Genetic mapping of cellular traits

Leopold Parts

Wellcome Trust Sanger Institute
Corpus Christi College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

September 2010

To people who like swimming at midnight, climbing trees, or hedgehogs. Or fractal snowflakes.

Acknowledgements

This will be long - I am very thankful.

There are many people and institutions that have made this document possible. Along the way, I have been given chances that have one after the other, led me to writing these words. I thank Prof. Urmas Varblane for my first job, Prof. Helge Loebler for my first year abroad, Prof. Jaak Vilo for bringing me to the field, Ewan Birney for allowing me to sniff science for the first time, and Manolis Kellis for letting me in on large exciting project. Over the last four years, Richard Durbin, my supervisor, has been a solid source of fair critique, thoughtful comments, and strong writing - I thank him for his time and support. I am grateful to Wellcome Trust for funding my studies.

I have worked with many good collaborators, and slept on the floor of the best ones. Most of the work in this thesis was fought through with Oliver Stegle, often into the late hours at the Cavendish with good coffee. Gianni Liti and Francisco Cubillos have allowed me to think about biology and the experiments - and actually see them happen. There are many others who have been a pleasure to work with.

I have been blessed through time with great company both for enthusiasm for science as well as good times. The people in rdgroup have been a source of advice, knowledge, and banter; other sports-playing, french-speaking, GO-mongering colleagues have made the campus a pleasant environment. Phd06 - Kiki, Mare, Matti, Sergi, Steve - have been great companions for going through the trenches of academia together.

Most of all, I am indebted to my family. Gratitudes in Estonian. Aitähh – emale, kes väikseid lapsi laborisse nuuskima lubas, isale,

kes millestki puudust ei lasknud tunda, vennale, kes hea töö eest jäätiseid jagas, õele, kes teismelisi tööl arvutisse lasi, õele, kes bioloogide toredust näitas, vanaemale, kes bussipeatuses väikelapse 200-ni lugemist kontrollis, ja kõigile teistele suurtele-vanadele, noortele-väikestele, sõpradele-sugulastele, kes ei lase meelest minna, kus kodu on.

Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text. This thesis does not exceed the length limit set by the Biology Degree Committee.

Leopold Parts
31 August 2010

Abstract

Many important traits are heritable, and have a strong genetic component. In simple cases, such as Mendelian diseases, the genetic cause can be found with linkage methods, and many trait genes have been mapped to date. More recently, association mapping studies have focused on complex traits that include prevalent human diseases, such as type 2 diabetes, hypertension, and others. Numerous genome-wide association studies have corroborated that no single gene explains all or even a large part of the heritable variability in such traits, and that individual effect sizes due to common variants are small. The effect of a single locus genotype on a global trait has to be mediated by cellular, tissue, and organ phenotypes. Thus, genetics of cellular traits is central to developing an understanding of the genetic basis of complex traits.

In this thesis, we address the problem of mapping cellular traits. First, we develop a statistical model based on Bayesian regression and factor analysis for association mapping with high-dimensional phenotypes. We show how accounting for global, non-genetic variance components in the phenotype data increases power to detect genetic associations. Applying the method on human gene expression variation data, we find that up to 30% of transcripts have a statistically significant association to a proximal locus genotype.

Second, we show how to infer intermediate phenotypes and use them for mapping genetic associations and interactions. We use a sparse factor analysis model to infer hidden factors, which we treat as intermediate cellular phenotypes that in turn affect gene expression in a yeast dataset. We find that the inferred phenotypes are associated with locus genotypes and environmental conditions, and can explain genetic associations to nearby genes. For the first time, we consider and find interactions between genotype and intermediate phenotypes inferred from gene expression levels, complementing and extending established results.

Third, we 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 baker's 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 protective alleles. Sequencing DNA from the pool before and after selection pinpoints genes responsible for the increased fitness. This novel method provides a rapid and fine scale QTL mapping strategy improving resolution and power.

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

Contents

Nomenclature

Acronyms

CEU	HapMap 2 ’European’ population - U.S. residents with Northern and Western European ancestry
CHB	HapMap 2 Chinese population - individuals from Beijing
EBV	Epstein-Barr virus
eQTL	Expression QTL
FDR	False discovery rate
FNR	False negative rate
FPR	False positive rate
fVBQTL	Fast VBQTL
GEO	Gene Expression Omnibus
GO	Gene Ontology
GWAS	Genome-wide association study
GxE interaction	Gene-environment interaction
iVBQTL	Iterative VBQTL

JPT	HapMap 2 Japanese population - individuals from the Tokyo area
KEGG	Kyoto Encyclopedia of Genes and Genomes
KL	Kullback-Leibler
LCL	Lymphoblastoid cell line
LOD	Log-odds
MCMC	Markov chain Monte Carlo
mRNA	Messenger RNA
MS	Mass spectrometry
NA	North American strain
PCA	Principal Components Analysis
PCAsig	PCA with significance testing
PEER	Probabilistic estimation of expression residuals
QTL	Quantitative Trait Locus
RIN	RNA integrity number
SNP	Single nucleotide polymorphism
SVA	Surrogate Variable Analysis
VBeQTL	eQTL on residuals of fVBQTL
VBQTL	Variational Bayesian QTL mapper
WA	West African strain
YRI	HapMap 2 Nigerian population - Yoruba people of Ibadan

List of Figures

List of Tables

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 population 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,

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

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

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 $i = 1 \ldots M$, producing a vector $\mathbf{y}_t = (y_{1,t}, \ldots, y_{M,t})^T$. 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 $\mathbf{1}_M = \underbrace{(1, 1, \ldots, 1)^T}_{M}$, 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 $\hat{X}$:

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

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,

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

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.

Chapter 2

Association mapping with high-dimensional traits

Collaboration note *This chapter contains work performed in collaboration with Dr. Oliver Stegle and Dr. John Winn for methods development, and Alexandra Nica for eQTL finding in the MuTHER dataset. Oliver first established the eQTL model used in this chapter (Stegle et al., 2008), we then expanded on this work jointly (Stegle et al., 2010). In particular, I reimplemented and extended the existing model to make it usable for large scale studies, applied it on various datasets, and analysed the results. This coauthored manuscript forms the backbone of the chapter. Alexandra performed the eQTL calling on the MuTHER dataset, I obtained the results presented here based on those calls. The combined results are presented in Nica et al. (2011)*

The basic principle behind association mapping with high-dimensional traits is same as for single traits. The additional complexities arise from covariance structure between the traits or individuals, which can confound the sought signal. In the following, we consider joint modelling of high-dimensional traits for mapping gene expression QTLs; the same methods can straightforwardly be extended to any high-dimensional trait.

2.1 Expression QTLs

DNA microarray technologies allow for quantification of expression levels of thousands of loci in the genome. These measurements enable exploring how a variable, such as clinical phenotype, tissue type, or genetic background, affects the transcriptional state of the sample. Recently, gene expression levels have been studied as quantitative genetic traits, investigating the effect of genotype as the primary variable. Studies have found and characterised large numbers of expression quantitative trait loci (eQTLs) in yeast (Brem et al., 2002) and other organisms (Schadt et al., 2005), exploring their complexity (Brem and Kruglyak, 2005), population genetics (Spielman et al., 2007; Stranger et al., 2007) and associations with disease (Chen et al., 2008; Emilsson et al., 2008).

An important issue in such studies is additional variation in expression data that is not due to the genetic state, as illustrated in Figure 2.1. Intracellular fluctuations, environmental conditions, and experimental procedures are factors that all can have a strong effect on the measured transcript levels (Brem and Kruglyak 2005, Leek and Storey 2007, Gibson 2008, Plagnol et al. 2008) and thereby obscure the association signal. When measured, correct estimation of the additional variation due to these *known factors* allows for a more sensitive analysis of the genetic effect. For example, in Emilsson et al. (2008), the authors reported finding additional human eQTLs when including the known factors of age, gender, and blood cell counts in the model. It is also standard procedure to correct for batch effects, such as image artefacts or sample preparation differences (Balding et al., 2003).

In practise it is not possible to measure or even be aware of all potential sources of variation, but nevertheless it is important to account for them. Unobserved, *hidden factors*, such as cell culture conditions (Pastinen et al., 2006) often have an influence on large numbers of genes. We and others have proposed methods to detect and correct for such effects (Leek and Storey 2007, Stegle et al. 2008, Kang et al. 2008). These studies demonstrated the importance of accounting for hidden factors, yielding a stronger statistical discrimination signal.

The challenge in modelling several confounding sources of variation (Figure 2.1) is to correctly estimate the contribution that is due to each one of them.

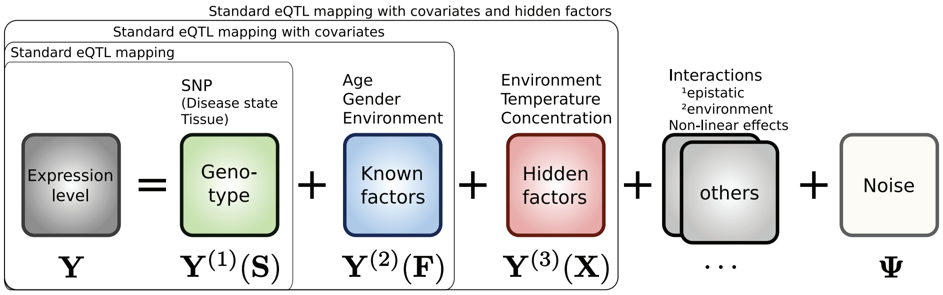


Figure 2.1: General additive model for sources of gene expression variability. The $G \times J$ matrix $\mathbf{Y}$ of measured gene expression levels of G genes from J individuals is modelled by additive contributions from components $\{\mathbf{Y}^{(m)}\}$ and observation noise $\mathbf{\Psi}$. Here, the components capture the signal due to primary effect of the genetic state $\mathbf{S}$, known factors $\mathbf{F}$ and hidden factors $\mathbf{X}$. Some examples of possible underlying sources of variation are given above the model boxes. The groupings represent some standard genetic association models commonly used.

There are open questions concerning how to ensure that only spurious signal is eliminated by methods that account for hidden factors (see for instance discussion in Kang et al. (2008)), and how to deal with situations when both known and hidden factors are present. The problem of identifying the correct causes of the signal is even harder in the presence of additional sources of variability. For example, when searching for epistatic or genotype-environment interactions, the primary effects of other known factors and hidden factors also need to be accounted for.

The key for correctly attributing expression variability is controlling the complexity of the statistical models for each source of variation. For example, the number of genotypes considered in an association scan can be enormous, and not all of them affect the expression level of every probe. Threshold values, obtained from likelihood ratio statistics or empirical p-value distributions, can be used to determine the significance of individual associations, thereby avoiding overfitting by controlling the model complexity (Lander and Botstein, 1989; Stranger et al., 2007). Similar measures are necessary for models of other sources of variability such as hidden factors.

In this chapter, we first present PEER (probabilistic estimation of expression residuals), a joint Bayesian framework for gene expression variability, and

VBQTL (variational Bayesian QTL mapper) is a specific configuration of this framework that accounts for the signal from genotype, known factors, and hidden factors (Chapter 2.2). While previous attempts have been specific to a narrow set of underlying sources, our approach is flexible and can be adapted to a particular study design. The probabilistic treatment allows uncertainty to be propagated between models, and yields a posterior distribution over model parameters. Complexity control is tackled at the level of individual models, where parameters are regularised in a Bayesian manner.

We then compare the performance of VBQTL with existing approaches for detecting expression QTLs (Chapter 2.3). A simulation experiment contrasts VBQTL with common approaches that use non-Bayesian techniques for distinguishing global hidden factor effects from genetic effects. This study highlights differences in the methodology to control model complexity with implications to eQTL detection power. The necessity and difficulty to account for variability that confounds the genetic signal is demonstrated. Results on datasets from a human outbred population and crosses of inbred yeast and mouse strains show that VBQTL identifies more significant associations than alternative methods.

Third, we apply VBQTL to perform a whole-genome eQTL scan on the HapMap phase 2, and MuTHER expression and genotype data, demonstrating the scalability of our framework to large numbers of samples and probes (Chapter 2.4). We find up to three times more *cis* eQTLs than a standard association mapping method, suggesting more extensive genetic control of gene expression by common variants than previously shown.

Finally, we explore applications of this model not centered on eQTL finding (Chapter 2.5). We consider interpreting the inferred hidden factors to understand the main gene expression variance components in different tissues and organisms. We also combine data from different tissues to assess the advantages of sharing information across multiple datasets for inference.

2.2 The PEER framework

Here, we present PEER, a general framework for modelling diverse sources of gene expression variability. The model underlying this framework assumes that gene expression levels are influenced by additive effects from independent sources, e.g. in the case of VBQTL these are contributions from genotype, known factors, and hidden factors (Figures 2.1, 2.2a). We cast the full model in a probabilistic setting, treating its parameters as random variables.

We perform Bayesian inference in the joint model, which is appealing for several reasons. First, it allows possible dependencies between the different sources of variation to be captured. The effects of the genotype, known and hidden factors are learned jointly, taking other parts of the model into account. Propagation of uncertainty leads to more accurate parameter estimates (Rattray et al., 2006), and avoids possible pathologies, for instance of maximum likelihood methods (MacKay, 2003). Second, Bayesian inference allows different models to be flexibly combined according to the needs of a particular study. Many existing approaches can be cast as special cases of this general framework, with some examples given in Figure 2.1. Finally, the Bayesian approach leads itself to efficient approximate inference schemes such as variational methods (Jordan et al., 1999), rendering the resulting algorithms applicable to large-scale and high-dimensional datasets. Also, variational learning allows an inference schedule to be specified by the user, leading to distinct algorithms with different computational complexity and properties (Chapter 2.2.2).

In the following, we present the mathematical model of VBQTL, and an outline of the inference procedure. We then describe alternative non-Bayesian models for expression QTL studies used in the experiments. An in-depth treatment of the framework including full details about the parameter estimation is provided in Appendix A.

2.2.1 Model

The observed gene expression matrix $\mathbf{Y} = \{y_{g,j}\}$ for genes $g \in \{1, \ldots, G\}$ and individuals $j \in \{1, \ldots, J\}$ is modelled by the sum of contributions $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, ..., \mathbf{Y}^{(M)}$ from M sources (in the VBQTL model, these include genotype, known and hidden factor effects), and Gaussian noise with precisions τ_g for each gene g

$$P(y_{g,j} \,|\, y_{g,j}^{(1)}, y_{g,j}^{(2)}, ..., y_{g,j}^{(M)}, \tau_g) = \mathcal{N}(y_{g,j} \,|\, y_{g,j}^{(1)} + y_{g,j}^{(2)} + ... + y_{g,j}^{(M)}, \frac{1}{\tau_g}), \tag{2.1}$$

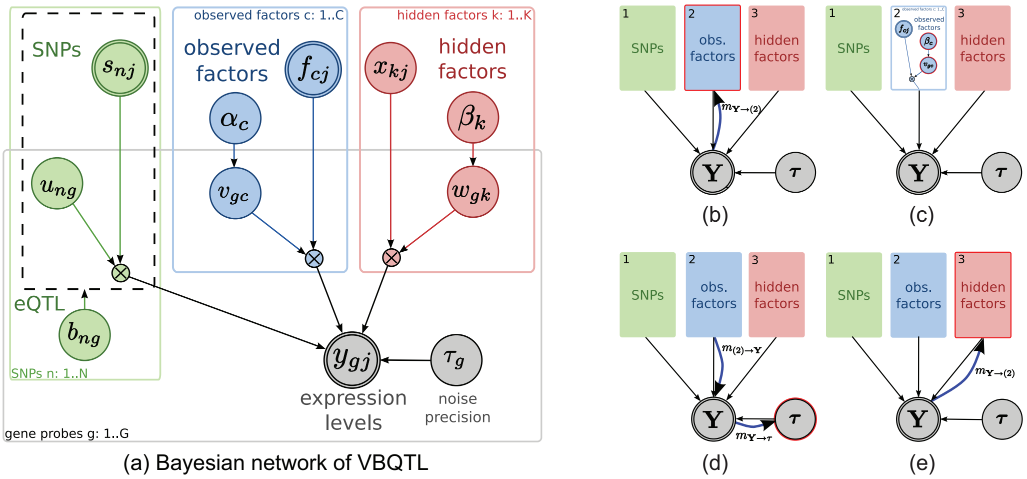


Figure 2.2: Bayesian network and outline of the inference schedule for VBQTL. **(a)** The Bayesian network for the model of gene expression variation used in VBQTL. The full model combines genetic (green), known factor (blue) and hidden factor (red) models to explain the observed gene expression levels $\mathbf{Y}$. The solid rectangles indicate that contained variables are duplicated for each gene probe (g), SNP (n) or factor (c, k) respectively. A similar rectangle for individuals (j) is omitted in this representation. The dashed rectangle indicates that the variable $b_{n,g}$ switches the contained part of the graph on or off representing the existence or lack of an association. Nodes with thick outlines ($s_{n,j}$, $f_{c,j}$ and $y_{g,j}$) are observed. **(b)-(e)** Update cycle of the known factors model introduced in section Inference. The red outline highlights the parts of the model that change in a step, and the thick blue arrows illustrate the flow of information. Details of these updates are discussed in the text.

with a gamma prior on the noise precisions $P(\tau_g) = \Gamma(\tau_g \,|\, a_\tau, b_\tau)$ (Figure 2.2a). The $\mathbf{Y}^{(i)}$ comprise the contribution of individual sources to the variability in the observed expression levels, and are themselves treated as random variables with different underlying models. In the VBQTL model used throughout the rest of the chapter, three different models for sources of variability are used:

1) Genotype effect model represents the probabilistic variant of the standard genetic association model, where some of the SNP genotypes have a linear effect on gene expression levels. The genetic component of the expression level $y^{(1)}_{g,j}$ of the gth gene probe in the jth individual is explained by linear effects of the genotypes of N SNPs $\mathbf{s}_j = \{s_{1,j}, \ldots, s_{N,j}\}$ (Figure 2.2a, green plate):

$$P(y^{(1)}_{g,j} \,|\, \mathbf{s}_j, \mathbf{b}_g, \mathbf{u}_g, \tau_g) = \mathcal{N}(y^{(1)}_{g,j} \,|\, \sum_{n=1}^{N} b_{n,g} \cdot (u_{n,g} s_{n,j}) \,, \frac{1}{\tau_g}) \tag{2.2}$$

$$P(b_{n,g}) = \text{Bernoulli}(b_{n,g} \,|\, p_{\text{ass}}) \tag{2.3}$$

$$P(u_{n,g}) = \mathcal{N}(u_{n,g} \,|\, 0, 1). \tag{2.4}$$

The weights $\mathbf{u}_g = \{u_{1,g}, \ldots, u_{N,g}\}$ control the magnitude of the effect of the SNP on the expression levels of genes g. The binary variables $\mathbf{b}_g = \{b_{1,g}, \ldots, b_{N,g}\}$ determine whether the SNP effect is significant ($b_{n,g} = \text{true}$) or not ($b_{n,g} = \text{false}$). The prior probability p_{ass} of an individual association controls the complexity of the model by influencing the a priori expected number of significant associations; this parameter corresponds to a significance threshold in a classical setting.

To reduce the computational cost, inference in the association model is approximated, only considering a single most relevant SNP-regulator per gene, with the other $b_{n,g}$ forced to 0. This bottleneck approximation ensures tractability of the joint association model for large-scale studies, avoiding the need to track the covariance between effects from multiple SNPs.

2) Known factor model accounts for the effect of known covariates $\mathbf{F}$ of individual samples, such as environmental conditions, gender, or a population indicator. The linear effects of C measured covariates in the jth individual, $\mathbf{f}_j = \{f_{1,j}, \ldots, f_{C,j}\}$, is taken into account using Bayesian regression (Figure 2.2a, blue plate):

$$P(y^{(2)}_{g,j}\,|\,\mathbf{f}_j, \mathbf{v}_g, \tau_g) = \mathcal{N}(y^{(2)}_{g,j}\,|\,\sum_{c=1}^{C} v_{g,c}\, f_{c,j}, \frac{1}{\tau_g}) \tag{2.5}$$

$$P(v_{g,c}\,|\,\alpha_c) = \mathcal{N}(v_{g,c}\,|\,0, \frac{1}{\alpha_c}) \tag{2.6}$$

$$P(\alpha_c) = \Gamma(\alpha_c\,|\,a_\alpha, b_\alpha). \tag{2.7}$$

Here, $\mathbf{v}_g = \{v_{g,1}, \dots, v_{g,C}\}$ is the corresponding weight vector for each gene g. The gamma prior on the inverse variance α_c for weights of each factor introduces automatic relevance detection (ARD) (Mackay, 1995; Neal, 1996), driving the weights of unused factors to 0 and thereby switching them off. This provides complexity control of the model by regularising the effective number of covariates.

3) Hidden factor model accounts for the effect of hidden factors (such as unmeasured covariates and global effects on expression levels) on the gene expression levels. We use a probabilistic variant of the classical factor analysis model for this task. It has been shown that this model captures hidden factors better than alternative linear models, such as probabilistic principal component analysis or independent component analysis (Stegle et al., 2008). Similarly to known factors, the expression level of gene g in individual j is modelled by linear effects from a chosen number of K hidden factors $\mathbf{x}_j = \{\mathbf{x}_{1,j}, \dots, \mathbf{x}_{K,j}\}$ (Figure 2.2a, red plate).

$$P(y^{(3)}_{g,j}\,|\,\mathbf{x}_j, \mathbf{w}_g, \tau_g) = \mathcal{N}(y^{(3)}_{g,j}\,|\,\sum_{k=1}^{K} w_{g,k}\, x_{k,j}, \frac{1}{\tau_g}) \tag{2.8}$$

$$P(w_{g,k}\,|\,\beta_k) = \mathcal{N}(w_{g,k}\,|\,0, \frac{1}{\beta_k}) \tag{2.9}$$

$$P(x_{k,j}) = \mathcal{N}(x_{k,j}\,|\,0, 1) \tag{2.10}$$

$$P(\beta_k) = \Gamma(\beta_k\,|\,a_\beta, b_\beta). \tag{2.11}$$

Note that in contrast to the known factor model, the factor activations $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_J\}$ are unobserved random variables that need to be inferred from the expression profiles. Again, the ARD prior switches unused factors off, thereby providing probabilistic complexity control (Stegle et al. (2008), Chapter 2.3).

2.2.2 Inference

Parameter inference in VBQTL is implemented using variational Bayesian learning (Jordan et al., 1999), a generalisation of the expectation maximisation algorithm. An approximate Q-distribution over model parameters is iteratively refined until convergence. In each iteration, approximate distributions of individual parameters are updated according to a specified schedule, taking the current state of all other parameter distributions into account (Figure 2.2b-e). Choosing an approximation that factorises over individual models, the variational update equations have an intuitive interpretation:

1. The current belief of the residual dataset for a particular active model is calculated, taking the prediction from all other models and the estimated noise precision into account (Figure 2.2b).

2. The parameters of the active ith model are updated based on their previous states and the new residual dataset (Figure 2.2c).

3. The distribution of the model contribution $\mathbf{Y}^{(i)}$ is recalculated using the updated parameter values. The global noise precisions τ_g are updated (Figure 2.2d) based on the first and second moments of all the contributions.

4. The same procedure is in turn applied to the remaining models in the schedule (Figure 2.2e) until convergence.

This iterative procedure, performing updates of local parameter distributions in turn, can be interpreted as a message passing algorithm, where sufficient statistics of parameter and data distributions are propagated across the graphical model (Winn and Bishop, 2006).

The initial values of parameters are determined from maximum likelihood solutions. A random initialisation via sampling from the prior is possible as well; we have not explored the implications of this alternative here. Details on inference and the individual parameter update equations are given in Appendix A.

In experiments, we compare two alternative inference schedules of VBQTL. In iterative VBQTL (iVBQTL), the parameters are learned using several iterations through all model components, first updating the genetic model, then known and hidden factors. An important property of iVBQTL is that hidden factors are estimated jointly with the genetic state and known factors. This choice of schedule and the iterative learning help to ensure that variability that is due to genetic associations is not explained away by other parts of the model (Chapter 2.3).

In cases where neither known nor hidden factors are correlated with the genetic state, their effect can be learned independently without running the risk of explaining away meaningful association signal. This motivates fast VBQTL (fVBQTL), which performs a single update iteration of the full model, first inferring the contribution from the known and hidden factors, and then from the genetic state. This simpler schedule can save significant computation time, since the factor effects can be precalculated, and only a single iteration of the computationally more expensive genetic association model is needed. In cases where the genetic state is approximately orthogonal to the known and hidden factors, this cheaper approximation performs equally with iVBQTL for finding genetic associations (Chapter 2.3).

2.2.3 Alternatives

We compared VBQTL with previous methods that account for confounding variance in the context of expression QTL mapping. Similarly to VBQTL, they model known and hidden factors in the expression levels. The differences between the alternative methods are in the hidden factor model used, which in turn vary in the complexity control approach employed as highlighted below. Thus these alternative models are named after the hidden factor estimation method.

Standard model The classical model explains the expression variability solely by the effects of known factors and SNP genotypes, without accounting for the hidden factors. The model is identical to that presented in Chapter 1.4.1.

PCA Principal components analysis (PCA) can be interpreted as decomposition of the gene expression matrix $\mathbf{Y} = (\mathbf{y}_1 ... \mathbf{y}_N)$ into a product $\mathbf{UDV}^T$, where $\mathbf{U}$ is the matrix of left singular vectors, $\mathbf{D}$ is a diagonal matrix of singular values $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N \geq 0$, and $\mathbf{V}$ is the matrix of right singular vectors. To apply

PCA, we used $\mathbf{U}$ as the weight matrix $\mathbf{W}$, and $\mathbf{D}\mathbf{V}^T$ as the latent factors $\mathbf{X}$. For the benchmark figures, illustrating the effect for different numbers of factors, we limited the number of learned factors to a given number K by setting $d_{i,i} = 0$ for $i > K$.

