Chapter 1

Introduction

Life is amazing in its complexity, yet robustness. Eloquent, intricate features are faithfully transmitted from parents to their children, generation after generation. For most species, the progeny start out as a single cell. Thus, all the information necessary to reproduce the traits of the parents, as well as the blueprints for the machinery to perform the reproduction, must be encoded in that tiniest of volumes of $10^{-13}\mathrm{m}^3$. Encoding all the heritable information in one cell, and robustly reproducing it is a miracle. I want to understand how this fascinating, important process works.

Transmission of heritable information is interesting in itself, but it is also central to many questions about human health. For example, the genetic background of the parents can determine not only that a human child will have ears and toes, but also a ten-fold higher risk of developing cancer forty years down the road (Liede et al., 2004). We need to identify, quantify, and understand the mechanisms of the genetic risks to be able to prevent or treat the onset of the disease. This requires an understanding of trait genetics in general.

There are two aspects to understanding the heritable component of a phenotype. First, where in the genome are specific traits encoded? This problem is that of mapping heritable traits, where great advances have been made in the last 50 years, which I will review below. Second, how does a specific region of the genome define the trait? This is a problem of identifying the effect of genetic information on organism features, and the area of functional genomics in general.

Understanding the effect of genotype on cellular traits is a prerequisite for understanding the genetics of tissue and organ phenotypes that ultimately explain global characteristics, such as human disease risk. In this Thesis, I address the questions of finding and interpreting genetic effects together by focusing on mapping cellular traits.

In the introductory chapter, I first outline the history, methods and progress in trait mapping so far, focusing ultimately on human studies. I then discuss the current state of mapping cellular traits using these methods, as well as some more specific approaches available only in unicellular organisms. Finally, I introduce the common statistical models and methods used for genetic mapping of low- and high-dimensional traits discussed. However, I will not address the vast literature on modelling variability in high-dimensional traits in general.

1.1 Mapping the genetic basis of heritable traits

The heritable information is encoded in the genome. It is instructive to understand how we have come to know this most basic trait mapping result to appreciate the current opportunities as well as outstanding questions.

We can only use biological assays which give us a readout we can visualise. Thus, until the development of high quality microscopy, cellular analyses were impossible, and budding geneticists used plants and domestic animals for their experiments. The hero of genetics, Gregor Mendel, worked in a quiet monastery in Brno on crossing peas in the 1860s, and produced a paper on segregation of traits in an obscure journal (Mendel, 1865), which would be unlikely to be read today, as it was not in English, and probably not peer reviewed. This paper, showing the existence of dominant and recessive alleles, as opposed to a continuous distribution of traits among the progeny, failed to make an impact.

In 1869, a doctor named Friedrich Miescher was working in Tübingen, and managed to isolate an acidic, phosphate-rich substance from the pus of the used bandages (Dahm, 2008). This was the first time DNA had been purified. Like Mendel's discoveries, its importance became known only later.

Mendel's work was rediscovered at the end of the century by a Dutchman Hugo de Vries. He spent his life replicating and extending Mendel's experiments,

crossing plants, and phenotyping the progeny in his Amsterdam estate. His, and William Bateson's series of papers and monographs (Bateson, 1909; de Vrijes, 1901) established the foundation of genetics at the brink of the last century.

Thomas Hunt Morgan and his student Alfred Sturtevant pursued visible, selectable phenotypes, and analysed their inheritance patterns in the fruit fly *Drosophila melanogaster*, and established that "genes" were actually on chromosomes, and arranged linearly (Morgan, 1910).

However, not all traits were readily visualisable under a microscope, so different assays had to be used to make progress on understanding where the heritable information lies. Radioactive isotopes had become widely available, with one application as a tag for specific molecules to give a readout for the abundance of that molecule. By 1952, it was established that DNA was the carrier of genetic information, in a classic paper by Hershey and Chase, who measured radioactivity in a viral infection experiment (Hershey and Chase, 1952). The DNA or proteins of the T2 phage were tagged with heavy isotopes, and the infected cells were tested for radioactivity readout, which confirmed DNA as the carrier of genetic material.

X-ray crystallography, another way of getting a visual readout of biological information, allowed arguably the greatest breakthrough of the last century, as Crick and Watson used Franklin's DNA diffraction pattern image to give the physical model of the DNA double helix (Watson and Crick, 1953). The central dogma of molecular biology and the genetic code (Crick, 1970; Gardner et al., 1962) were established shortly thereafter. This completed the basic understanding of the molecules involved in transmitting heritable information.

Once it was established that DNA is the carrier of genetic information, and stretches of nucleotide sequence determine the functional outcome according to the central dogma, the next big questions concerned gathering the genetic information. Southern, northern, and western blots were developed to visualise the size distribution, and sequence of DNA, RNA, and proteins (Alwine et al., 1977; Southern, 1975; Towbin et al., 1979). The ability to query the sequence of the heritable information, and the rapid development of methods to scale up capacity, resulted in an exponential increase in sequence data. The full genetic makeup of the first genome was first established for the bacteriophage lambda (Sanger et al.,

1982), followed by the first free living organism *H. influenzae* (Fleischmann et al., 1995), the first eukaryote *S. cerevisiae* (Goffeau et al., 1996), the first multicellular organism *C.elegans* (C. elegans Sequencing Consortium, 1998), and culminating with the human genome in 2001 (Lander et al., 2001).

The last decade has seen work building on the success story of decoding the human genome and those of model organisms. One example of this is the application of genotyping and gene expression arrays, that use the sequence at polymorphic sites or coding regions, to assay the genetic state or mRNA expression level at specific loci. We can already produce very large quantities of sequence data in a routine fashion. The per-base sequencing costs are decreasing, and technologies are constantly improving, with polony-, nanopore-, or ion capture based approaches yielding promising results. The hurdles of understanding the nature of heritable information, and measuring it, have been largely cleared. Now, combining the relatively cheap and accessible sequence or variation data capture with phenotype assays has enabled genetic mapping of many traits using methods I will outline next. I will not cover effects of other inherited state, such as methylation, chromatin state, etc. since downstream phenotypes are largely independent of them if the transcript levels are measured.

1.1.1 Linkage mapping

Genetic information is passed on in chromosomes. In the case of sexually reproducing eukaryotes, the child inherits a copy of each chromosome from both parents via a haploid zygote. During meiosis, the chromosomes recombine, forming the final haplotypes of the child that are made up of contiguous tracts of DNA coming from one parent (Alberts et al., 2007). If a trait cosegregates with one allele of a specific locus, the correlation of the locus genotype and individual phenotype can be used to map the trait.

