some unwanted variation from the data, emphasising differences between cells in different phases of the cell cycle.

5.5 Software and data integration

To be as effective as possible, contemporary bioinformatics tools ought to build on existing tools, operate in familiar environments and interact and integrate with other tools offering complementary functionality. The SCATER package achieves these goals as it is written in the R language, possibly the most popular environment for statistical computing (at least in the life sciences), and builds on many general statistical R packages and other Bioconductor packages focused on bioinformatics. The package is also able to integrate many different forms of data useful for scRNA-seq data, enabling better data organisation and thus more efficient and reproducible analyses.

There are two subsections in this section. The first subsection discusses SCATER's integration with other software. It describes: (1) the other software packages upon which SCATER builds directly and other packages that provide methods that could be integrated

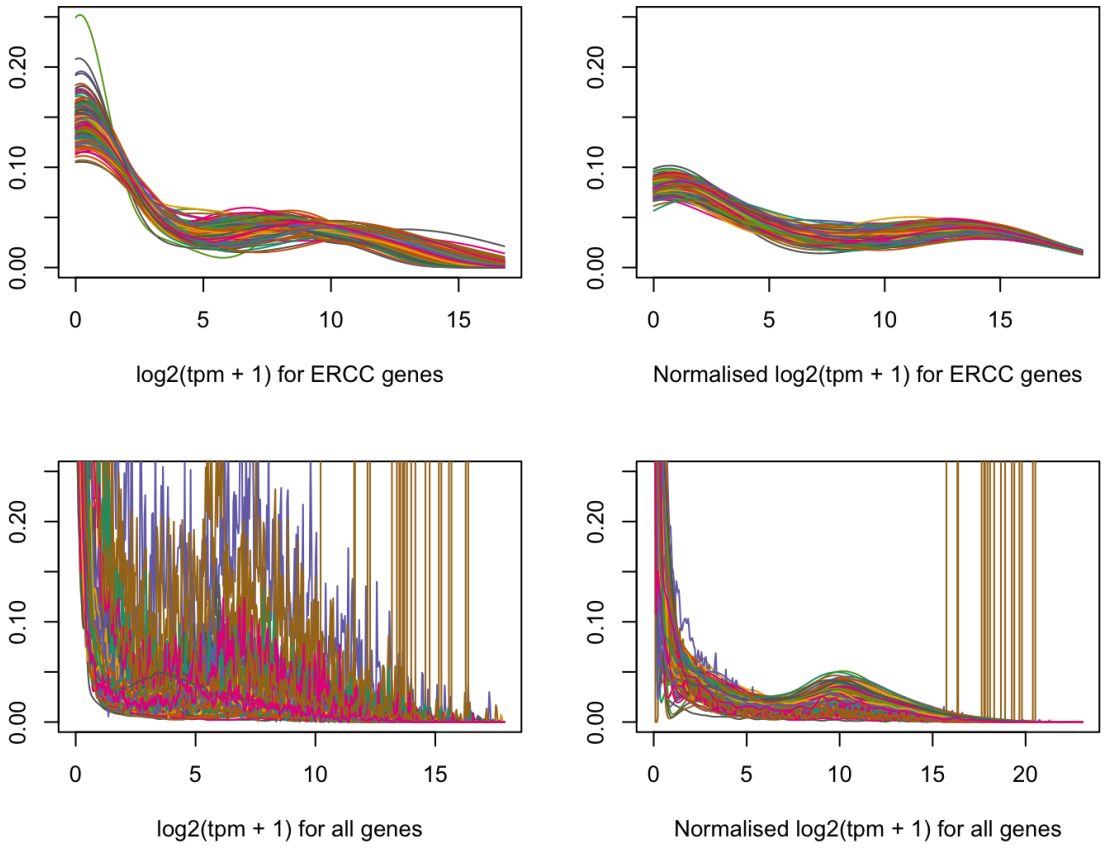


Figure 5.20: Density plots for ERCC genes (top row) and all genes (bottom row) when using log2-transformed TPM values (left column) or ERCC-TPM normalised log-transformed TPM values as expression values. Each curve represents the expression distribution for one primary mouse cell in the Cell Cycle Dataset.

with SCATER and be incorporated into a SCATER workflow; (2) the SCESet class and its advantages in further detail; (3) rapid quantification of transcript abundance using SCATER's wrappers to KALLISTO; and (4) details of SCATER's automated QC output. The second subsection describes SCATER's integration of scRNA-seq data with other data modalities.

5.5.1 Integration with other software

5.5.1.1 Building SCATER on R and Bioconductor

As previously mentioned, SCATER is an R package that builds on the R (R Core Team, 2015b) and Bioconductor (Gentleman et al., 2004; Huber et al., 2015) ecosystems of statistical and bioinformatic tools. The SCATER package uses many recent packages for best-practice functionality.

I build on numerous packages from the Comprehensive R Archive Network (CRAN) that provide general tools for statistical computing. In particular, I make use of the `ggplot2` (Wickham, 2009, 2011a), `ggthemes` (Arnold, 2015), `cowplot` (Wilke, 2015), and `Rtsne` (Kri-

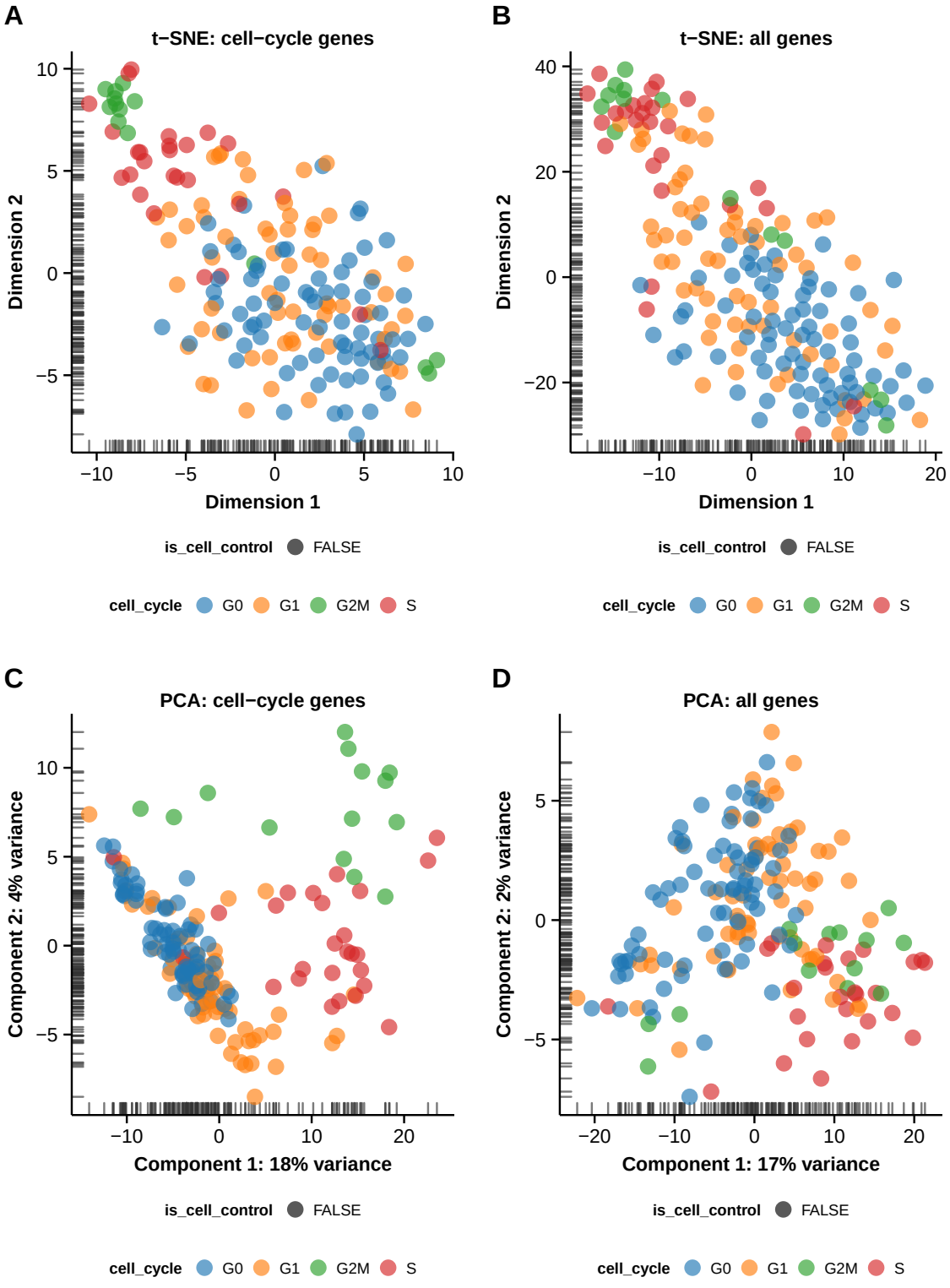


Figure 5.21: t-SNE plots using log2-transformed TPM values after ERCC TPM normalisation using (A) only cell cycle genes and (B) all genes. PCA plots after ERCC TPM normalisation using (C) only cell cycle genes and (D) all genes. Cells are coloured by observed cell cycle phase, determined by Hoechst and pyronin-Y staining.

jthe, 2015) for producing the flexible, attractive and user-modifiable plots shown earlier in this chapter. For reading in and manipulating data, SCATER uses `data.table` (Dowle et al., 2014), `dplyr` (Wickham & Francois, 2015), `plyr` (Wickham, 2011b), `reshape2` (Wickham, 2007), `rhdf5` (Fischer & Pau, 2015) and `rjson` (Couture-Beil, 2014). I use the `matrixStats` package (Bengtsson, 2015) for several functions that allow high-speed computations on matrices, and the `parallel` package (R Core Team, 2015b) to run computations on multiple threads.

For added functionality specific to bioinformatics problems, SCATER uses many packages from the Bioconductor project. Following good programming practice to maximise code reuse (Wilson et al., 2014), I build on the foundation of the Bioconductor architecture made available in the packages `Biobase` and `BiocGenerics` (Huber et al., 2015). In particular, as discussed in further detail in the next section, SCATER's SCESet class builds on the Bioconductor ExpressionSet class. In SCATER, functions from the `edgeR` package (Robinson et al., 2010) are used to compute normalisation factors and counts-per-million values, the `limma` package (Smyth, 2004; Ritchie et al., 2015) is used for highly efficient fitting of feature-wise linear models and SCESet objects can be converted to CellDataSet objects for use with the `monocle` package to produce pseudo-temporal orderings of single cells (Trapnell et al., 2014).

Even with extensive use of existing R and Bioconductor packages, SCATER adds a great deal of extra functionality. The package defines approximately 75 functions. About one third of these are user-level functions and the remainder are internal. These functions comprise approximately 3,500 lines of code with 2,500 lines of documentation. To help the user take advantage of the SCATER's functionality, the package is fully documented and comes with a "vignette" explaining and demonstrating the main workflows available.

Section 5.1.2.4 reviewed the major analysis methods for scRNA-seq data. Almost all of the existing analysis tools are implemented in R, which makes it easy to incorporate them into a SCATER workflow. Furthermore, almost all of the statistical modelling methods assume a tidy, pre-processed dataset, so quality control and normalisation remains necessary, even if one wishes to do sophisticated analyses with other tools downstream. The modular design of SCATER and the convenience of the flexible SCESet class as a container for scRNA-seq datasets makes SCATER a very useful foundation for all scRNA-seq analyses. The SCESet class has slots for cell-cell and gene-gene similarity matrices (see Figure 5.4), so output from the many cell and gene clustering methods can be stored in an SCESet object and plotted, as can reduced-dimension representations of cells (discussed in the previous section). The package also provides an implementation of rank-product methods for differential expression, which can be used for identifying DE genes in a single dataset, for meta-analysis or for obtaining a consensus DE ranking from a set of DE methods (as discussed in Section 5.1.2.4).

The SCATER package thus builds on a large number of excellent R software tools and, by acting as a hub in the scRNA-seq analysis network, can facilitate the use of many more existing packages for downstream analysis.

5.5.1.2 The SCESet class and its advantages

The cornerstone of the SCATER package is the SCESet class, which I developed as the data structure for scRNA-seq data upon which all of the methods for the package are built. As well as enabling the suite of tools available in SCATER, the SCESet offers some specific advantages over alternative possible data structures for scRNA-seq data. Below, I outline some aspects of the SCESet class not covered in Section 5.1.3.

In SCATER, single-cell expression data is organised in objects of the SCESet class (Figure 5.4). The class is derived from the Bioconductor ExpressionSet class, which provides a general, structured and common interface for expression data (microarray, RNA-seq, etc), which may be familiar to those who have analysed microarray or RNA-seq experiments with Bioconductor (Huber et al., 2015). The class requires a minimum of three input objects:

1. exprs, a numeric matrix of expression values, where rows are features, and columns are cells;

2. phenoData, an AnnotatedDataFrame object (a Bioconductor class), where rows are cells, and columns are cell attributes (such as cell type, culture condition, day captured, etc.);

3. featureData, an AnnotatedDataFrame object, where rows are features (e.g. genes), and columns are feature attributes, such as gene identifier, biotype, GC content, etc.

In addition, an SCESet object can contain much more information, as summarised in Figure 5.4 and discussed below.

The requirements for the SCESet class (as with other S4 classes in R and Bioconductor) are strict. The idea is that enforcing strictness when generating a valid class object ensures that downstream methods applied to the class will work reliably. Thus, the expression value matrix *must* have the same number of columns as the phenoData object has rows, and it must have the same number of rows as the featureData dataframe has rows. Row names of the phenoData object need to match the column names of the expression matrix. Likewise, row names of the featureData object need to match row names of the expression matrix. Further sanity checks are carried out in SCATER when generating or modifying SCESet class objects and an error is returned to the user if any of the checks fail.

A new SCESet object can be constructed directly from Fastq files using the wrapper functions to KALLISTO (as demonstrated in Section 5.3.1), or by using raw count or other

data matrices with the `newSCESet` function. If count data are supplied, then the exprs slot in the SCESet object will be generated as either $\log_2(\text{TPM}+1)$ or $\log_2(\text{CPM})$. Transcripts-per-million (TPM) is the preferred unit for RNA-seq expression (bulk or single-cell), as discussed above, but calculation of TPM requires feature lengths (specifically transcript lengths; Trapnell et al., 2013). If, for example, only gene-level counts are available, or if feature lengths are unavailable for transcript-level counts, then CPM can be used instead. In SCATER, values are computed using the `cpm` function from EDGER (Robinson et al., 2010). Although not generally recommended for analysis, FPKM values can be stored in and accessed from SCESet objects. Almost all of SCATER's QC, normalisation and visualisation methods work with count, TPM, CPM, and FPKM data stored in an SCESet object, so the full functionality of the package is available for whichever units the user wishes to employ for their scRNA-seq data.

The SCESet class in SCATER has several slots that are not present in the Bioconductor ExpressionSet class, but are necessary or useful for scRNA-seq data (see also Figure 5.4):

- logged: a logical scalar indicating whether or not the expression values in the exprs slot have been transformed to the log2-scale most commonly used for gene expression data.
- logExprsOffset: a numeric scalar providing an offset value applied to expression data when undergoing log2-transformation to avoid trying to take logarithms of zero values.
- lowerDetectionLimit: a numeric scalar giving the threshold above which observations are deemed to be "expressed", that is, have a non-negligible expression value.
- cellPairwiseDistances: a numeric matrix containing pairwise distances between cells. There is no restriction on how the distances are computed, only that the matrix has the correct size.
- featurePairwiseDistances: a numeric matrix containing pairwise distances between features (usually genes or transcripts). As with the cell distances, there are no restrictions other than the requirement that the matrix is the correct size.
- reducedDimension: a numeric matrix containing a reduced dimension representation of the cells. Rows represent cells, and columns represent coordinates for components in the reduced dimensional space. For example, this slot may contain the first 20 principal components for the cells.
- bootstraps: a numeric array containing bootstrap samples of the expression (possibly count) values. If KALLISTO is used to quantify transcript abundance there is the option to compute bootstrap samples of the estimated counts. This slot exists to store any bootstrap samples produced. It has row and column dimensions the same as the

assayData slot (which holds exprs, counts and other expression assay data), and the third dimension represents the different bootstrap samples.

The slots above are utilised in many settings but are not required for a valid SCESet object. The logged, logExprsOffset and lowerDetectionLimit values can be defined by the user or computed automatically on generation of an SCESet object.

The SCESet object acts as a container for all of the different types of information that we need to associate with the expression data itself, greatly simplifying the organisation of an scRNA-seq analysis. This container can contain many transformations of the expression data (for example, count data, TPM data, FPKM data, log2-transformed data, standardised expression values, and normalised expression values), feature information, and cell metadata, along with all of the information in the slots described above. Critically, complete subsetting methods are available for SCESet objects. This may sound trivial, but it means that one can select any subset of features and cells and know, with full confidence, that the correct feature information, cell metadata, and everything else, will be subsetted as well. Instead of an R environment containing many separate objects that need to be very carefully managed, the SCESet object allows the analyst to have just one data object.

Crucially, SCESet objects have many methods assigned to them. This approach of object-orientation is important from a programming perspective, enabling "safe" programming practices based on the strict class definition of the SCESet (Chambers & Lang, 2011). Methods for SCESet objects can be more stable, and more modular (and *should* be less likely to have bugs), since the method "knows" the structure of the data object and what information it can contain. The package contains many user-level functions that are designed to work on SCESet objects (Figure 5.3 shows the most important of these). In addition, there are many specific functions for accessing data from and assigning data to slots in an SCESet object (Supplementary Table A.3).

As well as the direct advantages of the SCESet class within the SCATER package, the SCESet class allows almost seamless interaction of SCATER with other Bioconductor packages. The SCESet class "inherits" the ExpressionSet class, which means that any functions in other packages that work on ExpressionSet objects will work on SCESet objects. For example, in the previous sections I produced expression density plots using the `matdensity` function from the QUANTRO package (see FIgure 5.20 and Supplementary Figures A.3 & A.4). Using the QUANTRO (Hicks & Irizarry, 2015) function on the SCESet object "just works" because the design of the class, and the whole SCATER package, is geared towards such seamless interoperability for SCATER and other packages, particularly those in the Bioconductor project.