PCAsig PCA with significance testing (PCAsig) model is an extension of PCA, where complexity is controlled by retaining only components that explain more variance than expected by chance. Significance testing of PCA components in the PCAsig model was performed analogously to SVA (Leek and Storey, 2007), but without enforcing uniformity of the p-values. We found the variance explained by each component i by calculating the statistic $d_i = \frac{\lambda_i^2}{\sum_{j=1}^{N} \lambda_j^2}$. We then permuted the columns of $\mathbf{Y}$ L times, calculating null statistics $d_{i1}, d_{i2}, ..., d_{iL}$ analogously. Given a cutoff value α, component i was deemed to be significant if the fraction of null statistics greater than d_i was less than α.

SVA Surrogate variable analysis (SVA) model is a further extension of PCAsig. After applying the PCAsig model, each retained significant component is tested for association with all the genes using a 5% FDR cutoff. For each component, PCA is applied on the subset of genes associated with it, and the first principal component (i.e. the mean of the gene expression values) is used as the surrogate variable. The SVA package was downloaded from http://www.genomine.org/sva, and applied to datasets with default parameters, using 100 permutations and varying only the significance cutoff. The model implementation uses a Python to R bridge provided by RPy (http://rpy.sourceforge.net), allowing to call the original code provided by the authors.

For a quantitative evaluation of the performance of each method, we considered the resulting residuals of the estimated effects from known and hidden factors. To detect eQTLs we applied standard statistical tests employing a linear model on the SNP genotype on these residual datasets (Chapter 1.4.1). For iVBQTL and fVBQTL, we inferred the posterior parameter distributions, and subtracted off the estimated effect of known and hidden factors. For other methods, we first subtracted off the standard linear regression fit of the known factors, and then learned and subtracted off the hidden factor effects on the residuals. All these alternative methods are also implemented in the general framework.

While VBQTL shares basic assumptions with these alternatives, there are a number of differences. First, it is a probabilistic model that operates with uncertainties in the parameter estimates as explained above. Second, the hidden factor model allows for non-orthogonal components, and provides probabilistic complexity control based on ARD. Third, the iVBQTL schedule takes the genetic signal into account when estimating the hidden factor effect. Finally, the VBQTL model estimates a global gene-specific noise level, while the non-Bayesian models either estimate noise levels implicitly (SVA) or assume noise-free observations (PCA, PCAsig).

2.3 Method comparison

We employed a simulated dataset to highlight the differences between alternative approaches to account for global factors in eQTL finding.

2.3.1 Comparison on simulated data

Simulation setup

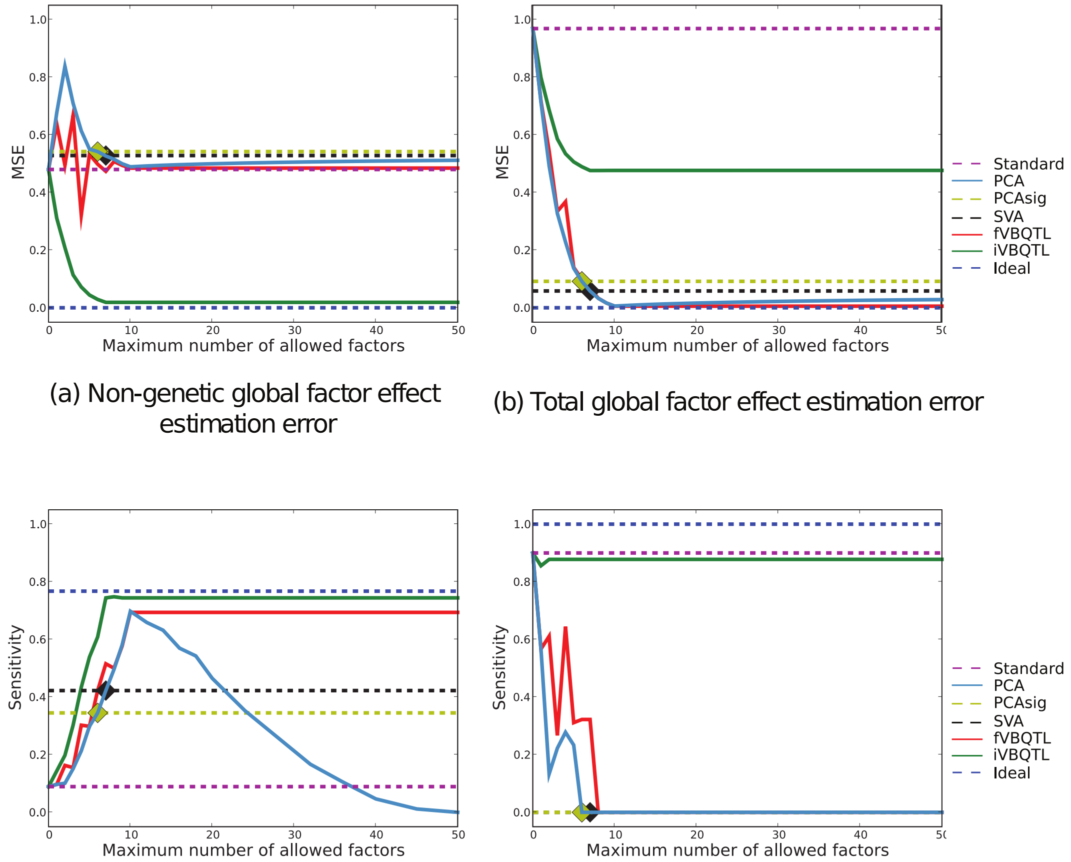
Our synthetic expression data combines linear effects from genetic associations (eQTLs), known, hidden, and genetic global factors, and gene-specific noise (Appendix A). We used three known and seven unknown global factors whose influence varies significantly to simulate effects with a range of magnitudes. These factors are meant to represent sources of confounding variation that are encountered in the study of the real datasets. We also introduced three global genetic factors giving rise to *trans* eQTL hotspots, mimicking the action of a genetic variant in a transcriptional regulator (e.g. transcription factor or pathway component). Such loci have been observed in several eQTL mapping studies (Brem et al., 2002; Schadt et al., 2005). We designated three genes with a simulated eQTL as such regulators, and simulated correlated expression levels for 15% of the genes for each. While the specific simulation scenario may be biased in the comparative performance of different methods, its underlying linear model is shared by all the considered approaches, and it gives intuition for the results on real datasets discussed later.

Complexity control determines the accuracy of the hidden factor model.

We assessed the ability of the considered methods to recover the simulated confounding variability. For those approaches that do infer hidden factor effects, we varied the corresponding complexity control parameters to investigate the influence on performance. For methods that take the number of components in the hidden factor model as a parameter (PCA, VBQTL), performance for one to 50 hidden factors was compared. For significance-testing based methods, we considered different significance cutoffs α in the range $[0.01, 0.5]$.

iVBQTL correctly captured the non-genetic global factor effects (Figure 2.3a), as it is the only method that models the genetic signal when learning hidden factors. All other methods treat the simulated transcription factor contributions as confounding variation and explain them away. This can be a desired effect when the genetic signal is not of primary interest, or a serious shortcoming when downstream eQTLs are sought.

Complexity control settings determined the performance of capturing the simulated global effects on expression levels. PCA was most accurate when the number of hidden factors was set to 10, since seven hidden factors and three transcription factors were simulated. For larger number of components PCA overfitted, and started explaining away genetic signal, resulting in the increase in error. For a small number of components, transcription factor effects were explained away first, which increased the error in estimating the hidden factors alone. However, the estimates of the total global effects improved. PCAsig and SVA found 6 and 7 significant hidden factors for the wide range of significance cutoffs, $\alpha \in [0.01, 0.5]$, respectively. They failed to detect some of the weaker hidden effects that continued to mask the genetic signal, and underfitted the data. Their performance was similar to PCA with the matching number of components. While the significance-testing based complexity control prevents these approaches from overfitting, only a single outcome is observed for a wide range of parameter settings, with the models settling to a rigid suboptimal solution.



(a) Non-genetic global factor effect estimation error

(b) Total global factor effect estimation error

(c) Immediate (cis) eQTLs

(d) Downstream (trans) eQTLs

Figure 2.3: Sensitivity of recovering simulated hidden factor effects and eQTLs for Bayesian and non-Bayesian methods. **(a)** Mean-squared error in estimating only the hidden factor contribution. Methods that do not explicitly retain the genetic factors explain them away as hidden global factors, resulting in high error comparable to not accounting for hidden factors at all (Standard). **(b)** Mean-squared error in estimating the contribution from hidden and genetic factors. **(c)** Sensitivity of recovering immediate SNP associations. **(d)** Sensitivity of recovering downstream associations. Seven hidden factors and three transcription factor effects were simulated. For eQTL sensitivity, standard eQTL finding on simulated data (Standard) and same data without the hidden effects (Ideal) are included as comparisons. PCAsig and SVA identified a constant number of hidden components (marked with a diamond shape), thus only a single result (dashed line) is given.

fVBQTL achieved the most accurate estimation of global variation. Notably, unlike PCA, its performance did not degrade for large numbers of hidden factors in the model, exhibiting good complexity control in this scenario.

Hidden factor effect estimation accuracy is mirrored in eQTL finding sensitivity.

We determined the sensitivity and specificity of the considered methods for detecting the immediate and downstream simulated genetic associations. The significance of an eQTL was tested using a two-sided t test on the correlation coefficient with a 0.1% Bonferroni corrected per-gene false positive rate in the genetic association model. The results when calling eQTLs using regression on ranks, or permutations to establish the empirical null distribution of LOD scores were almost identical. As a benchmark, the comparison includes eQTL finding using the standard method on both raw expression data (Standard), and an ideal case, where the simulated hidden factor effects are removed, but the simulated genetic factors maintained (Ideal).

The accuracy of the hidden factor effect estimation mirrored the immediate eQTL finding sensitivity (Figure 2.3c). The specificity was consistent with the chosen false positive rate for all methods (data not shown). fVBQTL and iVBQTL recovered more true *cis* eQTLs compared to other methods, approaching the performance of the ideal case, mirroring the accuracy of estimating hidden factor effects. PCA overfitted when the number of components used was greater than the true number of ten simulated global factors, explaining away genetic signal. While the PCA error for detecting global effects increased only marginally, the decrease in sensitivity for identifying eQTLs was severe. The overfitting in case of PCA, and underfitting in case of PCAsig and SVA both resulted in a loss of sensitivity to find the simulated *cis* associations. fVBQTL and iVBQTL did not suffer from either deficiency, capturing nearly all the associations possible in the ideal case.

All methods except iVBQTL and standard method explained away simulated *trans* eQTL hotspots (Figure 2.3d). This is due to the global factor effect estimation accuracy, where iVBQTL alone refrained from explaining the hotspots away as a global factor. The standard method found nearly all the original *trans* associations, actually outperforming methods that explain away confounding variability. Thus, in cases where there is true genetic signal with widespread downstream effects, its contribution needs to be taken into account to retain its relation to genotype, and avoid attributing it to a confounding global cause. This is straightforward in our framework, and is demonstrated by the good performance of iVBQTL in this scenario. iVBQTL retained the original associations, while explaining away non-genetic causes of expression variability, thus adding power to detect the weaker, masked eQTLs. This effect is also observed in the study of crosses of inbred strains below.

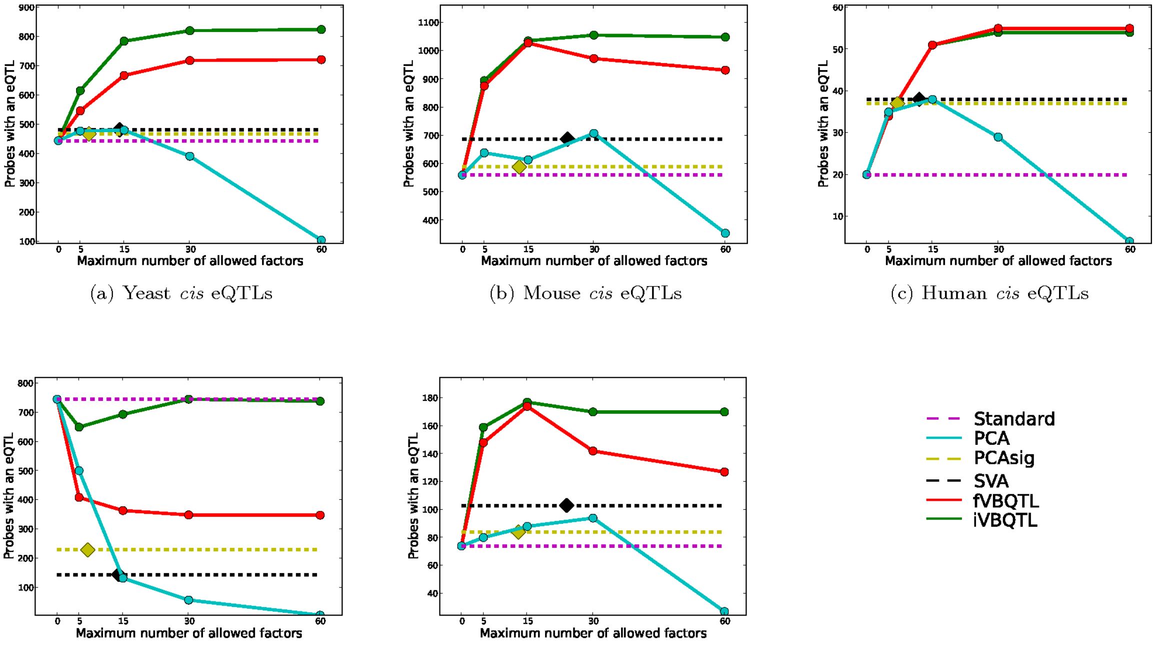
Taken together these results suggest that it is important to account for the confounding sources of variation in expression levels, while keeping the signal of the genetic state. Correct complexity control is required to avoid over- and underfitting in order to achieve optimal sensitivity for detecting true genetic associations.

2.3.2 Comparison on real data

Next, we compared the same methods for expression QTL finding on yeast (Brem and Kruglyak, 2005), mouse (Schadt et al., 2005), and human (Stranger et al., 2007) datasets. These represent common study designs of an outbred population (human), and a population of crosses between inbred strains (yeast, mouse). We considered 5, 15, 30, and 60 hidden factors for PCA and VBQTL, and $0.01, 0.1$, and 0.3 as significance cutoffs for SVA and PCAsig. Expression QTLs were detected using a two-sided t test analogously to the simulation scenario. Again, results for alternative genetic association tests were similar (data not shown).

Accounting for hidden factors helps to detect additional *cis* eQTLs in an outbred population

We applied the considered methods on the genotype and expression data from 90 individuals of the CEU (CEPH from Utah) HapMap phase 2 samples (Stranger et al., 2007; The International HapMap Consortium, 2005). The data consisted of genotypes of 55,000 SNPs and expression levels of 618 probes from chromosome 19 (results for three more chromosomes were similar, data not shown). The



(a) Yeast *cis* eQTLs

(b) Mouse *cis* eQTLs

(c) Human *cis* eQTLs

(d) Yeast *trans* eQTLs

(e) Mouse *trans* eQTLs

Figure 2.4: Number of probes with an eQTL found as a function of maximum number of hidden factors for three previously published datasets. Significance-testing based methods (PCAsig, SVA) identified the same number of factors for a wide range of cutoff values ($\alpha \in [0.01, 0.3]$), thus only a single count is given (dashed lines), together with the number of factors found (diamond shape). Other methods were applied with a maximum number of 5, 15, 30 and 60 hidden factors.

expression levels were measured in EBV-transformed lymphoblastoid cell lines of healthy individuals. The gender covariate was included as a known factor for all methods. We did not consider probes with overlapping SNPs. Following Stranger et al. (2007), an association was called to be in *cis* when the SNP was within 1Mb from the probe midpoint and in *trans* otherwise.

The standard method found the least gene probes with a *cis* association (20, Figure 2.4c), suggesting that strong confounding sources of variation are present in this dataset. The number of identified probes with a *trans* association was not significantly higher than expected by chance at the chosen FPR, which is in line with previous results (Stranger et al., 2007), and suggests little intrachromosomal *trans* regulation.

PCA, the simplest method for accounting for hidden factors, found additional associations when up to 30 principal components were used, but substantially fewer for 60 components. This is expected, since there are no more than 90 degrees of freedom in this dataset, and 60 principal components accounted for over 94% of the variance (Table B.6), and hence PCA is likely to explain away part of the genetic association signal for large numbers of components.

The significance-testing based methods, SVA and PCAsig both found additional associations compared to the standard method. It is remarkable that both found a constant number of significant hidden factors for the wide range $\alpha \in \{0.01, 0.1, 0.3\}$ of significance cutoffs considered, again exhibiting rigid complexity control. The performance of SVA with the 12 hidden factors found is close to performance of PCA with 15 components (both find 38 probes with an association). Similarly, PCAsig with the 7 significant components performs comparably to PCA with 5 components (37 vs. 35 probes with an association). This shows the intrinsic similarity of these methods to PCA, as was also observed in the simulation scenario.

fVBQTL and iVBQTL found more probes with an association (55 and 54) than all other methods, representing an almost threefold increase in the number of genes with a *cis* eQTL. Complexity control assured that the performance saturated for large enough number of factors and did not degrade as for PCA. None of the estimated hidden factors was significantly correlated to a SNP genotype, suggesting that individual genetic variants do not have global effects on many gene expression levels in this dataset.

It is important to note that the model performance depends on two aspects. First, the model complexity control, regulating the amount of variance explained, is important to ensure that genetic signal is not attributed to hidden factors. Overfitting in case of PCA for a large number of components is an example of such an effect. Second, while alternative hidden factor models explained similar amounts of variance, their performance differed due to the underlying model. For example, PCA and fVBQTL both explained about 70% of variance in the observed expression levels (Table B.6) yet fVBQTL identified additional associations. These findings are consistent with the simulation study results, and suggest that the additional associations found with Bayesian models are due to differences in the underlying model and complexity control.

Accounting for hidden factors adds power to detect *cis* associations in crosses between inbred mouse and yeast strains.

Next, we applied the methods to two datasets of inbred strain crosses. The yeast expression dataset (Brem and Kruglyak, 2005) (GEO (Barrett et al., 2009) accession GSE1990 with genotypes provided by authors) contained 7084 expression measurements and 2925 genotyped loci in 112 crosses of segregating yeast strains. The mouse expression data consisted of 23,698 expression measurements for 111 F_2 mouse lines, and genotypes at 137 genetic markers. An association was called to be in *cis* if the probe and the genotyped locus were from the same chromosome, and in *trans* otherwise.

The relative performance of different methods was similar to their ability to detect *cis* eQTLs in the outbred population dataset (Figures 2.4a, 2.4b). The absolute performance gain was significantly lower for all methods, however. This finding suggests that the genetic signal is stronger compared to confounding sources of variation, which is not unexpected from the study design. All factor methods identified additional associations compared to the standard method. PCA overfitted for larger numbers of principal components used, explaining away genetic association signal. SVA and PCAsig found the same number of significant hidden factors for a range of significance cutoffs considered, exhibiting little flexibility. Again, their performance was similar to extrapolation of PCA results with matching numbers of effective components. fVBQTL and iVBQTL found additional genetic associations in *cis* compared to the standard model and other methods for accounting for confounding variance, as observed in simulations and human dataset. Summary statistics for the method performance can be found in tables B.6 to B.8.

Iterative learning with iVBQTL overcomes difficulties in detecting *trans* associations for crosses of inbred strains.

All methods found additional *trans* associations in mouse, but fewer than the standard method in yeast (Figure 2.4d, 2.4e). In yeast, the more variance was explained by the hidden factors, the fewer *trans* eQTLs were found, suggesting that the global determinants of gene expression variation were correlated with the genetic state. Indeed, the inferred hidden factor levels were correlated with genotypes of "pivotal loci" that are associated with expression levels of hundreds of genes.

The effect of pivotal loci has been observed before, and interpreted in different ways (Kang et al., 2008; Leek and Storey, 2007). It could be that the additional variation is artefactual, and correlated to the genetic state by chance. In this case, all the original *trans* associations are spurious. The alternative explanation is that the genotype of these loci have real downstream effects on the expression profiles of very many genes. In this case the variance is not confounding the genetic signal, but in fact is a part of it, and hence should not be explained away.

Previous methods do not provide consistent ways of dealing with this issue. The SVA authors also suggest to remove the effect of the primary variable first. However, the authors do not consider accounting for the genetic effect in their application to the same yeast dataset (Leek and Storey, 2007). Kang et al. (2008) also explain away *trans* associations when applying their correction procedure. We provide a principled approach for dealing with this situation and show its merit. The iVBQTL scheduling takes the genetic state into account while learning the hidden factors, and as a consequence is more sensitive to genetic associations.

2.4 Expression QTL mapping in large human populations

After confirming that our method works on simulated data, and comparing performance on different small scale real datasets, we analysed several human large scale expression datasets in depth.

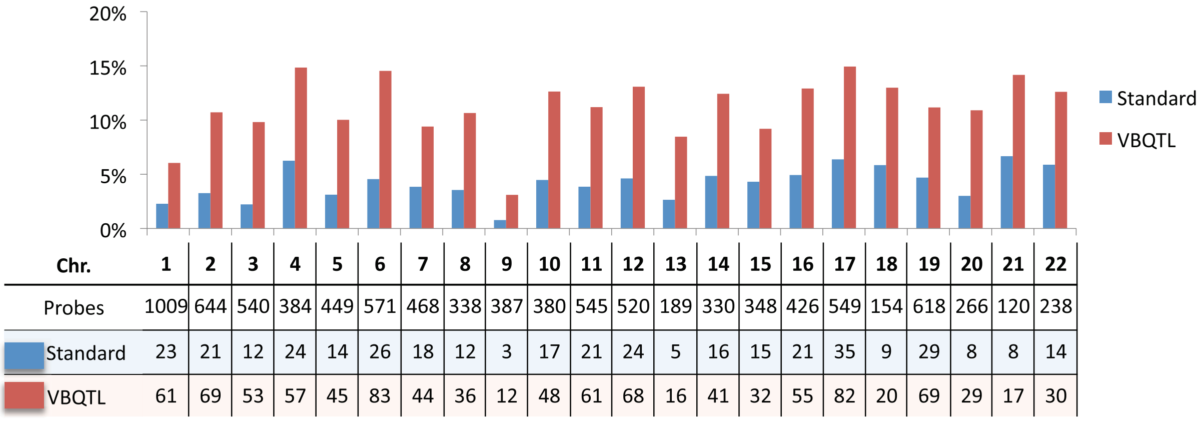
2.4.1 HapMap phase 2 dataset.

Motivated by the results of the initial study of a single human chromosome, we applied fVBQTL, learning 30 hidden factors, to the 10,000 most variable expression probes of the HapMap 2 dataset. We searched for *cis* eQTLs in the original expression data (standard eQTLs) as well as the residuals of fVBQTL (VBeQTLs), using a 2-tailed t test with 0.1% Bonferroni-corrected per-gene FPR to assess the significance of association.

VBQTL increases power threefold

On the CEU population, we found 1051 genes with a VBeQTL at false discovery rate (FDR) of 0.9%, and 382 genes with a standard eQTL at FDR of 2.6% (Figure 2.5). This result corresponds to nearly a threefold increase in the number of genes with an association, and is consistent across chromosomes. A similar increase in the number of associations was found for other populations (Table B.1).

We repeated this genome-wide experiment on pooled populations. Due to the increased sample size, it was possible to detect additional associations. We found 2696 genes with a VBeQTL compared to 1045 genes with a standard eQTL at the 0.1% FPR (Figure 2.6a). The VBeQTLs in the pooled sample cover 27% of all the considered probes, suggesting that the number of human genes whose expression levels are affected by common *cis*-acting genetic variation may be significantly higher than previously shown (Stranger et al., 2007; Williams et al., 2007). This additional abundance of associations suggests that detection of *cis* eQTLs has not been saturated and larger sample sizes may lead to evidence of even more extensive *cis* regulation by common polymorphisms.



Chr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Probes	1009	644	540	384	449	571	468	338	387	380	545	520	189	330	348	426	549	154	618	266	120	238
🟦 Standard	23	21	12	24	14	26	18	12	3	17	21	24	5	16	15	21	35	9	29	8	8	14
🟥 VBQTL	61	69	53	57	45	83	44	36	12	48	61	68	16	41	32	55	82	20	69	29	17	30

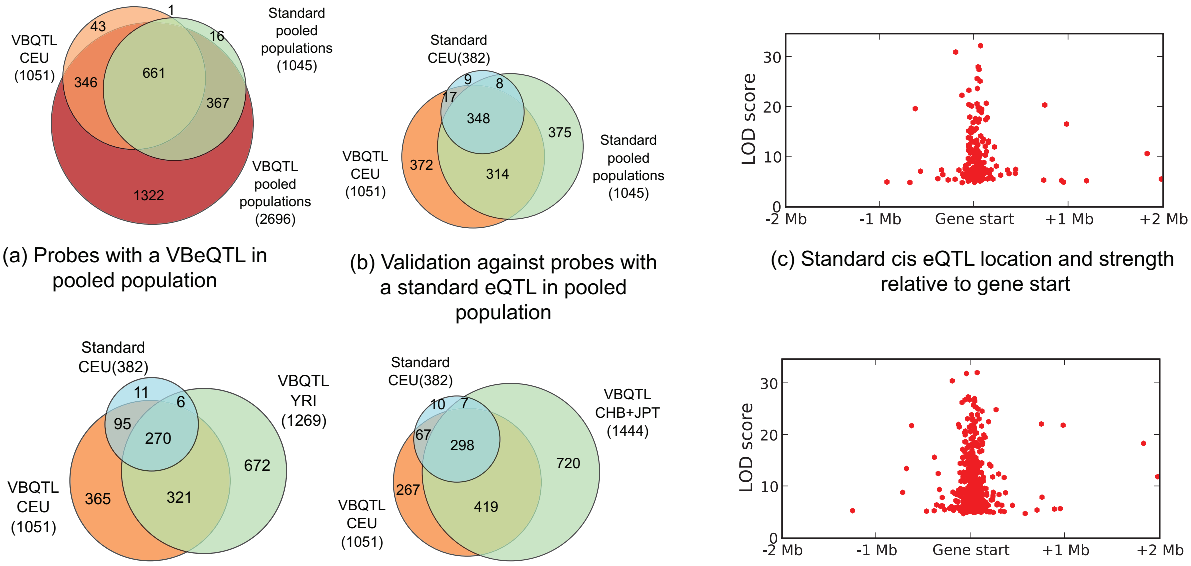
Figure 2.5: Fraction of tested genes with a *cis* association in individual chromosomes for the HapMap CEU population (FPR=0.1%).