Human pedigrees

Mendelian traits, such as the pea flower colour or leaf crumbliness, are single gene traits of full penetrance. The trait is determined by one gene only, and a specific genotype confers a certainty of observing the trait. These traits are well amenable

to linkage mapping approaches, as their segregation can be traced in individual pedigrees. One prominent example was observed in the progeny of European and Russian monarchs in the late 19th and early 20th centuries, whose ranks were thinned by haemophilia, an X-linked recessive disease (Ingram, 1976).

The original linkage mapping approach, established by Morgan and others, showed that genes lie on chromosomes, and traits in the fruit fly such as eye colour, wing defects, etc. were mapped in the early last century (Green, 2010). Linkage mapping approaches for cellular traits were further developed in the 1970's and 80's (Petes and Botstein, 1977). These ideas were soon expanded to humans, where restriction fragment length variants were postulated and shown to be polymorphic in the human population (Lander and Botstein, 1989). The application of these methods led to discovery of the genetic basis of Huntington disease (The Huntington's Disease Collaborative Research Group, 1993) as well as cystic fibrosis (Rommens et al., 1989).

Linkage to polymorphic sites has implicated many regions for disease risk, but for many rare conditions, the causative alleles are either private, or not polymorphic in human population. Some of such variants have been recently been identified. Most recently, a causal mutation was identified by whole-genome sequencing of a single family with four cases of a rare Charcot Marie Tooth disease (Lupski et al., 2010). Furthermore, disease genes have been mapped by whole-exome sequencing of a very small number of diseased families (Ng et al., 2010) or a single case in one proband (Sobreira et al., 2010).

While many genes have been mapped using linkage in human pedigrees, Mendelian traits constitute a minority of human diseases with a genetic component.

Designed crosses in model organisms

Traits have been mapped by linkage using segregation in general pedigrees, but pedigrees of controlled structure are often used. Controlled crosses of haploid or homozygous inbred lines produce progeny with predictable genotypes, that can be further crossed in more intricate designs. This approach is obviously feasible

only in model organisms, and has been used with great success. I will focus on a specific design, an F_n cross.

Typically, two phenotypically different parental strains are crossed to produce large numbers of progeny (F_1 generation). The children are then phenotyped and genotyped, and the genetic basis of the trait can be mapped using linkage. In case of a larger generation cross, additional rounds of crossing between the children are undertaken (F_n cross, $n > 1$). This has the effect of reducing the size of contiguous blocks of genetic material inherited from one of the original parents, and reduces linkage between nearby loci. Given a sufficiently dense genetic map, it allows mapping to considerably finer intervals (Darvasi and Soller, 1995).

A large body of work has focused on two F_1 crosses between haploid yeast strains. The Kruglyak lab has used a cross between a laboratory and a wineyard strain to study genetics of gene expression, proteome variation, small molecule response, and gene-environment interactions (Brem et al. 2002, Brem et al. 2005, Brem and Kruglyak 2005, Foss et al. 2007, Perlstein et al. 2007, Ehrenreich et al. 2009, Ehrenreich et al. 2010). Another cross was used in a series of studies by Steinmetz et al. to map and dissect QTLs, as well as study the recombinational landscape in yeast (Steinmetz et al. 2002, Sinha et al. 2006, Wei et al. 2007, Mancera et al. 2008, Zheng et al. 2010).

In diploid organisms, genetics is greatly simplified if individuals are inbred to homozygosity. Inbred diploid F_1 cross progeny (recombinant inbred lines) of two strains have been developed and used to map traits in *Caernohabitis elegans*, *Drosophila melanogaster*, mouse, and rat (Ayyadevara et al., 2003; Doroszuk et al., 2009; Voigt et al., 2008) and reviewed by Flint and Mackay (2009).

1.1.2 Association studies

Over two thousand Mendelian traits have been mapped by linkage in humans to date. However, this approach did not work for many conditions common in humans, such as diabetes, that cluster in families. While it was clear that these diseases are heritable, the genes could not be traced via transmission in pedigrees. A view emerged that many common traits are polygenic, with many loci contributing. A different mapping approach was needed.

Case-control GWAS

One alternative way to map disease genes is to compare frequencies of alleles in healthy individuals (controls) and ones having the disease (cases) at many loci in the genome to test for association of one allele to the disease. The important distinction is that there is no family structure present, and the individuals are assumed to be independent.

Initially, as genotyping was expensive, and numbers of assayable polymorphisms small, this was done for candidate genes, such as the HLA locus (Cudworth and Woodrow, 1976). As more human variation data became available, the power of association studies improved. The number of loci mapped using association has steadily risen. While only a handful of loci were reproducibly mapped by the late 1990's, the data from human genome project allowed the development of genotyping arrays to query the known common segregating sites.

Since genotyping tens of thousands to millions of markers became standard, a steady and impressive march of genome-wide association studies has produced hundreds of loci contributing to disease conditions (reviewed e.g. by Altshuler et al. (2008)). With even denser arrays with data from large scale human resequencing studies, and sample sizes nearing 100,000 individuals, we can expect this trend to continue in the next few years, and many more loci to be uncovered.

Association to quantitative traits in reference populations

The association approach can be successfully used outside the case-control paradigm as well. Instead of looking for differences in allele frequency between the healthy and diseased, one can measure a trait in a reference population of (nominally) healthy individuals, and look for an association between a trait value and the individual genotype. This approach has the most power when individuals vary considerably in the trait tested.

In model organisms, it is hard to obtain large numbers of individuals without complex population structure. A large scale project to generate many inbred lines from a random cross between 8 mouse strains is in progress (Threadgill et al., 2002). Recently, genotyping 191 inbred lines of *Arabidopsis thaliana*, a common grass, at 215,000 loci showed the potential of assaying all markers in

the genome (Atwell et al., 2010). They showed the marked differences between effects of population structure on trait mapping between plants and humans.

This hypothesis-free approach to mapping has informed us of many global human traits that have been found to be heritable, and associated with specific loci, such as height (Weedon et al., 2008), weight (Loos et al., 2008), telomere length (Codd et al., 2010; Glass et al., 2010), blood pressure (Newton-Cheh et al., 2009), etc.

Personal genomics companies are amassing genotype data of thousands of individuals, and also collecting additional information, both disease related and of general interest. While spending public money on studies on ear bud shape or the ability to roll one's tongue is not reasonable, it is possible to carry out such studies in these cohorts. Several companies have started weighing in on the scientific debate using summary statistics from their clients who have given appropriate consent, contributing to discussions on controversial results with their data (23andMe, 2010), or publishing their own work (Eriksson et al., 2010). Combining efforts of both private and public sectors to obtain genotype data for very large well-phenotyped cohorts will further increase the rate of discovering genetic associations to human traits.

1.1.3 Other approaches

In model organisms, other mapping approaches for genetic mapping are possible where variation in genotype is created in a more directed manner.