5.5.1.3 Rapid quantification of transcript abundance

Quantification of transcript (or gene) abundance from RNA-seq data has typically been done with an "align-and-count" strategy, as described earlier in the chapter. As RNA-seq datasets reach the scale of tens of millions of reads across hundreds or thousands of samples, however, the align-and-count strategy for abundance quantification becomes a significant bottleneck in the analysis of RNA-seq data. For single-cell RNA-seq data, typical datasets currently consist of millions of reads from hundreds of cells, but the largest published single-cell RNA-seq studies are already reaching the tens of thousands of cells (Macosko et al., 2015, for example). Such scale will inevitably become widespread as high-throughput technologies are further developed and adopted. Quantification strategies that require computational time measured in hours of core time per sequenced library are not suited to such large single-cell RNA-seq datasets.

An alternative to the "align-and-count" approach has recently emerged enabling extremely rapid transcript quantification using "pseudoalignment" (Bray et al., 2015) or "lightweight alignment" (Patro et al., 2015) strategies. As discussed in Section 5.1.2.1, these approaches use a probabilistic framework based on the idea of "transcript compatibiity" with sequenced reads to estimate transcript abundance. Instead of requiring each RNA-seq read to be mapped to a specific location in the transcriptome (or genome), these new methods seek to answer the question, with which transcripts is each read compatible? With apparently little or no loss of accuracy, these new approaches reduce the computational time required for transcript quantification by orders of magnitude compared with the align-and-count strategies, with low memory requirements. Results presented by Bray et al. (2015) indicate that transcript abundance quantification with KALLISTO could be over 200 times faster than the combination of TOPHAT and HTSEQ currently used in the standard pipeline by the Genomics Core at WTCHG, Oxford. My experience suggests that at least a 200-fold increase in speed seems achievable in practice on real data (data not shown as precise timings were not available from the WTCHG Genomics Core).

As demonstrated in Section 5.3.1, SCATER enables a user to conduct rapid transcript abundance quantification from raw read data within R. Currently, KALLISTO is supported in SCATER with the function `runKallisto` (parallelised using multiple threads, if available) to quantify abundance, and the results can be read into the R session, directly into an SCESet object using the function `readKallisto`. Support for SALMON will be added to SCATER in the next development cycle. As the results for the Simmons Data case study above showed, these methods in SCATER enable the user to go from millions of raw reads for hundreds of cells to transcript abundances in an SCESet object in R within a couple of hours, compared to waiting weeks for summarised count data from a bioinformatics core

facility, or using a week of cluster time. Remarkably, transcript abundance quantification for datasets of this size could even be done overnight on a laptop.

Importantly, there seems to be no downside in terms of data quality when quantifying transcript abundance with KALLISTO. On the Simmons Data, KALLISTO tends to provide higher counts than the align-and-count approach used in the WTCHG Genomics Core (Figure 5.22A). Across the full range of expression values, KALLISTO appears to extract more information out of the raw read data than the align-and-count approach. Agreement is very close for ERCC spike-in genes and mitochondrial genes, which are highly expressed (Figures 5.22D and 5.22E). For cyclin genes and eukaryotic translation factor genes, agreement between the two approaches is high for highly expressed genes, but there are large differences at low expression levels (Figures 5.22B and 5.22C). In particular, many observations with low expression according to the align-and-count method obtain counts many orders of magnitude higher with KALLISTO. Thus, KALLISTO is able to find transcript compatibility for more reads than can be assigned and counted for features with the align-and-count strategy, which discards multiply-mapping reads.

Rapid transcript quantification with KALLISTO and pre-processing and QC with SCATER enables rapid prototyping for scRNA-seq studies. The field of single-cell genomics is progressing remarkably quickly, so researchers in the field need to produce results within short time frames. However, technologies and protocols are also developing quickly and prone to all manner of failures and biases. The nature of single-cell genomics at present thus makes rapid prototyping crucial for successful scRNA-seq studies. For example, one might wish to run a pilot experiment using one 96-cell C1 chip, before proceeding (or not) with the study. The pilot may reveal experimental or technical weaknesses to be remedied before spending more time, money and effort on a larger experiment. If prototyping time, specifically data pre-processing, QC and analysis time (once reads are obtained from the sequencer), for an experiment can be reduced from weeks to days, then large benefits to researchers accrue.

5.5.1.4 Automated QC output

With extremely fast tools like KALLISTO and SALMON likely to become the standard tools for RNA-seq expression quantification in the near future, current RNA-seq quality control workflows need to change. Current RNA-seq pipelines, particularly in large bioinformatics "core" facilities, use the standard align-and-count approach to quantification, so map RNA-seq reads and typically produce QC metrics from the mapped read output. Tools like KALLISTO and SALMON will not necessarily produce equivalent output. Thus, quality control workflows for the future cannot rely on using mapped reads to define QC metrics. This expectation for the future of RNA-seq analysis drives SCATER's focus on generating

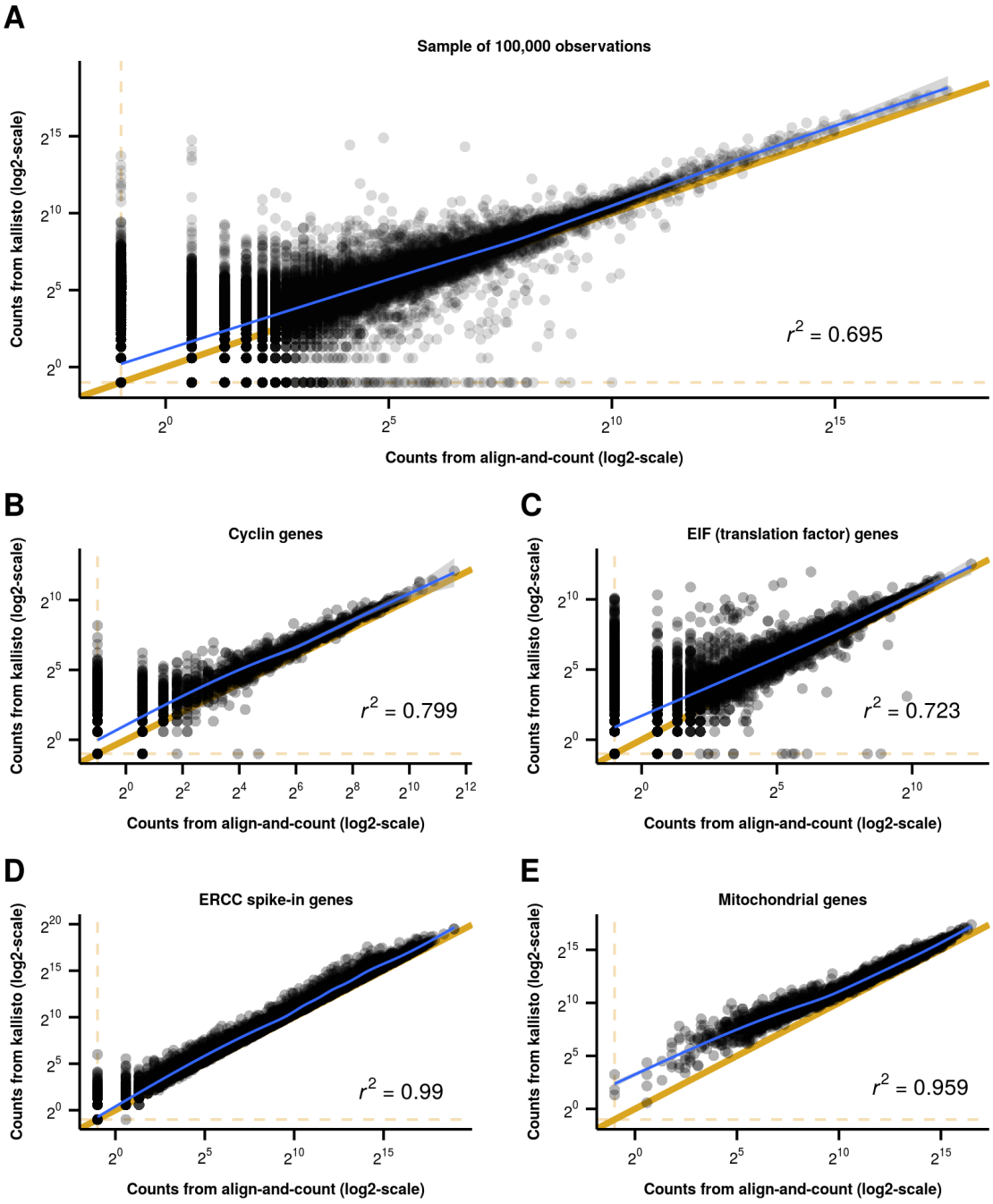


Figure 5.22: Scatterplots of gene counts for the batch 2 and batch 3 cells from the Simmons Data computed using KALLISTO (version 0.42.2) and the standard "align-and-count" approach from the WTCHG Genomics Core, which uses TOPHAT version 2.0.12 and HTSEQ version 0.6.1p1. The human transcriptome, with ERCC spike-in sequences added, from Ensembl Release 75 (Flicek et al., 2014) was used for both quantification approaches. Plots are shown for: (A) a random sample of 100,000 observations; (B) cyclin genes (involved in the cell cycle); (C) EIF family genes (eukaryotic translation factor genes); (D) ERCC spike-in control genes, and (E) mitochondrial genes. Raw counts are plotted with an offset of 0.5 on the log2-scale. Pearsons's r^2 values are shown on each plot, computed from the log-transformed counts. A smoothed fit from a generalised additive model is shown in blue on each plot.

QC metrics directly from the transcript- and gene-level abundance data, without needing read mapping information.

To make single-cell RNA-seq as seamless as possible, SCATER automatically computes a large number of QC metrics with the `calculateQCMetrics` function, as shown in Section 5.3.4. With that function, a large number of the feature-level QC metrics (Supplementary Table A.1) and cell-level QC metrics (Supplementary Table A.2) are automatically calculated. These metrics are very useful for pre-processing and quality control, as demonstrated throughout the Simmons Data case study in Section 5.3. The data analyst does not need to spend time ruminating on appropriate QC metrics to compute, or writing code to compute them. Instead, they can proceed to explore the dataset through the lens of the automatically computed metrics and the visualisation tools provided in the package (as seen in Sections 5.3 & 5.4). There is sufficient information in the feature-level expression data that quality control metrics derived from mapped reads are not required.

5.5.2 Integration with other data modalities

Integration of different data types with single-cell expression data is a key aspect of scRNA-seq studies, and one which will only become more important as single-cell genomics technologies develop further. The SCATER package currently has substantial capacity for integrating other data modalities with single-cell expression data, discussed below along with ideas for future work that would extend these capabilities further in exciting directions.

With expression data, possibly in several forms, for a large number of cells and features (transcripts or genes) stored in an SCESet object, it is usually necessary to add annotation information about the features. I showed in Section 5.3.2 how easily SCATER can integrate feature annotations with expression data. The `getBMFeatureAnnos` function allows a user to add a large range of feature information to the SCESet object. Thus, with a single function call, just about any desired feature information can be added to an SCESet object. The feature data can be obtained from an SCESet object with the accessor function `fData`. The Simmons Data case study showed that feature annotation information is useful throughout the QC process, as well as for more biologically-focused downstream analyses (not shown). Adding this information to the SCESet is convenient, as it then becomes securely "attached" to the expression data, even as subsetting of cells and features takes place.

As important as feature annotation information (if not more so) is incorporating cell metadata into an SCESet object. Any responsible experimentalist will record a large amount of experimental metadata for each cell, typically in the form of a spreadsheet or data frame. This information is crucial for quality control (the focus here) and for answering interesting scientific questions about the biology of the system under study (the focus almost

everywhere else). The accessor/assignment function `pData` makes it easy to add cell metadata ("phenotype data" in SCATER/Bioconductor parlance) to an SCESet object, as demonstrated in Section 5.3.3. Validity checks on the object make sure that the cell names of the phenotype data match the cell names of the expression data. This double-checking reduces the chance of the wrong metadata being assigned to a cell: an easy mistake, but one that can be hard to catch and have serious downstream consequences (Baggerly & Coombes, 2009; Hutson, 2010).

The SCATER package is therefore capable of integrating any kind of cell-level data with the expression data in an SCESet object. To take a particularly useful example, cell imaging data (obtainable with certain protocols; see Figure 5.2), such as Hoechst staining of cells for DNA content (Latt et al., 1975) or pyronin-Y staining for RNA content (Darzynkiewicz et al., 1986), can provide valuable information about cell viability or cell cycle (mentioned in passing for the Cell Cycle Data above). Any quantitative or qualitative information from cell imaging is, obviously, cell-level phenotype data (or metadata), so can naturally be added to the phenoData slot of an SCESet object to be safely integrated with the expression and other data for the experiment. This functionality will be particularly useful with the arrival of Fluidigm's new "Polaris" instrument, which enables more interaction with, live imaging of, and genomic assaying of, single cells (Fluidigm, 2015) and will provide even richer cell phenotype information than is currently available.

An important current trend for single-cell genomics appears to be assaying multiple data modalities from the same cells. It is already possible to sequence the genome and transcriptome of single cells (Dey et al., 2015; Li et al., 2015; Macaulay et al., 2015) in parallel, and singlecell epigenomic assays are developing rapidly, with single-cell bisulfite sequencing for assaying genome-wide methylation (Guo et al., 2013; Smallwood et al., 2014; Farlik et al., 2015) and a single-cell assay for transposase-accessible chromatin using sequencing (ATAC-seq; Cusanovich et al., 2015; Buenrostro et al., 2015a,b) already available. It seems likely that in the near future we will be able to measure genomic DNA, transcript and gene expression and multiple epigenetic data types in parallel in individual cells. Integrating such diverse data modalities into a single analytical framework is challenging, but this is a major goal for the future development of the SCATER package, discussed further in the context of future work below.

5.6 Discussion and conclusions

Single-cell RNA-seq is an exciting new technology for studying the transcriptome at the resolution of individual cells. However, with this new data type, pre-processing, quality control and data normalisation are crucial. This chapter introduced the SCATER package for pre-processing, quality control, normalisation and visualisation of scRNA-seq data. The

package fills a current gap in the scRNA-seq workflow between raw read data and clean, tidy gene- or transcript-level expression data ready for downstream analysis.

The integration of KALLISTO with SCATER enables extremely fast quantification of transcript abundance using raw read data, from within an R environment. Until early in 2015 the idea that one could process and analyse raw RNA-seq read data from hundreds of single cells in a matter of hours on a laptop was fanciful. However, the combination of SCATER and the new extremely rapid transcript-compatibility tools for transcript abundance quantification now make this possible. For researchers who have grown accustomed to transcript and gene abundance estimates for RNA-seq data to take on the order of weeks of computation time to generate (plus delays due to inevitable backlogs in data centres and bioinformatics core facilities), this increase in speed of processing and flexibility for the end user to run everything themselves expands the possibilities for scRNA-seq studies. The traditional "pipeline" of multiple bioinformatic tools needing to be connected with *ad hoc* glue scripts and careful checkpointing can be almost completely replaced with a SCATER workflow in R.

The SCATER package helps address the fundamental problems of normalisation and quality control for scRNA-seq data. The functionality of the package is a long way ahead of existing software in terms of the completeness and flexibility of the tools it provides and the workflow it enables. To achieve this, SCATER also solves the fundamental problem of how to organise scRNA-seq data in a sensible fashion. The SCESet class, tailored specifically for scRNA-seq data, underpins the SCATER package and enables the large suite of methods shown throughout this chapter. A thorough and consistent approach to accessing data from, and assigning data to, an SCESet object and seamless subsetting methods make the data analyst's life much easier. As I hope this chapter has demonstrated, SCATER provides a simple, but flexible and thorough, workflow for pre-processing, QC and normalisation of scRNA-seq data. The workflow makes use of many types of QC plots available in SCATER, which features deep visualisation capabilities for scRNA-seq data.

In the world of RNA-seq and, as is becoming evident, scRNA-seq—unlike in many areas of statistical genetics—data analysis is often not done by expert statisticians, genomicists and bioinformaticians. Instead, it is frequently done by lab biologists who wish to (or have no choice but to) analyse their own data, but have limited statistical, bioinformatic and programming experience. For such people, coming to terms with a single environment for data analysis (R/SCATER) is greatly preferable to having to learn multiple command line tools. The SCATER package is fully documented and has an accompanying vignette (a document combining code examples and explanation) to introduce users to the package's capabilities and demonstrate the basic workflow.

Big clusters and data centres are not necessarily required for "big data". Instead, a "distributed model" is possible for scRNA-seq, in which individual researchers analyse

their data locally on their laptop, desktop or server. Such a model benefits the researchers themselves as well as reducing demand on overworked core facilities. The SCATER package enables this model for scRNA-seq data analysis.

By providing a single workflow, in a single environment, that starts with raw read data and ends with data prepared for downstream analysis, SCATER can aid reproducible research. As demonstrated by this chapter, itself a "live" KNITR document with embedded R code, SCATER is well suited to analyses written as "literate programming" documents in LaTeX or R Markdown (RStudio, 2015). To provide just one example of many possibilities, SCATER could be used for automatic QC report generation. Essentially, a "targets" file could be supplied indicating the files containing raw read data and a QC report could be produced providing many of the diagnostic plots and analyses illustrated in this chapter. For the time being, hands-on QC and analysis is likely to be required, but such automatic report generation could be a useful starting point for core facilities or groups generating a large number of scRNA-seq datasets.

The SCATER package is open-source and I have made the source code available on GitHub (`github.com/davismcc/scater`). I plan to submit the package to Bioconductor for release in early 2016. The utility of SCATER is already being proved in practice. The package is currently used by a number of colleagues at the Wellcome Trust Centre for Human Genetics and the Weatherall Institute for Molecular Medicine in Oxford. Furthermore, SCATER forms the basis of Kieran Campbell's EMBEDDR package (`github.com/kieranrcampbell/embeddr`).

The SCATER package has now reached the stage of a stable beta release, but is under continuing development. A great deal of future work is planned to improve and broaden SCATER's capabilities.

To improve the computed QC metrics I would like to develop a way to convey library complexity in a single metric in a similar way to how PCR duplicate reads are currently used. There are other also other, more sophisticated, methods under development for identifying problematic cells from scRNA-seq data (Sarah Teichmann, personal communication), so I plan to integrate or incorporate such methods once they become available.

I would also like to increase the options for normalisation offered in SCATER. Single-cell quantile and rank normalisation methods require substantially more development, but I would like to add these type of methods to the package once they are mature. In the shorter term I plan to integrate more sophisticated normalisation routines such as BASiCS and GRM into the SCATER workflow. I also plan to incorporate testing for overdispersed or highly variable genes in the package in the near future.