Exploratory results indicate additional power to find *trans* eQTLs without explaining away eQTL hotspots (Table B.4). These should be interpreted with caution due to very stringent requirements for multiple testing correction, however.

Additional associations are due to increased sensitivity.

It is important to demonstrate that the additional associations found after removing the learned non-genetic factors are biologically meaningful. We provide evidence that the additional associations found in HapMap phase 2 data are real in three ways.

First, we investigated how many of the genes with a VBeQTL in each of the three populations individually were replicated using the standard method on a pooled data set containing all populations. Note that this will only validate weak associations that occur in multiple populations – we would not expect weak population-specific associations to be replicated in the pooled data set. However, we expect many of the associations to be replicated in multiple populations (Stranger et al., 2007). A total of 63% of all and 46% of the additional



(a) Probes with a VBeQTL in pooled population

(b) Validation against probes with a standard eQTL in pooled population

(c) Standard cis eQTL location and strength relative to gene start

(d) Replication with probes having VBeQTLs in YRI population

(e) Replication with probes having VBeQTLs in CHB+JPT populations

(f) VBeQTL location and strength relative to gene start

Figure 2.6: Validation of VBQTLs by comparison to standard eQTLs. **(a,b,d,e)** Venn diagrams depicting overlap of probes with a standard eQTL or VBeQTL in the CEU population and probes with an eQTL in other populations.

associations found in the CEU population were recovered using the standard method in the pooled population (Figure 2.6b). The remaining additional associations may be explained by even weaker signals that were recovered by applying fVBQTL, or as population-specific effects that do not stand out in the pooled sample. Analogous overlaps were found when excluding the CEU population from the pooled analysis (Table B.3).

Second, we evaluated to what extent the additional genes with a VBeQTL in a single population were replicated in other populations. For instance, 56% of genes with a CEU VBeQTL were replicated on the YRI population (Figure 2.6d), and 68% on the CHB+JPT population (Figure 2.6e). These overlaps are consistent with overlaps of standard eQTLs, and are similar for other populations (Table B.2), and alternative methods accounting for hidden factors.

Finally, we validated that the locations of the novel associations are distributed similarly to the original ones. We analysed the distribution of the position of additional *cis* associations around the gene start along with the association LOD scores. The additional VBeQTLs have very similar characteristics to the standard eQTLs, being concentrated around the gene start (Figure 2.6c, 2.6f), in line with results from Stranger et al. (2007).

2.4.2 The MuTHER study

The MuTHER (Multiple Tissue Human Expression Resource) project is a large scale collaboration that seeks to understand genotype, gene expression, methylation, and disease phenotype variation (Nica et al., 2011). Over 800 individuals (a mixture of monozygotic and dizygotic twins from the TwinsUK cohort (Spector and Williams, 2006)) have donated blood, fat, skin, and in some cases muscle, samples to the project. In the following, I will discuss some of the analysis aspects of the pilot gene expression data. These data include gene expression measurements from fat, skin, and LCLs for about 160 individuals and 27,000 probes. We sought to find expression QTLs in multiple tissues by applying the Bayesian factor analysis model of PEER to the tissue gene expression data.

Fitting hyperparameters to maximise consistency

In the studies of HapMap samples, we varied the number of latent factors as the only free parameter. Here, we also varied the ARD hyperprior, as well as noise prior, to sensitively adjust how much gene expression variability is explained by the hidden factors.

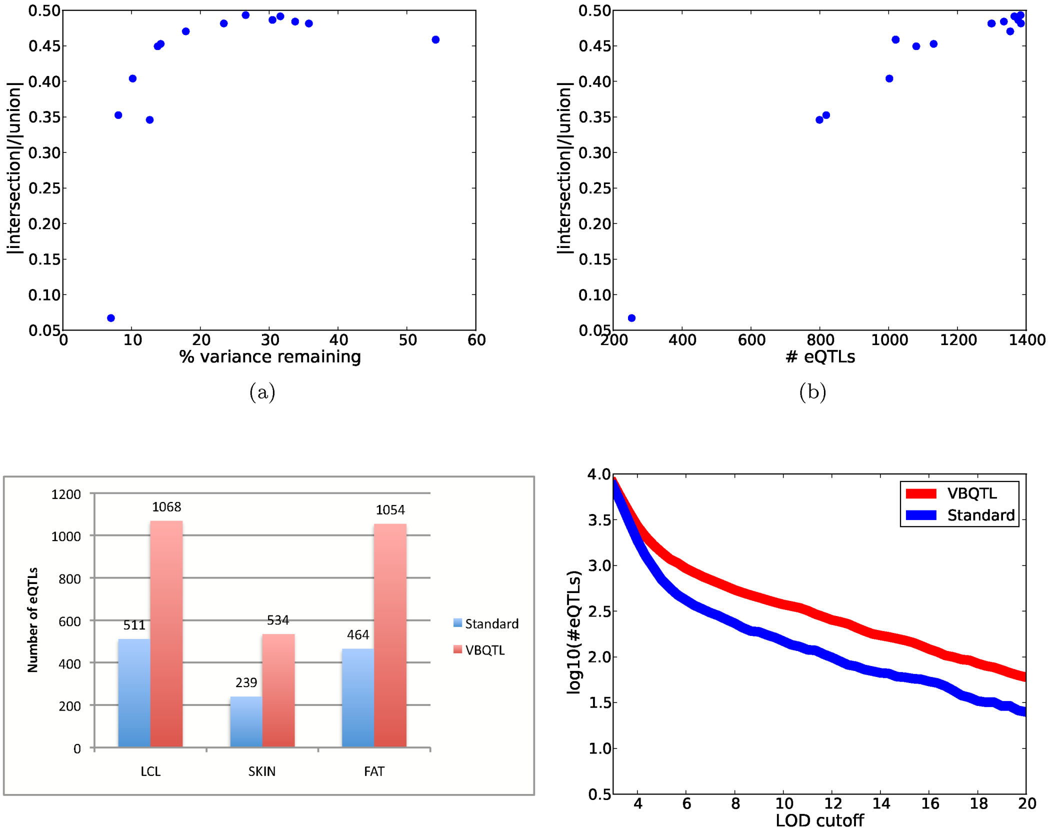
The parameters of the inverse variance prior have a natural interpretation in the context of exponential family models. The conjugate prior is $\Gamma(a_0, b_0)$ distributed, where a_0 and b_0 correspond to the sum and effective number of prior observations, respectively (Davison, 2003). We varied the prior mean $\frac{a_0}{b_0}$ from 10^{-6} to 10^{-2}, and the number of observations b_0 from $10^{-3}N$ to N (where N is the number of observations in data) for both weight and noise precision prior, and learned 120 latent factors.

To choose the best parameter settings, we used the fraction of overlap between eQTLs found in co-twins as the objective function to optimise. The study cohort has a natural structure of paired twins. We called eQTL sets $Q1$ and $Q2$ (Alexandra Nica, require 10^{-3} nominal Spearman Rank Correlation p-value) in the sets of "first" and "second" twins in a twin pair, and calculated the Jaccard index $J(Q1, Q2) = \frac{|Q1 \cap Q2|}{|Q1 \cup Q2|}$ between them, as well as the fraction of residual variance remaining for each parameter setting after subtracting off the factor analysis model contribution.

We found a broad peak of parameter settings that produced a similar fraction of variance explained and eQTL overlap (Figure 2.7a). This confirms that the method is robust to a wide range of parameter settings, spanning many orders of magnitude. Furthermore, the overlap of eQTLs between co-twins was a very good predictor of total findings (Figure 2.7b), motivating the choice of highest overlap fraction from another angle.

Many more QTLs are found

We found many eQTLs in the three tissues (Figure 2.7c). The properties and overlaps of these are discussed in other work (Nica et al., 2011). The relatively low number of discoveries in skin is likely due to poorer quality RNA. There is no relation between the overall expression level or the weight of RNA integrity



(a)

(b)

(c)

(d)

Figure 2.7: eQTL finding results on the MuTHER dataset. **(a)** Overlap of co-twin eQTLs as a function of variance explained by the factor analysis model **(b)** Correlation of co-twin eQTL overlap and total number of discoveries **(c)** Total number of discovered eQTLs in the three tissues with standard model and VBQTL (LOD>5) **(d)** Difference in number of discoveries between standard model and VBQTL as a function of significance cutoff.

number (RIN) on the expression level in a linear model and the frequency of eQTL discovery. In addition, we could interpret some of the broad variance components in the skin and fat tissue (Chapter 2.5 below).

As an additional quality control, we tested whether VBQTL increases discoveries at all significance cutoffs. If a lot of discoveries are made at lenient cutoffs, it would indicate a large fraction of likely false positives, and problems with the model. However, we found that VBQTL finds additional eQTLs only at relatively high cutoffs ($-\log_{10} p > 5$, Figure 2.7d), confirming that our approach does not indiscriminately amplify all signal.

2.4.3 The 1000 Genomes low coverage pilot

Some of the HapMap phase 2 unrelated individuals have been sequenced at low coverage genome wide as part of the 1000 Genomes Project (Consortium, 2010). It is interesting to test whether the availability of genotypes at all loci increases power to detect eQTLs.

We used the expression and genotype data for the 43 CEU, 42 YRI, and 59 CHB+JPT indviduals for whom we have the expression and genome sequence data. We filtered the HapMap 2 genotypes to 317,000 to 1,000,000 polymorphisms assayed by standard Illumina genotyping chips (designated 317K, 610K, and 1M), and also included the 1000 Genomes genotypes (1000G) at all loci called from sequencing data.

We then searched for eQTLs in a 50kb window centered around the expression probe independently for each population and genotype dataset. We used Spearman's Rank Correlation coefficient as a test statistic, and assessed its significance by performing 20 permutations of the entire analysis to obtain a genome-wide significance cutoff corresponding to 5% false discovery rate. Both standard eQTL model on original data (Standard) and same approach on residuals of the PEER factor analysis model (VBQTL) were assessed.

Consistent with previous experiences, we found additional eQTLs using expression residuals from PEER (Figure 2.8). More interestingly, we observed an increase in the number of discoveries using the full genetic background. For populations that are relatively well represented in the genotyping chips used (CEU,

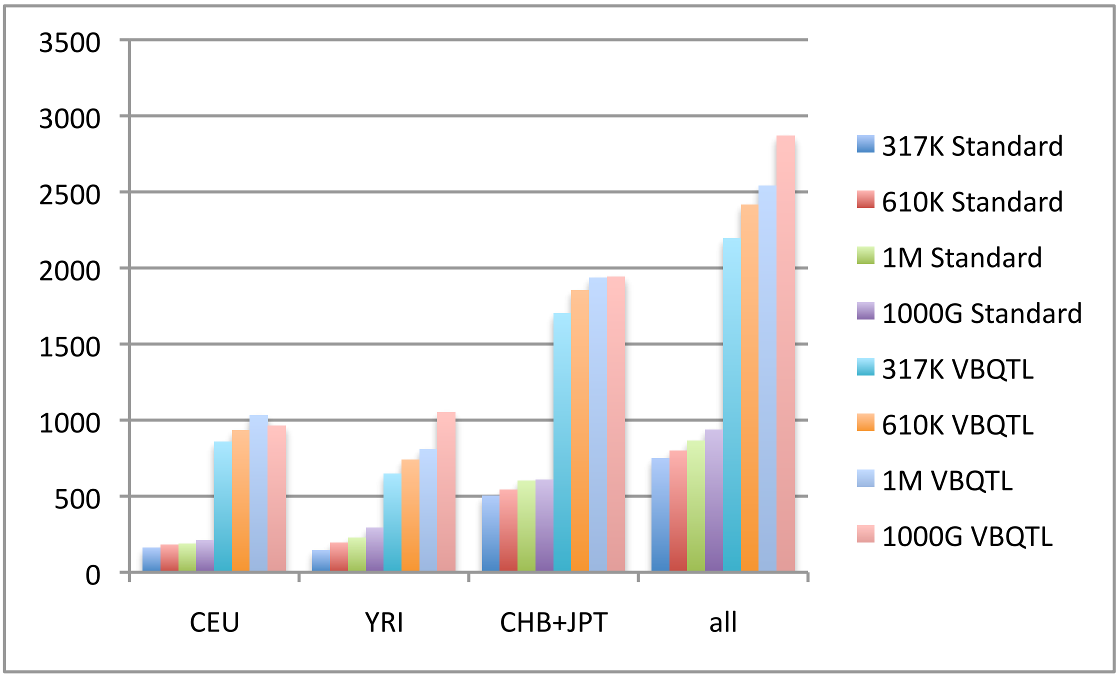


Figure 2.8: eQTL finding results on the HapMap2 dataset with 1000 Genomes genotypes. Number of eQTLs in three different populations as well as combined population significant at 5% FDR using the standard eQTL model (Standard), and residuals from PEER factor analysis model (VBQTL). Spearman's Rank Correlation was used as a test statistic, with genome-wide significance cutoff determined from permutations. Appropriate covariates for gender and population were included in the models.

CHB+JPT), the increases were low, while for a genetically more diverse population (YRI) as well as the full combined population with more power to assess significance of rarer alleles, the number of discoveries increased by 30 and 10%, respectively. We expect the genotypes of low frequency alleles to be even more beneficial in larger cohorts, where they are assayed in sufficient numbers to reliably test their effect on gene expression levels.

2.5 Interpretation of learned hidden factors.

The hidden factor models hypothesise a set of unobserved non-genetic factors that influence the measured gene expression levels. To gain insights into their interpretation we considered correlations to known effects such as gender, population or environment, and the sets of genes most influenced.

Human panels. We applied fVBQTL to expression data from individuals of all three HapMap populations, and tested for correlation between the inferred hidden factors and the population and gender indicator variables. The resulting correlation coefficients (Table B.5) indicate that many of the learned latent causes are correlated with population and that one is strongly correlated with gender. This implies that the hidden factor model can recapture variance in the gene expression levels due to true underlying properties of individuals. However, none of the global factors learned in one population was correlated with any single SNP genotype.

We could not attribute any variance to the same causes in the MuTHER LCL expression data, as all samples came from women in one population. However, in other tissues, we could link some of the largest variance component to a single trait. In the fat tissue, the individual body mass index was correlated with the second inferred factor (Pearson's $r^2 = 0.27$). This is not unexpected, as obesity-related traits, including body mass index, have been shown to be correlated to many gene expression levels (Emilsson et al., 2008). The strongest influence on the MuTHER skin tissue gene expression data was RNA integrity number (RIN), which was correlated with the first inferred factor (Pearson's $r^2 = 0.37$). Many samples had low quality total skin RNA, due to the aggressive extraction procedures needed to isolate RNA from the resistant skin tissue. Low RNA quality implies degradation of RNA molecules among other effects, which has a broad effect on many gene expression levels, and was captured with an inferred latent factor.

Crosses of yeast strains A recent study in yeast looked for changes in eQTLs when segregating strains were grown in different media (Smith and Kruglyak, 2008). We applied fVBQTL to the expression data of this study (GEO accession GSE9376), without including any information about the growth condition. The first hidden factor learned was highly correlated with the indicator variable for the growth condition ($r^2 = 0.96$), demonstrating that the VBQTL model can successfully recover a strong environmental effect if it is present.

The global factors identified can be further analysed for biological signals, looking for GO term over-representation in the genes that they affect. We used the ordered GO profiling method (Reimand et al., 2007) to find significantly enriched GO categories for the 30 genes most affected by each factor. Recent results (Biswas et al., 2008) show that related linear Gaussian models find biologically relevant factors in the yeast expression dataset. We replicated these findings with our model, yielding factors enriched in biological functions, including sugar, alcohol and amino acid metabolic processes. Similar analysis in human and mouse did not show significant over-representation of GO categories, providing no evidence that the main axes of variation in the expression levels for these experiments are due to variation in common biological function. This could be due to poor GO annotation of the genes, gene features not related to GO biological function, or more technical sources of global variation, such as cell culture conditions (Pastinen et al., 2006).

2.6 Discussion

We have presented VBQTL, a probabilistic model to dissect gene expression variation in the context of genetic association studies. The model is implemented in a Bayesian inference framework that allows uncertainty to be propagated between different parts of the model, and yields posterior distributions over parameter estimates for more sensitive analysis. In comparative eQTL mapping experiments, VBQTL outperformed alternative methods for eQTL finding on simulated and real data. In the most striking example, VBQTL found up to three times more eQTLs than a standard method, and 45% more compared to the best alternative in the HapMap 2 expression dataset.

Our approach advances the methodology for understanding phenotypic variation. The implementation of a flexible framework allows models for explaining the observed variability to be straightforwardly combined. Notably, non-Bayesian models can also be included, as we demonstrated with PCA, SVA, and linear regression models. VBQTL controls the model complexity at the level of all individual components of expression variability, thereby preventing from over- and underfitting. Our experimental results on simulation and real data showed how explaining away too much variability removes some signal of interest from the data, and failing to account for all sources of confounding variation decreases power to detect the relevant signal. When the variable of interest is correlated with many gene expression levels, its effect can be falsely explained away by the hidden factor model. We showed that in such settings the choice of an iterative schedule helps to ensure that variability is explained by the appropriate part of the model. There can be no silver bullet solution that provides perfect results in any scenario with no supervision. Instead, modelling assumptions must be made explicit, and incorporated in the analysis, as is elegantly done in the Bayesian setting.

VBQTL and other methods that account for hidden factors all found additional expression QTLs in the datasets studied compared to the standard method. It is remarkable that, with only 270 samples, and looking in one tissue type, we can find significant genetic associations to 27% of the expressed genes. The replication of the additional associations in different populations suggests that they are genuine. The increase in power is due to the hidden factor model, which explains away unwanted non-genetic variability, thereby allowing the genetic effects to stand out to a greater extent. The high number of additional associations suggests that association finding studies in human have not saturated, and we expect the fraction of genes with an eQTL will increase further as the number of samples grows. It may be that the expression of the majority of human genes varies as a result of segregating genetic variation. While previous studies have reported only 12% of heritable variation to be due to *cis* variants (Price et al., 2008), this does not contradict the presence of weak *cis* eQTLs for a large fraction of the genes.

In conclusion, we believe that VBQTL provides a principled and accurate way to study gene expression and other high-dimensional data. Increasingly complex models combining genetic and other effects can explain significantly more of the variance in observed phenotypes, as suggested by this study and others. Our general framework provides the flexibility to facilitate these richer models, for example, we have already started exploring interaction effects as an additional model of the framework. It will be interesting to see how these approaches can contribute to our understanding of human disease genetics, potentially involving intermediate phenotypes such as gene expression and other factors.

The software used in this study is freely available online at http://www.sanger.ac.uk/resources/software/peer/.

Chapter 3

Genetic mapping with inferred traits

Collaboration note

This chapter contains work performed in collaboration with Dr. Oliver Stegle and Dr. John Winn for methods development. Oliver developed and implemented the sparse factor analysis model used in this chapter (Stegle et al., 2009), we then expanded on this work jointly (Parts et al., 2011). In particular, I applied the factor analysis model to simulated and real data, and performed the analyses of the results, including all association and interaction mapping. The coauthored manuscript forms the backbone of the chapter.

Expressing RNA molecules is a highly regulated process that depends on activations of specific pathways and regulatory factors. Such state of the cell is hard to measure (Chapter 1.3.1), making it difficult to understand what drives the changes in the gene expression. To close this gap we apply a statistical model to infer the cell state variables, such as activations of transcription factors and molecular pathways, from gene expression data. We demonstrate how the inferred state helps to explain the effects of variation in the DNA and environment on the expression trait via both direct regulatory effects and interactions with the genetic state. Such analysis, exploiting inferred intermediate phenotypes, will aid understanding effects of genetic variability on global traits, and help to interpret the data from existing and forthcoming large scale studies.

3.1 Expression analysis with cellular traits

Gene expression levels are determined by the state of the cell, as well as genotypes of the gene regulatory regions. A correct model for gene expression should incorporate both effects.

Context-dependent genetic effects

Locus effects in isolation are not sufficient to account for gene expression variability (see also Chapters 1.2.2 and 1.3.2). Environment and intermediate cellular phenotypes (e.g. transcription factor or pathway activation) can and do have large effects on the measured transcript levels (Brem and Kruglyak, 2005; Gibson, 2008). To understand the genetics of gene expression, we must therefore analyse the consequences of genetic variants in the context of these other factors. Studies in segregating yeast strains have investigated epistatic interactions (Brem and Kruglyak, 2005; Storey et al., 2005), recovering interactions with genotypes of a few major transcriptional regulators. Large scale efforts to map functional epistasis between genes are currently underway with promising initial results (Costanzo et al., 2010). A recent study also searched for genotype-environment effects, and found many gene expression levels affected by an interaction between the environment and the genotype of a major transcriptional regulator (Smith and Kruglyak, 2008). However, much remains to be done in this area. While gene expression has been used as an intermediate phenotype to study the genetics of global traits (Schadt et al. 2005, Emilsson et al. 2008, Chen et al. 2008), genetics of gene expression itself has not been considered jointly with relevant cellular phenotypes such as pathway or transcription factor activations. This is an important gap. It is the state of the cell that determines how genetic variation can affect the gene expression levels, thus a joint analysis with the intermediate phenotypes should inform us about the mechanisms involved – a crucial step for understanding the causes of phenotypic variability.

Inferring unmeasured cellular traits

Despite their importance, the intermediate phenotypes are usually not measured, thus genetic effects cannot be analysed in their cellular context. Fortunately, statistical approaches have been developed that allow inferring unmeasured factors which influence expression levels from expression data alone. Methods such as principal components analysis (Alter et al., 2000), network components analysis (Liao et al., 2003), surrogate variable analysis (SVA, Leek and Storey, 2007), independent components analysis (Biswas et al., 2008), and the PEER framework (Chapter 2) can be used to determine a set of variables that explain a part of gene expression variability with (usually) a linear model. Their application has been shown to increase power to find expression quantitative trait loci (eQTLs) by explaining away confounding variation (Leek and Storey, 2007; Stegle et al., 2010), and to yield variance components of the expression data that may be interpretable (Stegle et al., 2010).

Our approach

Here, we perform a thorough joint genetic analysis of a gene expression dataset with intermediate phenotypes inferred from gene expression levels. We revisit the data of Smith and Kruglyak (Smith and Kruglyak, 2008), where the authors looked for gene-environment interactions affecting gene expression levels in a population of segregating yeast strains grown in two different carbon sources. First, we use a variant of a sparse factor analysis model (Rattray et al., 2009; Stegle et al., 2009) to infer intermediate phenotypes from the gene expression levels (Figure 3.1a). Importantly, this method uses prior information to guide the inference of which factors are affecting which genes, as opposed to unsupervised methods (e.g. PEER, SVA, ICA) that learn broad effects. We use Yeastract (Teixeira et al., 2006) transcription factor binding and KEGG (Kanehisa et al., 2002) pathway data as prior information in the model, which allows the inferred phenotypes to be interpreted as transcription factor and pathway activations. We then analyse the variation in the learnt activations, and find that growth condition and segregating locus genotypes have a strong influence (Figure 3.1b). Finally, for the first

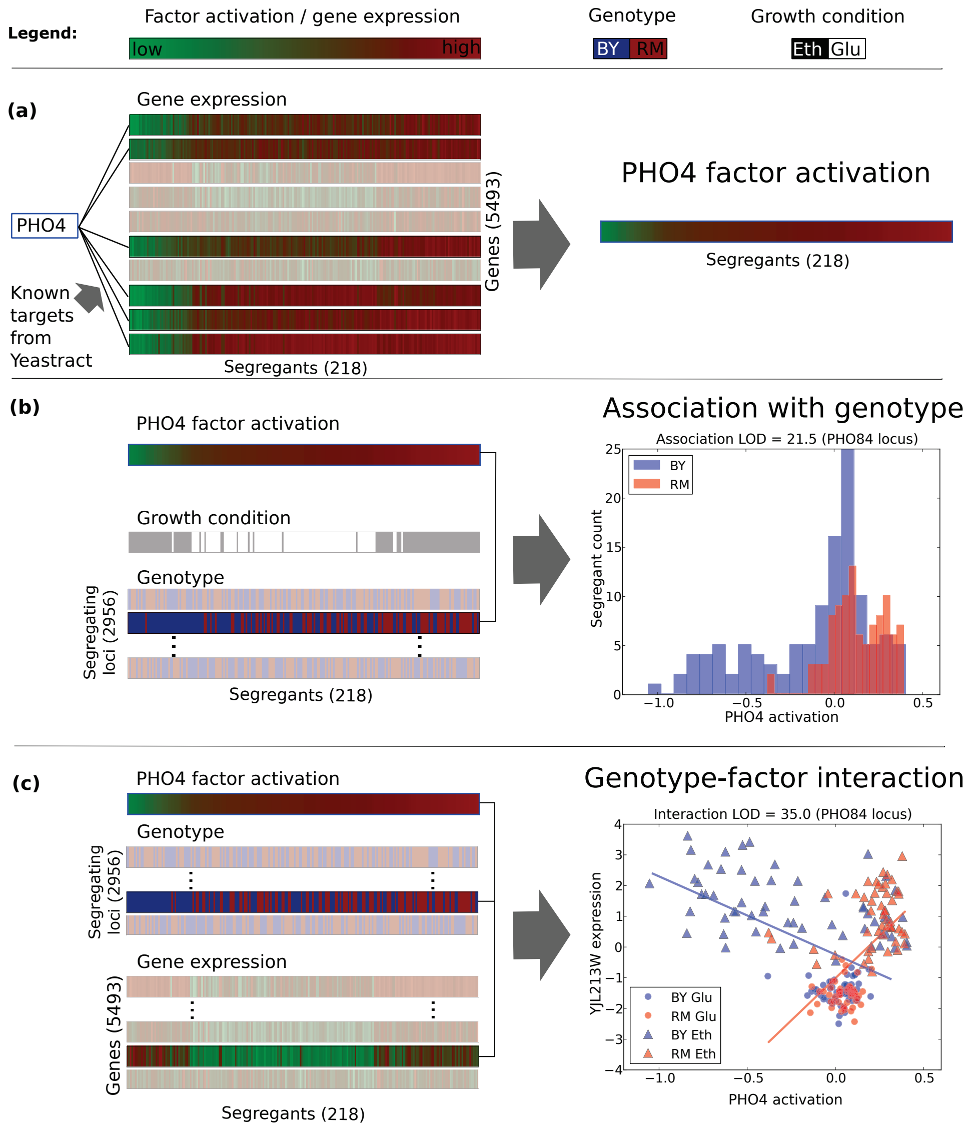


Figure 3.1: Analysing genetic effects in the context of intermediate phenotypes using *PHO4* as an example. **(a)** Intermediate phenotypes are learnt from expression levels using prior information from Yeastract database on the targets of the factor. The highlighted genes are known targets of *PHO4*. These activations are learned jointly for all factors. **(b)** The variation in intermediate phenotypes can be explained by locus genotypes or the growth condition of the segregants. For most loci (greyed out), the genotype is uncorrelated with the factor activation level. For the PHO84 locus at chrIII-46084, not greyed out and indicated by arrow, it is correlated. The plot at right shows the distribution of factor activations stratified by genotype at this locus. **(c)** Some genotypes show a statistical interaction with the inferred intermediate phenotype affecting gene expression levels, in this case YJL213W. See also Figure 3.3.

time, we consider genotype-dependent effects of the inferred intermediate phenotypes. We find genetic interactions with the inferred phenotypes that affect gene expression levels (Figure 3.1c), and identify hotspots in the genome that show an excess of these interactions. We show that many genotype-environment interactions are captured with the estimated intermediate phenotype, helping to interpret the environmental effect, and generate plausible, testable hypotheses for the mechanisms of several determined interactions. We propose that as pathway and transcription factor target annotations improve, our approach will produce even more useful intermediate traits that should be included in analysis and interpretation of high-throughput gene expression data.