Characterising mutants

Genetic manipulation of individuals allows modifying a single locus either randomly, or in a controlled fashion, while keeping the rest of the genetic background constant. This gives an opportunity to observe the effect of the modified locus on a trait, and can be used to validate a locus mapped via linkage or association.

One such modification is a gene knockout, usually taken to mean removing the gene product from the cell. This can be achieved either via excising the gene from the chromosome using recombination techniques, or introducing a mutation that renders its non-functional, such as a premature stop codon or splice acceptor.

Libraries of gene knockouts have been created and characterised for every gene for yeast (Giaever et al., 2002), and are underway for mouse.

Another, more fine grained approach available in yeast, is allele mutagenesis, where exactly one locus is modified (Storici et al., 2001). As the rest of the genetic context is kept entirely constant, this allows assessing the effect of an allele in isolation. A more crude, but rapid approach is that of reciprocal hemizygosity (Steinmetz et al., 2002). A diploid hybrid of two haploid strains is created, followed by construction of two strains, each with one of the parental alleles deleted. In this case, however, the effect of the allele is manifested in the context of the rest of the hybrid genome.

Artificial selection

Instead of phenotyping individual mutants, which consumes resources and time, artificial selection can be applied to an entire mutant library to separate the mutants based on a trait. Again, individual mutants can then be assigned a phenotype.

In bacteria, this approach has been used in transposon mutagenesis screens (Langridge et al., 2009). Specifically, a transposon insertion is introduced randomly into the genome for many individual bacteria to produce a library that can then be tested for resistance to different conditions. The readout of the frequency of an insertion at all loci can be made by amplifying and quantifying the sequence from the transposon insertion sites. Similar approaches are also pursued in eukaryotic models such as yeast and mouse cell lines (Daniel Jeffares, Stephen Pettitt, personal communication).

Besides assigning a phenotypic effect to individual yeast gene knockouts the yeast knockout collection can also be used in artificial selection experiments. This is possible due to the barcode sequences introduced for each individual knockout strain, which allow detecting lack and abundance of individual knockouts in the full mutant pool in response to stress (Scherens and Goffeau, 2004).

Most recently, artificial selection of phenotypic extremes was applied to a very large pool of yeast segregants obtained from an F1 cross between two haploid lines. Combining the power of analysing very large numbers of segregants with

the cross design demonstrated the abundance of loci contributing to trait makeup even in "simple" eukaryotes, and emphasised the marked differences in genetic architecture of traits (Ehrenreich et al., 2010). We had been working on a similar approach in parallel, with results presented in Chapter 4.

1.2 Genetic structure of traits

Our knowledge of the genetic basis of heritable traits has come a long way in the last 100 years, and much remains to be done. There are many tools at our disposal for mapping. When trying to understand traits relevant for human health, what can we expect to find in general, and what are the characteristics of our findings so far?

1.2.1 Independent locus effects

Most of the existing work has focused on effects of individual variants in isolation. There are two basic questions about effects of individual loci - how many loci contribute to a trait, and how many traits does a locus contribute to.

Number of loci contributing to a trait and their effect sizes

The early studies in model organisms using crosses of inbred strains or recombinant inbred lines found several loci that determined most of the phenotypic variability (reviewed by Flint and Mackay (2009)), spurring the quest for finding common genetic variants with similarly large effect sizes in human population. However, as more individuals were analysed in such crosses, more trait loci were found, suggesting that there is a large set of mutations with smaller effect sizes. The controlled crosses in model organisms allow reducing the sources of variability associated with possible confounders, and thus let genetic variability make up more of the total phenotypic variability, making the trait loci easier to map.

Consistent with the idea of many trait loci with small effects, the recent human GWA studies have yielded hundreds of loci, almost all of which have small effect sizes with odds ratios less than 1.4 (Hindorff et al., 2009; Manolio et al., 2009). There is a chance that many rare variants with large effect sizes exist (Cirulli

and Goldstein, 2010); this hypothesis remains to be tested by using genotyping arrays including these rare variants, as well as resequencing studies.

Number of traits affected by a locus

Perhaps surprisingly, there is an emerging view that many trait loci are pleiotropic, affecting more than one trait. This has been observed in model organisms for years, where in crosses of yeast strains, a small set of loci determine much of the phenotypic variability, as well as other models, such as mouse (Brem et al., 2002; Chen et al., 2008). Similarly, some results from human GWA studies have identified unexpected links between seemingly disconnected disorders (Barrett et al., 2008).

The existence of pleiotropic loci is consistent with the notion of hubs in gene networks - genes that are central in pathways and whose variation has large downstream effects (Babu et al., 2004; Luscombe et al., 2004). Such loci induce correlation between traits, and thus motivate modelling them jointly to capture this effect.

1.2.2 Context dependent locus effects

The functional impact of a genetic variant is determined by its cellular context. The state of the cell - abundance, localisation, and configuration of molecules - is a product of RNA and protein polypeptides produced from the DNA, as well as temperature and concentrations of other molecules that can be influenced by external factors. Thus, the effect of the variant can depend on either the sequence of the RNAs and proteins, or some other state not directly determined by the genome. Context dependent effects are the focus of Chapter 3 of this Thesis.

Epistatic interactions

Gene-gene, or epistatic interactions are non-independent contributions of two loci to a trait. Usually, this is taken to mean a deviation from a standard statistical additive/multiplicative model (statistical epistasis), but can also mean masking or enhancing a genotype effect in general (functional epistasis). Notably, there

is not necessarily a physical interaction between the two gene products (Phillips, 2008).

Most reports on interaction effects come from crosses or manipulations of homozygous lines in model organisms. The reasons for a general lack of strong evidence for interactions in human traits is the extensive linkage in pedigree studies, where the effect of a single locus cannot be isolated, and the huge multiple testing problem in association studies, where billions of statistical tests need to be performed in order to assess the significance of all pairwise interactions between common polymorphisms. However, some lone examples of epistatic interactions in humans can be found (Butt et al., 2003; Tiret et al., 1994). Most association studies do not report any interaction effects, or only consider them between the mapped trait loci (Cordell, 2009).

A convincing demonstration of interactions between variants in four yeast transcription factors highlighted the potential for epistatic interactions to explain phenotypic variability (Gerke et al., 2009). In *C. elegans*, knockdowns of a small number of "hub" genes were shown to enhance the phenotypic effect of other knockdowns (Lehner et al., 2006). Evidence of local adaptation and interactions between nearby loci are also evident in the fruit flies (Mackay, 2004). Intriguingly, interaction effects have been shown to be pervasive in a mouse, where single chromosomes were replaced between strains (Shao et al., 2008).