I plan to extend the plotting capabilities further, in particular by expanding the range of dimension reduction methods available. For example, I would like to add functionality to produce visualisations from zero-inflated factor analysis (Yau & Pierson, 2015) and

diffusion maps (Haghverdi et al., 2015) as easily as PCA and t-SNE. I plan on making adjustments to some of the existing plotting functions to ensure that all of the plots produced feel the same thematically. Along similar lines, I plan to revise naming conventions throughout the package to ensure that they are consistent and intuitive.

Developing the SCATER architecture and methods to handle multiple genomic, transcriptomic and epigenomic data modalities from the same cells is a major priority for the next development cycle. Current trends in the field suggest that integrating and cross-mapping different data types will be crucial for future single-cell analyses (Wills & Mead, 2015). Integrating genomic and epigenomic data types requires non-trivial extensions to the SCESet class, but is eminently achievable. Adding such functionality would ensure that SCATER remains an ideal foundation for a wide-range of analyses with single-cell genomics data.

The SCATER package provides a very useful set of software tools for pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data. Future development will improve the package further and expand its range to handle more types of single-cell genomic data. Using the foundation provided by SCATER my future work will focus again on developing downstream analysis methods and applying them to answer biologically-focused scientific questions.

Appendix A

Supplementary Material for Chapter 5

A.1 Supplementary tables

QC metric	Description
`mean_exprs`	The mean expression level of the gene/feature.
`exprs_rank`	The rank of the feature's mean expression level in the cell.
`n_cells_exprs`	The number of cells for which the expression level of the feature is above the detection limit (default detection limit is zero).
`total_feature_counts`	The total number of counts assigned to that feature across all cells.
`log10_total_feature_counts`	Total feature counts on the log10-scale.
`pct_total_counts`	The percentage of all counts that are accounted for by the counts assigned to the feature.
`is_feature_control`	Is the feature a control feature? Default is 'FALSE' unless control features are defined by the user. If more than one feature control set is defined (as above), then a column of this type is produced for each control set (e.g. here, `is_feature_control_ERCC` and `is_feature_control_MT`) as well as the column named `is_feature_control`, which indicates if the feature belongs to any of the control sets.

Table A.1: Feature-level QC metrics computed in SCATER with the `calculateQCMetrics` function. These QC metrics are added as columns to the "featureData" slot of the SCESet object so that they can be inspected and are readily available for other functions to use. As with the cell-level metrics, wherever "counts" appear in the above, the same metrics will also be computed for "exprs", "tpm" and "fpkm" values (if TPM and FPKM values are present in the SCESet object), with the appropriate term replacing "counts" in the name.

QC metric	Description
`total_counts`	Total number of counts for the cell (aka "library size")
`log10_total_counts`	Library size on the log10-scale
`total_features`	The number of features for the cell that have expression above the detection limit (default detection limit is zero)
`filter_on_total_counts`	Would this cell be filtered out based on its log10-total counts being (by default) more than 5 median absolute deviations from the median log10-total counts for the dataset?
`filter_on_total_features`	Would this cell be filtered out based on its total features being (by default) more than 5 median absolute deviations from the median total features for the dataset?
`counts_from_feature_controls`	Total number of counts for the cell that come from (one or more sets of user-defined) control features. Defaults to zero if no control features are indicated. If more than one set of feature controls are defined (for example, ERCC and MT genes are defined as controls), then this metric is produced for all sets, plus the union of all sets (so here, we get columns `counts_from_feature_controls_ERCC`, `counts_from_feature_controls_MT` and `counts_from_feature_controls`).
`log10_counts_from_feature_controls`	Just as above, the total number of counts from feature controls, but on the log10-scale. Defaults to zero (i.e. log10(0 + 1), offset to avoid negative infinite values) if no feature control are indicated.
`pct_counts_from_feature_controls`	Just as for the counts described above, but expressed as a percentage of the total counts. Defined for all control sets and their union, just like the raw counts. Defaults to zero if no feature controls are defined.
`filter_on_pct_counts_from_feature_controls`	Would this cell be filtered out on the basis that the percentage of counts from feature controls is higher than a defined threhold (default is 80%)? Just as with `counts_from_feature_controls`, this is defined for all control sets and their union.
`pct_counts_from_top_50_features`	What percentage of the total counts is accounted for by the 50 highest-count features? Also computed for the top 100 and top 200 features, with the obvious changes to the column names.
`counts_from_endogenous_features`	Total number of counts for the cell that come from endogenous features (i.e. not control features). Defaults to 'total counts' if no control features are indicated.
`log10_counts_from_endogenous_features`	Total number of counts from endogenous features on the log10-scale. Defaults to zero (i.e. log10(0 + 1), offset to avoid infinite values) if no control features are indicated.
`n_detected_feature_controls`	Number of defined feature controls that have expression greater than the threshold defined in the object (that is, they are "detectably expressed"; see `object@lowerDetectionLimit` to check the threshold). As with other metrics for feature controls, defined for all sets of feature controls (set names appended as above) and their union. So we might commonly get columns `n_detected_feature_controls_ERCC`, `n_detected_feature_controls_MT` and `n_detected_feature_controls` (ERCC and MT genes detected).
`is_cell_control`	Has the cell been defined as a cell control? If more than one set of cell controls are defined (for example, blanks and bulk libraries are defined as cell controls), then this metric is produced for all sets, plus the union of all sets (so we could typically get columns `is_cell_control_Blank`, `is_cell_control_Bulk`, and `is_cell_control`, the latter including both blanks and bulks as cell controls).

Table A.2: Cell-level QC metrics computed in SCATER with the `calculateQCMetrics` function. These QC metrics are added as columns to the "phenotypeData" slot of the SCESet object so that they can be inspected and are readily available for other functions to use. Furthermore, wherever "counts" appear in the above metrics, the same metrics will also be computed for "exprs", "tpm" and "fpkm" values (if TPM and FPKM values are present in the SCESet object), with the appropriate term replacing "counts" in the name.

Function	Description
`counts(object)`, `counts(object)<-`	Returns/assigns the matrix of read counts. If no counts are defined for the object, then the counts matrix slot is simply `NULL`.
`norm_counts(object)`, `norm_counts(object)<-`	Returns/assigns the matrix of normalised read counts.
`exprs(object)`, `exprs(object)<-`	Returns/assigns the matrix of feature expression values. Typically these should be log2(transcripts-per-million) values, although log2(counts-per-million) or log2(fragments-per-kilobase-per-million-mapped) could also be used. For many statistical analyses, expression values need to be appropriately normalised. The SCATER package will generally assume that the values in the 'exprs' slot are the values to use for expression.
`norm_exprs(object)`, `norm_exprs(object)<-`	Returns/assigns the matrix of normalised feature expression values.
`stand_exprs(object)`, `stand_exprs(object)<-`	Returns/assigns the matrix of standardised feature expression values.
`is_exprs(object)`, `is_exprs(object)<-`	Returns/assigns a logical matrix indicating whether each gene expression observation is above the defined 'lowerDetectionLimit' (default is 0). This can be determined on the count scale or the "expression" (i.e. 'exprs(object)') scale. Changing the threshold by which we decide whether or not observations count as "expressed" is straight-forward with the `calcIsExprs` function.
`tpm(object)`, `tpm(object)<-`	Returns/assigns the matrix of TPM values.
`norm_tpm(object)`, `norm_tpm(object)<-`	Returns/assigns the matrix of normalised TPM values.
`cpm(object)`, `cpm(object)<-`	returns/assigns the matrix of CPM values.
`norm_cpm(object)`, `norm_cpm(object)<-`	Returns/assigns the matrix of normalised CPM values.
`fpkm(object)`, `fpkm(object)<-`	returns/assigns the matrix of FPKM values.
`norm_fpkm(object)`, `norm_fpkm(object)<-`	Returns/assigns the matrix of normalised FPKM values.
`reducedDimension(object)`, `reducedDimension(object)<-`	Returns/assigns the matrix of reduced-dimenion coordinates for cells.
`redDim(object)`, `redDim(object)<-`	Returns/assigns the matrix of reduced-dimenion coordinates for cells.
`norm_tpm(object)`, `norm_tpm(object)<-`	Returns/assigns the matrix of normalised TPM values.
`pData(object)`, `pData(object)<-`	Returns/assigns the "AnnotatedDataFrame" with the phenotype (cell) metadata.
`fData(object)`, `fData(object)<-`	Returns/assigns the "AnnotatedDataFrame" with the feature metadata.
`cellPairwiseDistances(object)`, `cellPairwiseDistances(object)<-`	Returns/assigns the matrix containing cell-cell distances.
`featurePairwiseDistances(object)`, `featurePairwiseDistances(object)<-`	Returns/assigns the matrix containing feature-feature (i.e. gene-gene) distances.

Table A.3: Accessor and assignment functions for SCESet objects.

A.2 Supplementary figures

```
ggplot(fData(sce_simmons),
       aes(x = gene_biotype, y = n_cells_exprs, colour = is_feature_control)) +
    geom_violin(group = fData(sce_simmons)$gene_biotype, fill = "aliceblue",
                colour = "gray50", scale = "width") +
    geom_boxplot(colour = "gray30", width = 0.3, outlier.size = 0) +
    ggthemes::scale_colour_tableau() + theme_cowplot(8) +
    theme(axis.text.x = element_text(angle = 75, vjust = 0.5),
          legend.position = c(0, 1), legend.justification = c(0, 1))
```

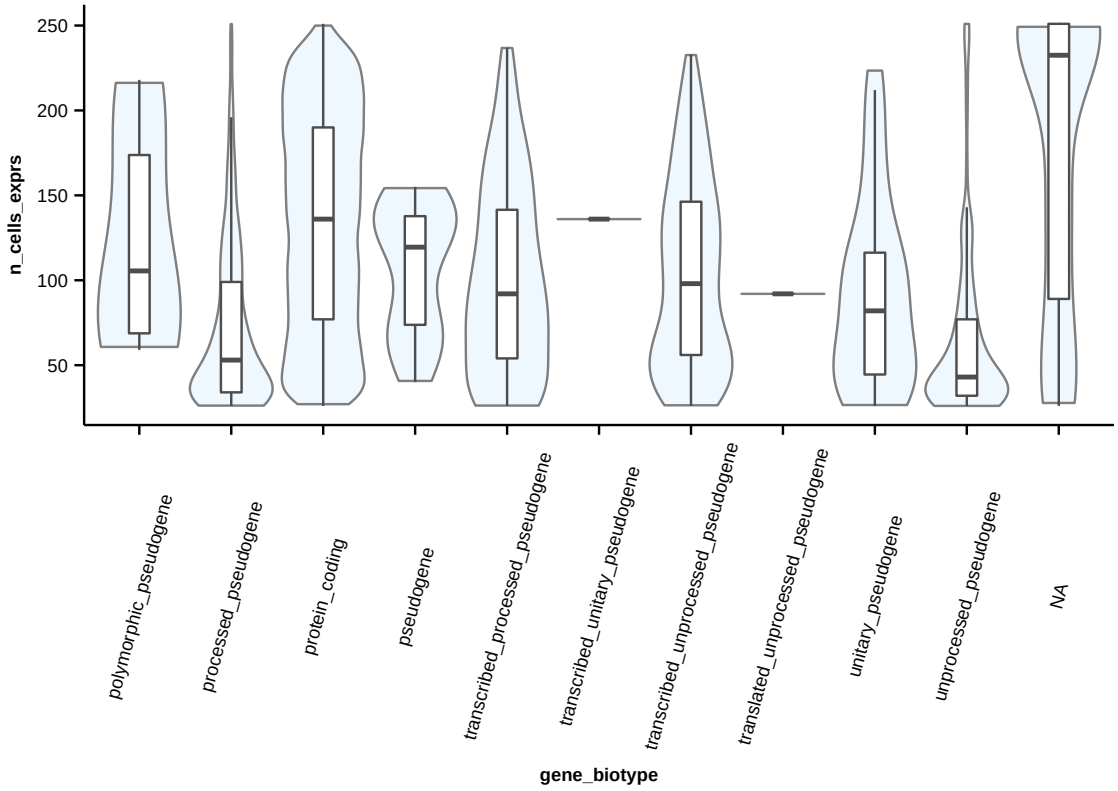


Figure A.1: For the Simmons Data, the number of cells expressing a feature is plotted against the gene biotype defined for each feature. For each biotype, a boxplot overlaid on a violin plot illustrates the distribution of number of cells expressing for the features in the category. Ensembl's automatic annotation system classifies genes and transcripts into biotypes, meaning the category of biological function that the gene is expected to have. For details, consult the Ensembl website (`www.ensembl.org/`). The "NA" biotype here is in fact the ERCC spike-in control features, which are not given a biotype from `getBMFeatureAnnos`, because they are not in the appropriate BIOMART database. The featureData slot of an SCESet object can be accessed easily with the `fData` function. It returns a data frame, which can be used directly as an input to GGPLOT, as shown. As well as providing a great deal of functionality, SCATER is highly interoperable with other R packages and methods.

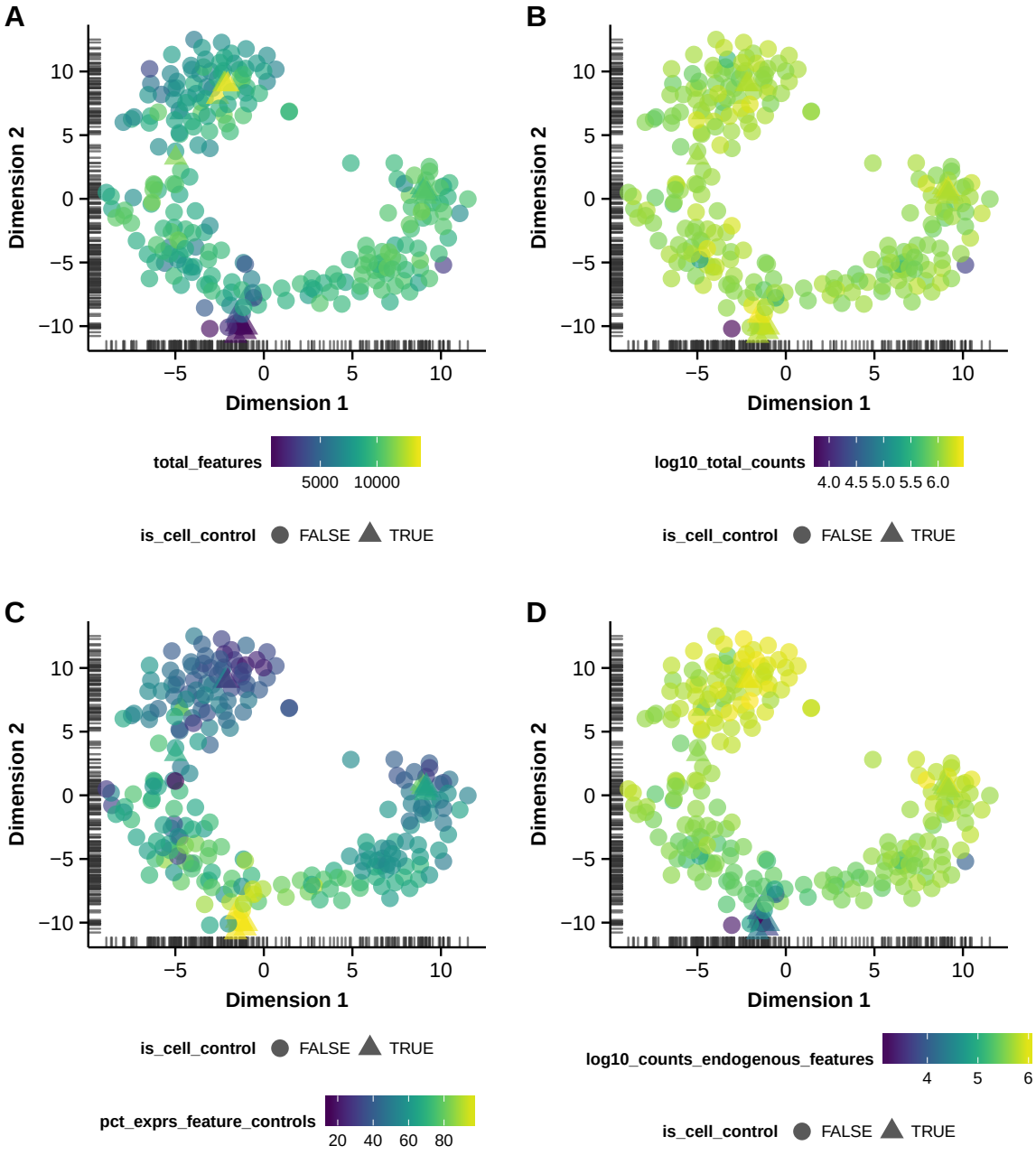


Figure A.2: Scatter plots of the first two t-SNE components of the Simmons Data with points coloured by (A) total features, (B) total counts (log-10 scale), (C) percentage of expression accounted for by feature controls, and (D) counts from endogenous features (log-10 scale).