3.2 Model of expression with unmeasured traits

We used a joint model of genotype and unmeasured trait effects on gene expression data, and used a two-stage inference procedure to estimate the individual effects.

3.2.1 Statistical model

The statistical model underlying our analysis assumes that the gene expression levels are influenced by effects of locus genotypes, intermediate factors, and interaction effects between them. These effects jointly influence expression variability in an additive manner, resulting in a generative model for expression $y_{g,j}$ of gene g in individual j of the form:

$$y_{g,j} = \mu_g + \underbrace{\sum_{n=1}^{N} \theta_{g,n} s_{n,j}}_{\text{SNP effect}} + \underbrace{\sum_{k=1}^{K} w_{g,k} x_{k,j}}_{\text{factor effect}} + \underbrace{\sum_{k=1}^{K} \sum_{n=1}^{N} \phi_{g,k,n} \left(s_{n,j} x_{k,j} \right)}_{\text{interaction term}} + \psi_{g,j}. \tag{3.1}$$

Here, μ_g is the mean expression level, $\psi_{g,j}$ the residual expression, and $\theta_{g,n}$ denote the weights of genotypes of SNPs $s_{n,j}$. The activations $\mathbf{x}_k = \{x_{k,1}, \ldots, x_{k,J}\}$ of K intermediate factors are modelled as unobserved latent variables that linearly influence gene g with weights $w_{g,k}$. Finally, the strength of interaction effects between factor k and SNP n is regulated by the interaction weights $\phi_{g,k,n}$.

On a second level of the model, the latent factor activations $\mathbf{x}_k$ may themselves be associated to the genetic state. Again assuming a linear model, these relations are cast as

$$x_{k,j} = \mu_k + \sum_{n=1}^{N} \underbrace{\beta_{k,n} s_{n,j}}_{\text{SNP effect}} + \epsilon_{k,j}, \tag{3.2}$$

where $\beta_{k,n}$ is the association weight and $\psi_{k,j}$ denotes the observation noise.

While appealing because of its generality, it is hard to perform joint parameter inference in the model implied by Equations (3.1) and (3.2). Here, we follow a two-step approach that yields tractable inferences and allows for statistical significance testing of the relevant factors contributing to the total gene expression variability (Equation (3.1)).

1. **Factor inference:** The latent factors $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_K\}$ and weights $\mathbf{W} = \{w_{g,k}\}$ are inferred from the expression levels alone, not taking the effects of SNPs $s_{n,j}$ via association and interaction into account.

2. **Association and interaction testing:** Significance of associations of factors to SNPs (Equation (3.2)) and SNP-gene-factor interaction terms (Equation (3.1)) are tested conditioned on the state of the inferred factors.

In this scheme, the factor inference is approximated as the contribution of direct SNP effects and interactions is not taken into account while learning. In this context, this approximation is well justified because of the relative effect sizes. The total variance explained by the interactions is small compared to the direct factor effects. If necessary on other datasets, this step-wise procedure could also be iterated, refining the state of the inferred factors given the state of associations and interactions.

3.2.2 Trait inference

Factors are inferred using a sparse Bayesian factor analysis model (Rattray et al., 2009; Stegle et al., 2009), presented here for completeness. Starting from the full model in Equation (3.1), the terms for direct genetic associations and interactions are dropped. The remaining factor model explains the expression profile

$\mathbf{y}_j = (y_{1,j}, \dots, y_{G,j})^{\mathrm{T}}$ of the G genes for segregant j by a product of activations $\mathbf{x}_j = (x_{1,j}, \dots, x_{K,j})^{\mathrm{T}}$ of the K factors, and the G times K weight matrix $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_G)$ and per-gene Gaussian noise $\boldsymbol{\psi} = (\psi_1, ...\psi_G)^{\mathrm{T}}$

$$y_{g,j} = \mathbf{w}_g \cdot \mathbf{x}_j + \psi_g. \tag{3.3}$$

The expression data $\mathbf{Y}$ is observed, and all other variables are treated as random with corresponding prior probabilities. The indicator variable $z_{g,k}$ encodes whether factor k regulates gene g ($z_{g,k} = 1$) or not ($z_{g,k} = 0$).

$$\begin{aligned} P(w_{g,k}|z_{g,k} = 0) =& \mathcal{N}(w_{g,k}\,; 0, \sigma_0) \\ P(w_{g,k}|z_{g,k} = 1) =& \mathcal{N}(w_{g,k}\,; 0, 1) \end{aligned} \tag{3.4}$$

The width σ_0 of the first Gaussian is small, driving the weight to zero. In experiments, we used $\sigma_0 = 10^{-4}$. This existing knowledge about whether a factor affects a gene, extracted from KEGG or Yeastract, is then encoded as a Bernoulli prior on the indicator variables $z_{g,k}$.

$$\pi_{g,k} = P(z_{g,k} = 1) = \left\{ \begin{array}{ll} \eta_0 & \text{no link} \\ 1 - \eta_1 & \text{link} \end{array} \right. . \tag{3.5}$$

The variable η_0 can be thought of as the false negative rate (FNR), the frequency at which prior information is incorrectly set to “no link”. Similarly, η_1 is the false positive rate (FPR) of the observed prior information. We used $\eta_0 = 0.06$ and $\eta_0 = 0.0001$ for Yeastract and KEGG factors, respectively, and $\eta_1 = 0.001$ for both. The ratio of the false positive and false negative rate is motivated by relatively high false positive rates in chromatin immunoprecipitation experiments, and confidence in the KEGG annotations.

Prior probabilities over factors $\mathbf{X}$ are standard Gaussian distributed, $x_{k,j} \sim \mathcal{N}(0, 1)$, and the per-gene noise precisions τ_g, $\psi_g \sim \mathcal{N}(0, \tau_g)$, are a priori Gamma distributed, $\tau_g \sim \mathrm{Gamma}(\tau_g \,|\, a_\tau, b_\tau)$. For the experiments this prior was set to be uninformative with $a_\tau = b_\tau = 0.001$.

Inference in the sparse factor analysis model is achieved using a hybrid of two deterministic approximations, variational learning (VB) (Jordan et al., 1999) and

Expectation Propagation (Minka, 2001), with exact details presented in (Rattray et al., 2009; Stegle et al., 2009).

Statistical identifiability of factors and model restarts

In general, factor analysis models are prone to suffering from intrinsic symmetries such as sign flips or factor permutations with impacts on the interpretability of obtained solutions. The informative sparsity prior of the factor analysis model (Equation (3.5)) substantially reduces these ambiguities, as it introduces constraints on possible factor configurations.

As an additional measure, our analysis explicitly takes the variability of factor solutions into account by analysing a set of inference solutions rather than a single point estimate. In the experiments, we performed 20 independent runs of the factor analysis model with parameters randomly initialised from their respective prior distributions, and used this whole ensemble to test for significant association and interaction effects.

3.2.3 Association and interaction testing

Following the generative model (Equation 3.1) we use standard association and interaction statistics (Lynch and Walsh, 1998) to test for associations between known variables (genotype of SNP n, environment indicator, or mRNA expression level) and the inferred factor activations. For completeness, we first present the model and test statistics used for both associations and interactions, followed by the significance testing approach. The derivation is developed explicitly using the SNP genotype as the known variable and factor activation as the dependent variable; tests for other covariates (or eQTL effects) are performed analogously.

Test statistics

We perform independent tests for association between the activation $\mathbf{x}_k$ of individual factor k and genotype $\mathbf{s}_n$ of SNP n, fitting a liner model of the form

$$x_{k,j} = \mu_k + \underbrace{\beta_{k,n} s_{n,j}}_{\text{SNP effect}} + \epsilon_{k,j}, \tag{3.6}$$

assuming Gaussian observation noise $\epsilon_{k,j} \sim \mathcal{N}(0, \sigma^2_{k,j})$. For each pair of SNP n and factor k, we calculate the association log-odds (LOD) score

$$L^a_{k,n} = \log P(\mathbf{x}_k \,|\, \beta_{k,n}) - \log P(\mathbf{x}_k \,|\, \beta_{k,n} = 0) \tag{3.7}$$

as a test statistic. The weight in the foreground model $\beta_{k,n}$, the mean μ_k and the noise level $\sigma^2_{k,n}$ are fit by maximum likelihood for every calculation.

Test statistics for the interaction terms are calculated analogously based on an independent interaction model. In short, we calculate the residuals of the factor analysis model and apply a standard interaction model between SNP n, factor k and gene g. This corresponds to the linear model

$$y_{g,j} = \mu_g + \overbrace{\underbrace{\theta_{g,n} s_{n,j}}_{\text{SNP effect}} + \underbrace{w_{g,k} x_{k,j}}_{\text{factor effect}}}^{\text{direct effects}} + \underbrace{\phi_{g,k,n}\,(s_{n,j} x_{k,j})}_{\text{interaction term}} + \underbrace{\left[\sum_{l \neq k} w_{g,l} x_{l,j}\right]}_{\text{remaining factor effect}} + \psi_{g,j}, \tag{3.8}$$

where the expression level of gene probe g for individual j is described by fitted effects of the tested SNP $s_{n,j}$, learned factor activation $x_{k,j}$ and the interaction term $s_{n,j} x_{k,j}$ with the residuals explained by 0-meaned Gaussian noise $\psi_{g,j}$. The log-odds test statistic for the interaction between factor k and SNP n to influence gene g follows as

$$L^i_{g,k,n} = \log P(\mathbf{y}_g \,|\, \phi_{g,k,n}) - \log P(\mathbf{y}_g \,|\, \phi_{g,k,n} = 0). \tag{3.9}$$

The respective mean variable μ_g, weights $\theta_{g,n}$, $w_{g,k}$ (but not $w_{g,k'}$ where $k' \neq k$), and $\phi_{g,k,n}$, as well as noise variance $\psi_{g,j}$ are fitted independently using maximum likelihood for each factor, gene, SNP triplet. The contribution from all remaining factors is not refit to preserve the sparsity pattern learnt from the factor inference. To reduce the number of effective tests, we used the strongest interaction LOD score $\hat{L}^i_{g,n} = \max_k L^i_{g,k,n}$ across factors, thus performing tests for every SNP and gene pair. This approach corresponds to the assumption that at most a single factor is interacting with a given gene-SNP pair. The consistency of the strongest interacting factor is informative of the identifiability of the interaction effect (see below).

Incorporating several random initialisations

For all our analysis of intermediate phenotypes, we generated factor inference results from $R = 20$ random initialisations of the model parameters to capture the variability in the model and avoid overfitting of inferred factor activations to local optima (See Statistical identifiability of factors below). Thus, we designed a significance testing scheme based on Q-values (Storey and Tibshirani, 2003) that employs the full set of runs, taking the uncertainty in the factor posterior distributions into account. We present this approach below for associations. Testing interactions is analogous except for the specifics of permutations highlighted in the text. In case of analyses where the multiple restarts are not used (e.g. eQTLs), we calculated Q-values from the single instance. In all cases, the null distribution of LOD scores was obtained by combining all calculated null statistics in the random restart.

Q-value calculation For every run $r = 1, \ldots, R$ of the factor analysis model, we evaluated the test statistics of factor associations ($L^a_{k,n}$) for every pair of factor k and SNP s. This analysis was then repeated on 20 permuted datasets in each run with the genotypes shuffled with respect to the factor activations, while keeping individual segregants grown in two conditions paired. For interaction LOD scores, the factor activations and gene expression levels were not permuted with respect to each other. From this empirical null distribution of LOD scores in run r (across all SNPs and factors), we calculated Q-values $q^r_{n,r}$ (local FDR) for each candidate association (Storey and Tibshirani, 2003) between SNP n and inferred posterior of factor k in this run.

Combining Q-values The Q-values from all runs were then combined into an overall Q-value $q_{k,n} = R^{-1} \sum_{r=1}^{R} q^r_{k,n}$, which was used to assess significance at a given FDR threshold.

From a probabilistic viewpoint, averaging Q-values over multiple restarts of the model can intuitively be thought of as integrating out the uncertainty from the factor inference. For example, for an association test assessing the significance of the weight $\beta_{k,n}$, we are truly interested in the probability of an association being absent (Bayesian Q-value, see for example (Storey, 2003)) given uncertain inference of the factor activation $P(\mathbf{x}_k \,|\, \mathbf{Y}, \boldsymbol{\pi})$. Conditioned on the observed data $\mathbf{Y}$ and prior $\boldsymbol{\pi}$ this probability follows as

$$P(\beta_{k,n} = 0 \,|\, \mathbf{Y}, \boldsymbol{\pi}, \mathbf{s}_n) = \int_{\mathbf{x}_k} P(\beta_{k,n} = 0 \,|\, \mathbf{x}_k, \mathbf{s}_n) P(\mathbf{x}_k \,|\, \mathbf{Y}, \boldsymbol{\pi}). \tag{3.10}$$

In general this integral is not analytically tractable. Assuming we have instead a number of R samples $\mathbf{x}_k^r$ from the factor posterior, the integral can be approximated by

$$\approx \frac{1}{R} \sum_{r=1}^{R} P(\beta_{k,n} = 0 \,|\, \mathbf{x}_k^r, \mathbf{s}_n) \tag{3.11}$$

in a Monte Carlo fashion. Finally, identifying the null probabilities as Bayesian Q-values we get

$$= \frac{1}{R} \sum_{r=1}^{R} q_{k,n}^r. \tag{3.12}$$

Note that the restarts from the factor analysis model are not exactly samples from its posterior but nevertheless characterise the posterior uncertainty sufficiently well (See also Simulation study below). Full MCMC sampling is computationally infeasible due to the size of the regulatory network; for a comparison of MCMC sampling and deterministic inference as employed here, see Stegle et al. (2009).

3.3 Phenotype inference

We inferred intermediate phenotypes on expression levels of 5493 genes from 109 yeast segregants grown in two environmental conditions (Chapter 3.3.2, Smith and Kruglyak (2008)). We performed the inference 20 times with different random initialisations of the parameters.

We considered three alternative types of prior information. First, we hypothesised the factors to be transcription factor activation levels, and used data for 167 transcription factors from Yeastract (Teixeira et al., 2006) to assign a prior probability of a factor affecting a gene expression level. Second, we hypothesised the factors to be pathway activations, and used KEGG database information (Kanehisa et al., 2002) for 63 pathways for the prior probability of a link between a pathway activation and a gene. Third, for comparison, we employed an uninformative prior, where 30 factors were *a priori* equally likely to affect all genes. The datasets are described in more detail in Chapter 3.3.2 We call the inferred factor activations Yeastract factors, KEGG factors, and freeform factors, respectively.

3.3.1 Factor analysis model performance

In-depth comparison of inference approaches for the sparse factor analysis model used is given in other work (Stegle et al., 2009); the model was found to accurately recover factor activations in a setup similar to this study.

One way to further assess the reproducibility of the factor inference is to consider the correlation between the posterior means of individual factor activations. We called the inferred activation of factor k in u-th run $\mathbf{x}_k^u = (x_{k,1}^u, ..., x_{k,J}^u)$ reproducible if its Pearson correlation $\rho(\mathbf{x}_k^u, \mathbf{x}_k^v) > 0.7$ for at least 16 of the 20 different v. 72 of 167 (31%) Yeastract and 19 of 63 (30%) KEGG factors were reproducible. While we explicitly took the variability between runs into account in further analyses, these numbers are instructive for developing intuition about the model.

3.3.2 Datasets used

For completeness, we provide specific details of the datasets used.

Gene expression data from (Smith and Kruglyak, 2008) (GEO accession number GSE9376) was downloaded using PUMAdb (http://puma.princeton.edu). In line with (Smith and Kruglyak, 2008), we considered spots good data if the intensity was well above background and the feature was not a nonuniformity outlier. Transcripts with more than 20% of missing values were discarded. All other missing expression values were replaced with the averages across the corresponding growth condition.

The remaining expression data consisted of 5493 probe measurements for 109 crosses of BY (laboratory) and RM (wild) strains grown in either glucose or ethanol, resulting in a total of 218 individuals. Strain genotypes were kindly provided by R. Brem. Each of the 109 segregant strains was genotyped at 2956 loci to give a crude map of genetic background.

Transcription factor binding data was downloaded from Yeastract (Teixeira et al., 2006) (Version 1.1438) and contained binary indicators of binding between 174 transcription factors and 5914 genes. We considered the 3000 most variable probes whose corresponding genes were included in the binding matrix, and transcription factors that influenced at least 5 genes. After further discarding probes for which there were no data available, the remaining Yeastract prior dataset consisted of binding data for 167 transcription factors affecting 2941 genes.

Similarly, pathway information was downloaded from the KEGG database (Kanehisa et al., 2002). Only pathways with at least 5 genes were included in the network prior. This filtering procedure retained 63 pathways controlling 1263 genes. The results of Smith and Kruglyak (2008) were not used as a source of information for either of the prior datasets.

3.4 Association analysis with inferred phenotypes

First, we looked for the causes and consequences of variability in the inferred intermediate phenotypes.

3.4.1 Genotype and environment

Although the factors were inferred jointly from the expression data alone, many factor activations were significantly associated with a locus (SNP) genotype or indicator variable encoding growth in ethanol or glucose as a carbon source ("environment", Tables B.9 to B.11). Thirty two Yeastract factors were associated with a SNP genotype at false discovery rate (FDR) less than 5% and 26 with the environment. Similarly, 7 KEGG factors were associated with a SNP genotype, and one with the environment while 24 freeform factors were significantly associated with a SNP genotype and one with the environment. Some of the genotype associations were due to pleiotropic effects of single loci, while others were private to a locus-factor combination (Tables B.12 to B.14).

Many of these individual associations to Yeastract and KEGG factors can be interpreted by considering the role of the inferred factors and functional annotations of genes at associated loci. We now give some examples to further corroborate the use of factor activations as intermediate phenotypes. All associations are significant at 5% FDR, with corresponding Q-values q (minimal FDR for which the association is significant (Storey and Tibshirani, 2003)) and average log-odds scores L given.

Yeastract factors.

Loci associated with Yeastract factor activations encode genes functionally related to the corresponding transcription factor. The *PHO84* (an inorganic phosphate transporter) locus was associated with the *PHO4* (a major regulator of phosphate-responsive genes) transcription factor activation ($q < 0.03, L = 15.5$). The association implicates genetic variation in the transporter as a determinant of the transciptional activation of phosphate-responsive genes through *PHO4* activation. The mechanism of action is likely a switch in transcriptional response when *PHO84*, a high affinity phosphate transporter, is rendered ineffective by a mutation (Wykoff et al., 2007).

The *SUM1* (transcriptional repressor of middle sporulation-specific genes) factor activation was associated with the genotype of the *RFM1* (repression factor of middle sporulation) locus ($q < 10^{-5}, L = 115.2$). This is intriguing since *RFM1* recruits the *HST1* histone deacetylase to some of the promoters regulated by *SUM1* (McCord et al., 2003; SGD project), suggesting that genetic variation in the *RFM1* gene indirectly alters the effect of *SUM1* on individual genes.

There is also a straightforward eQTL that regulates the *HAP1* (heme activation protein) gene expression ($q < 10^{-5}, L = 80.6$), as well as factor activation ($q < 10^{-5}, L = 38.7$). This is a *cis* effect, since the locus is proximal to the gene, and manifests itself as a *trans* eQTL hotspot by affecting expression levels of some of the 170 known *HAP1* targets. Thirty four of the 84 (40%) significant *trans* eQTLs are also known targets of *HAP1*. Our data suggest that the other 50 may either be previously undiscovered targets of *HAP1*, or downstream effects of some of its direct targets.

The *THI2* thiamine metabolism transcription factor activation was associated with the genotype of the *THI5* locus ($q < 10^{-5}, L = 52.2$). This suggests a regulatory role of *THI5* upstream of *THI2* in thiamine biosynthesis for the previously poorly characterised *THI5* gene. This illustrates how our inference allows generating hypotheses for the function for genes that are implicated in a cellular pathway, but not annotated with a specific role.

KEGG factors.

Associations to KEGG pathways tend to capture the effect of a pathway component genotype. For example, two amino acid metabolism pathways are associated with locus genotypes of genes in the pathway. The inferred activation of lysine biosynthesis pathway was associated with genetic variation in the *LYS2* locus ($q < 10^{-4}, L = 25.6$), and the activation of arginine and proline metabolism pathway with the *ARG8* locus ($q < 10^{-5}, L = 46.7$), both members of the respective pathways. We thus hypothesise that variants in these genes directly affect the activation of the corresponding pathways. Also, the nitrogen metabolism pathway was associated with the *ASP3* (cell-wall L-asparaginase) gene cluster locus genotype. ($q < 10^{-5}, L = 119.9$). The *ASP3* genes are part of the pathway, and are present in four copies in the reference strain S288c, conferring increased resistance to nitrogen starvation stress. The inferred state of the pathway thus likely corresponds to the *ASP3* copy number via the locus genotype proxy.

Furthermore, the fatty acid metabolism pathway activation was associated with the *OAF1* (oleate-activated transcription factor) locus genotype ($q < 0.01$, $L = 67.1$), which is a known regulator of the pathway (Smith et al., 2007). We thus hypothesise that genetic variants in *OAF1* between the two strains are responsible for differences in fatty acid metabolism in the segregants, as has also been proposed in earlier work (Lee et al., 2009).

Finally, the environment is strongly associated to the very wide metabolic pathways category ($q < 10^{-5}, L = 393.2$). This KEGG entry comprises 619 genes, and captures the effect of the growth condition of the segregants on their metabolic state.

Freeform factors.

The freeform factors capture broad variance components in the data, with each factor's activation contributing to very many probe expression levels. Regardless of the unsupervised inference of the activations, they still show strong associations to environment and locus genotypes. However, due to this global nature of the factors, the associations are less straightforwardly amenable to interpretation. The first factor is associated with the environment ($q < 10^{-5}, L = 289.5$), and accounts for any mean shifts in gene expression levels between segregants grown in glucose and ethanol (Table B.11). Several of the other factors are associated with genotypes of "pivotal loci" described before (Brem and Kruglyak, 2005; Smith and Kruglyak, 2008; Yvert et al., 2003). It may be possible to make suggestions about the functionality via methods such as overrepresentation of GO categories within sets of genes with large weights for a factor, such as a recent study that performed a similar association analysis with unsupervised factors (Biswas et al., 2008). Our approach of using existing data for guidance is stronger compared to unsupervised methods as we use evidence of which gene is affected by the factor, thus improving statistical identifiability, and do not rely on an *ad hoc* choice of number of factors. This yields interpretable results that are more useful for generating hypotheses for the consequence of genetic or environmental variation.

Response to small molecule stress has been measured in the same segregants to map drug response loci (Perlstein et al., 2007). This study found eight QTL hotspots, six of which are within 20kb of loci that also show several associations to our inferred intermediate phenotypes (Tables B.12 to B.14), corroborating their pleiotropic effect.

3.4.2 mRNA and protein levels

Twenty five of 167 Yeastract factors were associated with the probe expression level measuring the mRNA abundance of the corresponding transcription factor gene (Table B.9, Figure 3.2). Twenty of the 25 (80%) were also significantly associated with a SNP genotype or environment. While statistically significant, these associations do not explain the majority of the factor variability, as only four

Yeastract factors were correlated with their probe expression level with Pearson $r^2 > 0.5$.

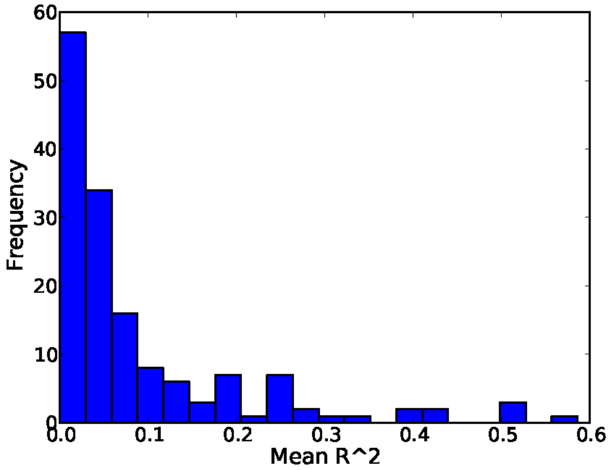


Figure 3.2: Pearson's correlation of Yeastract factors and their corresponding probe expression levels.