GxE interactions

The gene-environment (GxE) interactions can be thought of as an environment-specific genetic effect. In humans, they are usually found by observing a prevalence of a trait in a specific environment, and then conducting an association or linkage study that conditions on it (Hunter, 2005). This has worked for several traits, but the success, as measured by the number of identified interactions, has not been on the scale of genome-wide association studies. Still, several highlights are worth noting. For example, a CCR5 (cell surface receptor) null mutant in humans interacts with HIV exposure, as HIV requires the receptor in order to enter the cell (Smith et al., 1997). More commonly, people with fair skin (a

heritable trait) are more prone to developing skin cancer in response to extensive sunlight (Rees, 2004). Recently, genetic variants have been associated with correct warfarin dosage (Takeuchi et al., 2009).

Gene-environment interactions are easiest studied in model organisms, where the genotype can be held constant, and then exposed to a variety of environments. Any gene mapped by screening a library or strain collection in different environments can be considered as part of such an interaction.

1.2.3 Missing heritability in human traits

Comparison of correlation of traits in monozygotic and dizygotic twins has shown that almost all medically relevant human traits, from physiological, such as height, weight, and heart rate to psychological, such as anxiety, depression, and boredom susceptibility, are heritable (Boomsma et al., 2002; Visscher et al., 2008). However, millions of pounds spent on genetic mapping studies have made us appreciate that independent effects of common alleles do not explain a substantial part of heritable variability in humans. Indeed, as most of the variants identified using GWAS have modest effect sizes, they explain only a small part of the heritable variation, although some recent results claim improvements on this (Yang et al., 2010). Leaving aside possible effects of epigenetics (Flintoft, 2010) as well as problems with accurate heritability estimation (Visscher et al., 2008) for now, we are still left with a gap in our knowledge. We know the information for passing on traits we care about is there - but where, and how can we find it? The answers lie in more accurate models and better assays; this Thesis seeks both.

1.3 Genetics of cellular traits

Most common human traits have a complex basis. Human clinical conditions are a constellation of symptoms, each based on deviations in tissue traits of individual organs. It is a little optimistic to assume that the genotype of a variant, whose effect is dependent on genetic background as well as environment, can carry substantial information on a global label of the organism, sweeping all the

underlying complexity into a binary disease state. Instead, it is much more feasible that we are able to map traits that are closer to DNA, such as molecular or cellular quantitative traits. In a fairly homogenous tissue, the cellular characteristics can be extrapolated to the entire tissue, and the tissue phenotype is already the appropriate level of abstraction for a disease symptom. Thus, I believe that mapping cellular and tissue traits is the correct way to proceed to map physiological human traits in general.

The effect of a single locus genotype on a global trait has to be mediated by cellular, tissue, and organ phenotypes. There is a very limited number of ways that the genotype of a variant can have any impact at all. A cell is a collection of molecules undergoing reactions; a change in the amount or properties of one of these molecules can have an effect on the kinetic parameters and equilibrium of some of the reactions. A variant could nudge the rate of a reaction a little, corresponding to a small effect size, or shift the balance of the reaction. If most of the associations found in human GWA studies are due to small cellular changes that have correspondingly small effects on tissues and organs, mapping cellular traits will offer no advantage compared to mapping more global traits. However, if there are substantial changes in cellular properties, we can hope to map these large effects by measuring the cellular traits that are affected by the balance of the particular reaction. Note that these effects can be dampened out by other fluctuations or compensatory mechanisms at a higher level to produce a weak effect on phenotype (Raj et al., 2010).

Perhaps most importantly, DNA sequence variation can result in protein sequence variation. This in turn can have an effect on secondary and tertiary structure of the protein (Ng and Henikoff, 2006), binding affinities to DNA (Zheng et al., 2010) or other proteins (Moreira et al., 2007), signalling and sorting properties etc. - in short, the activity of the protein. Thus, the activity of proteins (or other functional genetic elements) in the cell in its natural environment is the trait we would like to assay.

DNA sequence variation can also affect the affinity of proteins or nucleic acids binding to the nearby sequence. This can have an effect on chromatin state (McDaniell et al., 2010), propensity for epigenetic modifications (Gibbs et al., 2010), and amount of transcripts produced from a proximal or distal locus (Stranger

et al., 2007). Thus, binding affinities of proteins to DNA sequence, and their functional consequences, such as abundance of mRNA and protein molecules produced from it, are also phenotypes we would like to measure.

1.3.1 Assaying cellular traits

Cells are small, and making measurements from them is hard. Ideally, we would like to assay the quantity we are interested in, in a single cell, in a physiologically relevant environment, in high throughput, over time, controlling for all possible confounders, quickly, and at no cost. Reality, however, does not allow all these to be satisfied. For trait mapping we do need to phenotype many individuals, so the assay must be relatively high-throughput and low-cost; the rest of the desiderata can be sacrificed to a greater or lesser extent.

Single cells and cell populations

Ideally, we would like to measure cellular traits from single cells, but this imposes several hurdles that the assay needs to clear. Firstly, as most cells are small, the number of molecules in a cell is limited, thus the assay must be very sensitive. For example, measuring gene expression levels, or sequencing DNA requires on the order of a few μg of DNA or total RNA, whereas only 10 pg of RNA is present in the cell. Some recent developments address this, allowing quantities to be measured from even individual cells (Kurimoto et al., 2007). Microfluidics and microwell approaches promise to deliver assays on a chip that really do use individual cells, but do not, however, yet assay all mRNAs (Marcus et al., 2006). Secondly, it must be possible to isolate individual cells. This is not possible for many cell types, thus researchers often resort to *in situ* experiments, where individual cells are highlighted by clever genetic engineering techniques (Yuste, 2005). If possible, simply visualising the desired trait in the cell would be best. However, this requires good microscopes, and usually human inference for image analysis - but advances are made both on the level of microscopy (Yuste, 2005) as well as phenotyping by image analysis (Iyer-Pascuzzi et al., 2010).

For assays that require larger amounts of material, cell populations have to be used, which introduces confounding factors. Firstly, the readout is then a popu-

lation average, and it depends on the measured trait, how well that characterises the entire distribution of the trait. For example, gene expression in individual cells has been shown to be bursty (Paulsson, 2004), thus averaging over the population can give a medium gene expression level, whereas in each individual cell, the gene is either more highly expressed, or not present at all. Secondly, the population might not be homogenous. If human tissue is used, for example, mixtures of cell types are almost always present, and deconvoluting the signal post hoc becomes difficult, though not impossible (Clarke et al., 2010). Finally, even in a population comprised of just one cell type, activity profiles of molecules greatly vary with the cell cycle (Alberts et al., 2007). Thus, when analysing an unsynchronised cell population, one is dealing with average measurements across the cell cycle.