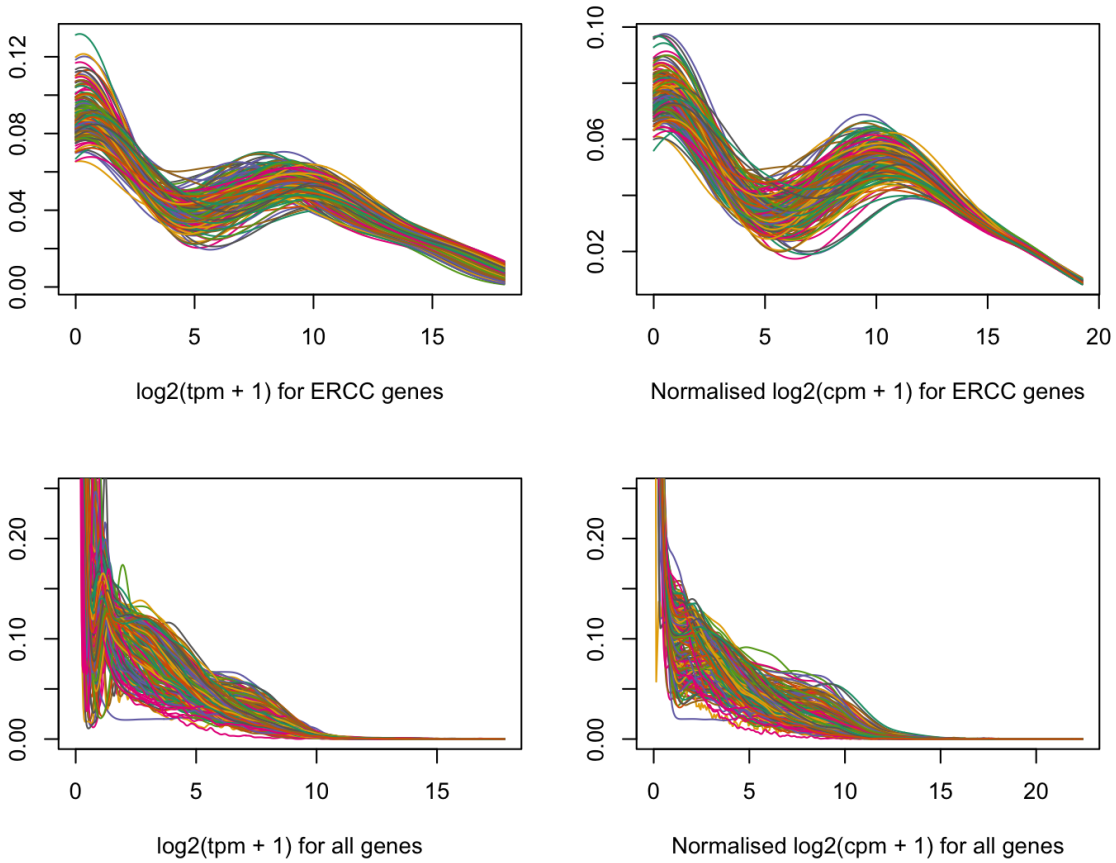


Figure A.3: Density plots for expression data in the Simmons Data for ERCC genes (top row) and all genes (bottom row) when using log2-transformed TPM values (left column) or ERCC-count normalised log-transformed CPM values as expression values. Each curve represents the distribution of expression values for one cell.

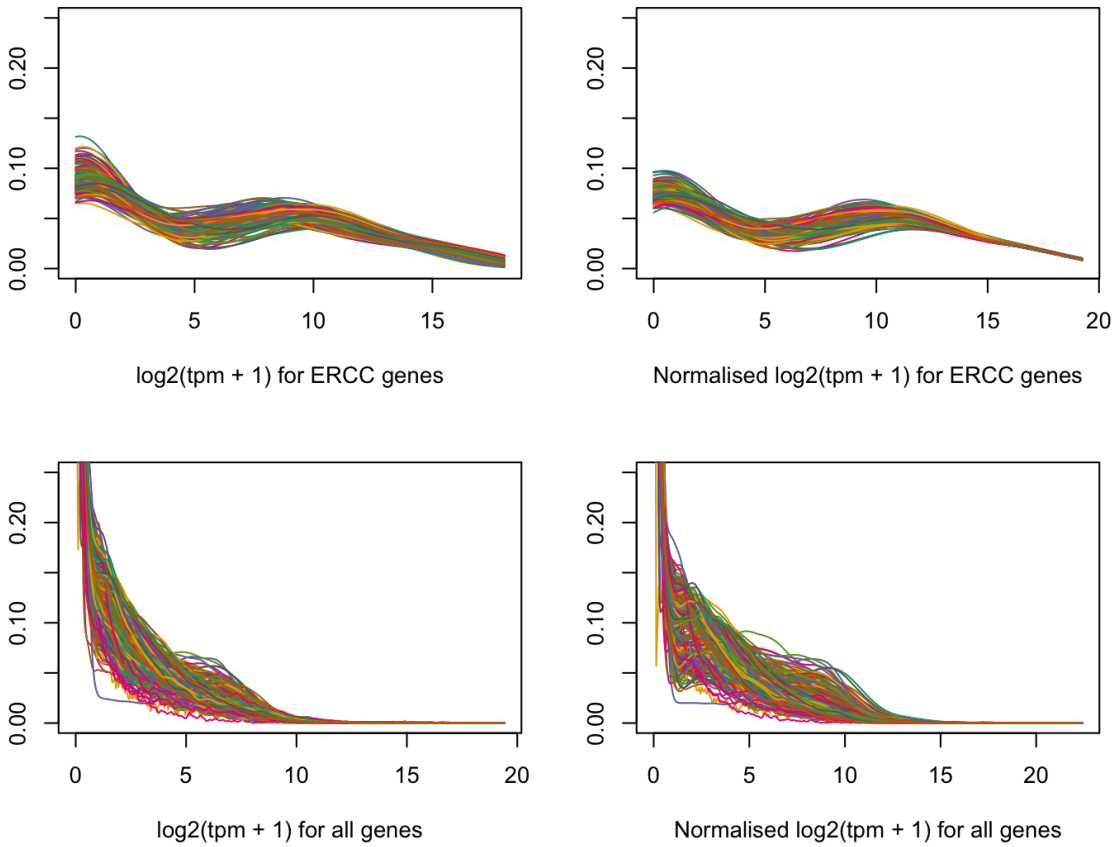


Figure A.4: Density plots for the Simmons Data for ERCC genes (top row) and all genes (bottom row) when using log2-transformed TPM values (left column) or ERCC-TPM normalised log-transformed TPM values as expression values. Each curve represents the distribution of expression values for one cell.

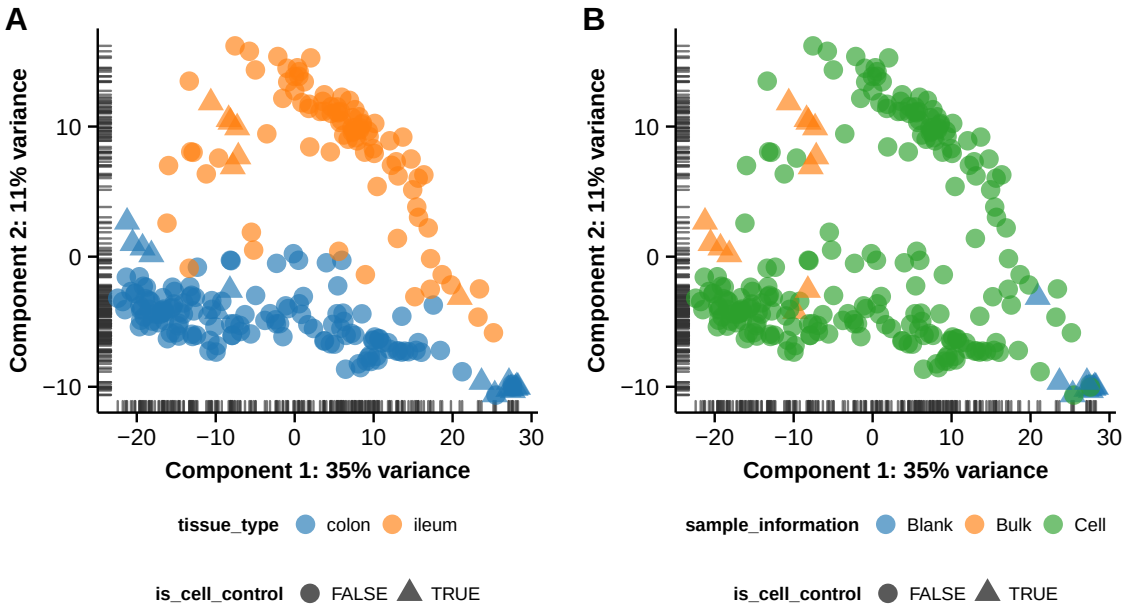


Figure A.5: Scatter plots of the first two principal components from PCA of the cells of the Simmons Data are plotted with points coloured by (A) tissue type and (B) sample type.

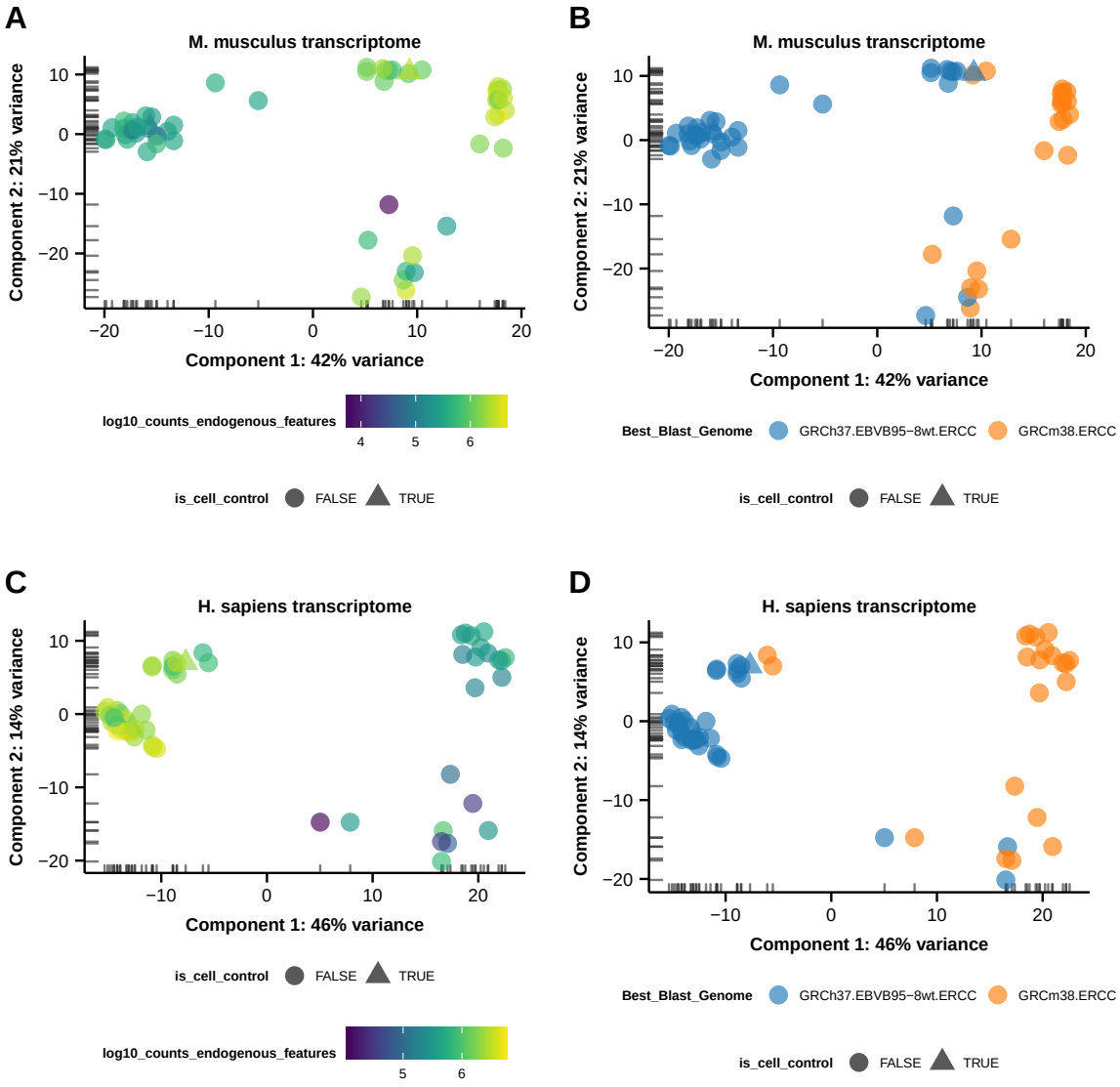


Figure A.6: Scatter plots of the first two components from a principal components analysis of the cells from Chip 12 of the Cell Cycle Data are plotted with points coloured by (A,C) log-10 counts from endogenous genes (i.e. non control genes) tissue type and (B,D) best genome hit from blasting 100 random reads from the library against the non-redundant BLAST database, when using expression quantities using the mouse transcriptome (A,B) and the human transcriptome (C,D).

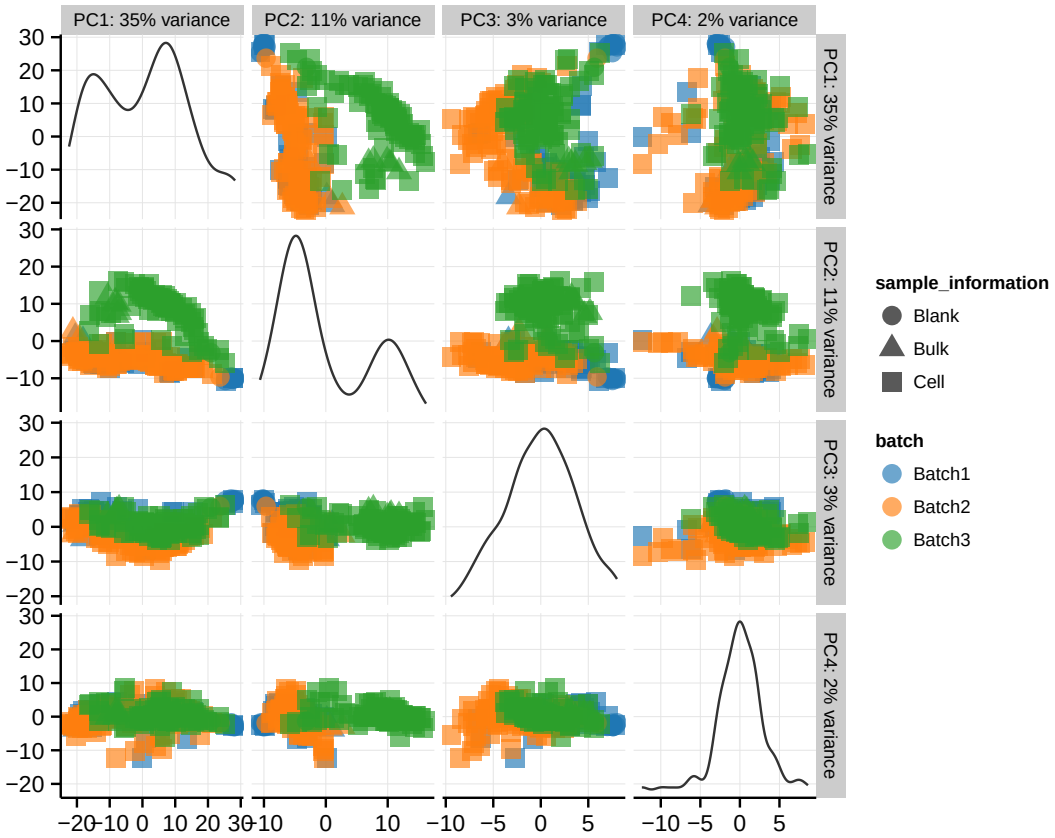


Figure A.7: Paired scatter plots of the first four principal components of the Simmons Data. Points (cells) in the plot are coloured by experimental batch. Boxes on the diagonal in the scatter plot matrix show the density for each component.

"Session information" providing details about the R environment used to conduct the analysis:

```
sessionInfo()

R version 3.2.2 (2015-08-14)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.11.1 (El Capitan)

locale:
[1] en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8

attached base packages:
[1] parallel  methods   stats     graphics  grDevices utils     datasets
[8] base

other attached packages:
 [1] Rtsne_0.10          cowplot_0.5.0       scater_0.1.10
 [4] ggplot2_1.0.1       Biobase_2.30.0      BiocGenerics_0.16.1
 [7] dplyr_0.4.3         reshape2_1.4.1      scales_0.3.0
[10] RColorBrewer_1.1-2  gplots_2.17.0       knitr_1.11

loaded via a namespace (and not attached):
 [1] nlme_3.1-122               bitops_1.0-6
 [3] matrixStats_0.15.0         doParallel_1.0.10
 [5] GenomeInfoDb_1.6.1         tools_3.2.2
 [7] doRNG_1.6                  nor1mix_1.2-1
 [9] R6_2.1.1                   irlba_2.0.0
[11] KernSmooth_2.23-15         DBI_0.3.1
[13] lazyeval_0.1.10            colorspace_1.2-6
[15] gridExtra_2.0.0            base64_1.1
[17] preprocessCore_1.32.0      formatR_1.2.1
[19] pkgmaker_0.22              rtracklayer_1.30.1
[21] labeling_0.3               caTools_1.17.1
[23] quadprog_1.5-5             genefilter_1.52.0
[25] Rsamtools_1.22.0           stringr_1.0.0
[27] digest_0.6.8               illuminaio_0.12.0
[29] siggenes_1.44.0            GEOquery_2.36.0
[31] XVector_0.10.0             limma_3.26.3
[33] highr_0.5.1                ggthemes_2.2.1
[35] RSQLite_1.0.0              VGAM_1.0-0
[37] quantro_1.4.0              combinat_0.0-8
[39] BiocParallel_1.4.0         mclust_5.1
[41] gtools_3.5.0               RCurl_1.95-4.7
[43] magrittr_1.5               futile.logger_1.4.1
[45] Matrix_1.2-3               Rcpp_0.12.2
[47] munsell_0.4.2              S4Vectors_0.8.3
[49] proto_0.3-10               viridis_0.3.1
[51] stringi_1.0-1              edgeR_3.12.0
[53] MASS_7.3-45                SummarizedExperiment_1.0.1
[55] zlibbioc_1.16.0            plyr_1.8.3
[57] bumphunter_1.10.0          grid_3.2.2
[59] minfi_1.16.0               gdata_2.17.0
[61] lattice_0.20-33            Biostrings_2.38.2
```

```
[63] splines_3.2.2              multtest_2.26.0
[65] GenomicFeatures_1.22.6     annotate_1.48.0
[67] locfit_1.5-9.1             beanplot_1.2
[69] igraph_1.0.1               GenomicRanges_1.22.1
[71] corpcor_1.6.8              rngtools_1.2.4
[73] mixOmics_5.2.0             codetools_0.2-14
[75] biomaRt_2.26.1             stats4_3.2.2
[77] futile.options_1.0.0       XML_3.98-1.3
[79] evaluate_0.8               lambda.r_1.1.7
[81] foreach_1.4.3              gtable_0.1.2
[83] reshape_0.8.5              assertthat_0.1
[85] xtable_1.8-0               monocle_1.4.0
[87] survival_2.38-3            HSMMSingleCell_0.104.0
[89] iterators_1.0.8            ellipse_0.3-8
[91] GenomicAlignments_1.6.1    AnnotationDbi_1.32.0
[93] registry_0.3               IRanges_2.4.4
[95] cluster_2.0.3              fastICA_1.2-0
[97] rgl_0.95.1429
```

Bibliography

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., & McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Achim, K., Pettit, J.B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., & Marioni, J.C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology*, 33(5):503–509.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., & Moreno, R.F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656.

Adzhubei, I.a., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., & Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–9.

Airy, G.B. (1861). *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*. Macmillan & Company, Cambridge and London.