The general lack of correlation between factor activation and the corresponding measured expression level for the remaining transcription factors is perhaps not surprising. Presumably what matters for the factor activation is protein activity level, not mRNA abundance. Previous studies have found poor correlation between mRNA and protein expression levels (Foss et al., 2007; Gygi et al., 1999). Also, alternative mechanisms for activation exist. Many Yeastract factors without significant correlation to transcript levels have been shown to be activated not via increase in expression, but other means. For example, *PHO4* is activated by multiple phosphorylation events (Komeili and O'Shea, 1999). Simlarly, nuclear localisation and therefore activation of *ACE2* and *MSN2* are controlled by phosphorylation state (Goerner et al., 1998; O'Conallain et al., 1999). We predict most of the other transcription factors to also be activated by non-transcriptional means.

The protein level of one of the Yeastract factors, *GIS2*, has been assayed quantitatively in a previous study (Foss et al., 2007) for 87 of the 109 segregants we considered in a similar growth condition. For this transcription factor, the inferred factor activation was better correlated to the protein level than the corresponding probe expression level for 15 of the 20 random initialisations. This example gives further support to treating the inferred factors as meaningful intermediate quantitative traits.

3.4.3 eQTL hotspots

As observed before (Brem et al., 2002; Smith and Kruglyak, 2008; Yvert et al., 2003) some segregating loci show significant associations with up to 271 (*IRA2*, regulator of the RAS-cAMP pathway locus) probe expression levels, forming *trans* eQTL hotspots. There are five such hotspots with at least 30 associations each. On average, 32% of the genes associated with a *trans* eQTL hotspot (FDR<5%) are explained by a transcription factor associated with the hotspot locus genotype targeting the gene (Table B.15). In 94% of these cases, the association with the inferred factor activation is stronger than with the locus genotype, and for three of the five hotspots, many additional associations with factor targets are recovered. For example, the *PHO84* locus is associated with the *PHO4* Yeastract factor activation ($q < 0.03, L = 15.5$), as well as 31 probe expression levels in *trans*. Eleven of these are also significantly associated with the *PHO4* factor activation, all showing a stronger association. *PHO4* itself is significantly associated with 454 probes, greatly expanding the range of plausible effects of the *PHO84* locus. This shows that using inferred intermediate phenotypes can reveal additional associations that otherwise would not be statistically significant.

3.5 Interaction analysis with inferred phenotypes

Beyond understanding the causes of variability in the inferred traits, we are also interested in their genotype-dependent effects on gene expression levels.

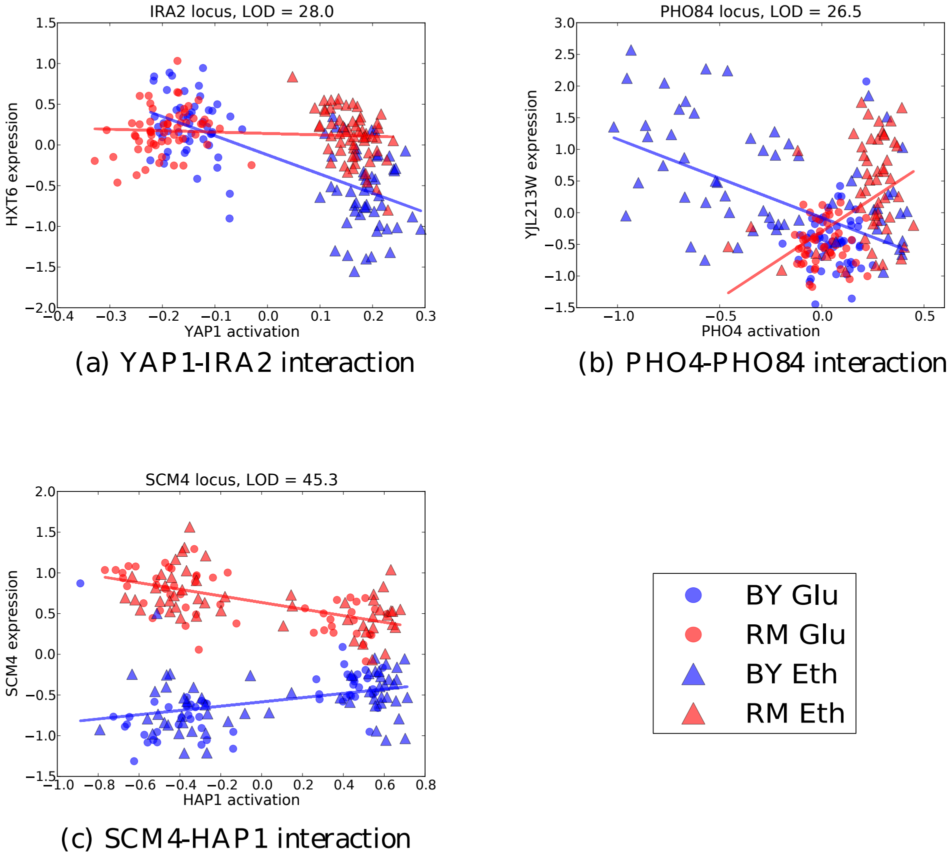
3.5.1 Discovering interactions

We scanned the genome for genotype-factor interactions that affect gene expression levels (Figure 3.1c) using a standard linear interaction model, and recovered three broad classes of interactions (Figure 3.3). We tested each locus-gene pair independently for interaction with any inferred factor using 20 permutations, and information from all the random restarts of the model. If a single factor was observed with the strongest interaction score for a locus-gene pair in at least half the multiple restarts, we interpreted it as the true interacting factor; in other cases, we did not designate a factor to an interaction effect. We give examples of interactions we find below, highlighting how they add to the understanding of the propagation of the genotype effect.

The largest set of interactions was found at the *IRA2* locus. Many Yeastract factors, such as *MIG1*, *HAP4*, *YAP1* and *MSN2* show high interaction LOD scores with this locus (Figure 3.3a). All these corresponding transcription factors act in glucose response, nutrient limitation or stress conditions, which is consistent with the role of *IRA2* in environmental stress response by mediating cAMP levels in the cell. Their factor activations are associated with the environment (Table B.9), and the interactions thus recapitulate gene-environment interactions. While all these factor activations are correlated due to the strong association with the environment, making it hard to identify the true interacting factor, we can narrow the factor down to a few that exhibit strong LOD scores. Identifiability of the interacting factor is hard in general for factors that capture large effects, or have target sets that largely overlap with other factors. The inferred factors do capture the true underlying sources of variability, which is even more useful in settings where not all sources of variability are measured. Also, even having measured the relevant growth condition, we can further interpret the interactions as transcription factor activation having an effect in a specific genetic background in some cases, a more specific claim.

The *PHO4* factor activation is associated with ($q < 0.03, L = 15.5$) and interacts with the *PHO84* locus on chromosome XIII to influence 245 genes (Figure 3.3b). At the same time, the activation also interacts with the environment variable to influence gene expression levels. Notably, the statistical interaction for the *PHO4* expression, *PHO84* genotype and the same gene expression levels also has LOD scores greater than 11. Thus these interactions are not artifactual, but can be traced back to measured quantities for all interacting variables.

We also recovered epistatic interactions that failed the stringent multiple testing criteria on their own, but showed a stronger signal via the intermediate fac-



(a) YAP1-IRA2 interaction

(b) PHO4-PHO84 interaction

(c) SCM4-HAP1 interaction

Figure 3.3: Three broad classes of interaction effects between locus genotype and transcription factor activation affecting gene expression (for details see text). Each marker shows the gene expression and factor activation for one individual segregant of either BY (blue) and RM (red) background at the locus, and grown in ethanol (triangles) or glucose (circles) as a carbon source. Maximum likelihood fits for expression data for the BY and RM segregants are plotted as solid lines; an interaction effect corresponds to a difference in slope in the two genetic backgrounds. **(a)** Genotype-environment interaction mediated by the inferred *YAP1* transcription factor activation. **(b)** Interaction between the *PHO84* locus and *PHO4* transcription factor activation, which is associated both with the *PHO84* locus genotype and the *PHO4* probe expression level. **(c)** Epistatic interaction between HAP1 and its target, SCM4, mediated by the HAP1 activation.

tor. For example, *HAP1* factor activation interacts with ($q < 0.01, L = 38.6$) the *SCM4* (suppressor of *CDC4* mutation) locus genotype to influence *SCM4* expression level (Figure 3.3c), while the epistatic interaction LOD score is only 7.9. As *SCM4* has a *HAP1* binding site in its promoter region, it is plausible that genetic variants could directly inhibit *HAP1* binding. This effect would only be observable in case *HAP1* is active, which in turn is controlled by the *HAP1* locus genotype ($q < 10^{-5}, L = 38.7$). This is an example of an epistatic interaction that is mediated by an intermediate phenotype of transcription factor activity.

In total, we found 2,397 genes with a gene-Yeastract factor interaction effect ($q < 0.05$). We also found 2,211 genes that show genetic interactions with KEGG factors and 2,250 with freeform factors. We noted several interaction "peaks" in the genome, such as the *IRA2* locus, where the locus genotype interacts with several genes via one or multiple factors (Figure 3.4). These coincide with *trans* eQTL peaks and gene-environment interaction peaks observed before (Smith and Kruglyak, 2008; Yvert et al., 2003), and have been annotated for potential causal genes.

3.5.2 Recovering interactions

We found 10,049 locus-environment interactions affecting 676 gene expression levels (Figure 3.4) using the same model and testing approach as for inferred factor interactions (FDR < 5%). Of these, we recovered 4605 interactions (46%) affecting 505 genes (75%) with the Yeastract factors, 6464 interactions (64%) affecting 572 genes (85%) with the KEGG factors, and 3065 interactions (31%) affecting 420 genes (62%) with the freeform factors. All environment-associated Yeastract factors had a strong interaction LOD scores with the *IRA2* locus, affecting hundreds of genes. These interactions recapitulate the gene-environment interaction reported and validated in the original analysis of the data (Smith and Kruglyak, 2008). It is reassuring that we are able to recover these interactions with the inferred intermediate phenotypes, and to expand their repertoire as well as provide hypotheses for their mechanism.

Preliminary results from an ongoing screen for gene-gene interactions have shown epistatic interactions for 95,445 gene pairs (Costanzo et al., 2010). Three

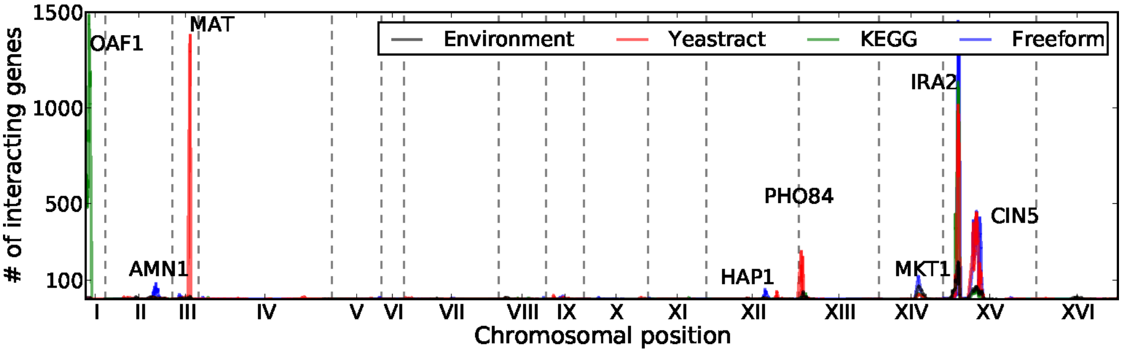


Figure 3.4: Number of genes affected by a genotype-factor interaction for each locus for Yeastract factors (blue), KEGG factors (red), freeform factors (green), and environment (gray).

hundred and sixty eight knockouts of a Yeastract factor gene and an interaction peak gene were tested in this large-scale assay, with 40 epistatic interactions found. We could find interactions for 22 of the tested pairs, and recovered one of the 40 interactions of Costanzo et al. (2010). Our screen is for genetic interactions that are different from the synthetic lethal screen of Costanzo et al. Consistent with this, we find some overlap, but not more or less than expected by chance.

3.6 Discussion

Our genetic analysis of the gene expression data from (Smith and Kruglyak, 2008) has shown that inferred intermediate phenotypes are valuable for generating hypotheses about plausible connections between genetic and gene expression variation. Using these inferred cellular phenotypes, we identified loci associated with transcription factor and pathway activations, thus giving the genetic effect a straightforward mechanistic interpretation, and often suggesting a candidate gene responsible for the change. For the first time, we considered and found statistical interaction effects with inferred intermediate phenotypes.

Our work is a step towards interpreting and understanding effects of genetic variants by putting them into cellular context. Conventional analysis, relating genotype and expression levels, is restricted to observed measurements, often producing only statistical associations instead of a plausible mechanistic view. Going beyond this, our approach yields phenotypic variables at an intermediate level which can be used in the analysis. We showed that these provide additional interpretability and in some settings increase statistical power. Besides standard association and interaction effects between genotype and gene expression, our approach allows more rich hypothesis spaces to be explored, where the dependent variable we model is not a global organism phenotype such as disease label, or a very specific measurement like a single gene expression level. We have shown that this analysis is both feasible, and gives interesting results.

The idea of looking for associations and interactions with inferred intermediate phenotypes will be even more useful in forthcoming studies that include other cellular measurements. The inferred transcription factor or pathway activations allow interpreting the variability in these measured phenotypes as a result of changes in regulator activity or pathway state, bridging the gap between individual molecule measurements, and states of protein complexes, cellular machines, and pathways. We believe that the inferred intermediate phenotypes can be much more informative about the state of the cell and organism than individual locus genotypes and gene expression levels, and will also show stronger associations to downstream cellular and tissue phenotypes.

The intermediate activation phenotype has lower dimensionality compared to the space of genotypes and gene expression levels, which helps against multiple testing issues present in genome-wide scans for epistatic interactions. We were able to infer association and interaction effects, including proxies for epistasis, while finding epistatic interactions by testing all locus pairs is usually hindered by the billions of tests performed (Brem and Kruglyak, 2005; Cordell, 2009; Storey et al., 2005). The incorporation of prior information to infer interpretable factors is a flexible way to reduce the number of tests by capturing relevant parts of the data variation in a few factors, and can also add power if the factor is a better proxy for the true interacting variable.

The inferred transcription factor activations did not mostly correlate with their expression level. This is expected, as the activity of a protein depends on the protein level, localisation, posttranslational modification state, and existence of binding partners to carry out its function. Expression level alone is often a poor proxy for a measure of protein activity.

A range of prior work has applied linear or generalised linear models to infer unobserved determinants of gene expression levels. For example, broad hidden factors have been inferred from gene expression that are likely to be due to confounding sources and hence can safely be explained away, thereby increasing the power of eQTL studies (Leek and Storey, 2007; Stegle et al., 2010). Although methodologically related, this work has a completely different aim. Also, unsupervised sparse linear models have been applied to infer hidden determinants in gene expression which are subsequently analysed for association to the genetic state (Biswas et al., 2008). This approach is closely related to the "freeform factors" included in this analysis for comparison. Overall, we show that factor learning taking prior knowledge into account adds statistical identifiability of the actual factors thereby providing interpretability. Other interesting approaches perform feature selection to capture relevant properties of the segregating sites in order to pinpoint the causative allele (Lee et al., 2009), or build a predictive (network) model of gene expression, followed by analysing its cliques and subnetworks (Zhu et al., 2008), but neither explicitly model unobserved phenotypes. A very recent paper proposed an integrated Bayesian ANOVA model that explains the gene expression profile by modules (Zhang et al., 2010). These modules in turn are modelled as a function of the genotype, taking direct and epistatic regulation into account. Importantly, both these related approaches infer gene expression determinants in an unsupervised fashion, and hence the interpretation of these association signals can be difficult and remains as a retrospective analysis step. Finally, a methodologically related sparse factor analysis model employing prior information has been applied to a narrower dataset with an aim to explain *trans* eQTL hotspots (Sun et al., 2007). However, the study does not consider the idea of genetic effects in the phenotypic context, or look for interaction effects, which is a primary focus of this work.

There has been speculation that a significant proportion of heritable variability that cannot be attributed to associations with single loci is due to interaction effects. This hypothesis is intuitively appealing, since we expect some genetic variants only to have an effect in a specific context. We have found an abundance of such statistical interactions, and have shown how some of them help to understand and interpret yeast gene expression regulation. Often, they recapitulate epistatic or gene-environment interactions, but nevertheless add a plausible mechanism of action. It will be especially interesting and important to see how these methods work on large, extensively genotyped and phenotyped human cohorts that are becoming available in the near future.

An open source Python implementation of the statistical models and the analysis pipeline is available from ftp://ftp.sanger.ac.uk/pub/rd/PEER

Chapter 4

Genetic mapping using artificial selection

Collaboration note

This chapter contains work performed in collaboration with many people, most notably Dr. Gianni Liti and Francisco Cubillos. The results have been submitted for publication (Parts et al., 2010), and the manuscript forms the backbone of this chapter. I am including brief validation results from some of their experiments for completeness, detailed acknowledgements are given below.

I conceived and developed the project with Gianni. Gianni and Richard Durbin designed the intercross approach. Gianni, Francisco, and Kanika Jain performed the genotyping, crossing, selection, and validation experiments. Michael Quail prepared the sequencing libraries. Jared Simpson assembled the parental strains. Jonas Warringer performed the phenotyping. I analyzed the sequencing and genotyping data. Amin Zia and Alan Moses performed individual allele analysis.

One approach to understanding the genetic basis of traits is to study their pattern of inheritance among offspring of phenotypically different strains (Mackay et al., 2009; Nordborg and Weigel, 2008; Rockman, 2008). Previously, such analysis has been limited by low mapping resolution, high labour costs, and large sample size requirements for detecting modest effects. We present a novel approach to map trait loci using artificial selection. We subject a large pool of haploid or diploid twelfth generation progeny between two budding yeast strains to heat stress for extended time. Sequencing total DNA from the pool before and during selection reveals the genetic architecture of heat resistance in this cross. Many regions, some contained within a single gene, change in allele frequency, show evidence of negative epistatic interactions, and exhibit dominant, recessive, and additive effects.

4.1 Trait mapping with natural genetic variation

A central challenge of modern genetics is to identify genes and pathways responsible for variation in quantitative traits. In the last decade, efforts of large international collaborations have revealed numerous loci that influence disease risk in humans by genotyping and phenotyping very large cohorts of individuals (Chapter 1.1.2). However, the effects of single alleles are generally modest, and explain only a small proportion of the heritable variability. Studies in model organisms, where causality can be addressed by reverse genetic tools, can help understand the genetic complexity of such traits (Chapter 1.1.1).

4.1.1 Shortcomings of existing approaches

Mapping the effect of naturally occurring genetic variation on traits is not straightforward even in model organisms. Designed crosses often use manipulated laboratory strains (Ehrenreich et al., 2009), and produce segregants that have to be laboriously genotyped and phenotyped. It is also costly to develop and maintain outbred populations of sufficient size (Valdar et al., 2006). Recently, analysis of a very large pool of recombinant yeast strains has been used to identify quantitative trait loci (QTLs) for multiple traits (Ehrenreich et al., 2010; Segrè et al., 2006; Wenger et al., 2010) without characterizing individual segregants.

4.1.2 Leveraging artificial selection

While Ehrenreich et al. (2010) found many QTL regions, the problem of finding all responsible loci, and localising the trait genes within QTL peaks remains. Furthermore, such analyses in yeast have previously been limited to haploid samples. Here, we present a precise and sensitive approach to QTL mapping, extending the approach of Ehrenreich et al. (2010), and identify trait loci and genes in both haploid and diploid populations. We used a three step process (Figure 4.1). First, we generated intercross lines between two phenotypically different yeast strains. We then applied selective pressure to the pool by growing it in a restrictive condition (40°C heat or 400 μg/ml paraquat) to enrich for individuals with beneficial alleles. Finally, we sequenced the pool before and at multiple timepoints during selection to directly assess the changes in population allele frequencies throughout the genome.

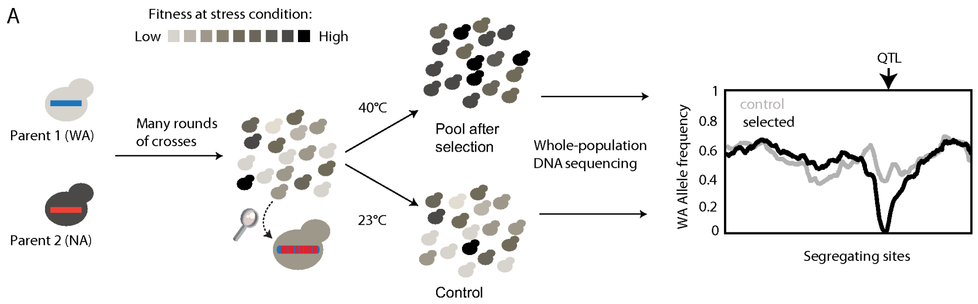


Figure 4.1: A 3-step QTL mapping strategy by crossing two phenotypically different strains for many generations to create a large segregating pool of individuals of various fitness, and growing the pool in a restrictive condition that enriches for beneficial alleles that can be detected via sequencing total DNA from the pool.

Methods used throughout the chapter are outlined in Chapter 4.4. The technical aspects of the experimental approaches designed and performed by collaborators are available elsewhere (Parts et al., 2010).

4.2 Very large segregating yeast population

We generated up to 12 generations of advanced intercross lines (F12 AILs, 12 generations of random mating between YPS128, a North American oak tree bark (NA) strain, and DVGBP6044, a West African palm wine (WA) strain. The resulting haploid or heterozygous diploid intercross pools consisted of 10-100 million random segregants, with a segregating site every 170 bases on average.

We sought to characterise the properties of this mapping pool to assess the increase in number of recombinations, and confirm that the alleles present in the parents are still segregating after many generations of intercross.

4.2.1 Recombination rate

Using many rounds of crosses should expand the genetic map due to reduction of linkage between nearby loci (Figure 4.2, Darvasi and Soller (1995)). To confirm this, we genotyped 30 markers in 96 individual segregants from each of three generations in three regions to assess the change in recombination fraction between adjacent markers.

The genetic distance (measured in 100 times the average number of recombination events) between two chromosome XIII loci separated by 204kb increased from 88 in F1 to 125 in F6 and 180 in F12. We further sequenced two segregants from the F6 pool at low coverage and observed 64 and 68 recombination events, a 125% increase compared to an average of 30 events detected in 96 F1 segregants (Figure 4.2b, Cubillos et al. (2011)).

We observed fewer recombination events than expected if an independent set of crossovers occurred every generation. There are several explanations to this. First, it is known that the recombination rate is not uniform, but accentuated in specific regions (recombination hotspots, Tsai et al. (2010)). Therefore, multiple recombinations can occur at the same site, leading to underdetection of recombination events. Second, it is possible that there is recombination preference in the heterozygous diploids with homozygous regions. We are not able to detect such events. Finally, we have conservatively filtered out very closely spaced events (2kb, Chapter 4.4) as well as subtelomeric events, introducing a further bias.

Some of these issues can be addressed by using a different cross with higher recombination frequency, or mutant strains that exhibit alternative recombination patterns.

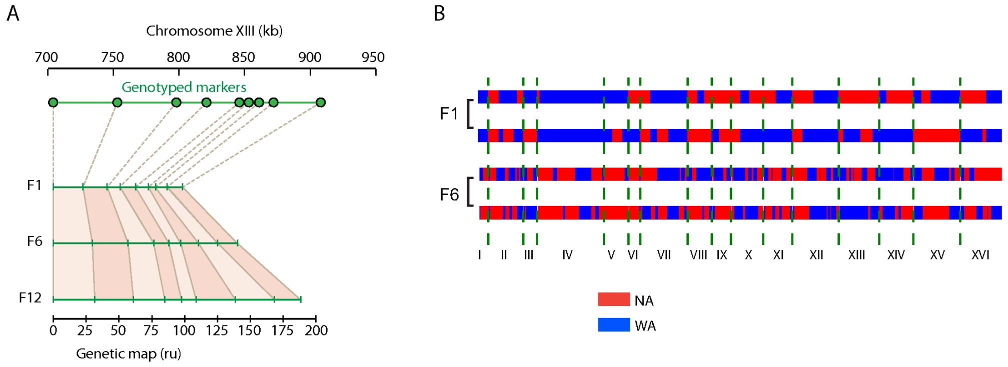


Figure 4.2: Recombination landscape after mulptiple rounds of intercrosses. a) Expansion of the genetic map, measured in recombination units (ru) of 100 times the average number of recombination events from first to twelfth generation (bottom of panel) of a 200-kb chromosome XIII locus genotyped at 9 markers (top of panel). b) Genetic background of two segregants from first (F1) and 6th (F6) generation cross shows a sharp increase in recombination events.

4.2.2 Parental allele frequency

Sequencing total DNA from pools before selection shows that more than 99% of the mappable genome is segregating in the F6 generation with minor allele frequency greater than 10%, and 97% in the F12 generation (Figure 4.3a). A small fraction of the genome is strongly selected for during the intercross rounds, due to alleles favoring sporulation, mating, or resistance to selection steps used in the cross (Chapter 4.4). This allowed us to map 6 regions responsible for these traits as a byproduct of our approach (Table B.17).

4.3 Mapping trait loci using selection

After establishing a large segregating population, we employed it for trait mapping.