Primary tissues, proxy tissues, and cell lines

For human studies, it is not straightforward to obtain required tissues. Most organs are not readily accessible, and should not be physically damaged for healthy humans to get a sample. However, some easily replenishable tissues whose biopsy or collection does not have side effects, such as peripheral blood, but also hair, skin, fat, and in some cases, even muscle, are sampled for studies in healthy subjects (Nica et al., 2011). Some tissues are naturally left over during procedures in hospitals; fat and skin samples from plastic surgeries are a prime example. Tissues can also be collected from diseased people pre and post mortem, however, these tissues may no longer reflect standard homeostatic conditions. Some initiatives are proposing using road accident victims who are also organ donors as sources of research tissue. Of course, in most model organisms, these issues are not as relevant, since (subject to appropriate ethical controls and plenty of funding) sufficient amounts of tissue can be obtained from sacrificed individuals.

If the desired tissue cannot be sampled in the required quantity, a proxy tissue, or a cell line can be used. A proxy tissue is a more available tissue that is still informative about the desired trait. Peripheral blood is often used, as it is most easily available (Scherzer et al., 2007). Cell lines, an immortalised clonal cell population, are an alternative if bulk quantities are needed. While

there are doubts about whether they are useful for inference about naturally occurring traits, they have been successfully used in many studies. For example, Epstein-Barr virus transformed lymphoblastoid cell lines from the genetically very well characterised HapMap populations (The International HapMap Consortium, 2005), have been used in genetics of gene expression studies (Stranger et al., 2007), as well as many others (McDaniell et al., 2010).

Available high-throughput assays

Given the desires and constraints, which high-throughput assays can we use in studies into genetics of cellular traits?

Gene expression microarrays have been used for profiling mRNA levels for over a decade with great success (Montgomery and Dermitzakis, 2009), with arrays also being developed for other types of RNAs, such as microRNAs (Krichevsky et al., 2003). They are relatively cheap, well established, give a readout of thousands of traits, and their data will be used extensively throughout this Thesis. Recently, RNAseq has been used as a competing and complementary technique (Montgomery et al., 2010; Pickrell et al., 2010).

Mass spectrometry based approaches are nearly feasible for large sample sizes (Foss et al., 2007; Garge et al., 2010) to measure protein levels in a cell population. However, early studies have not shown a very dense coverage of the proteome, and posttranslational modifications further obfuscate the signal. Accurately measuring protein levels and their activities in the cell remains a challenging task (Choudhary and Mann, 2010).

Most recently, availability of relatively cheap, very high throughput sequencing with appropriate pulldown techniques has spurred studies into protein-DNA and protein-RNA binding events. X-seq (Medip-Seq, Chip-Seq, MethylC-Seq, DNAse-seq, CLIP-seq, Hit-Seq) and other approaches (3C/4C/5C, IClip, etc.) have allowed locus-specific quantification of various types of binding events in the cell population, reviewed e.g. in Hawkins et al. (2010).

I will not cover the many imaging approaches, but note that though they are powerful in principle, they require sophisticated machine vision algorithms (or many hours of good eyesight) for phenotyping, and have begun to yield exciting results (Fuchs et al., 2010; Hutchins et al., 2010; Neumann et al., 2010).

1.3.2 Genetics of gene expression

Many of the variants that have been identified in genome-wide association studies do not change coding sequences (Mackay et al., 2009), suggesting that many functional variants regulate gene expression, and so the genetics of gene expression is central to understanding of the genetic basis of complex traits. It has become possible to assay transcript levels on a large scale and treat them as quantitative traits, enabling research into the genetic makeup of these basic cellular phenotypes (Montgomery and Dermitzakis, 2009).

Gene expression has a genetic basis

Work began in a simple model organism, baker's yeast, where segregating strains from an F1 cross were first used to address genetics of gene expression (Brem et al., 2002; Yvert et al., 2003). A year later, explorations in mice, maize, and humans followed (Schadt et al., 2003). These studies showed that variation in gene expression levels is heritable, and found numerous statistical associations between both loci proximal to the expression probe (*cis*), as well as distal (*trans*). Usually, a locus within up to a 1 megabase window around the probe is considered to be in *cis*, and posited to have a direct, sequence-specific effect (Stranger et al., 2005). All other loci are considered to be in *trans*, and their mechanism of action to take place via a gene whose protein product is affected by variation at the locus. However, enhancers, insulators, and other regulators proximal to the gene can act at distances larger than one megabase, so the choice is arbitrary. The *trans* eQTLs were often clustered together into "eQTL hotspots"; however, contrary to expectation, they were not necessarily linked to variation in transcription factors (Yvert et al., 2003). Notably, some of these hotspots in yeast were found to be associated to many traits (Perlstein et al., 2007), demonstrating pleiotropic effects beyond gene expression.

Results from early human association studies

Promptly after the first linkage-based genetics of gene expression studies in model organisms, more in-depth human studies using association mapping followed. Most of them used EBV-transformed lymphoblastoid cell lines (LCLs). Following

the demonstration that variation in expression of up to 31% of the genes is heritable in human families (Monks et al., 2004), association mapping studies were conducted in the HapMap CEPH population. These started with a limited number of genes (Morley et al., 2004), then, using the availability of new arrays and HapMap genetic variation data, were carried out genome-wide (Stranger et al., 2005). Due to limited sample size, these were only powered to find strong effects, yet identified associations for up to 10% of human genes. Almost all of these associations were in *cis* with only a handful of *trans* findings, as there is a huge multiple testing burden of associating tens of thousands of transcript levels to genotypes of millions of loci.

Tissue- and population specificity of associations

Upon establishing that many eQTLs exist in both humans and model organisms, attention turned to whether they are universal across tissues and populations. Associations specific to tissue or population are examples of gene-environment or possibly epistatic interactions, as their effect is only evident conditional on the physiology of the underlying tissue, or the ethnic background that comprises both genetic and environmental context.

An assortment of tissues is more readily available in model organisms, so tissue specificity of eQTLs was first addressed there. An early study in mouse recombinant inbred lines reported almost no sharing of eQTLs (Cotsapas et al., 2006), while later studies have found more sharing (van Nas et al., 2010).

In humans, there is a relatively small number of tissues that can be straightforwardly assayed. Nevertheless, recent and current attempts have given an indication of extensive tissue-specificity of associations in humans (Dimas et al., 2009; Nica et al., 2011). Whether the lack of overlap is a *bona fide* effect, a consequence of statistical power limits, or a simple statement of tissue-specific gene expression is not yet clear. Dimas et al. (2009) compared associations in primary T cells, primary fibroblasts, and LCLs, finding that 70-80% are cell type specific. Furthermore, the cell type specific eQTLs were more distal from the transcription start site, but still in *cis*, suggesting that they might affect tissue-specific enhancer

activity. However, this signal is indistinguishable from a low level of false positives, and demonstration of either diminuation of the signal further downstream, or functional studies are needed to back up this theory. Preliminary results for eQTL finding from a different set of tissues and larger set of samples are briefly presented in Chapter 2.