Allan, W. (1933). Heredity in diabetes. *Annals of Internal Medicine*, 6(10):1272–1274.

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

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Altshuler, D., Hirschhorn, J.N., Klannemark, M., Lindgren, C.M., Vohl, M.C., Nemesh, J., Lane, C.R., Schaffner, S.F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., et al. (2000). The

common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics*, 26(1):76–80.

American Diabetes Association (2010). Diagnosis and classification of diabetes mellitus. *Diabetes Care*, 33 Suppl 1:S62–9.

Amir, E.A.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., & Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, 31(6):545–552.

Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10).

Anders, S., McCarthy, D., Chen, Y., Okoniewski, M., Smyth, G., Huber, W., & Robinson, M. (2013). Count-based differential expression analysis of RNA sequencing data using r and bioconductor. *Nature Protocols*, 8(9):1765–1786.

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

Anderson, C., Pettersson, F., Clarke, G., Cardon, L., Morris, A., & Zondervan, K. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564–1573.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Arber, W. & Linn, S. (1969). DNA modification and restriction. *Annual Review of Biochemistry*, 38:467–500.

Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziuszko, O., Vitezic, M., Freeman, T.C., Alhendi, A.M.N., et al. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, 347(6225):1010–1014.

Arnold, J.B. (2015). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*. R package version 2.2.1.

Ashurst, J., Chen, C., Gilbert, J., Jekosch, K., Keenan, S., Meidl, P., Searle, S., Stalker, J., Storey, R., Trevanion, S., Wilming, L., & Hubbard, T. (2005). The vertebrate genome annotation (Vega) database. *Nucleic Acids Research*, 33(Database issue):D459–D465.

Astbury, W.T. (1947). X-ray studies of nucleic acids. *Symoposia of the Society for Experimental Biology*, 1:66–76.

Astle, W. & Balding, D.J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471.

Auer, P. & Doerge, R. (2010). Statistical design and analysis of RNA sequencing data. *Genetics*, 185(2):405–416.

Avery, O.T., MacLeod, C.M., & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *The Journal of Experimental Medicine*, 79(2):137–158.

Babbs, C., Roberts, N., Luis, S., Simon, M., Ahmed, M., Brown, J., Sabry, M., WGS500 Consortium, Bentley, D., Gil, M., Donnelly, P., Gileadi, O., Ponting, C., Higgs, D., & Buckle, V. (2013). Homozygous mutations in a predicted endonuclease are a novel cause of congenital dyserythropoietic anemia type I. *Haematologica*, 98(9):1383–1387.

Baggerly, K.A. & Coombes, K.R. (2009). Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, 3(4):1309–1334.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755.

Banting, F.G., Best, C.H., Collip, J.B., Campbell, W.R., & Fletcher, A.A. (1922a). Pancreatic extracts in the treatment of diabetes mellitus. *The Canadian Medical Association Journal*, 12(3):141–146.

Banting, F.G., Best, C.H., Collip, J.B., Macleod, J.J.R., & Noble, E.C. (1922b). The effects of insulin on experimental hyperglycemia in rabbits. *American Journal of Physiology*, 62:559–580.

Banting, F.G., Campbell, W.R., & Fletcher, A.A. (1923). Further clinical experience with insulin (pancreatic extracts) in the treatment of diabetes mellitus. *British Medical Journal*, 1(3236):8–12.

Banting, F.G. & Gairns, S. (1924). Factors influencing the production of insulin. *American Journal of Physiology – Legacy Content*, 68(1):24–30.

Bateson, P. (2002). William Bateson: A biologist ahead of his time. *Journal of Genetics*, 81(2):49–58.

Bendall, S.C., Davis, K.L., Amir, E.A.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., & Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, 157(3):714–725.

Bendall, S.C., Simonds, E.F., Qiu, P., Amir, E.A.D., Krutzik, P.O., Finck, R., Bruggner, R.V., Melamed, R., Trejo, A., Ornatsky, O.I., Balderas, R.S., Plevritis, S.K., et al. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696.

Bengtsson, H. (2015). *matrixStats: Methods that Apply to Rows and Columns of Matrices (and to Vectors)*. R package version 0.14.2.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., & Wheeler, D.L. (2005). GenBank. *Nucleic Acids Research*, 33(Database issue):D34–8.

Berget, S.M., Moore, C., & Sharp, P.A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(8):3171–3175.

Berk, A.J. & Sharp, P.A. (1977). Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell*, 12(3):721–732.

Bhargava, V., Head, S.R., Ordoukhanian, P., Mercola, M., & Subramaniam, S. (2014). Technical variations in low-input RNA-seq methodologies. *Scientific Reports*, 4:3678.

Bhargava, V., Ko, P., Willems, E., Mercola, M., & Subramaniam, S. (2013). Quantitative transcriptomics using designed primer-based amplification. *Scientific Reports*, 3:1740.

Bloom, J.S., Ehrenreich, I.M., Loo, W.T., Lite, T.L.V.o., & Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237.

Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 107(21):9546–9551.

Boycott, K.M., Vanstone, M.R., Bulman, D.E., & MacKenzie, A.E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10):681–691.

Bray, N., Pimentel, H., Melsted, P., & Pachter, L. (2015). Near-optimal RNA-Seq quantification. *arXiv*, 1505.02710.

Breitling, R., Armengaud, P., Amtmann, A., & Herzyk, P. (2004). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573(1-3):83–92.

Breitling, R. & Herzyk, P. (2005). Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. *Journal of Bioinformatics and Computational Biology*, 3(5):1171–1189.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., & Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095.

Brenner, S., Jacob, F., & Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190:576–581.

Broad Institute (2013). Picard tools. `http://picard.sourceforge.net/`. Accessed: 2013-2-23.

Brouzes, E., Medkova, M., Savenelli, N., Marran, D., Twardowski, M., Hutchison, J.B., Rothberg, J.M., Link, D.R., Perrimon, N., & Samuels, M.L. (2009). Droplet microfluidic technology for single-cell high-throughput screening. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14195–14200.

Browning, S.R. & Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.

Browning, S.R. & Browning, B.L. (2011). Population structure can inflate snp-based heritability estimates. *The American Journal of Human Genetics*, 89(1):191–193.

Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., & Mountain, J.L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*, 96(1):37–53.

Buenrostro, J.D., Wu, B., Chang, H.Y., & Greenleaf, W.J. (2015a). ATAC-seq: A method for assaying chromatin accessibility Genome-Wide. *Current Protocols in Molecular Biology*, 109:21.29.1–9.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., & Greenleaf, W.J. (2015b). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490.

Buettner, F., Moignard, V., Göttgens, B., & Theis, F.J. (2014). Probabilistic PCA of censored data: accounting for uncertainties in the visualization of high-throughput single-cell qPCR data. *Bioinformatics*, 30(13):1867–1875.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., & Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160.

Bullard, J., Purdom, E., Hansen, K., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, 273(5278):1058–1073.

Burdett, T., Hall, P.N., Hasting, E., Hindorff, L.A., Junkins, H.A., Klemm, A.K., J, M., Manolio, T.A., Morales, J., Parkinson, H., & Welter, D. (2015). The NHGRI-EBI Catalog of published genome-wide association studies. `www.genome.gov/gwastudies`. Accessed: 2015-7-16, version 1.0.8.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282(5396):2012–2018.

Cardon, L.R. & Bell, J.I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics*, 2(2):91–99.

Ceroni, F., Simpson, N.H., Francks, C., Baird, G., Conti-Ramsden, G., Clark, A., Bolton, P.F., Hennessy, E.R., Donnelly, P., Bentley, D.R., Martin, H., IMGSAC, et al. (2014). Homozygous microdeletion of exon 5 in ZNF277 in a girl with specific language impairment. *European Journal of Human Genetics*, 22(10):1165–1171.

Chambers, J.M. & Lang, D.T. (2011). Object-oriented programming in R. *R News*, 1(3):17–19.

Chargaff, E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6(6):201–209.

Chatrchyan, S., Khachatryan, V., Sirunyan, A.M., Tumasyan, A., Adam, W., Aguilo, E., Bergauer, T., Dragicevic, M., Erö, J., Fabjan, C., Friedl, M., Frühwirth, R., et al. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30–61.

Chauvenet, W. (1863). *A Manual of Spherical and Practical Astronomy: Theory and use of astronomical instruments, method of least squares*. A Manual of Spherical and Practical Astronomy. J. B. Lippincott & Company.

Chen, H., Guo, J., Mishra, S.K., Robson, P., Niranjan, M., & Zheng, J. (2015a). Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development. *Bioinformatics*, 31(7):1060–1066.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., & Zhuang, X. (2015b). RNA imaging. spatially resolved, highly multiplexed RNA profiling in single cells. *Sciencexpress*, 348(6233):aaa6090.

Chikina, M., Zaslavsky, E., & Sealfon, S.C. (2015). CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics*, 31(10):1584–1591.

Chow, L.T., Roberts, J.M., Lewis, J.B., & Broker, T.R. (1977). A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. *Cell*, 11(4):819–836.

Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S., Lu, X., & Ruden, D. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

Coffey, A., Kokocinski, F., Calafato, M., Scott, C., Palta, P., Drury, E., Joyce, C., Leproust, E., Harrow, J., Hunt, S., Lehesjoki, A., Turner, D., Hubbard, T., & Palotie, A. (2011). The GENCODE exome: sequencing the complete human exome. *European Journal of Human Genetics*, 19(7):827–831.

Cohen, S.N., Chang, A.C., Boyer, H.W., & Helling, R.B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 70(11):3240–3244.

Collins, F.S. & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795.

Consortium, .G.P., Abecasis, G., Auton, A., Brooks, L., Mark, D., Durbin, R., Handsaker, R., Kang, H., Marth, G., & Gil, M. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Cooper, G., Stone, E., Asimenos, G., Program, N.C.S., Green, E., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(7):901–913.

Corbeil, R.R. & Searle, S.R. (1976a). A comparison of variance component estimators. *Biometrics*, 32(4):779–791.

Corbeil, R.R. & Searle, S.R. (1976b). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics: a journal of statistics for the physical, chemical, and engineering sciences*, 18(1):31–38.

Cossins, J., Belaya, K., Hicks, D., Salih, M., Finlayson, S., Carboni, N., Liu, W., Maxwell, S., Zoltowska, K., Farsani, G., Laval, S., Seidhamed, M., et al. (2013). Congenital myasthenic syndromes due to mutations in ALG2 and ALG14. *Brain : a journal of neurology*, 136(Pt 3):944–956.

Costantini, F. & Lacy, E. (1981). Introduction of a rabbit beta-globin gene into the mouse germ line. *Nature*, 294(5836):92–94.

Couture-Beil, A. (2014). *rjson: JSON for R*. R package version 0.2.15.

Cragg, J.G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5):829–844.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Crosetto, N., Bienko, M., & van Oudenaarden, A. (2015). Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics*, 16(1):57–66.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., et al. (2015). Ensembl 2015. *Nucleic Acids Research*, 43(Database issue):D662–9.

Curtis, W.S. (1929). Diabetes in twins. *Journal of the American Medical Association*, 92(12):952–956.

Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., & Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914.

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Developmental Biology*, 278(2):274–288.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group (2011). The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158.

Danna, K. & Nathans, D. (1971). Specific cleavage of simian virus 40 DNA by restriction endonuclease of *Hemophilus influenzae*. *Proceedings of the National Academy of Sciences of the United States of America*, 68(12):2913–2917.

Danna, K.J., Sack, Jr, G.H., & Nathans, D. (1973). Studies of simian virus 40 DNA. VII. a cleavage map of the SV40 genome. *Journal of Molecular Biology*, 78(2):363–376.

Dar, R.D., Razooky, B.S., Singh, A., Trimeloni, T.V., McCollum, J.M., Cox, C.D., Simpson, M.L., & Weinberger, L.S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 109(43):17454–17459.

Darzynkiewicz, Z., Kapuscinski, J., Carter, S.P., Schmid, F.A., & Melamed, M.R. (1986). Cytostatic and cytotoxic properties of pyronin Y: relation to mitochondrial localization of the dye and its interaction with RNA. *Cancer Research*, 46(11):5760–5766.

Davison, D. (2009). shellfish: Parallel PCA and data processing for genome-wide SNP data. `http://www.stats.ox.ac.uk/~davison/software/shellfish/shellfish.php`. Accessed: 2014-9-25.

de Vries, H. (1900). Sur la loi de disjonction des hybrides. *Comptes Rendus de l'Académie des Sciences*, 130.

Deloukas, P., Schuler, G.D., Gyapay, G., Beasley, E.M., Soderlund, C., Rodriguez-Tomé, P., Hui, L., Matise, T.C., McKusick, K.B., Beckmann, J.S., Bentolila, S., Bihoreau, M., et al. (1998). A physical map of 30,000 human genes. *Science*, 282(5389):744–746.

Dembélé, D. & Kastner, P. (2014). Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinformatics*, 15:14.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38.

Dempster, E. & Lerner, I. (1950). Heritability of threshold characters. *Genetics*, 35(2):212–236.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D., Lagarde, J., Veeravalli, L., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research*, 22(9):1775–1789.

Dey, S.S., Kester, L., Spanjaard, B., Bienko, M., & van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology*, 33(3):285–289.

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I.W., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J., Hughes, T.E., et al. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–1336.

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 Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium, Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., Ferreira, T., Horikoshi, M., et al. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3):234–244.

Dimont, E., Shi, J., Kirchner, R., & Hide, W. (2015). edgeRun: an R package for sensitive, functionally relevant differential expression discovery using an unconditional exact test. *Bioinformatics*, 31(15):2589–2590.

Ding, B., Zheng, L., Zhu, Y., Li, N., Jia, H., Ai, R., Wildberg, A., & Wang, W. (2015). Normalization and noise reduction for single cell RNA-seq experiments. *Bioinformatics*, 31(13):2225–2227.

Djebali, S., Davis, C., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G., et al. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–108.

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

Dobzhansky, T. (1937). *Genetics and the Origin of Species*. Columbia University Press, New York, USA.

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

Dorfman, R. (1938). A note on the δ-method for finding variance formulae. *The Biometric Bulletin*, 1:129–137.

Dowle, M., Short, T., Lianoglou, S., with contributions from R Saporta, A.S., & Antonyan, E. (2014). *data.table: Extension of data.frame*. R package version 1.9.4.

Dueck, H., Khaladkar, M., Kim, T.K., Spaethling, J.M., Francis, C., Suresh, S., Fisher, S., Seale, P., Beck, S.G., Bartfai, T., Kuhn, B., Eberwine, J., & Kim, J. (2015). Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation. *Genome Biology*, 16(1):122.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., et al. (1999). The DNA sequence of human chromosome 22. *Nature*, 402(6761):489–495.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440.

Efroni, I., Ip, P.L., Nawy, T., Mello, A., & Birnbaum, K.D. (2015). Quantification of cell identity from single-cell gene expression profiles. *Genome Biology*, 16(1):9.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., & Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450.

Eilbeck, K., Lewis, S., Mungall, C., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The sequence ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5):R44.

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., et al. (2015). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426.

Eisinga, R., Breitling, R., & Heskes, T. (2013). The exact probability distribution of the rank product statistics for replicated experiments. *FEBS Letters*, 587(6):677–682.

El-Naggar, A.K. (2004). Concurrent flow cytometric analysis of DNA and RNA - springer. In R.G.H. T. S. Hawley, editor, *Flow Cytometry Protocols*, volume 26 of *Methods in Molecular Biology*, pages 371–384. Humana Press Inc., Totowa, NJ.

Elowitz, M.B., Levine, A.J., Siggia, E.D., & Swain, P.S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186.

ENCODE (2012). ENCODE cell types 2007-2012. `http://genome.ucsc.edu/ENCODE/cellTypes.html`. Accessed: 2014-8-30.

ENCODE Project Consortium, Bernstein, B., Birney, E., Dunham, I., Green, E., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Ensembl (2013a). Ensembl.

Ensembl (2013b). Ensembl gene set.

Ensembl (2013c). Variant effect predictor.

Ernst, J. & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216.

Exome Aggregation Consortium (ExAC) (2015). ExAC browser. `http://exac.broadinstitute.org/about`. Accessed: 2015-1-8.

Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., & Tress, M.L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, 23(22):5866–5878.

Fajans, S.S., Bell, G.I., & Polonsky, K.S. (2001). Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *New England Journal of Medicine*, 345(13):971–980.

Falconer, D.S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76.

Falconer, D.S. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Annals of Human Genetics*, 31(1):1–20.

Falconer, D.S. & Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*. 4th edn. Longman, Harlow.

Fan, H.C., Fu, G.K., & Fodor, S.P.A. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222):1258367.

Farlik, M., Sheffield, N.C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., & Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Reports*, 10(8):1386–1397.

Fazeli Farsani, S., van der Aa, M.P., van der Vorst, M.M.J., Knibbe, C.A.J., & de Boer, A. (2013). Global trends in the incidence and prevalence of type 2 diabetes in children and adolescents: a systematic review and evaluation of methodological approaches. *Diabetologia*, 56(7):1471–1488.

Feigelman, J., Theis, F.J., & Marr, C. (2014). MCA: Multiresolution correlation analysis, a graphical tool for subpopulation identification in single-cell gene expression data. *BMC Bioinformatics*, 15:240.

Finak, G., McDavid, A., Chattopadhyay, P., Dominguez, M., De Rosa, S., Roederer, M., & Gottardo, R. (2014). Mixture models for single-cell assays with applications to vaccine studies. *Biostatistics*, 15(1):87–101.

Fischer, B. & Pau, G. (2015). *rhdf5: HDF5 interface to R*. R package version 2.12.0.

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

Flannick, J., Kang, H.M., Gaulton, K.J., Agarwala, V., Ma, C., McCarthy, D.J., Moutsianas, L., Burtt, N.P., Fontanillas, P., Blackwell, T.W., Locke, A.E., Pearson, R.D., et al. (2015). Whole-genome sequencing of 2,657 individuals and the genetic architecture of type 2 diabetes. *(submitted)*.

Flannick, J., Thorleifsson, G., Beer, N.L., Jacobs, S.B.R., Grarup, N., Burtt, N.P., Mahajan, A., Fuchsberger, C., Atzmon, G., Benediktsson, R., Blangero, J., Bowden, D.W., et al. (2014). Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nature Genetics*, 46(4):357–363.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., & Merrick, J.M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science*, 269(5223):496–512.

Flemming, W. (1882). *Zellsubstanz, Kern und Zelltheilung*. Verlag von F. C. W. Vogel, Leipzig.