4.3.1 QTLs in haploid pool

We sequenced DNA from the F12 haploid pool to an average genome coverage of 25x to 150x (Table B.16) after 0 (T0), 96 (T1), 192 (T2) and 288 (T3, 144 generations) hours of growth at 40°C. There were 19 regions where the inferred allele frequency of the T2 pool changed by at least 10% in each of two biological replicates compared to both the initial pool and the control experiment (Figure 4.3a-c, Table B.18). The NA version of the locus was selected in about two thirds (12/19) of the cases, consistent with it being the more heat resistant strain(Cubillos et al., 2011), however, several antagonistic WA alleles were also selected for. These changes are specific to the heat stress condition, as the same pool exposed to oxidative stress (paraquat, 1.5 mM) yielded a different set of QTLs (not shown). In addition, all the mitochondrial genes were greatly reduced in copy number upon selection (Table B.19). On further testing, 189/189 segregants from F12 selected pool exhibited the petite phenotype when grown in non-fermentable carbon source (glycerol and ethanol), indicating loss of mitochondrial genome.

4.3.2 QTL validation

We validated three of the mapped QTLs. Conventional linkage analysis of 96 F1 segregants also indicated a strong QTL at the right end of chromosome XIII (variance explained 66%). No other strong QTLs were seen in this cross. This region corresponds to the most rapid change in allele frequency with the NA allele fixing early in the selection at T1 (Figure 4.3d). The only other QTL that reached fixation was the GTPase activating protein *IRA1*, a negative regulator of the RAS signalling pathway. Interestingly, three additional genes of the same pathway (*IRA2*, *GPB1* and *GPB2*), as well as some of its targets (*CDC25*, *BCY1*, *CYR1*, mitochondrial genome) were contained in intervals with sharp increase in

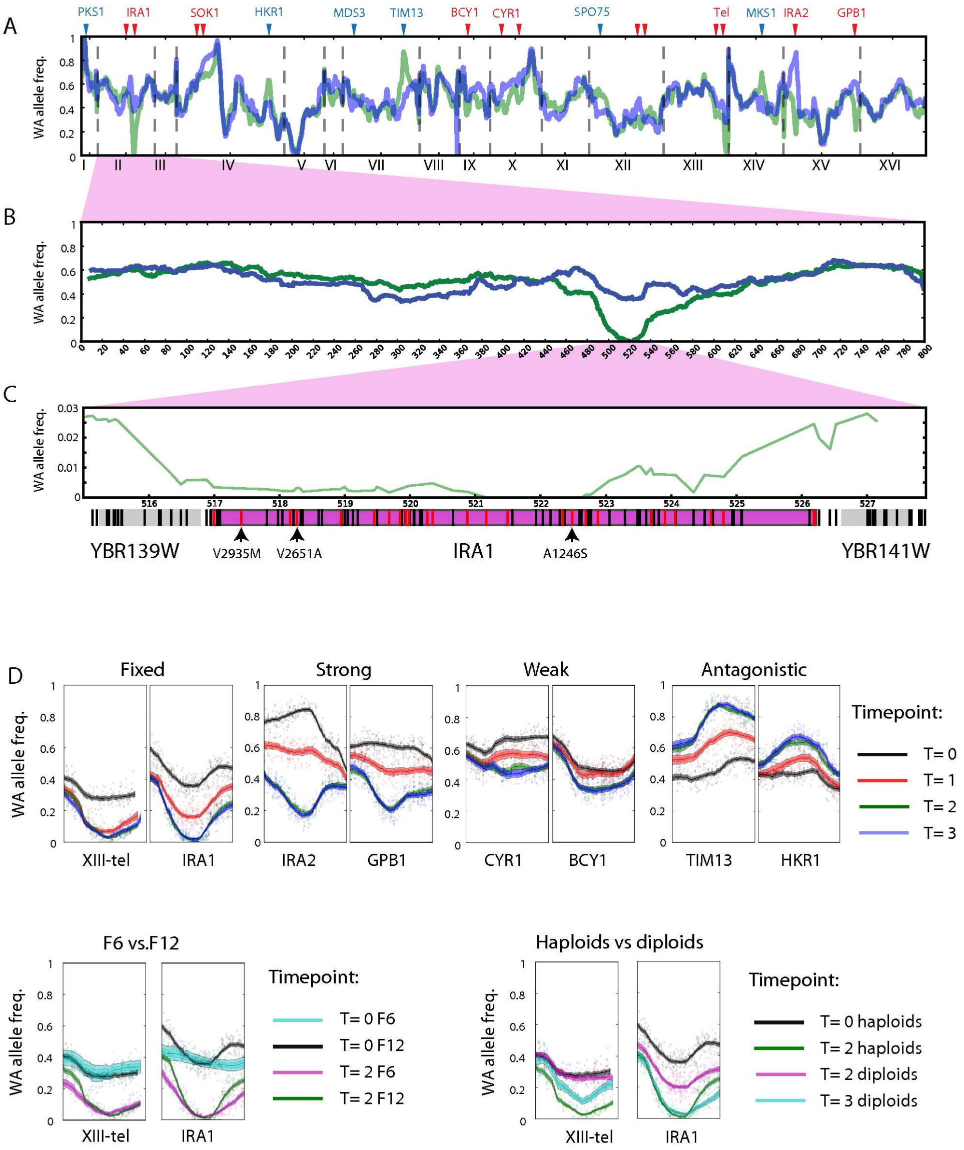


Figure 4.3: Changes in allele frequencies pinpoint QTLs. a-c) WA allele frequency of whole genome (a), chromosome II (b), and *IRA1* region (c) of the F12 pool before (blue) and after selection (green). Lines in gene regions in (c) denote segregating sites (black) and non-synonymous segregating sites (red). The sites with intolerable mutations (SIFT analysis) are highlighted with arrows and designated with the amino acid change. d) Individual examples of mapped QTLs that show differences in strength, beneficial allele, effect of recombination and ploidy. Each window spans 80kb and is centered on the locus with the largest allele frequency change in F12 T2 across two replicas. Shaded regions indicate 90% and 95% confidence intervals.

NA allele frequencies in the F6 or F12 pools, confirming the involvement of the entire RAS pathway in the heat resistance phenotype.

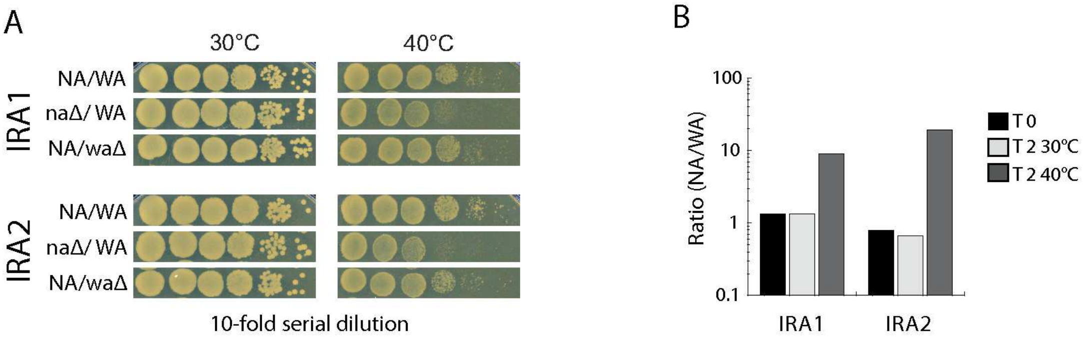


Figure 4.4: *IRA1* and *IRA2* are high temperature growth QTLs. a) Reciprocal hemizygosity confirms that *IRA1* and *IRA2* are high temperature growth QTLs. WA/NA hybrids were individually deleted for the IRA alleles and used to assess their contribution to high temperature growth. Plate spotting assay using 10-fold serial dilution demonstrates better growth of the hybrid when the NA allele is present. b) Competition experiment on hybrids with IRA reciprocal hemizygous deletions (as a) that resembles the selective step applied to the pool. This assay shows that hybrids carrying the NA allele are selected when cells were grown at 40C for 192 hours (T2).

We validated by reciprocal hemizygosity (Steinmetz et al., 2002) that *IRA1* and *IRA2* alleles affect high temperature growth. The effect was evident from a plating assay, growth curves, and competition experiments (Figure 4.4). These genes affect both growth rate (doubling time) and efficiency (final density) with *IRA1* having a stronger effect compared to *IRA2*, consistent with the difference in their final allele frequency. Interestingly, *IRA1* and *IRA2* do not have a pleiotropic effect on growth, even at environmental conditions where RAS activity has a strong influence. The clear identification of the *IRA1* and *IRA2* alleles as a cause of low performance at high temperatures shows that our method can directly map causative genes without any a priori information and without requiring further fine-mapping.

4.3.3 Mapping resolution

The advantage of reduced linkage is evident from narrow mapped intervals, in several cases localising to within single genes (Figure 4.3c-d). For example, in case of *IRA1*, we could map the selected variant down to a small region of the gene (Figure 4.3c), that also harbours the strongest candidate sequence variant between the two strains from SIFT (Ng and Henikoff, 2003) analysis. This resolution is in contrast to that from previous studies based on crosses between strains, including (Ehrenreich et al., 2010), which typically map to large regions containing many genes. An additional advantage of the intercross rounds is the ability to unlink independent QTLs at one locus (Figure 4.3d). There is a risk that long term culturing under stress conditions will select for new adaptive mutations that might rise to high frequencies and dominate the pool. However, as the pool did not become clonal, it is unlikely that haplotypes harbouring strongly adaptive mutations had risen to high frequency during selection (see simulations below).

4.3.4 Lack of fixation

While the alleles with strongest fitness effect, such as at *IRA1* gene (chrII:522kb, Figure 4.3b) and chrXIII subtelomeric region (Figure 4.3d) reached fixation in the pool upon selection, weaker ones required extended selective pressure to rise in frequency (Figure 4.3d), demonstrating the advantage of using extended selection. Three of the 19 QTLs (16%) had reached their T3 allele frequency at T1, but 17 by T2 (89%). Thus, only a minority of two loci were still changing in allele frequency after T2. This indicates that all the remaining haplotypes in the pool have nearly equal fitness in this stress condition, or are so rare even by T3 that change in their frequency does not have a major effect on the average pool genotype. It also suggests that we have saturated for individual loci with independent effects that are present in the founding strains.

4.3.5 Negative epistatic interactions

The fact that for 17 QTLs both alleles remained segregating in the pool after up to 288 hours under selection suggests that these segregating loci cannot have independent additive effects, as otherwise their beneficial version would continue to rise in frequency after T2. Thus, the pool after selection is a mixture of haplotypes with alternative QTL genotype combinations of similar fitness. This suggests an abundance of negative epistatic interactions, as otherwise the beneficial combination would keep rising in frequency. To test this explanation, we genotyped 192 segregants from the F12 pool after 240 hours of selection (T2.5), and looked for scarcity and abundance of specific allele combinations at the 11 strongest QTLs. None of the two-locus combinations was significantly different from the expectation under independence (lowest one- tailed p=0.09, Fisher's exact test) after correcting for multiple testing. Some evidence for lower than expected deleterious allele combination counts was observed when pooling the counts over all pairs (one-tailed p=0.19, Fisher's exact test). This pattern is consistent with complex control and interactions involving multiple genes.

4.3.6 QTLs in diploid pool

Importantly for drawing comparisons with human studies, we were able to map all the 19 heat resistance QTLs in the pool of heterozygous diploid individuals. The effect of selection was weaker for the diploid pool, as allele frequencies had not reached their equilibrium levels by T2, and continued to change until T3 (Figure 4.3d). Diploid pool allele frequency after selection indicates that the chrXIII QTL is consistent with a dominant effect with the final frequency of the homozygous deleterious genotype being removed from the pool, and the *IRA1* QTL with a recessive effect with the beneficial allele being fixed. While it was expected that we could map the recessive QTLs, it is surprising that the final allele frequencies for the other 17 loci were nearly identical for haploid and diploid segregants, consistent with the selected alleles having additive effects as observed for most human GWAS hits.

4.3.7 Comparison with F1 segregant analysis

The heat growth QTLs we found by linkage analysis of 96 F1 segregants partially overlapped the ones identified using our novel approach (Cubillos et al., 2011). However, the linkage analysis lacked power to detect the weak effects, as only the strongest chrXIII subtelomere QTL was detected with high confidence, and a chromosome IV QTL with borderline significance. This shows the additional power of our method. Furthermore, data from growth curves suggests that some of the QTL effects may not be detectable by phenotyping the segregants at 40°C, and phenotyping at higher temperatures will be more informative of individual effect sizes.

4.4 Data analysis

4.4.1 Sequencing data handling

Sequencing reads from the intercross pools before and after selection were mapped to the S288c reference genome obtained from the SGRP project website (Liti et al., 2009a) using BWA (Li and Durbin, 2010), with option ' -n 8' to allow mapping of divergent reads from the other strains. Pileup files comprising the genotypes of mapped reads were created for segregating sites inferred from both low-coverage capillary sequencing (Liti et al., 2009a) and the parental strain shotgun sequence mapping to the S288c assembly. For allele frequency inference, sites that were not segregating in the initial population, corresponding to likely false positive variant calls, were filtered out, as well as sites that were noted as heterozygous in either parental strain, indicative of copy number variation. Furthermore, for allele frequency inference, we filtered the variants to have minimum distance of at least 200 bases to ensure that any single read does not contribute disproportionately due to spanning many variants. The mapping pipeline is available upon request.

4.4.2 Segregant analysis

To analyse the genetic background of two individual F6 segregants, we mapped the sequencing reads to the genome as described earlier, and classified every segregating site to stem from one of the two parental strains, or a no-call. A site was called to be from one parent, if it was covered by at least 15 sequencing reads with base and mapping qualities at least 30, and 80% of them had the parental allele. We conservatively refrained from making a call at low-coverage variants, subtelomeric regions up to 30kb, and variants with ambiguous mapping data. We called a recombination event if a region of at least 2kb from one parent was followed by a region of at least 2kb from the other, and at least 5 calls were made in both regions. This results in a conservative estimate of recombination events, as it discards non-crossovers, and recombination in subtelomeric regions.

4.4.3 Copy number and missing sequence.

We mapped all reads to artificial chromosomes, each containing exactly one gene with 100 flanking bases, and recorded their average sequencing coverage every 100 bases. We used that to infer a copy number for each gene as the average gene coverage normalised by the average sequencing coverage. We also mapped the reads to the assembled contigs from parental sequence data that did not map to the S288c reference; no large allele frequency changes were observed.

4.4.4 Allele frequency inference

Under a simple model, there is an unobserved WA allele frequency f_l at each locus l; we want to infer the posterior distribution of f_l after observing the sequence data. We assume all reads to come from different segregants after filtering segregating sites to be distant, thus every segregant i has one allele a_i observed at some locus l_1 distance $d_{l,i}$ away from l. We take $d_{l,i}$ to be infinity if the loci are on different chromosomes. For that segregant, there is an unobserved allele $b_{l,i}$ at locus l, and the probability that these loci are linked, with no recombination event occurring during the intercross between them, is $q_{l,i} = 1 - \exp(-d_{l,i}\rho)$, where ρ is the recombination rate. We took $\rho = 30(1 + \frac{g-1}{2})$, where g is the number of intercross rounds, as there are on average 30 crossovers per tetrad, and every intercross after the first one has a 50/50 chance of introducing a switch between parental haplotypes. The likelihood of the allele frequency at locus l is thus $P(D|f_l) = \prod_i P(a_i|f_l)$, where

$$\begin{aligned}
P(a_i \,|\, f_l) &= P(a_i, b_{l,i} = \text{WA} \,|\, f_l) + P(a_i, b_{l,i} = \text{NA} \,|\, f_l) = \\
&= P(a_i \,|\, b_{l,i} = \text{WA}) P(b_{l,i} = \text{WA} \,|\, f_l) + P(a_i \,|\, b_{l,i} = \text{NA}) P(b_{l,i} = \text{NA} \,|\, f_l) = \\
&= q_{l,i}^{a_i = \text{WA}} (1 - q_{l,i})^{a_i = \text{NA}} f_l + q_{l,i}^{a_i = \text{NA}} (1 - q_{l,i})^{a_i = \text{WA}} (1 - f_l) \simeq \\
&\simeq q_{l,i} f_l^{a_i = \text{WA}} (1 - f_l)^{a_i = \text{NA}}
\end{aligned}$$

Here, we have discarded likelihood terms that require a recombination event, as we will filter $q_{l,i}$ to be large. We approximate the posterior of f_l with a Beta distribution with an uninformative prior, and find the maximum likelihood parameters of the distribution from for segregants for which $q_{l,i} > 0.95$ (0.75 for Fig. 2A-B for smoothness). This inference procedure corresponds to a smoothing approach within a fixed window with the width determined by the recombination rate, and has the effect of discriminating against extreme allele frequencies. The posterior confidence intervals were obtained from the approximated Beta distribution.

4.4.5 QTL inference

We inferred QTLs in the F12 selected pool by comparing the inferred allele frequencies before and after selection. The allele frequencies in the control experiment, propagating the cells without selection, were nearly identical to those before selection. We called QTLs by testing for inequality of the inferred approximate posterior allele frequencies before and after selection. As a simple cutoff, we called a QTL if the inferred allele frequency changed in the same direction by at least 10% in both biological replicas and 25% in total, a change larger than exhibited for the control experiments at permissive temperature of 23 degrees after 192 hours at any locus in either replica. A single QTL was called in any 20kb window, corresponding to the variant with the largest combined allele frequency change over the two replicas.

4.4.6 Linkage analysis in F1 segregants.

We used results from Cubillos et al. (2011) for F1 segregant QTL mapping. In short, we used standard marker regression for the 200 genotyped markers and

3 heat growth phenotypes to map QTLs significant at 5% false discovery rate (FDR) using a standard linear model and 1000 permutations.

4.4.7 Epistatic interaction tests

We used a standard linear model (Chapter 1.4.1) to assess the significance of an epistatic interaction term between two genotyped loci that affects any of the three growth phenotypes assayed for the segregants. No significant interactions were found at 5% false discovery rate (FDR, fraction of expected false positives in all calls), possibly due to the difference of the phenotyping temperature effect in solid and liquid media, or lack of power.

We tested for scarcity and abundance of two-locus genotype combinations for 11 genotyped trait loci in 189 segregants of F12 population after 120 hours in heat stress (Table S9). For each pair of segregating loci, we compiled a contingency table of genotype counts, and applied two-tailed Fishers exact test to calculate the p-value of independence of the loci. We calculated the false discovery rate (FDR, fraction of expected false positive calls) at a range of p-value cutoffs for the set of pairwise tests, and did not find individual interactions at FDR $< 10\%$. As an alternative, we pooled all allele combination counts for beneficial/beneficial (BB), beneficial/deleterious (BD), and deleterious/deleterious (DD) genotype combinations, to test for relative abundance of BB and DD combinations. For the 11 genotyped QTLs, there were $n_{11} = 607$ DD genotypes, $n_{12} = 3959$ BD genotypes, and $n_{22} = 5739$ BB genotypes for the $N = n_{11} + n_{12} + n_{22}$ genotypes. We then compiled a contingency table with the observed and expected combination counts calculated from the fraction of genotyped beneficial alleles, and calculated the chi-squared test p-value. This test is appropriate as the sample size is large. We also repeated the test for QTLs found within individual pathways; no p-values were significant at 10% cutoff.

4.5 Simulation experiments

We now expand on the argument for lack of adaptive mutations dominating the pool, as well as allele frequency changes under simplifying conditions in a simulation scenario. We simulated data from a simple generative model to explore the potential of adaptive mutations to dominate a haploid pool, as well as effects of more than one allele in haploid and diploid pools.

4.5.1 Adaptive mutations.

First, we provide three lines of computational evidence for lack of new adaptive mutations with large effect on intercross pool allele frequency during selection. Second, we demonstrate how interaction effects can account for lack of fixation, and ploidy can be responsible for the difference in response time to selection.

Firstly, the fitness requirement of adaptive mutations to dominate the pool is too high. A single adaptive mutation begins at very low initial frequency, $f_1 = \frac{1}{N}$, where we take N, the total number of segregants in the pool to be 10^7. The doubling times for the segregants range from 1.5 hours in permissive condition (or for fit segregants in restrictive condition) to 2 hours for unfit segregants in restrictive condition. Let us assume an adaptive mutation rises to the same frequency as the total frequency of haplotypes with beneficial alleles at the two loci that reach fixation - the *IRA1* and chrXIII subtelomeric loci (initial frequency $f_0 = 0.25$) all of which have doubling times $t_0 \sim 1.5$ hours. Over $T = 288$ hours of selection, the following identity must then hold for the doubling time t_1 of the adaptive mutation: $f_1 2^{\frac{T}{t_1}} \geq f_0 2^{\frac{T}{t_0}}$, or $t_1 \leq \frac{T}{\log_2 f_0 - \log_2 f_1 + \frac{T}{t_0}}$. Plugging in numbers for f_1, f_0, T, this gives $t_1 \leq 1.34 = 0.9 t_0$ Thus, in order to rise to appreciable frequencies in the very large pool, the haplotype with the adaptive mutation must grow 10% faster in restrictive condition than the segregants do in the permissive condition. If such mutations were possible, they would be more likely to rise during the many months of intercross rounds, not during the span of four days. However, in this case, the allele, not the haplotype, will be selected for, as further intercross rounds separate the adaptive mutation from the haplotype on which it arose.

Dominating adaptive mutations would drive the pool allele frequencies to extremes. In the very long run, the haplotype with the adaptive mutation will be the only one left in the pool, as no recombination happens during selection. As the frequency of the adaptive mutation rises in the pool, the pool loses heterozygosity and genetic complexity, and the frequency of the NA allele at all segregating loci will be driven to 0 or 1. If a haplotype with an adaptive mutation is present at high frequency in the pool, we would expect to see an allele frequency change from the initial pool at all loci towards the genotype of that haplotype, which we do not observe.

Adaptive mutations would continue to rise in frequency after 192 hours. We do not observe global allele frequency changes after 192 hours. However, as outlined above, haplotypes with adaptive mutations should continue to rise in frequency in the pool. These three lines of evidence point to little contribution from adaptive mutations to the final segregant pool allele frequency makeup. Adaptive mutations for sporulation, mating, or growth can arise during intercross, and could be traced. However, for QTL mapping, we are conditioning our analysis on all the segregating sites present in the pool at the beginning of selection, regardless of whether they were present in the parental strains.

4.5.2 Effects of selection on allele frequency.

We simulated allele frequency changes under simple assumptions for various scenarios. While standard (e.g. Hartl and Clark (2006)), the results give intuition for allele frequency changes observed.

Haploid individuals. We fixed the initial allele frequency of any locus to be 0.5 for simplicity, and calculated its change over generations in a deterministic way. For a one locus trait, the individuals with genotype '1' were assumed to have a fitness advantage s, which changed the rate at which they survived to the next generation, with the frequency $f_{l,t}$ of locus l at generation t was taken to be $\frac{(1+s)f_{l,t-1}}{(1+s)f_{l,t-1}+(1-f_{l,t-1})} = \frac{1+s}{1+sf_{l,t-1}}f_{l,t-1}$ If $s > 0$, f_l increases, and if $s < 0$, it decreases in a near-geometric manner. For these one locus haploid pools, the beneficial allele asymptotically approaches fixation, with the speed depending on the magnitude of the selection coefficient (Figure 4.5).

In case two loci are contributing, the calculation remains almost unchanged, but now the effect of selection is assumed to act only on the '11' genotype. In this case, if $s > 0$, the haplotypes with '11' genotype are fitter than the others, and again are driven to fixation. However, if $s < 0$, the ’11’ genotype is selected against, and will be purged from the pool in the long run. Both alleles will still be present at each locus (Figure 4.5A). We hypothesize such interactions to be responsible for the lack of fixation upon over 100 generations of selection. The usual intuition behind this is that fitness depends on functioning of a specific pathway. While any single mutation does not alter the functionality of the pathway, there are many possible combinations of genotypes that render it defective. These combinations are selected against, producing a change in allele frequency, but not fixation of any allele.

Diploid individuals. As the diploid individuals propagate clonally just like haploids, we have to trace the frequency of the genotypes, not alleles, since there is no further mixing of the haplotypes between individuals. We can therefore treat a one locus trait in diploids, identically to a two-locus trait in haploids, and find that for traits where the beneficial allele behaves in an additive or recessive way, selection drives the frequency of beneficial allele to fixation, and for dominant beneficial alleles, the homozygous non-beneficial allele combination is selected against (Figure 4.5). We observed QTLs with final allele frequencies as well as their speed of change consistent with both recessive (*IRA1*) and dominant (chrXIII subtelomere) beneficial alleles (main text). However, when the QTL acts in an additive manner, the allele frequency change is identical to that of the haploid pool.

If interaction effects are responsible for the allele frequency change, the effect can again be dominant, additive, or recessive. The differences to a one-locus model are slower effect of selection, as the fittest haplotype has lower initial frequency, and less extreme final allele frequency in case the interaction effect is dominant, as there are fewer genotype combinations selected against (Figure 4.5B).