Population specificity has been more straightforward to address due to the availability of LCLs from the HapMap populations. Indeed, studies in 4 (Stranger et al., 2007) and 9 (in preparation) populations have shown both increase in power to detect weaker effects due to larger sample size, as well as eQTLs specific to populations.

Remaining challenges

Many genetic associations to gene expression levels have been identified, and no doubt many more will follow. There is still no consensus on the extent of genetic regulation of gene expression - how many genes are regulated by an eQTL? Are these associations different in tissues and human populations? Which are the characteristics of the alleles that confer the change? How many *trans* associations are there in human populations? Do they correspond to transcription factors or other regulators? How much of the variability in phenotypes determined by gene expression can we attribute to mRNA level variability? Some of these issues will be addressed in this Thesis.

1.3.3 Genetics of other high-dimensional cellular traits

While this Thesis is mainly concerned with high-dimensional gene expression data, genetics of other high-dimensional cellular phenotypes have also been explored.

Nucleotide sequence traits

The arrangement, modifications, and binding events of nucleotide sequence can be assayed via selecting for the specific binding or modification events. The abundance of such events can then in some cases be associated with the genotype. This has been successfully done for DNA methylation (Gibbs et al., 2010), chromatin

structure (McDaniell et al., 2010), and protein binding traits (Hawkins et al., 2010).

Protein traits

Proteins carry out most of the cellular functions, thus protein traits, especially ones describing their activity, are some of the potentially most useful ones. Protein levels can be measured individually if they are engineered to carry a tag that provides a readout, in parallel using protein binding microarrays, or globally using mass spectometry (Vogel et al., 2010). The complexities of protein level quantification arise from many potential posttranslational modifications that can alter the mass-charge ratio of individual peptides for tandem MS experiments (Choudhary and Mann, 2010), or affinities for molecules used for pulldown.

1.4 Quantitative genetic models

Once the data for trait mapping are collected, the goal is to extract information about the underlying biological processes. This is not straightforward. The state of the cell is an extremely high-dimensional, time-varying function; an assay projects all this complexity into a low-dimensional visualisable readout. Inverting that readout to infer something about the state of the cell is a challenge.

Explicitly or implicitly, analysis of high throughput data always consists of formalising a quantitative model, a view of how we believe the world works, and allowing the gathered data to tell us either about specifics of the model, or whether the model is appropriate for describing the state of affairs at all. The most natural, and often most fruitful approach is that of probabilistic modelling. In fact, many claim it is the only logically sound way of making scientific inferences (Jaynes, 2003). I will not digress on this debate here, simply note that most data analysis problems can be cast as instances or useful approximations of Bayesian inference in probabilistic models.

Using a probabilistic model for data analysis consists of two steps. First, one must specify the model - a coherent set of functions that give a probability for each possible outcome for all states of the variables of interest. Probabilistic

modelling treats both observed and unobserved quantities as random variables with corresponding prior distributions. Then, upon observing some subset D of the variables to be equal to d, the posterior distribution of all the remaining, unobserved variables X can be inferred using Bayes rule,

$$P(X = x \mid D = d) = \frac{P(X = x, D = d)}{P(D = d)} = \frac{P(D = d \mid X = x)P(X = x)}{\int_{\mathcal{X}} P(D = d \mid X = x')P(X = x')dx'} \tag{1.1}$$

where we only need to be able to evaluate the prior $P(X)$ and data likelihood $P(D \mid X)$.

Using the posterior distribution is intuitively appealing. It combines all the available information in a principled manner into a single quantitative model of our knowledge. The probability of each possible state of the world can be obtained after observing any amount of the data, as long as we are able to perform the calculations. The posterior distribution also provides a measure of uncertainty that can be used in decision making or further analysis. In practice, many variables are often not endowed with a prior distribution, and are instead treated as parameters. In this case, the inference can provide only a point estimate of the parameter value via optimising the probability of the observed data, and the variability in the estimate has to be assessed by other means, such as bootstrapping.

Finding the best way to carry out these two steps of probabilistic modeling has kept scientists busy to provide more accurate quantitative descriptions of processes, and methods to make useful, tractable inferences. I will now describe some of the most often used models for trait mapping for low- and high-dimensional data, some of which will be extended or used in the later chapters, and then discuss ways of performing inference in them in the next section.

1.4.1 Single trait

Single traits t taking value $y_{i,t}$ in individual i have been modelled using random variables since the 1930's, when Ronald Fisher established the field of quantitative genetics (Fisher, 1939), many ideas of which are still used today.

Notation

In this section, I will use standard lower case letters (e.g. $y_{i,t}$) as random variables, boldface letters (e.g. $\mathbf{y}_t$) as vectors of random variables, capital letters (e.g. M) as constants denoting dimensions, and boldface capital letters (e.g. $\mathbf{\Sigma}_t$) as matrices of random variables. Greek letters usually denote parameters of specific distributions.

Single individual

Most often, quantitative traits are assumed to be normally distributed with mean μ_t and variance σ_t^2

$$y_{i,t} \sim \mathcal{N}(\mu_t, \sigma_t^2) \tag{1.2}$$

so that

$$P(y_{i,t} = y \,|\, \mu_t, \sigma_t^2) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(y-\mu_t)^2}{2\sigma_t^2}\right).$$

In this framework, there are two standard ways to introduce the effects of the genetic background. First, one can model the effect of genotype s of locus n in individual i as a *fixed effect* with weight $w_{n,t}$ that gives a fixed contribution to trait t. The standard way to encode the genotype of a locus with two alleles in a diploid individual is to assume independent effects of both haplotypes, and either encode the alleles as $(-0.5, 0.5)$ or $(0, 1)$ to give three possible genotypes $(-1, 0, 1)$ or $(0, 1, 2)$. Dominance and recessive models can be introduced via an additional weight on the heterozygous term, but these are not used in this work. In case of more alleles, a count vector must be introduced over them; the treatment remains unchanged in the context of these linear models, and simply scales w or offsets μ. In any case, the model becomes

$$y_{i,t} \sim \mathcal{N}(\mu_t + s_{i,n} w_{n,t}, \sigma_t^2). \tag{1.3}$$

An alternative is to say that the genetic bacground has a *random effect* on the trait, with magnitude $\sigma_{t,G}^2$, and the trait value y is a sum of contributions from the random genetic variance $\sigma_{t,G}^2$ and the private variability σ_t^2.

$$y_{i,t} \sim \mathcal{N}(\mu_t, \sigma_t^2) + \mathcal{N}(0, \sigma_{t,G}^2) = \mathcal{N}(\mu_t, \sigma_t^2 + \sigma_{t,G}^2). \tag{1.4}$$