Flicek, P., Ahmed, I., Amode, M., Barrell, D., Beal, K., Brent, S., Denise, C., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., et al. (2013). Ensembl 2013. *Nucleic Acids Research*, 41(Database issue):D48–D55.

Flicek, P., Amode, M., Barrell, D., Beal, K., Brent, S., Denise, C., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., et al. (2012). Ensembl 2012. *Nucleic Acids Research*, 40(Database issue):D84–D90.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C.G., et al. (2014). Ensembl 2014. *Nucleic Acids Research*, 42(Database issue):D749–55.

Florez, J.C., Burtt, N., de Bakker, P.I.W., Almgren, P., Tuomi, T., Holmkvist, J., Gaudet, D., Hudson, T.J., Schaffner, S.F., Daly, M.J., Hirschhorn, J.N., Groop, L., & Altshuler, D. (2004). Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. *Diabetes*, 53(5):1360–1368.

Fluidigm, I. (2015). Fluidigm | products | polaris. `https://www.fluidigm.com/products/polaris`. Accessed: 2015-7-13.

Fluidigm Corporation (2015). C1: The first automated solution for single-cell genomics, now capable of even more. `https://www.fluidigm.com/products/c1-system`. Accessed: 2015-6-7.

Ford, C.E., Jones, K.W., Polani, P.E., de Almeida, J.C., & Briggs, J.H. (1959). A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *The Lancet*, 1(7075):711–713.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, R.D., Weidman, J.F., et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, 270(5235):397–403.

Frayling, T.M. (2007). Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nature Reviews Genetics*, 8(9):657–662.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R.B., Elliott, K.S., Lango, H., Rayner, N.W., Shields, B., Harries, L.W., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889–894.

Frischmeyer, P., van Hoof, A., Kathryn, O., Guerrerio, A., Parker, R., & Dietz, H. (2002). An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science (New York, N.Y.)*, 295(5563):2258–2261.

Frumkin, D., Wasserstrom, A., Itzkovitz, S., Harmelin, A., Rechavi, G., & Shapiro, E. (2008). Amplification of multiple genomic loci from single cells isolated by laser micro-dissection of tissues. *BMC Biotechnology*, 8:17.

Fu, G.K., Hu, J., Wang, P.H., & Fodor, S.P.A. (2011). Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proceedings of the National Academy of Sciences of the United States of America*, 108(22):9026–9031.

Fu, Q., Hajdinjak, M., Moldovan, O.T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., Viola, B., Prüfer, K., et al. (2015). An early modern human from romania with a recent neanderthal ancestor. *Nature*, advance online publication.

Fujita, P., Rhead, B., Zweig, A., Hinrichs, A., Karolchik, D., Cline, M., Goldman, M., Barber, G., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T., et al. (2011). The UCSC genome browser database: update 2011. *Nucleic Acids Research*, 39(Database issue):D876–D882.

GA4GH Project (2015). Home | global alliance for genomics and health. `http://genomicsandhealth.org/`. Accessed: 2015-1-8.

Garrod, A.E. (1902). The incidence of alkaptonurua: A study in chemical individuality. *The Lancet*, 160(4137):1616–1620.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., Berney, T., Montanya, E., Mohlke, K.L., Lieb, J.D., & Ferrer, J. (2010). A map of open chromatin in human pancreatic islets. *Nature Genetics*, 42(3):255–259.

Genome Reference Consortium (2015). Human genome overview. `http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/`. Accessed: 2015-1-8.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10).

Gest, H. (2004). The discovery of microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, fellows of the Royal Society. *Notes and Records of the Royal Society London*, 58(2):187–201.

Gibbs, R.A. (2011). Bringing genomics and genetics back together. *Science*, 331(6017):548.

Gilmour, A.R., Thompson, R., & Cullis, B.R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51(4):1440–1450.

Gingeras, T. (2007). Origin of phenotypes: genes and transcripts. *Genome Research*, 17(6):682–690.

Gloyn, A.L., Weedon, M.N., Owen, K.R., Turner, M.J., Knight, B.A., Hitman, G., Walker, M., Levy, J.C., Sampson, M., Halford, S., McCarthy, M.I., Hattersley, A.T., & Frayling, T.M. (2003). Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes*, 52(2):568–572.

Goddard, M.E. & Hayes, B.J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics*, 10(6):381–391.

Goddard, M.E., Lee, S.H., Yang, J., Wray, N.R., & Visscher, P.M. (2011). Response to Browning and Browning. *American Journal of Human Genetics*, 89(1):193–195.

Golan, D., Lander, E.S., & Rosset, S. (2014). Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences of the United States of America*.

Goldstein, D.B. (2009). Common genetic variation and human traits. *New England Journal of Medicine*, 360(17):1696–1698.

Gordon, J.W. & Ruddle, F.H. (1981). Integration and stable germ line transmission of genes injected into mouse pronuclei. *Science*, 214(4526):1244–1246.

Grant, S.F.A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., Styrkarsdottir, U., Magnusson, K.P., et al. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genetics*, 38(3):320–323.

Grarup, N., Sandholt, C.H., Hansen, T., & Pedersen, O. (2014). Genetic susceptibility to type 2 diabetes and obesity: from genome-wide association studies to rare variants and beyond. *Diabetologia*, 57(8):1528–1541.

Green, E., Guyer, M., & Institute, N.H.G.R. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333):204–213.

Griffith, F. (1923). The influence of immune serum on the biological properties of pneumococci. *Ministry of Health Reports on Public Health and Med. Subjects*, 18.

Grindberg, R.V., Yee-Greenbaum, J.L., McConnell, M.J., Novotny, M., O'Shaughnessy, A.L., Lambert, G.M., Araúzo-Bravo, M.J., Lee, J., Fishman, M., Robbins, G.E., Lin, X., Venepally, P., et al. (2013). RNA-sequencing from single nuclei. *Proceedings of the National Academy of Sciences of the United States of America*, 110(49):19802–19807.

Grün, D., Kester, L., & van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods*, 11(6):637–640.

GTEx Consortium (2013). The Genotype-Tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585.

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

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., & Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.*, 23(12):2126–2135.

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

Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B.J., Diogo, D., Stahl, E.A., Gregersen, P.K., Worthington, J., Klareskog, L., Raychaudhuri, S., Plenge, R.M., Pasaniuc, B., & Price, A.L. (2013). Quantifying missing heritability at known GWAS loci. *PLoS Genetics*, 9(12):e1003993.

Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., Schizophrenia Working Group of the Psychiatric Genomics Consortium, SWE-SCZ Consortium, et al. (2014). Partitioning heritability of regulatory and Cell-Type-Specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5):535–552.

Guthrie, R. (1961). Blood screening for phenylketonuria. *JAMA*, 178(8):863–863.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in europe. *Nature*, 522(7555):207–211.

Habegger, L., Balasubramanian, S., Chen, D.Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M., & Gerstein, M. (2012). Vat: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics*, 28(17):2267–2269.

Haeckel, E.H. (1866). *Generelle Morphologie der Organismen allgemeine Grundzuge der organischen Formen-Wissenschaft, mechanisch begrundet durch die von Charles Darwin reformirte*

Descendenz-Theorie von Ernst Haeckel: AllgemeinN. Enlgle Entwickelungsgeschichte der Organismen kritische Grundzuge der mechanischen Wissenschaft von den entstehenden Formen der Organismen, begrundet durch die Descendenz-Theorie, volume 2. Verlag von Georg Reimer, Berlin.

Haghverdi, L., Buettner, F., & Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, advance online access.

Hara, K., Shojima, N., Hosoe, J., & Kadowaki, T. (2014). Genetic architecture of type 2 diabetes. *Biochemical and Biophysical Research Communications*.

Hardcastle, T. & Kelly, K. (2010). baySeq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11.

Hardy, G. (1908). Mendelian proportions in a mixed population. *Science (New York, N.Y.)*, 28(706):49–50.

Harismendy, O., Notani, D., Song, X., Rahim, N.G., Tanasa, B., Heintzman, N., Ren, B., Fu, X.D., Topol, E.J., Rosenfeld, M.G., & Frazer, K.A. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature*, 470(7333):264–268.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C., Chrast, J., Lagarde, J., Gilbert, J., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S., & Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, 7 Suppl 1:S4.1–S4.9.

Harrow, J., Frankish, A., Gonzalez, J., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., et al. (2012). GENCODE: the reference human genome annotation for the ENCODE project. *Genome Research*, 22(9):1760–1774.

Hart, R.K., Rico, R., Hare, E., Garcia, J., Westbrook, J., & Fusaro, V.A. (2015). A Python package for parsing, validating, mapping and formatting sequence variants using HGVS nomenclature. *Bioinformatics*, 31(2):268–270.

Harte, R., Farrell, C., Loveland, J., Suner, M., Wilming, L., Aken, B., Barrell, D., Frankish, A., Wallin, C., Searle, S., Diekhans, M., Harrow, J., & Pruitt, K. (2012). Tracking and coordinating an international curation effort for the CCDS project. *Database : the journal of biological databases and curation*, 2012.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports*, 2(3):666–673.

Hatem, A., Bozdağ, D., Toland, A., & Çatalyürek, U.V. (2013). Benchmarking short sequence mapping tools. *BMC bioinformatics*, 14.

Hattersley, A.T. & McCarthy, M.I. (2005). What makes a good genetic association study? *The Lancet*, 366(9493):1315–1323.

Hayashi, T., Shibata, N., Okumura, R., Kudome, T., Nishimura, O., Tarui, H., & Agata, K. (2010). Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its "index sorting" function for stem cell research. *Development Growth and Differentiation*, 52(1):131–144.

Hayes, B., Visscher, P., & Goddard, M. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research*, 91(1):47–60.

Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., & Teichmann, S.A. (2011). RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology*, 7:497.

Heimans, J. (1962). Hugo de Vries and the gene concept. *American Naturalist*, pages 93–104.

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, 343(6172):747–751.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9(2):226–252.

Hershey, A.D. & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology.*, 36(1):39–56.

Heskes, T., Eisinga, R., & Breitling, R. (2014). A fast algorithm for determining bounds and accurate approximate p -values of the rank product statistic for replicate experiments. *BMC Bioinformatics*, 15(1):367.

Hicks, S.C. & Irizarry, R.A. (2015). quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome Biology*, 16(1):117.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.*, 6(9):807–828.

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

Hirschhorn, J.N. & Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108.

Hivert, M.F., Vassy, J.L., & Meigs, J.B. (2014). Susceptibility to type 2 diabetes mellitus–from genes to prevention. *Nature Reviews Endocrinology*, 10(4):198–205.

Hogan, B. & Williams, J. (1981). Integration of foreign genes into the mammalian germ line: genetic engineering enters a new era. *Nature*, 294(5836):9–10.

Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L., & Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827.

Hoppe, P.S., Coutu, D.L., & Schroeder, T. (2014). Single-cell technologies sharpen up mammalian stem cell research. *Nature Cell Biology*, 16(10):919–927.

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

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3*, 1(6):457–470.

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

Hu, H., Huff, C.D., Moore, B., Flygare, S., Reese, M.G., & Yandell, M. (2013). Vaast 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology*, 37(6):622–634.

Hu, H., Roach, J.C., Coon, H., Guthery, S.L., Voelkerding, K.V., Margraf, R.L., Durtschi, J.D., Tavtigian, S.V., Shankaracharya, Wu, W., Scheet, P., Wang, S., et al. (2014). A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. *Nature biotechnology*, 32(7):663–669.

Huber, W., Carey, J., V., Gentleman, R., Anders, S., Carlson, M., Carvalho, S., B., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121.

Hudson, T.J., Stein, L.D., Gerety, S.S., Ma, J., Castle, A.B., Silva, J., Slonim, D.K., Baptista, R., Kruglyak, L., Xu, S.H., Hu, X., Colbert, A.M., et al. (1995). An STS-based map of the human genome. *Science*, 270(5244):1945–1954.

Hug, H. & Schuler, R. (2003). Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J. Theor. Biol.*, 221(4):615–624.

Hunt, K.A., Mistry, V., Bockett, N.A., Ahmad, T., Ban, M., Barker, J.N., Barrett, J.C., Blackburn, H., Brand, O., Burren, O., Capon, F., Compston, A., et al. (2013). Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature*, 498(7453):232–235.

Hutson, S. (2010). Data handling errors spur debate over clinical trial. *Nature Medicine*, 16(6):618–618.

Illumina, Inc (2013a). HiSeq 2000/1000. http://www.illumina.com/systems/hiseq_2000_1000.ilmn. Accessed: 30 June 2013.

Illumina, Inc (2013b). HiSeq 2500/1500. http://www.illumina.com/systems/hiseq_2500_1500.ilmn. Accessed: 30 June 2013.

Illumina, Inc (2013c). Off-Line Basecaller (OLB). http://support.illumina.com/sequencing/sequencing_so Accessed: 30 June 2013.

Illumina, Inc (2014a). GenomeStudio software. `http://bioinformatics.illumina.com/informatics/sequencing-microarray-data-analysis/genomestudio.ilmn`. Accessed: 2014-9-9.

Illumina, Inc (2014b). HumanOmni2.5-8 BeadChip kit: Illumina. `http://products.illumina.com/products/humanomni25-8_beadchip_kits.ilmn`. Accessed: 2014-9-28.

Ingram, V.M. (1956). A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature*, 178(4537):792–794.

International Diabetes Federation (2013). *IDF Diabetes Atlas*. Sixth Edition. International Diabetes Federation, Brussels, Belgium.

International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., Dermitzakis, E., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58.

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

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, 21(7):1160–1167.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P., & Linnarsson, S. (2012). Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nature Protocols*, 7(5):813–828.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., & Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166.

Jackson, D.A., Symons, R.H., & Berg, P. (1972). Biochemical method for inserting new genetic information into DNA of simian virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 69(10):2904–2909.

Jacob, H.J., Abrams, K., Bick, D.P., Brodie, K., Dimmock, D.P., Farrell, M., Geurts, J., Harris, J., Helbling, D., Joers, B.J., Kliegman, R., Kowalski, G., et al. (2013). Genomics in clinical practice: lessons from the front lines. *Science Translational Medicine*, 5(194):194cm5.

Jacobs, P.A. & Strong, J.A. (1959). A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature*, 183(4657):302–303.

Jaffee, S.R. & Price, T.S. (2007). Gene–environment correlations: a review of the evidence and implications for prevention of mental illness. *Molecular Psychiatry*, 12(5):432–442.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., & Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779.

Janss, L., de Los Campos, G., Sheehan, N., & Sorensen, D. (2012). Inferences from genomic models in stratified populations. *Genetics*, 192(2):693–704.

Jason, O., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., Wei, Z., Wang, K., & Lyon, G.J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine*, 5(3):28.

Jennrich, R.I. & Sampson, P.F. (1976). Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, 18(1):11–17.

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

Jimenez-Sanchez, G., Childs, B., & Valle, D. (2001). Human disease genes. *Nature*, 409(6822):853–855.

Johannsen, W. (1909). *Elemente der exakten Erblichkeitslehre*. Verlag von Gustav Fischer, Jena. Reprinted in 2015 by Cambridge University Press, Cambridge.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T.L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, 36(Web Server issue):W5–9.

Jolliffe, I. (2014). Principal component analysis. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd.

Jones, A.M. (1989). A double-hurdle model of cigarette consumption. *Journal of Applied Econometrics*.

Juliá, M., Telenti, A., & Rausell, A. (2015). Sincell: an R/Bioconductor package for statistical assessment of cell-state hierarchies from single-cell RNA-seq. *Bioinformatics*, advance access.

Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., & Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *The American Journal of Human Genetics*, 91(5):839–848.

Junker, J.P., Noël, E.S., Guryev, V., Peterson, K.A., Shah, G., Huisken, J., McMahon, A.P., Berezikov, E., Bakkers, J., & van Oudenaarden, A. (2014). Genome-wide RNA tomography in the zebrafish embryo. *Cell*, 159(3):662–675.

Kærn, M., Elston, T.C., Blake, W.J., & Collins, J.J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464.

Kahn, S.E., Cooper, M.E., & Del Prato, S. (2014). Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. *The Lancet*, 383(9922):1068–1083.

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

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

Keays, K.M., Owens, G.P., Ritchie, A.M., Gilden, D.H., & Burgoon, M.P. (2005). Laser capture microdissection and single-cell RT-PCR without RNA purification. *Journal of Immunological Methods*, 302(1-2):90–98.

Kennedy, S.R., Loeb, L.A., & Herr, A.J. (2012). Somatic mutations in aging, cancer and neurodegeneration. *Mechanisms of Ageing and Development*, 133(4):118–126.

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

Kharchenko, P.V., Silberstein, L., & Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742.

King, H., Aubert, R.E., & Herman, W.H. (1998). Global burden of diabetes, 1995–2025: Prevalence, numerical estimates, and projections. *Diabetes Care*, 21(9):1414–1431.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., & Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201.

Klein, C.A., Seidl, S., Petat-Dutter, K., Offner, S., Geigl, J.B., Schmidt-Kittler, O., Wendler, N., Passlick, B., Huber, R.M., Schlimok, G., Baeuerle, P.A., & Riethmüller, G. (2002). Combined transcriptome and genome analysis of single micrometastatic cells. *Nature Biotechnology*, 20(4):387–392.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., & Teichmann, S.A. (2015). The technology and biology of Single-Cell RNA sequencing. *Molecular Cell*, 58(4):610–620.

Kornberg, A. (1974). *DNA synthesis*. Freeman, San Francisco.

Krebs, J.E., Goldstein, E.S., & Kilpatrick, S.T. (2014). *Lewin's genes XI*. Jones & Bartlett Learning.

Krijthe, J. (2015). *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.10.

Kumar, P., Henikoff, S., & Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–81.

Kumar, R.M., Cahan, P., Shalek, A.K., Satija, R., Jay DaleyKeyser, A., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J.J., Ferrante, T.C., Regev, A., Daley, G.Q., & Collins, J.J. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, 516(7529):56–61.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lander, E.S. & Schork, N.J. (1994). Genetic dissection of complex traits. *Science*, 265(5181):2037–2048.

Lappalainen, T., Sammeth, M., Friedländer, M., 't Hoen, P., Monlong, J., Rivas, M., Mar, G., Kurbatova, N., Griebel, T., Ferreira, P., Barann, M., Wieland, T., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

Larance, M. & Lamond, A.I. (2015). Multidimensional proteomics for cell biology. *Nature Reviews Molecular Cell Biology*, 16(5):269–280.