4.6 Discussion

We have presented an accurate, sensitive, quick approach for QTL mapping in yeast. It is straightforward to apply our method to any selectable trait. We

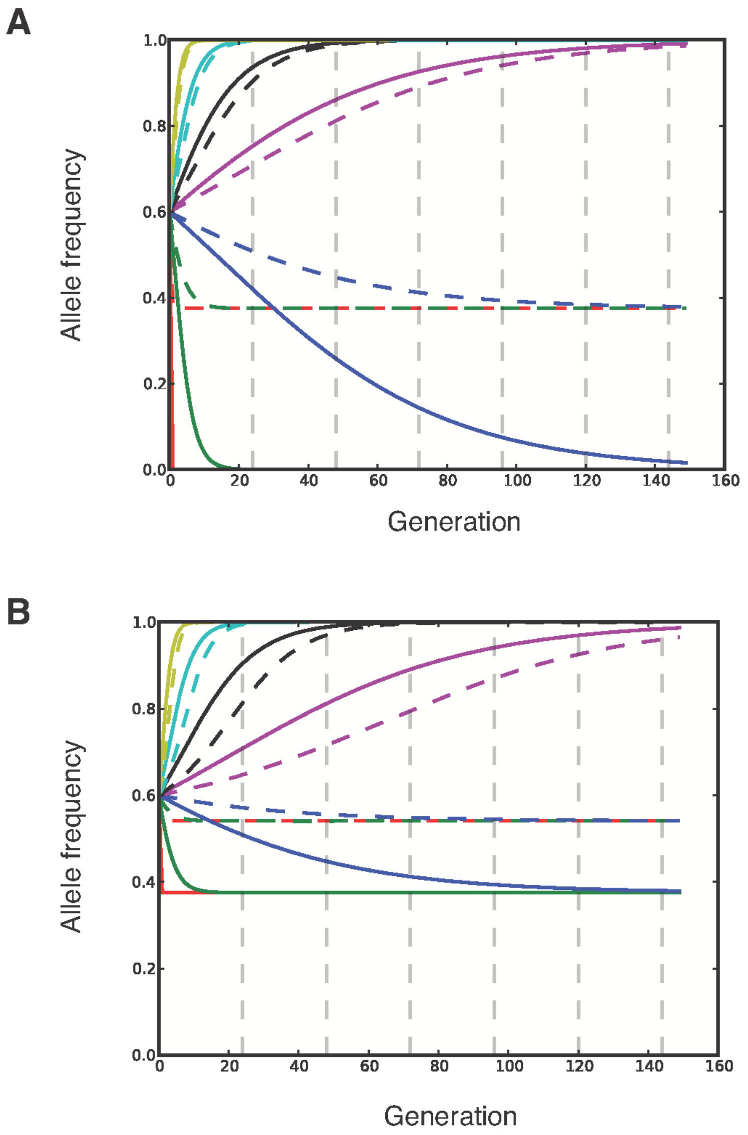


Figure 4.5: Haploid (solid lines) and diploid (dashed lines) pool allele frequency changes for 1-locus (a) and 2-locus effects (b). Initial allele frequency of a locus is 0.6. Individual lines correspond to different fitness modifiers, from top to bottom: $+1, +0.3, +0.1, +0.03, -0.03, -0.3, -1$.

expect to be able to extend these mapping populations to include more of the genetic diversity in the species by crossing a larger number of parental strains. As we were also able to map the trait loci in the diploid pool, there is a potential to establish an outbred yeast population that can be used as a model for natural diploid genome-wide association studies as carried out in humans.

Chapter 5

Additional gene mapping studies

Collaboration note *This chapter contains work performed in collaboration with Gemma Langridge and Dr. Keith Turner for bacterial transposon mutant mapping (Langridge et al., 2009), and Francisco Cubillos and Dr. Gianni Liti for yeast linkage analysis (Cubillos et al., 2011; Liti et al., 2009b).*
Gemma and Keith developed the transposon mutant library, performed the experiments, generated raw data, and did the high-level analysis; I contributed the statistical analyses of the data. Similarly, Francisco and Gianni designed and developed the yeast grid of crosses, performed the experiments, and high-level analysis; I contributed the statistical analyses and parts of interpretation of the data.

5.1 Gene mapping with one million bacterial transposon mutants

One trait mapping approach available in prokaryotic and simple eukaryotic organisms is generating a very large number of random mutants, and then examining which mutants survive selection (Chapter 1.1.3). This is related to the work in Chapter 4 on standing variation, but can access a wider variety of alleles. A version of this approach based on transposon insertions was recently developed for the bacterium *Salmonella enterica serovar Typhi* (*S. Typhi*, Langridge et al. (2009)).

5.1.1 Transposon insertion library for *Salmonella Typhi*

S. Typhi causes typhoid fever, and is responsible for hundreds of thousands of deaths in the developing world every year (Crump et al., 2004). One approach to fighting this disease agent is to map the genes essential for its survival in the permissive condition, restrictive conditions associated with its lifecycle in the human host, and under stress from therapeutic agents. To this end, our collaborators created a transposon insertion library with on the order of 1,000,000 mutants, each harbouring one transposon insertion. This large mutant library was then grown in a permissive condition and with added 10% ox bile to simulate gall bladder environment, followed by DNA extraction from the pool, amplification of DNA from the junction between transposon sequence and genomic DNA, and high-throughput sequencing. Mapping the sequencing reads to the genome results in a list of sites where some mutant had a transposon inserted.

Here, we focus on two mapping tasks. First, we look for essential genes that do not allow insertions, followed by study of genes essential for growth in bile, which is important for its persistence in the human host.

5.1.2 Mapping essential genes

To test whether a gene was essential, we quantified how unlikely it was to harbour a transposon insertion. Genes with no observed insertions are likely to be essential, while genes with many insertions are obviously not. For every gene g of length L_g, we calculated the insertion frequency $f_g = \frac{I_g}{L_g}$, where I_g is the observed number of insertions.

We noted that the distribution of f was bimodal with modes at 0 and roughly 0.05, and heavy-tailed (Figure 5.1). The mode at 0 corresponds to the essential genes that do not allow for any insertions, and the mode at 0.05 to all the other genes. Under the assumption of uniform incorporation of the transposon, we would expect the number of insertions in a gene to follow a Poisson distribution. However, the distribution is considerably more dispersed, indicating presence of unknown biases and potential sequence-specificity. Standard approaches to deal with overdispersion, such as using a negative binomial distribution, or a normal distribution with variance proportional to mean, did not give a substantially better fit, and were not straightforward to interpret.

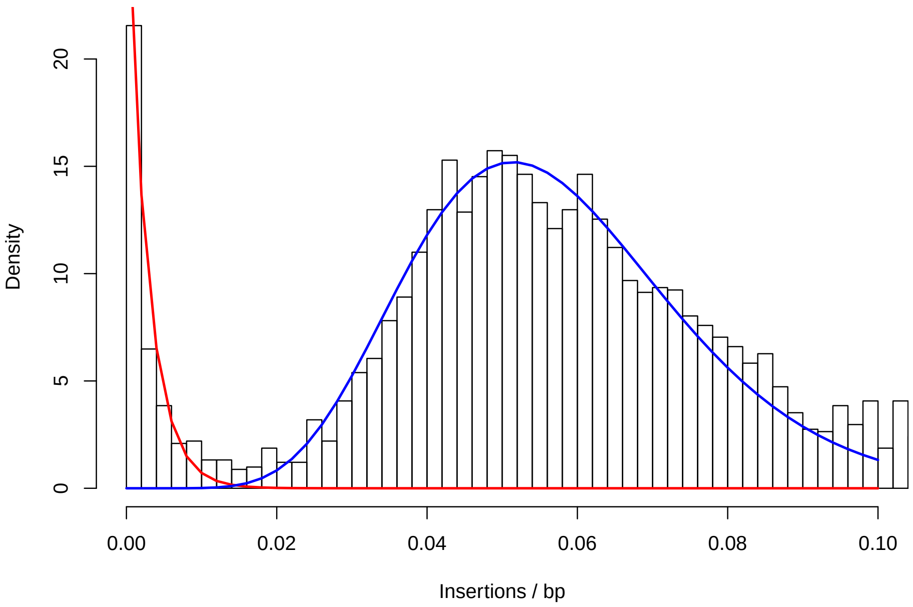


Figure 5.1: Histogram of per-base insertion frequency of individual genes. The blue line corresponds to the Gamma fit to the right mode (non-essential model), while the red line corresponds to the Gamma fit to the left mode (essential model).

Instead of modelling the generative process, we modelled the data directly. We fit Gamma distributions for the two modes of the distribution of the insertion site count in every condition using the R MASS library. For each f_g, we calculated the probabilities corresponding to the right tail of the essential model and left tail of the non-essential model. This represents our belief of observing an insertion index that is at least as extreme for both individual models. For every gene g, we calculated the base 2 logarithm of the likelihood ratio (L_g) between the two model fits, and classified the gene as essential ($L_g < -2$, essential model at least

4 times more likely), non-essential ($L_g > 2$, non-essential model at least 4 times more likely), or uncertain ($|L_g| < 2$).

We found 349 essential genes at false discovery rate less than 0.07. The analysis of these genes is presented in Langridge et al. (2009).

5.1.3 Mapping condition-specific essential genes

Next, we looked for genes essential for growth in bile. These genes should not be essential in general, but insertions in them should be observed less than in the permissive condition. There were three timepoints for growth in bile, we compared the data from each to the data from permissive condition. We analyse number of mapped reads instead of insertion events to avoid many more comparisons of very low frequency (1-2 insertions) events. From the raw data, it was clear that several genes had reduced insertion frequencies as assessed by the number of sequenced reads (Figure 5.2a).

For each pair of conditions (A, B), we calculated the $\log_2$ fold change ratio $S_{g,A,B}$ in the number of observed reads $R_{g,A}$, $R_{g,B}$ for every gene g as $S_{g,A,B} = \frac{R_{g,A}+100}{R_{g,B}+100}$. The correction of 100 reads in the numerator and denominator smooth out the high scores for genes with very low numbers of observed reads, and corresponds to a prior belief that if there is an insertion present, there should be an abundance of reads mapping to it.

Again, we modelled the data directly. We fit a normal model to the mode of distribution of $S_{A,B}$ over all genes, and calculated p-values for each gene according to the fit (Figure 5.2b). This procedure results in an ordered gene list. We chose an arbitrary cutoff, and considered a gene to be condition-specific if the fold-change between the conditions was greater than 4, which corresponds to p-value of 10^{-5}, and false discovery rate of 2.5×10^{-4}. These genes are analysed in depth in Langridge et al. (2009).

5.2 Gene mapping with grid of yeast crosses

Baker's yeast *Saccharomyces cerevisiae* has been successfully used in linkage studies over the last decade, focusing mainly on two F_1 crosses (Ehrenreich et al.

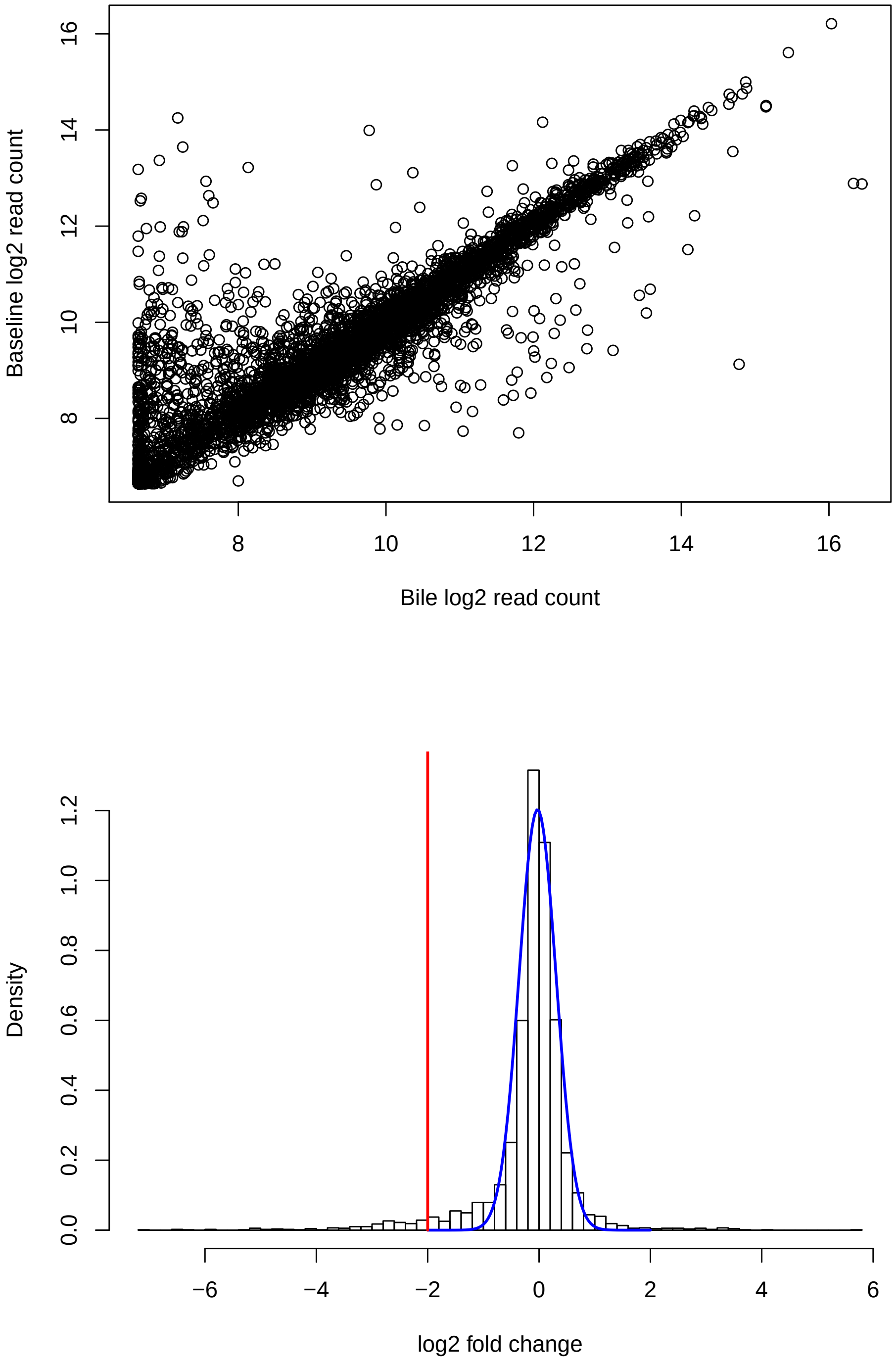


Figure 5.2: **(a)** Scatter plot of $\log_2$ read counts in two conditions. 100 is added to each gene's counts for smoothing. **(b)** Histogram of gene read count $\log_2$ fold change. The blue line corresponds to normal fit to the mode, red line is the cutoff used to determine condition-specificity.

(2009); Mancera et al. (2008); Steinmetz et al. (2002); Zheng et al. (2010), Chapter 1.1.1). However, a single cross gives restricted information about context-dependent allele effects, and is limited to variants present only in the two parental strains.

5.2.1 Grid of yeast crosses

To explore the effects of alleles in different genetic contexts, our collaborators generated a grid of crosses between all five clean (non-mosaic) lineages of *S. cerevisiae* sequenced as part of the *Saccharomyces* Genome Resequencing Project (Liti et al., 2009a). One of the strains (of Malaysian origin) was effectively reproductively isolated, and thus not included for further analysis. The remaining six crosses between the four strains captured 64% of the segregating sites identified by Liti et al. (2009a).

Ninety-six F1 segregants were isolated from 24 meiotic events for each cross. Every segregant was genotyped at 171 evenly spaced markers, followed by quantitative characterisation of growth curves in different conditions. Three growth environments were shared between all crosses, while the rest of the 32 tested environments were cross specific.

5.2.2 Recombination analysis

First, we characterised the global recombination landscape in the six crosses. We called a recombination event between two consecutive genotyped loci in one haploid segregant if the two observed alleles came from different parents.

We determined the average recombination rate ρ_k in each cross k as the number of observed recombination events divided by the genome size. We then used a Poisson model with mean ρ_k to assess the significance of hot- and coldspots in each cross k. A hotspot was deemed significant if the probability of observing as least as many recombination events under the model was less than $\alpha = 0.005$ (FDR$<$ 10%) in at least one cross. Similarly, coldspots were called significant, if the probability of observing up to that many recombination events was less than 0.005.

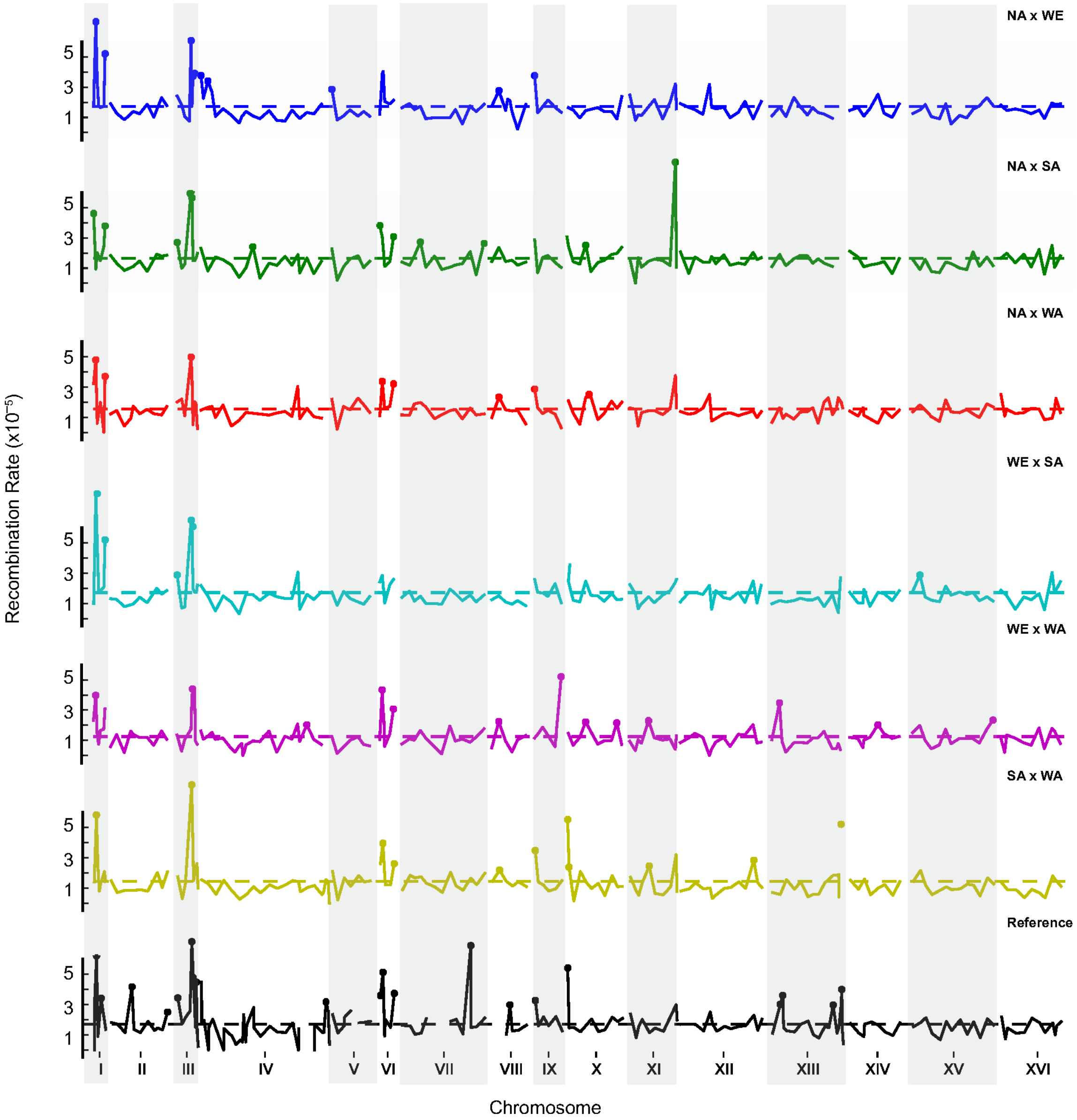


Figure 5.3: Observed recombination rate in each of six crosses, as well as a reference cross from previous work. Recombination hotspots are highlighted with a filled circular marker.

In addition, we used data from Mancera et al. (2008), who established a high-resolution crossover map in another cross using 56 meiotic events. For each marker we genotyped, we took the closest called genotype from their data for every segregant, and repeated our analysis on this dataset.

We found 32 hotspots (Figure 5.3) and 48 coldspots. Nine of the hotspots were not recovered with the high-resolution data form Mancera et al. (2008); seven of the nine were cross-specific, and the remaining two strain-specific, present in all crosses with one strain. We found ten of sixteen centromeric regions to be recombination coldspots in some cross, consistent with their reduced rate of meiotic recombination (Choo, 1998) while no other coldspots were shared between more than three crosses. These results suggest that hotspots, but not coldspots, are mostly conserved. In-depth analysis of these data is given in Cubillos et al. (2011).

5.2.3 Linkage mapping

We then mapped QTLs in all six crosses to determine the regions linked to growth phenotypes in different conditions. Linkage analysis was performed with the rQTL software (Broman et al., 2003) using the non- parametric (Kruskal-Wallis) test for QTLs and normal model for variance explained. LOD$>$ 2.63 was used as cutoff (FDR$<$ 5%) giving less than one QTL by chance per trait. We used the same approach to find strain-specific QTLs by performing one against all tests, pooling data from all crosses with a strain. We also searched for epistatic interactions using the normal model (Chapter 1.4.1), taking LOD$>$ 5.8 (FDR$<$ 5%) as a cutoff.

We found a plethora of QTLs. Two hundred and thirty-three marker-trait pairs were significant. Many of them were specific to a cross or strain, and other combinations were represented as well (Figure 5.4a). We found additional QTLs by pooling the genotypes across crosses (Figure 5.4b), and also detected putative epistatic interactions. Again, more analyses of the QTLs are provided in Cubillos et al. (2011).

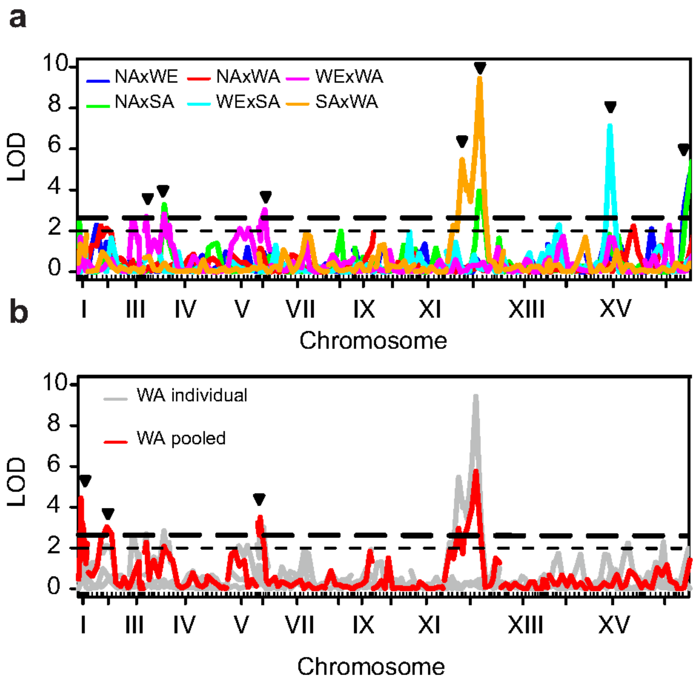


Figure 5.4: Paraquat growth rate QTLs found in each cross independently (a) and in a one-against-all test for the WA strain (b). The phenotyping approach and conditions used are described in Cubillos et al. (2011).

Chapter 6

Conclusions

6.1 Conclusions and discussion

I have spent the last four years trying to understand the genetic basis of cellular traits, focusing mostly on genetics of gene expression, and trait mapping by selection.

6.1.1 Abundance and importance of eQTLs

We found genetic associations to 30% of the transcript levels, and preliminary results suggest that this number increases to 85% in larger human cohorts. Thus, gene expression level, the most basic cellular phenotype, is certainly influenced by genotype. While it is not surprising that common genetic variation in gene regulatory regions does influence the structure and protein binding affinities of DNA, still only a small amount of variance is explained by genotype.

We (Chapter 3) and others (Foss et al., 2007) have found evidence for little correlation between mRNA and protein levels in the cell. This suggests that small scale gene expression variation is not amplified to protein levels, and rather dampened. This supports the view that many of the eQTLs we find may not have functional consequences in the tissue the gene expression was assayed. However, in other tissues, the effects may be larger.

Strong QTLs, however are candidates for causal regulatory effects, including human GWAS hits (Nica et al., 2010; Nicolae et al., 2010). It could also be that some of the weak eQTL alleles tag rare variants of large effect, that are not detected with the non-parametric methods commonly used. Furthermore, weak eQTLs in one tissue could be strong in another, where the genes are expressed at higher levels. While a comprehensive resource of human average tissue-specific expression levels has recently been published (Lukk et al., 2010), there is no comprehensive data available yet on the variability of expression levels in all tissues, and their genetic basis. Some projects, e.g. the MuTHER resource (Nica et al., 2011) are beginning to fill this gap.

Some dimensions of eQTL mapping remain understudied. First, there is a question of identifying the causative nucleotide(s). Functional annotation of the loci can help find variants that are in known or predicted protein binding regions, and therefore predicted to be functional. Further correlating the existence of a binding site with expression from the haplotype, which can be possible in mRNA-seq experiments, will give more evidence for the functional impact. Second, from studying cell populations, it is not clear whether the change in expression levels is due to a shift in the mean level in every cell, or a large change in a smaller number of cells. The problem of differentiating between small effects of full penetrance and large effects of low penetrance is fascinating and important, and pertinent to most cellular traits. I hope that large condition-specific effects are at play, as these will be easier to model, once appropriate assays have been undertaken.

6.1.2 Abundance and importance of interactions

We have developed methods to detect genotype-specific effects on cellular traits (Chapter 3). We found that modelling unmeasured cellular phenotypes lowers the dimensionality of the hypothesis space, eases the multiple testing burden, and yields interpretable genetic associations and interactions.

There is a gap between intuition developed in model organisms and findings from studying human cohorts. In models, epistasis has been found almost everywhere, while in humans, it remains elusive. This is a secondary problem. The real cellular mechanism for an epistatic effect is not DNA-DNA interaction, but instead, an interaction of two traits. Thus, the question is not where are epistatic effects, but rather which traits we need to measure or infer to understand variation in our favourite trait. The genotype then comes into play only as a source of variability for the interacting trait.

We have explored one direction of such interactions, genotype-specific transcription factor and pathway effects. We rely heavily on existing annotations for structuring our model, and have to use inferred phenotypes, as the traits we are really interested in are not measured. Thus, obvious extensions to existing work would include more detailed prior information, as well as modelling other measured traits. I believe that genotype-specific effects are also pervasive in humans, and will be detected using inferred intermediate phenotypes, or assays of further cellular phenotypes.

6.1.3 Trait mapping using artificial selection

We have established a method to map any selectable trait in yeast to narrow intervals, and found many loci to be contributing to heat resistance.