A population of independent individuals

For linkage and association mapping, phenotype data is gathered from many individuals $i = 1 \dots M$, producing a vector $\mathbf{y}_t = (y_{1,t}, \dots, y_{M,t})^T$. The model is then generalised to an $M-$dimensional Gaussian

$$\mathbf{y}_t \sim \mathcal{N}(\mu_t \mathbf{1}_M, \mathbf{\Sigma}_t) = \frac{|\mathbf{\Sigma}_t|^{-\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}} \exp\left(-\frac{(\mathbf{y}_t - \mu_t \mathbf{1}_M)^T \mathbf{\Sigma}_t^{-1} (\mathbf{y}_t - \mu_t \mathbf{1}_M)}{2}\right) \tag{1.5}$$

where $\mathbf{1}_M = \underbrace{(1, 1, \dots, 1)^T}_{M}$, and $\mathbf{\Sigma}_t = \text{diag}(\sigma_t^2, ..., \sigma_t^2)$. Here, as it is assumed that all individuals are independent, there is no information shared between them, and the covariance matrix is diagonal. Reintroducing the fixed effect corresponds to adding a weighted contribution of the genotype vector, $w_{n,t}\mathbf{s}_n$ to the mean.

Dealing with dependence between individuals

Even if we assume the individuals to be independent given the genotypes, sharing alleles at a trait locus induces a correlation between the individual phenotypes. Random genetic effects shared between individuals add an independent variance component $\sigma_{t,G}$ to corresponding off-diagonal elements of $\mathbf{\Sigma}_t$. Both of these operations induce covariance between individuals. In general, the covariance matrix could have any form. However, the problem with these, more accurate, models is the complexity of inference. Analytical solutions do not exist to obtain single best point estimates of the parameters, and iterative approaches have to be used.

In some recent applications, mixed linear models, combining both fixed genotype effects, and random genetic effects to capture known population or family structure, have been successfully used (Atwell et al., 2010; Kang et al., 2010). In human GWAS, a PCA-based correction of the genotype vector has been used to correct for stratification in the population structure (Patterson et al., 2006).

Covariates

While the fixed genetic effect is our primary interest for trait mapping, there are other measured variables that influence the trait. These confounders must be included in the model. In the current linear modeling framework, it is straightforward to do so via introducing a fixed effect of each of C observed factors, $\mathbf{f}_1, \ldots, \mathbf{f}_C$, with corresponding weights $v_{1,t}, ..., v_{C,t}$, giving

$$\mathbf{y}_t \sim \mathcal{N}(\mu_t \mathbf{1}_M + \sum_{c=1}^{C} \mathbf{f}_c v_{c,t}, \mathbf{\Sigma}_t) \tag{1.6}$$

This fixed effects model forms the basis for the vast majority of linkage and association studies. It is worth noting the many implicit assumptions present in the parametric form, the additive and linear influence of covariates etc., that can and do introduce artifacts or reduce power for mapping when they are not correct.

Nonlinear and interaction effects

Both gene-gene and gene-environment interaction effects can be included in these linear models, by introducing additional additive terms to the mean that combine the multiplicative effects of the epistatic, or gene- environment interactions. For example, in case of a statistical interaction between known factor c' and the genotype at locus n, the model becomes

$$\mathbf{y}_t \sim \mathcal{N}(\mu_t \mathbf{1}_M + w_{n,t}\mathbf{s}_n + \sum_{c=1}^{C} \mathbf{f}_c v_{c,t} + w_{c',n,t}\mathbf{s}_n \mathbf{f}_{c'}, \mathbf{\Sigma}_t) \tag{1.7}$$

The final term quantifies the departure from the independent linear effect of the different sources of variability.

Non-normal traits

Traits can also be binary, as is the case for case control studies, or ordinal, such as from count data, or arising from waiting times, etc. I will not delve into the details of modelling these special cases, as they are not used in this Thesis. The rich

generalised linear modelling framework (Nelder and Wedderburn, 1972) allows standard inference of many types of data via the use of suitable link functions and transformations. In short, the input space is transformed into some vector space by calculating a sufficient statistic, and the parameters are transformed into the same space to calculate the dot product of the two, which is then used to score specific combinations of data and parameter values.

1.4.2 High-dimensional traits

The standard statistical models for single traits are not optimal for use in high-dimensional traits, as these traits are usually not independent. There is additional information present in their covariance structure. A correct model would capture those dependencies, and allow the effects of genotype to stand out as the remaining variance to be explained. If the dependencies are not observed, and thus cannot be included in the model as covariates, they have to be included in the model and estimated from the data. In a similar vein to single trait modelling, the covariance between multiple traits can be modeled either by a linear ("fixed") effect to the mean by use of hidden variable models, or a random effect influencing the covariance matrix. Alternatively, a qualitative description of the trait correlations is possible.

Hidden variable models

Linear hidden variable models for traits $t = 1 \dots G$ observed in individuals $i = 1 \dots N$ hypothesise a smaller set of $K << N$ hidden factors $\mathbf{X} = (\mathbf{x}_1, \dots \mathbf{x}_K)$ that capture much of the variability in the trait:

$$\mathbf{y}_t \sim \mathcal{N}(\mu_t \mathbf{1}_M + \sum_{k=1}^{K} \mathbf{x}_k w_{k,t}, \mathbf{\Sigma}_t). \tag{1.8}$$

These factors, in contrast to the covariates, are unobserved, and have to be estimated from the data. Chapter 2 explores some of these models in greater detail.

Random effect model

An alternative way to include the trait covariance in the model is to introduce an additional term that needs to be estimated:

$$\mathbf{y}_t \sim \mathcal{N}(\mu_t \mathbf{1}_M + \mathbf{Z}_t \mathbf{b}, \mathbf{\Sigma}_t), \tag{1.9}$$

The design matrix $\mathbf{Z}_t$ indicates which traits are influenced by which of the random effects $\mathbf{b}$, and $b_i \sim \mathcal{N}(0, \sigma^2_{G,t})$.

Network models

A different approach to treating covariance between traits is looking at individual correlations. There are many papers that establish a trait graph by introducing a node for each trait, and an edge between nodes if some measure of correlation or mutual information is satisfied, followed by analysing statistical properties of the graph, its cliques, or an arbitrary subset of nodes (e.g. Zhang et al. (2010); Zhu et al. (2008)). Such models have produced many "hairball" cover images for journals, and (sometimes) accessible visualisation of high dimensional data (Freeman et al., 2007). However, without an explicit generative model, they are hard to interpret, and not used for this Thesis.

1.5 Inference for trait models

Once we have mathematically described how we believe the world works by establishing a quantitative model, and observed some data, we are ready to perform inference of the model unknowns.

1.5.1 Frequentist inference

In frequentist inference, dominant most of the 20th century, the standard practice is to construct estimators for parameters of interest, and test for their significance with respect to the expectation under a background model.