Latt, S.A., Stetten, G., Juergens, L.A., Willard, H.F., & Scher, C.D. (1975). Recent developments in the detection of deoxyribonucleic acid synthesis by 33258 Hoechst fluorescence. *Journal of Histochemistry and Cytochemistry*, 23(7):493–505.

Law, C.W., Chen, Y., Shi, W., & Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Ferrante, T.C., Terry, R., Turczyk, B.M., Yang, J.L., Lee, H.S., Aach, J., Zhang, K., & Church, G.M. (2015). Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*, 10(3):442–458.

Lee, S., Goddard, M., Visscher, P., & van der Werf, J. (2010). Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. *Genetics, Selection, Evolution*, 42.

Lee, S., Teresa, D., Ripke, S., Yang, J., (PGC-SCZ), S.P.G.A.S.C., (ISC), I.S.C., of Schizophrenia Collaboration (MGS), M.G., Sullivan, P., Goddard, M., Keller, M., Visscher, P., & Wray, N. (2012a). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, 44(3):247–250.

Lee, S., Wray, N., Goddard, M., & Visscher, P. (2011). Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics*, 88(3):294–305.

Lee, S., Yang, J., Chen, G., Ripke, S., Stahl, E., Hultman, C., Sklar, P., Visscher, P., Sullivan, P., Goddard, M., & Wray, N. (2013). Estimation of SNP heritability from dense genotype data. *American Journal of Human Genetics*, 93(6):1151–1155.

Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., & Wray, N.R. (2012b). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542.

Leek, J., Scharpf, R., Bravo, H., Simcha, D., Langmead, B., Johnson, W., Geman, D., Baggerly, K., & Irizarry, R. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739.

Leek, J. & Storey, J. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):1724–1735.

Leek, J.T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research*, 42(21):0.

Lejeune, J., Gautier, M., & Turpin, R. (1959). [study of somatic chromosomes from 9 mongoloid children]. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 248(11):1721–1722.

Levene, H. (1961). Robust tests for equality of variances. *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*, pages 279–292.

Levene, P.A. (1919). The Structure of Yeast Nucleic Acid: IV. Ammonia Hydrolysis. *Journal of Biological Chemistry*, 40(2):415–424.

Levit, S.G. & Pessikova, L.N. (1934). The genetics of diabetes mellitus. *Trudy Medical Genetics Institute Gorky*.

Levy, S., Sutton, G., Ng, P., Feuk, L., Halpern, A., Walenz, B., Axelrod, N., Huang, J., Kirkness, E., Denisov, G., Lin, Y., Jeffrey, M., et al. (2007). The diploid genome sequence of an individual human. *PLoS biology*, 5(10).

Li, B. & Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323.

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

Li, W., Calder, R.B., Mar, J.C., & Vijg, J. (2015). Single-cell transcriptogenomics reveals transcriptional exclusion of ENU-mutated alleles. *Mutation Research*, 772:55–62.

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., & Abecasis, G.R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.*, 21(6):940–951.

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual Reviews Genomics and Human Genetics*, 10:387–406.

Liao, Y., Smyth, G.K., & Shi, W. (2013). The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108.

Liao, Y., Smyth, G.K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740.

Linn, S. & Arber, W. (1968). Host specificity of DNA produced by *Escherichia coli, X*. in vitro restriction of phage fd replicative form. *Proceedings of the National Academy of Sciences of the United States of America*, 59(4):1300–1306.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835.

Lise, S., Clarkson, Y., Perkins, E., Kwasniewska, A., Sadighi Akha, E., Schnekenberg, R., Suminaite, D., Hope, J., Baker, I., Gregory, L., Green, A., Allan, C., et al. (2012). Recessive mutations in SPTBN2 implicate β-III spectrin in both cognitive and motor development. *PLoS genetics*, 8(12).

Listgarten, J., Lippert, C., & Heckerman, D. (2013). FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature Genetics*, 45(5):470–471.

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

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S., & Hirschhorn, J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, 33(2):177–182.

Loman, N., Misra, R., Dallman, T., Constantinidou, C., Gharbia, S., Wain, J., & Pallen, M. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 30(5):434–439.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., et al. (2013). The Genotype-Tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585.

Love, M.I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

Lund, S., Nettleton, D., Davis, M., & Smyth, G. (2012). Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*, 11(5).

Lunter, G. & Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–9.

Maarten Altelaar, A.F., Munoz, J., & Heck, A.J.R. (2012). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, 14(1):35–48.

MacArthur, D., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J., Montgomery, S., Albers, C., Zhang, Z., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science (New York, N.Y.)*, 335(6070):823–828.

Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., Shirley, L.M., Smith, M., Van der Aa, N., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519–522.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21.

Makowsky, R., Pajewski, N.M., Klimentidis, Y.C., Vazquez, A.I., Duarte, C.W., Allison, D.B., & de los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. *PLoS Genetics*, 7(4):e1002051.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Lucia, A., Hunter, D.J., Mccarthy, M.I., Ramos, E.M., Cardon, L.R., Cho, J.H., Guttmacher, A.E., Kong, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Maquat, L. (2002). Skiing toward nonstop mRNA decay. *Science*, 295(5563):2221–2222.

Marioni, J., Mason, C., Mane, S., Stephens, M., & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517.

Martin, H.C., Kim, G.E., Pagnamenta, A.T., Murakami, Y., Carvill, G.L., Meyer, E., Copley, R.R., Rimmer, A., Barcia, G., Fleming, M.R., Kronengold, J., Brown, M.R., et al. (2014). Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. *Human Molecular Genetics*, 23(12):3200–3211.

Mathers, C., Stevens, G., & Mascarenhas, M. (2009). *Global health risks: mortality and burden of disease*. World Health Organization, Geneva.

Mathers, C.D. & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine*, 3(11):e442.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195.

Maxam, A.M. & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564.

Mayr, E. & Provine, W.B. (1998). *The evolutionary synthesis: perspectives on the unification of biology*. Harvard University Press, Cambridge, USA.

Mazutis, L., Gilbert, J., Ung, W.L., Weitz, D.A., Griffiths, A.D., & Heyman, J.A. (2013). Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols*, 8(5):870–891.

McCarthy, D.J., Chen, Y., & Smyth, G. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.

McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M., Gaulton, K., the WGS500 Consortium, Cazier, J.B., & Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Med.*, 6(3):26.

McCarthy, M.I. (2010). Genomics, type 2 diabetes, and obesity. *New England Journal of Medicine*, 363(24):2339–2350.

McDavid, A., Dennis, L., Danaher, P., Finak, G., Krouse, M., Wang, A., Webster, P., Beechem, J., & Gottardo, R. (2014). Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS Computational Biology*, 10(7):e1003696.

McDavid, A., Finak, G., Chattopadyay, P.K., Dominguez, M., Lamoreaux, L., Ma, S.S., Roederer, M., & Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*, 29(4):461–467.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M.A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303.

McKusick-Nathans Institute of Genetic Medicine (2015). OMIM - online mendelian inheritance in man. `http://omim.org/`. Accessed: 2015-7-8.

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

Mendel, G. (1866). Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brunn 4: 3*, 44.

Meselson, M. & Yuan, R. (1968). DNA restriction enzyme from *E. coli*. *Nature*, 217(5134):1110–1114.

Meuwissen, T.H.E., Hayes, B.J., & Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Mewes, H.W., Albermann, K., Bähr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., Pfeiffer, F., & Zollner, A. (1997). Overview of the yeast genome. *Nature*, 387(6632 Suppl):7–65.

Mikkelsen, T.S., Xu, Z., Zhang, X., Wang, L., Gimble, J.M., Lander, E.S., & Rosen, E.D. (2010). Comparative epigenomic analysis of murine and human adipogenesis. *Cell*, 143(1):156–169.

Minkowski, O. (1893). Untersuchungen über den diabetes mellitus nach exstirpation des pankreas. *Archiv fuer experimentalische Pathologie und Pharmakologie*, 31(2-3):85–189.

Miyashita, S.I., Groombridge, A.S., Fujii, S.I., Minoda, A., Takatsu, A., Hioki, A., Chiba, K., & Inagaki, K. (2014). Highly efficient single-cell analysis of microbial cells by time-resolved inductively coupled plasma mass spectrometry. *Journal of Analytical Atomic Spectrometry*, 29(9):1598–1606.

Morgan, T., Sturtevant, A., Muller, H., & Bridges, C. (1915). *The Mechanism Of Mendelian Heredity*. Henry Holt And Company; New York.

Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H.M., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, 44(9):981–990.

Morrow, J.F., Cohen, S.N., Chang, A.C., Boyer, H.W., Goodman, H.M., & Helling, R.B. (1974). Replication and transcription of eukaryotic DNA in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 71(5):1743–1747.

Mortazavi, A., Williams, B., Kenneth, M., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628.

Mott, R., Yuan, W., Kaisaki, P., Gan, X., Cleak, J., Edwards, A., Baud, A., & Flint, J. (2014). The architecture of parent-of-origin effects in mice. *Cell*, 156(1-2):332–342.

Murray, J.C., Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V.C., Sunden, S., & Duyk, G.M. (1994). A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science*, 265(5181):2049–2054.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., & Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64.

Nakamura, T., Yabuta, Y., Okamoto, I., Aramaki, S., Yokobayashi, S., Kurimoto, K., Sekiguchi, K., Nakagawa, M., Yamamoto, T., & Saitou, M. (2015). SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression. *Nucleic Acids Research*, 43(9):e60.

National Center for Biotechnology Information (2013). RefSeq: NCBI reference sequence database.

National Human Genome Research Institute (2015). DNA sequencing costs. `http://www.genome.gov/sequencingcosts/`. Accessed: 2015-7-16.

NHLBI GO Exome Sequencing Project (ESP) (2015). Exome variant server. http://evs.gs.washington.edu/EVS/. Accessed: 2015-1-8.

Nielsen, E.M.D., Hansen, L., Carstensen, B., Echwald, S.M., Drivsholm, T., Glümer, C., Thorsteinsson, B., Borch-Johnsen, K., Hansen, T., & Pedersen, O. (2003). The E23K variant of kir6.2 associates with impaired post-OGTT serum insulin response and increased risk of type 2 diabetes. *Diabetes*, 52(2):573–577.

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F., & C, O. (1965a). RNA codewords and protein synthesis, VII. on the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America*, 53(5):1161–1168.

Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottman, F., & C, O. (1965b). RNA codewords and protein synthesis, VII. on the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America*, 53(5):1161–1168.

Nirenberg, M.W., Jones, O.W., Leder, P., Clark, B.F.C., Sly, W.S., & Pestka, S. (1963). On the coding of genetic information. *Cold Spring Harbor Symposia on Quantitative Biology*, 28:549–557.

Nobel Media AB (2014). The nobel prize in physiology or medicine 1923. http://www.nobelprize.org/nobel_prizes/medicine/laureates/1923/index.html. Accessed: 2014-8-25.

Oakley, W.G. (1962). Man may be the captain of his fate, but he is also the victim of his blood sugar.

Olson, M., Hood, L., Cantor, C., & Botstein, D. (1989). A common language for physical mapping of the human genome. *Science*, 245(4925):1434–1435.

Olson, M.V. (1993). The human genome project. *Proceedings of the National Academy of Sciences of the United States of America*, 90(10):4338–4344.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2):256–278.

Pacanowski, M.A., Hopley, C.W., & Aquilante, C.L. (2008). Interindividual variability in oral antidiabetic drug disposition and response: the role of drug transporter polymorphisms. *Expert Opinions in Drug Metabolism and Toxicology*, 4(5):529–544.

Palles, C., Cazier, J., Howarth, K., Domingo, E., Jones, A., Broderick, P., Kemp, Z., Spain, S., Guarino, E., Guarino Almeida, E., Salguero, I., Sherborne, A., et al. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, 45(2):136–144.

Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., NISC Comparative Sequencing Program, Black, B.L., Visel, A., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44):17921–17926.

Pasquali, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J.J., Morán, I., Gómez-Marín, C., van de Bunt, M., Ponsa-Cobas, J., Castro, N., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics*, 46(2):136–143.

Patro, R., Duggal, G., & Kingsford, C. (2015). Salmon: Accurate, versatile and ultrafast quantification from RNA-seq data using Lightweight-Alignment. *bioRxiv*.

Patro, R., Mount, S.M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464.

Patterson, N., Price, A., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2(12).

Pearson, K. (1901). LIII. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2.

Phipson, B. & Oshlack, A. (2014). DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biology*, 15(9):465.

Picelli, S., Björklund, A.K., Faridani, O.R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–1098.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using smart-seq2. *Nature Protocols*, 9(1):171–181.

Pincus, G. & White, P. (1933). On the inheritance of diabetes mellitus. *Proceedings of the National Academy of Sciences of the United States of America*, 19(6):631–635.

Pincus, G. & White, P. (1934a). On the inheritance of diabetes mellitus: II. further analysis of family histories. *The American Journal of the Medical Sciences*, 188(2):159.

Pincus, G. & White, P. (1934b). On the inheritance of diabetes mellitus. III. the blood sugar values of the relatives of diabetics. *The American Journal of the Medical Sciences*, 188(6):782.

Pirinen, M., Donnelly, P., & Spencer, C.C.A. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Annals of Applied Statistics*, 7(1):369–390.

Pollak, M.R., Chou, Y.H., Cerda, J.J., Steinmann, B., La Du, B.N., Seidman, J.G., & Seidman, C.E. (1993). Homozygosity mapping of the gene for alkaptonuria to chromosome 3q2. *Nature Genetics*, 5(2):201–204.

Poulsen, P., Kyvik, K.O., Vaag, A., & Beck-Nielsen, H. (1999). Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42:139–145.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., Li, H., Mallick, S., et al. (2014). The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49.

Pruitt, K. & Brown, G. (2013). RefSeq frequently asked questions (FAQ).

Pruitt, K., Brown, G., Tatusova, T., & Maglott, D. (2002 [Updated 2012 Apr 6]). The reference sequence (RefSeq) database. In J. McEntyre & J. Ostell, editors, *The NCBI Handbook [Internet]*, page Chapter 18. National Center for Biotechnology Information (US), Bethesda (MD). Available from: http://www.ncbi.nlm.nih.gov/books/NBK21091/.

Pruitt, K., Harrow, J., Harte, R., Wallin, C., Diekhans, M., Maglott, D., Searle, S., Farrell, C., Loveland, J., Ruef, B., Hart, E., Suner, M., et al. (2009a). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, 19(7):1316–1323.

Pruitt, K., Harrow, J., Harte, R., Wallin, C., Diekhans, M., Maglott, D., Searle, S., Farrell, C., Loveland, J., Ruef, B., Hart, E., Suner, M.M., et al. (2009b). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, 19(7):1316–1323.

Pruitt, K., Tatusova, T., Brown, G., & Maglott, D. (2012). NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, 40(Database issue):D130–D135.

Przeworski, M. (2011). The golden age of human population genetics. *Science*, 331(6017):547.

Purcell, S., Neale, B., Kathe, T., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., & Sham, P. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.

Python Software Foundation (2013). *Python Programming Language*. Python Software Foundation, Beaverton, USA.

Quaas, R.L. & Pollak, E.J. (1980). Mixed model methodology for farm and ranch beef cattle testing programs. *Journal of animal science*, 51(6):1277–1287.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43(3-4):353–360.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2015a). The comprehensive R archive network. `https://cran.r-project.org/`. Accessed: 2015-7-21.

R Core Team (2015b). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raj, A. & van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226.

Raj, A. & van Oudenaarden, A. (2009). Single-molecule approaches to stochastic gene expression. *Annual Reviews Biophysics*, 38:255–270.

Ramsköld, D., Luo, S., Wang, Y.C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., Schroth, G.P., & Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C.O., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752.

Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.a., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J., Pohl, A., Pheasant, M., et al. (2010a). The UCSC Genome Browser database: update 2010. *Nucleic Acids Research*, 38(Database issue):D613–9.

Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.a., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J., Pohl, A., Pheasant, M., et al. (2010b). The UCSC genome browser database: update 2010. *Nucleic Acids Research*, 38(Database issue):D613–9.

Rheinberger, H.J. (1995). When did carl correns read gregor mendel's paper? a research note. *Isis*, pages 612–616.

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., WGS500 Consortium, Wilkie, A.O.M., McVean, G., & Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46(8):912–918.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517.

Risso, D., Ngai, J., Speed, T.P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., & Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Roberts, A. & Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73.

Robinson, M., McCarthy, D., & Smyth, G. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140.

Robinson, M. & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3).

Robinson, M. & Smyth, G. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics (Oxford, England)*, 23(21):2881–2887.

Robinson, M. & Smyth, G. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics (Oxford, England)*, 9(2):321–332.

Royer-Pokora, B., Kunkel, L.M., Monaco, A.P., Goff, S.C., Newburger, P.E., Baehner, R.L., Cole, F.S., Curnutte, J.T., & Orkin, S.H. (1986). Cloning the gene for an inherited human disorder–chronic granulomatous disease–on the basis of its chromosomal location. *Nature*, 322(6074):32–38.

RStudio (2015). R Markdown—Dynamic documents for R. `http://rmarkdown.rstudio.com/`. Accessed: 2015-7-26.

Rubin, G.M. & Spradling, A.C. (1982). Genetic transformation of drosophila with transposable element vectors. *Science*, 218(4570):348–353.

Saadatpour, A., Guo, G., Orkin, S.H., & Yuan, G.C. (2014). Characterizing heterogeneity in leukemic cells using single-cell gene expression analysis. *Genome Biology*, 15(12):525.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., & Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491.

Sandhu, M.S., Weedon, M.N., Fawcett, K.A., Wasson, J., Debenham, S.L., Daly, A., Lango, H., Frayling, T.M., Neumann, R.J., Sherva, R., Blech, I., Pharoah, P.D., et al. (2007). Common variants in WFS1 confer risk of type 2 diabetes. *Nature Genetics*, 39(8):951–953.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., & Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):687–695.

Sanger, F. & Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., & Reich, D. (2014). The genomic landscape of neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357.

Santos, A., Wernersson, R., & Jensen, L.J. (2015). Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Research*, 43(Database issue):D1140–4.

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T., & Ueda, H.R. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biology*, 14(4):R31.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., & Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Res.*, 22(9):1748–1759.

Schiffels, S. & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925.

Schrijver, I., Aziz, N., Farkas, D., Furtado, M., Gonzalez, A., Greiner, T., Grody, W., Hambuch, T., Kalman, L., Kant, J., Klein, R., Leonard, D., et al. (2012). Opportunities and challenges associated with clinical diagnostic genome sequencing: a report of the association for molecular pathology. *The Journal of Molecular Diagnostics*, 14(6):525–540.