Maximum likelihood estimation

The standard estimator used is based on maximising the data likelihood under the model. Treating the unobserved variables X as parameters, and optimising the likelihood L or log-likelihood l of the observed data d with respect to them gives the maximum likelihood estimates $\hat{X}$:

$$\hat{X} = \text{argmax}_{x \in \mathcal{X}} \text{L}(x; d) = \text{argmax}_{x \in \mathcal{X}} P(D = d | X = x) \tag{1.10}$$

This approach provides a point estimate of X, which has some nice properties such as consistency. The associated uncertainty (termed observed information (Davison, 2003)) can be obtained by considering the second derivative of the likelihood (or log-likelihood) function. If the likelihood is flat, the uncertainty is high; if the likelihood is highly peaked, the estimate is precisely determined. However, maximum likelihood inference is prone to undesired failure modes (some examples are given in MacKay (2003)). It is still used in a variety of settings as it is usually quick to calculate compared to alternatives, and behaves well in many cases.

Significance testing

Claims about the interesting state of the world can be made by assessing how surprised we are to see the data if the world was in fact boring. This entails calculating a test statistic T (which can be any function of the data), and assessing the probability of observing a value at least as extreme from a null distribution of test statistics. The most classical approach is to consider a nested model, where the significance of an additional parameter is assessed by the change of the log-likelihood function. Notably,

$$T = 2 \left(l(x_1, x_2, ..., x_N; d) - l(x_1 = 0, x_2, ..., x_N; d) \right) \tag{1.11}$$

is approximately χ^2_1 distributed. Thus significance of the statistic can be quickly assessed by determining how frequently it is observed from the parametric form of this distribution.

It is worth mentioning that the classical significance testing approach does not explicitly include an alternative model. Conceptually, this is problematic for me. I do not care about the surprise level of one model fit; instead, I want to know what the best model describing my data is. I will not cover model selection here, noting that the Bayesian approach of specifying a prior over models and inferring the posterior probability of each is an appealing strategy.

Non-parametric approaches

Testing for significance of a genetic association or interaction using a standard linear model is vulnerable to violation of any of the numerous model assumptions. For example, when outliers are present, they are highly penalised by the normal distribution of errors, and can give a disproportionately high test statistic. The problem is the non-uniform distribution of p-values under the null hypothesis.

One alternative is to use a non-parametric model, that does not depend on specific parametrised distributions. Examples of this are Spearman Rank Correlation and Mann-Whitney U tests that use ranks of the data in place of actual values.

An alternative approach does not rely on an analytical null distribution of the test statistic. Instead, the null is constructed empirically using permutations. In the context of association studies, one can assume the individuals are exchangeable, and so permute the trait values between individuals, and calculate the test statistics on the permuted values. These statistics then serve as the null distribution against which the unpermuted test statistic is compared. I will use both this and the maximum likelihood approaches in this work.

1.5.2 Bayesian inference

An alternative to significance testing is to infer the posterior distribution of the unobserved variables. In simple cases, we can do this exactly in an analytic way. If this is not feasible, we are left with a choice between the correct posterior to an approximated model, or an approximate posterior to a more realistic model.

Exact inference

In simple cases, it is possible simply to calculate the posterior $P(X \mid D)$. However, the evidence $P(D)$ that appears in the denominator of the Bayes rule involves integrating over all possible parameter settings, which is prohibitive in more involved models. In some cases where the model is conveniently structured, the inference can be broken up into iterative steps (e.g. Baum-Welsch estimation of hidden Markov Model states, with biological examples in Durbin et al. (1999), or expectation maximisation in general). However, unless the parameter optimisation problem is convex, there is no guarantee of optimal parameter finding.

Approximate inference

If exact inference cannot be performed, some of the distributions have to be approximated. Frequently, conjugate prior distributions are chosen for computational convenience, and in general, approximations to any part of the model can be chosen arbitrarily. However, some approaches, such as variational (mean-field) approximations (Bishop, 2007), are more founded. In variational approximation, the KL-divergence between the approximation and the true posterior is minimised. The only other underlying assumption of the variational approach is a specific factorisation of the joint probability density. The specific forms of local marginals can either be fixed, or derived from integrating out all the other approximate marginals from the joint distribution; a task that is easier compared to the full problem due to the factorisation structure. Variational methods will be used in Chapter 2.

Other approaches

A non-approximate alternative to Bayesian inference is Markov Chain Monte Carlo (MCMC) methodology. The parameter estimates are iteratively sampled from a proposal distribution subject to balancing constraints, and form a Markov Chain whose asymptotic distribution is the true posterior (Davison, 2003). These computationally intensive methods are not used in this Thesis.

1.6 Major open questions

Given the history of trait mapping, all the questions answered, all the challenges remaining, all the technologies becoming available, and all the computational tools at our disposal, what are the areas ripe for advancing our knowledge?

I believe we are primed to make great advances in finding the genetic cause of all heritable traits. We can assay genotype at an unprecedented rate, and have amassed vast cohorts of human subjects and model organism strains. We can assay global as well as cellular phenotypes. All the data is or will be there. I want to use them to establish where are the trait loci, what are the functional implications of their variation, and how does this influence disease risk in humans.

1.7 Contributions of this Thesis

In this Thesis, I attack the problem of mapping cellular traits.

In the second chapter, I develop a statistical model based on Bayesian regression and factor analysis for association mapping with high-dimensional phenotypes. I show how accounting for global, non-genetic variance components in the phenotype data increases power to detect genetic associations. Application of the method on human gene expression variation data demonstrates that up to 30% of transcripts have a statistically significant association to a proximal locus genotype, three times more than were found with a standard model.

In the third chapter, I consider mapping genetics and interactions of inferred intermediate phenotypes. I apply a sparse factor analysis model to infer hidden factors, which are treated as intermediate cellular phenotypes that in turn affect gene expression in a yeast dataset. I find that the inferred phenotypes are associated with locus genotypes and environmental conditions, and can explain genetic associations to genes in trans. For the first time, interactions between genotype and intermediate phenotypes inferred from gene expression levels are considered and detected, which complements and extends established results.

Then, I take another angle at mapping cellular traits. I develop a novel approach to map trait loci rapidly and in narrow intervals using massively parallel sequencing. We created advanced intercross lines between two phenotypically

different wild isolates of bakers yeast with sequenced reference genomes. We then applied selective pressure on the intercross pool by growing it in a restrictive condition to enrich for individuals with alleles that confer a positive fitness effect. Sequencing DNA from the pool before and after selection pinpoints genes responsible for the increased fitness, or protective against reduced fitness under stress. This novel method provides a rapid and fine scale QTL mapping strategy improving resolution and power.

Finally, I conclude the Thesis by exploring mapping cellular traits in a series of short studies in different organisms.