Schwarz, J., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, 7(8):575–576.

Scialdone, A., Natarajan, K.N., Saraiva, L.R., Proserpio, V., Teichmann, S.A., Stegle, O., Marioni, J.C., & Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*.

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

Searle, S., Frankish, A., Bignell, A., Aken, B., Derrien, T., Diekhans, M., Harte, R., Howald, C., Kokocinski, F., Lin, M., Tress, M., Baren, M.V., et al. (2010). The GENCODE human gene set. *Genome Biology*, 11.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J.J., Gennert, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., Schwartz, S., Fowler, B., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369.

Shapiro, E., Biezuner, T., & Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630.

Sharma, V., Fenwick, A., Brockop, M., Simon, M., Goos, J., Hoogeboom, A., Brady, A., Jeelani, N., Lynch, S., Mulliken, J., Murray, D., Phipps, J., et al. (2013). Mutations in TCF12, encoding a basic helix-loop-helix partner of TWIST1, are a frequent cause of coronal craniosynostosis. *Nature Genetics*, 45(3):304–307.

Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

Shi, Q., Qin, L., Wei, W., Geng, F., Fan, R., Shin, Y.S., Guo, D., Hood, L., Mischel, P.S., & Heath, J.R. (2012). Single-cell proteomic chip for profiling intracellular signaling pathways in single tumor cells. *Proceedings of the National Academy of Sciences of the United States of America*, 109(2):419–424.

Shiroguchi, K., Jia, T.Z., Sims, P.A., & Xie, X.S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1347–1352.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., & Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18):8794–8797.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885.

Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., & Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–820.

Smith, H.O. & Nathans, D. (1973). Letter: A suggested nomenclature for bacterial host modification and restriction systems and their enzymes. *Journal of Molecular Biology*, 81(3):419–423.

Smith, H.O. & Wilcox, K.W. (1970). A restriction enzyme from *Hemophilus influenzae. I.* purification and general properties. *Journal of Molecular Biology*, 51(2):379–391.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article3.

Soumillon, Cacchiarelli, M., Semrau, D., van Oudenaarden, S., Mikkelsen, A., & Tarjei, S. (2014). Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*.

Speed, D. & Balding, D.J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.*, 24(9):1550–1557.

Speed, D., Hemani, G., Johnson, M., & Balding, D. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, 91(6):1011–1021.

Spudich, J.L. & Koshland, Jr, D.E. (1976). Non-genetic individuality: chance in the single cell. *Nature*, 262(5568):467–471.

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507.

Stegle, O., Teichmann, S.A., & Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145.

Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S., Baker, A., Snorradottir, S., et al. (2007). A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature Genetics*, 39(6):770–775.

Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F., & Huang, Y. (2014). Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 111(19):7048–7053.

Sturtevant, A. (1913a). Factors in Drosophila. *The Journal of Experimental Zoology*, 14(3):43.

Sturtevant, A. (1913b). The linear arrangement of six sex-linked factors in Drosophila as shown by their mode of association. *Molecular and General Genetics*, 10(1):293–294.

Sutton, W.S. (1903). The chromosomes in heredity. *The Biological Bulletin*, 4(5):231–250.

Tabor, H.K., Risch, N.J., & Myers, R.M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*, 3(5):391–397.

Tang, F., Barbacioru, C., Nordman, E., Li, B., Xu, N., Bashkirov, V.I., Lao, K., & Surani, M.A. (2010). RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nature Protocols*, 5(3):516–535.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., & Surani, M.A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382.

Taniguchi, Y., Choi, P.J., Li, G.W., Chen, H., Babu, M., Hearn, J., Emili, A., & Xie, X.S. (2010). Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538.

Taschner, P.E.M. & den Dunnen, J.T. (2011). Describing structural changes by extending HGVS sequence variation nomenclature. *Human Mutation*, 32(5):507–511.

Taylor, J.C., Martin, H.C., Lise, S., Broxholme, J., Cazier, J.B., Rimmer, A., Kanapin, A., Lunter, G., Fiddy, S., Allan, C., Aricescu, A.R., Attar, M., et al. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*, 47(7):717–726.

Teslovich, T.M., Mahajan, A., Flannick, J., Fuchsberger, C., Fontanillas, P., Morris, A.P., Rivas, M.A., Cingolani, P., Perry, J.R.B., Tajes, J.F., Kang, H.M., Sim, X., et al. (2015). Variation in protein-coding sequence and predisposition to type 2 diabetes. *(submitted)*.

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

The Perl Foundation (2013). *Perl Programming Language*. The Perl Foundation, Walnut, USA.

Thomas, D. (2010). Gene–environment-wide association studies: emerging approaches. *Nature reviews. Genetics*, 11(4):259–272.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., & Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31(1):46–53.

Trapnell, C., Pachter, L., & Salzberg, S. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–1111.

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., & Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–375.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., & Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2):124–130.

Trynka, G., Westra, H.J., Slowikowski, K., Hu, X., Xu, H., Stranger, B.E., Klein, R.J., Han, B., & Raychaudhuri, S. (2015). Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within Complex-Trait loci. *The American Journal of Human Genetics*, 97(1):139–152.

Tschermak, E. (1900). *Über künstliche Kreuzung bei Pisum sativum*. E. Tschermak.

Tsui, L.C. & Dorfman, R. (2013). The cystic fibrosis gene: a molecular genetic perspective. *Cold Spring Harbor Perspectives on Medicine*, 3(2):a009472.

Vallejos, C.A., Marioni, J.C., & Richardson, S. (2015). BASiCS: Bayesian analysis of Single-Cell sequencing data. *PLoS Computational Biology*, 11(6):e1004333.

Van der Aa, N., Zamani Esteki, M., Vermeesch, J.R., & Voet, T. (2013). Preimplantation genetic diagnosis guided by single-cell genomics. *Genome Med.*, 5(8):71.

Van der Maaten, L. (2009). Learning a parametric embedding by preserving local structure. In *International Conference on Artificial Intelligence and Statistics*, pages 384–391. machine-learning.wustl.edu.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85.

Van der Maaten, L. & Hinton, G. (2012). Visualizing non-metric similarities in multiple maps. *Machine Learning*, 87(1):33–55.

van Hoof, A., Frischmeyer, P., Dietz, H., & Parker, R. (2002). Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science (New York, N.Y.)*, 295(5563):2262–2264.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O., & Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science*, 280(5369):1540–1542.

Ver Hoef, J.M. (2012). Who invented the delta method? *American Statistician*, 66(2):124–127.

Vertex Pharmaceuticals Incorporated (2015). Kalydeco. `http://www.kalydeco.com/`. Accessed: 2015-7-23.

Visscher, P.M., Hill, W.G., & Wray, N.R. (2008). Heritability in the genomics era–concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., McCulloch, L.J., Ferreira, T., et al. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics*, 42(7):579–589.

von Mering, J. & Minkowski, O. (1889). Diabetes mellitus nach pankreasexstirpation. *Zeitschrift für klinische Medizin*, 10(393).

von Mering, J. & Minkowski, O. (1890). Diabetes mellitus nach Pankreasexstirpation. *Archiv für Experimentelle Pathologie und Pharmakologie*, 26(5-6):371–387.

Wang, K. (2013). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16).

Waterfield, T. & Gloyn, A.L. (2008). Monogenic β-cell dysfunction in children: clinical phenotypes, genetic etiology and mutational pathways. *Pediatric Health*, 2(4):517–532.

Watson, J.D., Crick, F.H.C., & Others (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738.

Weber, J.L. & May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *The American Journal of Human Genetics*, 44(3):388–396.

Weinberg, W. (1908). On the demonstration of heredity in man. In Boyer, SH, trans (1963), editor, *Papers on human genetics*. Prentice Hall, Englewood Cliffs, NJ.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., & Others (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., & Lathrop, M. (1992). A second-generation linkage map of the human genome. *Nature*, 359(6398):794–801.

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

Wellcome Trust Sanger Institute (2012). Human and vertebrate analysis and annotation (havana). Available at: http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/. Accessed: 25 October 2012.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & Parkinson, H. (2014a). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(Database issue):D1001–6.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & Parkinson, H. (2014b). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(Database issue):D1001–6.

White, P., Joslin, E.P., & Pincus, G. (1934). The inheritance of diabetes. *Journal of the American Medical Association*, 103(2):105–106.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Wickham, H. (2011a). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3.

Wickham, H. (2011b). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Wickham, H. & Francois, R. (2015). *dplyr: A Grammar of Data Manipulation*. R package version 0.4.2.

Wildeman, M., van Ophuizen, E., den Dunnen, J.T., & Taschner, P.E. (2008). Improving sequence variant descriptions in mutation databases and literature using the mutalyzer sequence variation nomenclature checker. *Human Mutation*, 29(1):6–13.

Wilke, C.O. (2015). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 0.5.0.

Wills, Q.F., Livak, K.J., Tipping, A.J., Enver, T., Goldson, A.J., Sexton, D.W., & Holmes, C. (2013). Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature Biotechnology*, 31(8):748–752.

Wills, Q.F. & Mead, A.J. (2015). Application of single-cell genomics in cancer: promise and challenges. *Human Molecular Genetics*.

Wilson, G., Aruliah, D.A., Brown, C.T., Chue Hong, N.P., Davis, M., Guy, R.T., Haddock, S.H.D., Huff, K.D., Mitchell, I.M., Plumbley, M.D., Waugh, B., White, E.P., & Wilson, P. (2014). Best practices for scientific computing. *PLoS Biology*, 12(1):e1001745.

Wilson, N.K., Kent, D.G., Buettner, F., Shehata, M., Macaulay, I.C., Calero-Nieto, F.J., Sánchez Castillo, M., Oedekoven, C.A., Diamanti, E., Schulte, R., Ponting, C.P., Voet, T., et al. (2015). Combined Single-Cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell*, 16(6):712–724.

Winckler, W., Weedon, M.N., Graham, R.R., McCarroll, S.A., Purcell, S., Almgren, P., Tuomi, T., Gaudet, D., Boström, K.B., Walker, M., Hitman, G., Hattersley, A.T., et al. (2007). Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. *Diabetes*, 56(3):685–693.

Witte, J.S., Visscher, P.M., & Wray, N.R. (2014). The contribution of genetic variants to disease depends on the ruler. *Nature Reviews Genetics*, 15(11):765–776.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., Amin, N., Buchkovich, M.L., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186.

World Health Organisation (2014). About diabetes. `http://www.who.int/diabetes/action_online/basics/en/`. Accessed: 2014-8-25.

Xie, Y. (2013). *Dynamic Documents with R and knitr*, volume 29. CRC Press.

Xu, C. & Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., & Others (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500(7464):593–597.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural and Molecular Biology*, 20(9):1131–1139.

Yang, J., Benyamin, B., Brian, M., Gordon, S., Henders, A., Nyholt, D., Madden, P., Heath, A., Martin, N., Montgomery, G., Goddard, M., & Visscher, P. (2010a). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.

Yang, J., Lee, S., Goddard, M., & Visscher, P. (2011a). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1):76–82.

Yang, J., Manolio, T., Pasquale, L., Boerwinkle, E., Caporaso, N., Cunningham, J., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M., Hill, W., Landi, M., et al. (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, 43(6):519–525.

Yang, J., Wray, N., & Visscher, P. (2010b). Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genetic Epidemiology*, 34(3):254–257.

Yau, C. & Pierson, E. (2015). ZIFA: Dimensionality reduction for zero-inflated single cell gene expression analysis. Preprint.

Yu, X. & Sun, S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC bioinformatics*, 14.

Zaitlen, N. & Kraft, P. (2012). Heritability in the genome-wide association era. *Human Genetics*, 131(10):1655–1664.

Zamboni, N., Saghatelian, A., & Patti, G.J. (2015). Defining the metabolome: Size, flux, and regulation. *Molecular Cell*, 58(4):699–706.

Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R.B., Rayner, N.W., Freathy, R.M., Barrett, J.C., Shields, B., et al. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341.

Zeisel, A., Manchado, A.B.M.n., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., & Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*.

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

Zhou, X., Carbonetto, P., & Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, 9(2):e1003264.

Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., Bennett, D.A., Houmard, J.A., et al. (2013). Genome-wide

chromatin state transitions associated with developmental and environmental cues. *Cell*, 152(3):642–654.

Zimmet, P., Alberti, K.G., & Shaw, J. (2001). Global and societal implications of the diabetes epidemic. *Nature*, 414(6865):782–787.

Zuk, O., Hechter, E., Sunyaev, S.R., & Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1193–1198.

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., & Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4):E455–64.

Glossary

B-lymphocyte a lymphocyte not processed by the thymus gland, and responsible for producing antibodies (a.k.a. B-cell). 125

epithelial of the thin tissue formin the outer layer of a body's surface and lining the alimentary canal and other hollow structures. 126

Epstein-Barr Virus a herpesvirus causing glandular fever and associated with certain cancers; one of the most common viruses in humans. 125

exon a segment of a DNA or RNA molecule containing information coding for a protein or peptide sequence. 22

fibroblast a cell in connective tissue which produces collagen and other fibres. 126

glucagon a hormone formed in the pancreas which promotes the breakdown of glycogen to glucose in the liver. 66

GRM genetic relatedness matrix. 83, 88, 89, 93–95, 114, 131, 134, 141, 160, 162, 184–186

GWAS genome-wide association study. 10–14, 67–70, 73, 76, 77, 90, 109, 113, 167, 193, 221

heritability degree of trait variance in a population attributable to genetic factors. 70–72, 141

hormone a regulatory substance, either produced in an organism or synthetically, and transported in tissue fluids such as blood or sap to stimulate specific cells or tissues into action. 64

insulin a hormone produced in the pancreas by the islets of Langerhans, which regulates the amount of glucose in the blood. 64–66

intron a segment of a DNA or RNA molecule which does not code for protein sequence; in humans, introns are transcribed from genomic DNA, but spliced out of the transcript before translation of the RNA sequence into protein sequence. 22

keratinocyte an epidermal cell that produces keratin. 126

lymphoblast an immature lymphocyte that can by activated by an antigen and differentiate to form mature lymphocutes while making clones of its original naïve cells. 125

myoblast embryonic stem cell which becomes a muscle cell or fibre. 126

pancreas a large gland behind the stomach which secretes digestive enzymes into the duodenum. Embedded in the pancreas are the islets of Langerhans, which secrete into the blood the hormones insulin and glucagon. 65, 66

pancreatic islet or islet of Langerhans is a region of the pancreas that contains its endocrine (i.e. hormone-producing) cells. Pancreatic beta cells (β cells) produce insulin. 65

Acronyms

BMI body mass index. 69

cDNA complementary DNA. 201, 204–206

ChIP chromatin immunoprecipitation. 125, 126, 144, 198

CPM counts-per-million. 207, 208, 230, 232, 240, 263, 275

DE differential expression. 213–218, 261

DNA deoxyribonucleic acid. 5–11, 15, 16, 18, 23, 76, 195, 196, 226, 247, 269

EST expressed-sequence tag. 7

FACS fluorescence-activated cell sorting. 201, 226

FPKM fragments per kilobase per million mapped. 207, 208, 215, 230, 232, 263, 264, 275

GAM generalized additive model. 215

GoT2D Genetics of Type 2 Diabetes. 64, 71, 75–80, 89, 92–94, 98, 100, 101, 112, 113, 117, 119, 120, 122, 123, 141, 153, 157, 167, 182, 183, 186, 188, 189, 194

HGP Human Genome Project. 7–9, 12

HGVS Human Genome Variation Society. 23, 24

HSC haematopoietic stem cell. 225, 226, 245, 256

IVT in vitro transcription. 201

LD linkage disequilibrium. 112–114, 134, 160, 168, 169, 171, 184

LMM linear mixed model. 63, 71, 73, 74, 76, 77, 82–84, 86–89, 92, 94–97, 103, 104, 106, 112–114, 120, 122–124, 128, 131, 142, 153, 160, 168, 169, 171, 182, 184, 186, 190, 191, 193

LoF loss-of-function. 30–35, 39, 48, 49, 58–60

MAF minor allele frequency. 10, 11, 77, 78, 82, 83, 85, 93–95, 97–111, 113–116, 118–120, 122, 123, 129, 133–148, 150, 151, 154–156, 158, 159, 161, 164–178, 180–190, 192–194

mRNA messenger RNA. 5, 7, 195, 200, 201, 203, 204

ncRNA noncoding ribonucleic acid. 144, 145, 149, 152

PCA principal components analysis. 212, 237, 249–251, 255–257, 272

PCR polymerase chain reaction. 201, 214, 215, 219, 271

REML residual (or restricted) maximum likelihood. 74, 88, 90–92, 95, 102, 111, 118, 122, 131, 141, 162, 163, 182, 183, 190

RNA ribonucleic acid. 5, 18, 23, 195, 203–206, 208–210, 226, 247, 255, 256, 269

RNA-seq RNA sequencing. 196, 197, 199, 203–205, 207, 208, 210–214, 217–219, 225, 228, 230, 245, 262, 265, 266, 269, 270, 272

scRNA-seq single-cell RNA sequencing. 198–222, 224–226, 228, 240, 242, 249, 254, 258, 259, 261–264, 266, 268–271

SNP single nucleotide polymorphism. 10, 13, 64, 70, 75–79, 94, 98, 101, 114, 129, 142, 153, 165, 166, 195

T1D type 1 diabetes. 65

T2D type 2 diabetes. 63, 65–71, 75–78, 80, 82, 89, 90, 92, 94, 97–101, 103, 104, 109, 111, 114, 115, 117–120, 122–125, 127, 128, 133, 136, 138, 141–143, 145, 147, 149, 151–153, 158, 162–164, 167, 169, 172, 173, 177, 178, 181, 183, 191–194

TFBS transcription factor binding site. 125, 130, 144–147, 149

TPM transcripts-per-million. 204, 207, 208, 229, 230, 232, 240, 241, 256, 263, 264, 275

t-SNE t-distributed stochastic neighbour embedding. 212, 213, 237, 240, 245, 249–251, 254–257, 272

UMI unique molecule identifier. 203

UTR untranslated region. 22, 27, 33, 125, 144, 145, 149, 209

VC variance component. 97, 101, 103, 135, 140–142, 177–180, 184, 185

VCF variant call format. 29, 30

VE variance explained. 75, 85, 92, 122, 168, 176, 178, 181–183

WES whole-exome sequence. 78, 79

WGS whole-genome sequence. 78, 79