It is not surprising that many loci are responsible for variation in one trait. While simple characteristics can be determined by one specialised protein such as efflux pumps of specific molecules, most cellular traits are determined by the action of entire pathways. Thus, variation in any part of the pathway that also effects its activity will be a source of variability for the phenotype. Any allele that affects a pathway component and has downstream effects will be under selection if pathway activation is selected for.

The lack of fixation of individual alleles is also explained by selection for pathway activation. Once the activation is perturbed enough to produce a fit individual, genotypes of other alleles have no fitness effect, and are under no selective pressure. Alternatively, once enough deleterious alleles are present to abolish the pathway activation and produce an unfit individual, the additional alleles will not influence the fitness of the individual further. Such effects correspond to negative epistatic interactions, and we are validating whether they explain our observations (see Future work below).

It has been reassuring to observe genetic complexity in a simple model organism. Hopefully, much of what we learn about the fine scale structure of complex yeast traits mapped by artificial selection will also translate to higher eukaryotes.

6.2 Future work

The development of the PEER framework (Chapter 2) and the interaction model (Chapter 3) is complete. What remains to be finished is an interface more accessible to the general user. To this end, we are reimplementing both general as well as sparse factor analysis models as an R package. Further work along modelling latent phenotypes will use the MuTHER dataset, and combine information from genotype, gene expression, small RNA expression, methylation, lipid level, metabolite, DEXA scan, and clinical questionnaire data, building towards a generative model of human cellular and molecular physiology, and its relation to disease.

There are many possibilities to extend work on mapping by artificial selection (Chapter 4). There are experiments to be done, analysis to undertake, and models to develop.

Modelling. The first priority is establishing and implementing inference for a correct generative model of allele frequencies and QTLs. Currently, we employ a simplistic smoothing approach. Instead, we have a HMM-like model in mind, where binary indicators designate QTL locations, local recombination rates are modelled as random variables with informed prior distributions, and allele frequencies are inferred taking the above into account. This model would yield a more finescale QTL map, inform of the required sequencing depth for accurate mapping, as well as provide estimates of local recombination rates. The exact QTL locations can be used in designing further genotyping experiments.

Analyses. We have generated a rich dataset, and many questions are not yet answered. We are looking to analyse population genetics and signatures of selection for all QTL regions, run SIFT analyses to detect intolerable alleles, and perform additional computational experiments to assess the effect of candidate causative alleles on mRNA expression levels or protein structure and function.

Finally, we are in a position to assess local recombination rates both from the model proposed above, as well as long insert libraries that we can scan for read pairs with evidence of material from more than one parental strain.

Experiments. We are in the process of extending this work in several directions.

Firstly, we are genotyping 1,000 segregants from the pool after selection at all the QTL loci to find epistatic interactions. Secondly, we have generated a four-way cross of the strains used in the grid of crosses experiment (Chapter 5.2). We are mapping QTLs in this genetically more diverse population, and genotyping and phenotyping 384 of the segregants from the intercross pool. We need to perform a few follow-up experiments for replicating diploid results, and understand how much information is shared between haploid and diploid screens. Finally, it may be interesting to assess the dynamics of the allele frequency change at higher resolution, and in longer term.

Most importantly, we can map any selectable trait to high resolution. This opens a wide range of possible experiments. Most interesting ones pertain to general cell biology that would also transfer to higher eukaryotes, such as DNA damage response, oxidative stress response, ageing, and cell adhesion. We are looking to focus on specific biological questions that can be answered in this model.

We have measured mRNA levels from the pool before and after selection, both steady state, as well as in response to heat shock after 15 minutes. We may be looking to supplement these experiments with protein level measurements of a few key proteins to assess the role of their abundance in the heat resistance trait to trace the phenotypic effect of the alleles. Furthermore, we may attempt to rescue the heat resistance phenotype in some segregants with low fitness by introducing a plasmid that modifies *RAS* activity. Finally, we could generate allele replacement strains for the individual QTLs to assess the effect of the alleles in isolation and specific combinations.

References

23andMe. SNPwatch: Uncertainty Surrounds Longevity GWAS. 2010. URL http://spittoon.23andme.com/2010/07/07/snpwatch-uncertainty-surrounds-longevity-gwas/. 8

Alberts, B., Johnson, A., Lewis, J., Raff, M., et al. *Molecular Biology of the Cell*. Garland Science, 5 edition, 2007. 4, 16

Alter, O., Brown, P.O., and Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.*, 97:10101–10106, 2000. 67

Altshuler, D., Daly, M.J., and Lander, E.S. Genetic mapping in human disease. *Science (New York, N.Y.)*, 322(5903):881–888, 2008. 7

Alwine, J.C., Kemp, D.J., and Stark, G.R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A*, 74(12):5350–4, 1977. 3

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–31, 2010. 8, 24

Ayyadevara, S., Ayyadevara, R., Vertino, A., Galecki, A., et al. Genetic loci modulating fitness and life span in Caenorhabditis elegans: categorical trait interval mapping in CL2a x Bergerac-BO recombinant-inbred worms. *Genetics*, 163(2):557–70, 2003. 6

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M., et al. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, 14(3):283–91, 2004. 11

Balding, D., Bishop, M., and Cannings. *Handbook of Statistical Genetics*. Wiley J. and Sons Ltd., N.Y., second edition, 2003. 34

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet*, 40(8):955–62, 2008. 11

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37(Database issue):D885–90, 2009. 51

Bateson, W. *Mendel's Principles of Heredity*. Cambridge University Press, 1909. 3

Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition, 2007. 30

Biswas, S., Storey, J.D., and Akey, J.M. Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. *BMC Bioinformatics*, 9(1):244, 2008. 62, 67, 80, 88

Boomsma, D., Busjahn, A., and Peltonen, L. Classical twin studies and beyond. *Nat Rev Genet*, 3(11):872–82, 2002. 13

Brem, R.B. and Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1572–1577, 2005. 6, 34, 48, 51, 66, 80, 87

Brem, R.B., Storey, J.D., Whittle, J., and Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, 436(7051):701–703, 2005. 6

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science*, 296(5568):752–755, 2002. 6, 11, 18, 34, 44, 82

Broman, K.W., Wu, H., Sen, S., and Churchill, G.A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7):889–90, 2003. 116

Butt, C., Zheng, H., Randell, E., Robb, D., et al. Combined carrier status of prothrombin 20210A and factor XIII-A Leu34 alleles as a strong risk factor for myocardial infarction: evidence of a gene-gene interaction. *Blood*, 101(8):3037–41, 2003. 12

C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, 282(5396):2012–8, 1998. 4

Chen, Y., Zhu, J., Lum, P.Y.Y., Yang, X., et al. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, 2008. 11, 34, 66

Choo, K.H. Why is the centromere so cold? *Genome Res*, 8(2):81–2, 1998. 116

Choudhary, C. and Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nat Rev Mol Cell Biol*, 11(6):427–39, 2010. 17, 21

Cirulli, E.T. and Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425, 2010. 10

Clarke, J., Seo, P., and Clarke, B. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics (Oxford, England)*, 26(8):1043–1049, 2010. 16

Codd, V., Mangino, M., van der Harst, P., Braund, P.S., et al. Common variants near TERC are associated with mean telomere length. *Nat Genet*, 42(3):197–9, 2010. 8

Consortium. 1000 Genomes Project. 2010. URL http://www.1000genomes.org. 59

Cordell, H.J. Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics*, 10(6):392–404, 2009. 12, 87

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., et al. The Genetic Landscape of a Cell. *Science*, 327(5964):425, 2010. 66, 85, 86

Cotsapas, C.J., Williams, R.B.H., Pulvers, J.N., Nott, D.J., et al. Genetic dissection of gene regulation in multiple mouse tissues. *Mamm Genome*, 17(6):490–5, 2006. 19

Crick, F. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970. 3

Crump, J.A., Luby, S.P., and Mintz, E.D. The global burden of typhoid fever. *Bull World Health Organ*, 82(5):346–53, 2004. 110

Cubillos, F.A., Zörgö, E., Parts, L., Fargier, P., et al. Assessing the complex architecture of polygenic traits in diverged yeast populations. *Molecular Ecology*, 2011. 93, 95, 99, 102, 109, 116, 117

Cudworth, A.G. and Woodrow, J.C. Genetic susceptibility in diabetes mellitus: analysis of the HLA association. *Br Med J*, 2(6040):846–8, 1976. 7

Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum Genet*, 122(6):565–81, 2008. 2

Darvasi, A. and Soller, M. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, 141(3):1199–1207, 1995. 6, 93

Davison, A.C. *Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003. 28, 30, 57

de Vrijes, H. *Die mutationstheorie. Versuche und beobachtungen über die entstehung von arten im pflanzenreich.* Leipzig,Veit comp., 1901. 3

Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, 325(5945):1246–50, 2009. 19

Doroszuk, A., Snoek, L.B., Fradin, E., Riksen, J., et al. A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans. *Nucleic Acids Res*, 37(16):e110, 2009. 6

Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999. 30

Ehrenreich, I.M., Torabi, N., Jia, Y., Kent, J., et al. Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature*, 464(7291):1039–1042, 2010. 6, 10, 91, 92, 98

Ehrenreich, I., Gerke, J., and Kruglyak, L. Genetic dissection of complex traits in yeast: insights from studies of gene expression and other phenotypes in the BYxRM cross. In *Cold Spring Harb Symp Quant Biol*, volume 74, pp. 145–153. 2009. 6, 91, 112

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, 2008. 34, 61, 66

Eriksson, N., Macpherson, J.M., Tung, J.Y., Hon, L.S., et al. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet*, 6(6):e1000993, 2010. 8

Fisher, R.A. *The Genetical Theory of Natural Selection*. Oxford University Press, USA, 1939. 22

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512, 1995. 4

Flint, J. and Mackay, T.F.C. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res*, 19(5):723–33, 2009. 6, 10

Flintoft, L. Complex disease: Adding epigenetics to the mix. *Nature Reviews Genetics*, 11(2):94–95, 2010. 13

Foss, E.J., Radulovic, D., Shaffer, S.A., Ruderfer, D.M., et al. Genetic basis of proteome variation in yeast. *Nature Genetics*, 39(11):1369–1375, 2007. 6, 17, 81, 118

Freeman, T.C., Goldovsky, L., Brosch, M., van Dongen, S., et al. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol*, 3(10):2032–42, 2007. 27

Fuchs, F., Pau, G., Kranz, D., Sklyar, O., et al. Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Mol Syst Biol*, 6:370, 2010. 17

Gardner, R.S., Wahba, A.J., Basilio, C., Miller, R.S., et al. Synthetic polynucleotides and the amino acid code. VII. *Proc Natl Acad Sci U S A*, 48:2087–94, 1962. 3

Garge, N., Pan, H., Rowland, M.D., Cargile, B.J., et al. Identification of quantitative trait loci underlying proteome variation in human lymphoblastoid cells. *Mol Cell Proteomics*, 9(7):1383–99, 2010. 17

Gerke, J., Lorenz, K., and Cohen, B. Genetic Interactions Between Transcription Factors Cause Natural Variation in Yeast. *Science*, 323(5913):498–501, 2009. 12

Giaever, G., Chu, A.M., Ni, L., Connelly, C., et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418(6896):387–91, 2002. 9

Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet*, 6(5):e1000952, 2010. 14, 20

Gibson, G. The environmental contribution to gene expression profiles. *Nature reviews. Genetics*, 9(8):575–581, 2008. 34, 66

Glass, D., Parts, L., Knowles, D., Aviv, A., et al. No correlation between childhood maltreatment and telomere length. *Biol Psychiatry*, 68(6):e21–2; author reply e23–4, 2010. 8

Goerner, W., Durchschlag, E., Martinez-Pastor, M., Estruch, F., et al. Nuclear localization of the C2H2 zinc finger protein MSN2P is regulated by stress and protein kinase A activity. *Genes & development*, 12(4):586, 1998. 81

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., et al. Life with 6000 genes. *Science*, 274(5287):546, 563–7, 1996. 4

Green, M.M. 2010: A century of Drosophila genetics through the prism of the white gene. *Genetics*, 184(1):3–7, 2010. 5

Group, T.H.D.C.R. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–83, 1993. 5

Gygi, S., Rochon, Y., Franza, B., and Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, 19(3):1720, 1999. 81

Hartl, D.L. and Clark, A.G. *Principles of Population Genetics, Fourth Edition*. Sinauer Associates, Inc., 4th edition, 2006. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0878933085. 105

Hawkins, R.D., Hon, G.C., and Ren, B. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11(7):476–486, 2010. 17, 21

Hershey, A.D. and Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol.*, 36:39–56, 1952. 3

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–7, 2009. 10

Hunter, D.J. Gene-environment interactions in human diseases. *Nat Rev Genet*, 6(4):287–298, 2005. 12

Hutchins, J.R.A., Toyoda, Y., Hegemann, B., Poser, I., et al. Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science*, 328(5978):593–9, 2010. 17

Ingram, G.I. The history of haemophilia. *Journal of Clinical Pathology*, 29(6):469–479, 1976. 5

Iyer-Pascuzzi, A.S., Symonova, O., Mileyko, Y., Hao, Y., et al. Imaging and analysis platform for automatic phenotyping and trait ranking of plant root systems. *Plant Physiol*, 152(3):1148–57, 2010. 15

Jaynes, E.T. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. 21

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37:183–233, 1999. 37, 41, 71, 127

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30(1):42, 2002. 67, 76, 77

Kang, H.M.M., Sul, J.H.H., Service, S.K., Zaitlen, N.A., et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010. 24

Kang, H.M.M., Ye, C., and Eskin, E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–1925, 2008. 34, 35, 52

Komeili, A. and O'Shea, E. Roles of phosphorylation sites in regulating activity of the transcription factor Pho4. *Science*, 284(5416):977, 1999. 81

Krichevsky, A.M., King, K.S., Donahue, C.P., Khrapko, K., et al. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA*, 9(10):1274–81, 2003. 17

Kurimoto, K., Yabuta, Y., Ohinata, Y., and Saitou, M. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat Protoc*, 2(3):739–52, 2007. 15

Lander, E.S. and Botstein, D. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics*, 121(1):185–199, 1989. 5, 35, 123

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. 4

Langridge, G.C., Phan, M.D., Turner, D.J., Perkins, T.T., et al. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Res*, 19(12):2308–16, 2009. 9, 109, 110, 112

Lee, S.I.I., Dudley, A.M., Drubin, D., Silver, P.A., et al. Learning a prior on regulatory potential from eQTL data. *PLoS genetics*, 5(1):e1000358+, 2009. 79, 88

Leek, J.T. and Storey, J.D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet*, 3(9):e161–1735, 2007. 34, 43, 52, 67, 88

Lehner, B., Crombie, C., Tischler, J., Fortunato, A., et al. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet*, 38(8):896–903, 2006. 12

Li, H. and Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010. 100

Liao, J.C., Boscolo, R., Yang, Y., Tran, L.M., et al. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci.*, 100(26):15522–15527, 2003. 67

Liede, A., Karlan, B.Y., and Narod, S.A. Cancer risks for male carriers of germline mutations in BRCA1 or BRCA2: a review of the literature. *J Clin Oncol*, 22(4):735–42, 2004. 1

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., et al. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341, 2009a. 100, 114

Liti, G., Haricharan, S., Cubillos, F.A., Tierney, A.L., et al. Segregating YKU80 and TLC1 alleles underlying natural variation in telomere properties in wild yeast. *PLoS Genet*, 5(9):e1000659, 2009b. 109

Loos, R.J.F., Lindgren, C.M., Li, S., Wheeler, E., et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet*, 40(6):768–75, 2008. 8

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., et al. A global map of human gene expression. *Nat Biotechnol*, 28(4):322–4, 2010. 119

Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*, 362(13):1181–91, 2010. 5

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., et al. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–12, 2004. 11

Lynch, M. and Walsh, B. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, 1 edition, 1998. 72, 124

Mackay, D.J.C. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995. 40, 126

MacKay, D.J.C. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003. 28, 37

Mackay, T.F.C. The genetic architecture of quantitative traits: lessons from Drosophila. *Curr Opin Genet Dev*, 14(3):253–7, 2004. 12

Mackay, T.F.C., Stone, E.A., and Ayroles, J.F. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*, 10(8):565–577, 2009. 18, 90

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., et al. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–85, 2008. 6, 114, 116

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009. 10

Marcus, J.S., Anderson, W.F., and Quake, S.R. Microfluidic single-cell mRNA isolation and analysis. *Anal Chem*, 78(9):3084–9, 2006. 15

McCord, R., Pierce, M., Xie, J., Wonkatal, S., et al. RFM1, a Novel Tethering Factor Required To Recruit the Hst1 Histone Deacetylase for Repression of Middle Sporulation Genes. *Molecular and Cellular Biology*, 23:2009–2016, 2003. 78

McDaniell, R., Lee, B.K., Song, L., Liu, Z., et al. Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. *Science*, 328(5975):235–239, 2010. 14, 17, 21

Mendel, G. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden*, 3(47), 1865. 2

Minka, T.P. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*, volume 17, p. 362–369. 2001. 72

Monks, S.A., Leonardson, A., Zhu, H., Cundiff, P., et al. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet*, 75(6):1094–105, 2004. 19

Montgomery, S.B. and Dermitzakis, E.T. The resolution of the genetics of gene expression. *Human molecular genetics*, 18(R2):R211–215, 2009. 17, 18

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777, 2010. 17

Moreira, I.S., Fernandes, P.A., and Ramos, M.J. Hot spots–a review of the protein-protein interface determinant amino-acid residues. *Proteins*, 68(4):803–12, 2007. 14

Morgan, T. Sex limited inheritance in Drosophila. *Science*, 32(1):120–122, 1910. 3

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., et al. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–747, 2004. 19

Neal, R.M. *Bayesian Learning for Neural Networks*. Springer, 1996. 40, 126

Nelder, J.A. and Wedderburn, R.W.M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. 26

Neumann, B., Walter, T., Hériché, J.K., Bulkescher, J., et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464(7289):721–7, 2010. 17

Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet*, 2009. 8

Ng, P.C. and Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucl. Acids Res.*, 31(13):3812–3814, 2003. 98

Ng, P.C. and Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*, 7:61–80, 2006. 14

Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet*, 2010. 5

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., et al. Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. *PLoS Genet*, 6(4):e1000895+, 2010. 118

Nica, A.C., Parts, L., Glass, D., Nisbet, J., et al. The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genet*, 7(2):e1002003, 2011. 16, 19, 33, 56, 57, 119

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., et al. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet*, 6(4):e1000888+, 2010. 118

Nordborg, M. and Weigel, D. Next-generation genetics in plants. *Nature*, 456(7223):720–723, 2008. 90

O'Conallain, C., Doolin, M., Taggart, C., Thornton, F., et al. Regulated nuclear localisation of the yeast transcription factor Ace2p controls expression of chitinase (CTS1) in Saccharomyces cerevisiae. *Mol Gen Genet.*, 262:275–282, 1999. 81

Parts, L., Cubillos, F., Jain, K., Warringer, J., et al. Revealing the genetic structure of a trait by sequencing a population under selection, 2010. (submitted). 90, 92

Parts, L., Stegle, O., Winn, J., and Durbin, R. Joint Genetic Analysis of Gene Expression Data with Inferred Cellular Phenotypes. *PLoS Genet*, 7(1):e1001276, 2011. 65

Pastinen, T., Ge, B., and Hudson, T.J. Influence of human genome polymorphism on gene expression. *Hum Mol Genet*, 15 Spec No 1, 2006. 34, 62

Patterson, N., Price, A.L., and Reich, D. Population Structure and Eigenanalysis. *PLoS Genetics*, 2(12):e190+, 2006. 24

Paulsson, J. Summing up the noise in gene networks. *Nature*, 427(6973):415–8, 2004. 16

Perlstein, E.O., Ruderfer, D.M., Roberts, D.C., Schreiber, S.L., et al. Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nature genetics*, 39(4):496–502, 2007. 6, 18, 80

Petes, T.D. and Botstein, D. Simple Mendelian inheritance of the reiterated ribosomal DNA of yeast. *Proc Natl Acad Sci U S A*, 74(11):5091–5, 1977. 5

Phillips, P.C. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics*, 9(11):855–867, 2008. 12

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772, 2010. 17

Plagnol, V., Uz, E., Wallace, C., Stevens, H., et al. Extreme Clonality in Lymphoblastoid Cell Lines with Implications for Allele Specific Expression Analyses. *PLoS ONE*, 3(8):e2966+, 2008. 34

Price, A.L., Patterson, N., Hancks, D.C., Myers, S., et al. Effects of cis and trans Genetic Ancestry on Gene Expression in African Americans. *PLoS Genetics*, 4(12):e1000294+, 2008. 64

Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913–918, 2010. 14

Rattray, M., Liu, X., Sanguinetti, G., Milo, M., et al. Propagating uncertainty in microarray data analysis. *Briefings in Bioinformatics*, 7(1):37–47, 2006. 37

Rattray, M., Stegle, O., Sharp, K., and Winn, J. Inference algorithms and learning theory for Bayesian sparse factor analysis. *Journal of Physics: Conference Series*, 197:012002, 2009. 67, 70, 72

Rees, J.L. The genetics of sun sensitivity in humans. *Am J Hum Genet*, 75(5):739–51, 2004. 13

Reimand, J., Kull, M., Peterson, H., Hansen, J., et al. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*, 35(Web Server issue), 2007. 62

Rockman, M.V. Reverse engineering the genotype–phenotype map with natural genetic variation. *Nature*, 456(7223):738–744, 2008. 90

Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 245(4922):1059–65, 1989. 5

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., et al. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol*, 162(4):729–73, 1982. 3

Schadt, E.E., Lamb, J., Yang, X., Zhu, J., et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005. 34, 44, 48, 66

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, 2003. 18

Scherens, B. and Goffeau, A. The uses of genome-wide yeast mutant collections. *Genome Biol*, 5(7):229, 2004. 9

Scherzer, C.R., Eklund, A.C., Morse, L.J., Liao, Z., et al. Molecular markers of early Parkinson's disease based on gene expression in blood. *Proc Natl Acad Sci U S A*, 104(3):955–60, 2007. 16

Segrè, A.V., Murray, A.W., and Leu, J.Y. High-Resolution Mutation Mapping Reveals Parallel Experimental Evolution in Yeast. *PLoS Biol*, 4(8):e256, 2006. 91

SGD project. Saccharomyces Genome Database. World Wide Web electronic publication, 2009. URL `http://www.yeastgenome.org/`. 78

Shao, H., Burrage, L.C., Sinasac, D.S., Hill, A.E., et al. Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. *Proceedings of the National Academy of Sciences*, 105(50):19910–19914, 2008. 12

Sinha, H., Nicholson, B.P., Steinmetz, L.M., and McCusker, J.H. Complex genetic interactions in a quantitative trait locus. *PLoS Genetics*, 2, 2006. 6

Smith, E.N. and Kruglyak, L. Gene-Environment Interaction in Yeast Gene Expression. *PLoS Biology*, 6(4):e83+, 2008. 62, 66, 67, 75, 76, 77, 80, 82, 85, 86

Smith, J., Ramsey, S., Marelli, M., Marzolf, B., et al. Transcriptional responses to fatty acid are coordinated by combinatorial control. *Molecular Systems Biology*, 3, 2007. 79

Smith, M.W., Dean, M., Carrington, M., Winkler, C., et al. Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort (SFCC), ALIVE Study. *Science*, 277(5328):959–65, 1997. 12

Sobreira, N.L.M., Cirulli, E.T., Avramopoulos, D., Wohler, E., et al. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet*, 6(6):e1000991, 2010. 5

Southern, E.M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol*, 98(3):503–17, 1975. 3

Spector, T.D. and Williams, F.M.K. The UK Adult Twin Registry (TwinsUK). *Twin Res Hum Genet*, 9(6):899–906, 2006. 56

Spielman, R.S., Bastone, L.A., Burdick, J.T., Morley, M., et al. Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics*, 39(2):226–231, 2007. 34

Stegle, O., Sharp, K., Winn, J., and Rattray, M. A comparison of inference in sparse factor analysis models. *Journal of Machine Learning Research*, 2009, in preparation. 65, 67, 70, 72, 75, 76

Stegle, O., Kannan, A., Durbin, R., and Winn, J. Accounting for Non-genetic Factors Improves the Power of eQTL Studies. In *Research in Computational Molecular Biology*, pp. 411–422. 2008. 33, 34, 40

Stegle, O., Parts, L., Durbin, R., and Winn, J. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLOS Computational Biology*, 6(5):e1000770, 2010. 33, 67, 88

Steinmetz, L.M., Sinha, H., Richards, D.R., Spiegelman, J.I., et al. Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, 416(6878):326–330, 2002. 6, 9, 97, 114

Storey, J. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 31(6):2013–2035, 2003. 74

Storey, J. and Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440, 2003. 74, 78

Storey, J.D., Akey, J.M., and Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*, 3(8):e267, 2005. 66, 87

Storici, F., Lewis, L.K., and Resnick, M.A. In vivo site-directed mutagenesis using oligonucleotides. *Nat Biotechnol*, 19(8):773–6, 2001. 9

Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., et al. Genome-Wide Associations of Gene Expression Variation in Humans. *PLoS Genet*, 1(6):e78, 2005. 18, 19

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., et al. Population genomics of human gene expression. *Nature Genetics*, 39(10):1217–1224, 2007. 14, 17, 20, 34, 35, 48, 50, 53, 54, 56, 124

Sun, W., Yu, T., and Li, K. Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics*, 23(17):2290, 2007. 88

Takeuchi, F., McGinnis, R., Bourgeois, S., Barnes, C., et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet*, 5(3):e1000433, 2009. 13

Teixeira, M.C., Monteiro, P., Jain, P., Tenreiro, S., et al. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Research*, 34:D3–D5, 2006. 67, 75, 77

The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005. 17, 48

Threadgill, D.W., Hunter, K.W., and Williams, R.W. Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm Genome*, 13(4):175–8, 2002. 7

Tiret, L., Bonnardeaux, A., Poirier, O., Ricard, S., et al. Synergistic effects of angiotensin-converting enzyme and angiotensin-II type 1 receptor gene polymorphisms on risk of myocardial infarction. *Lancet*, 344(8927):910–3, 1994. 12

Towbin, H., Staehelin, T., and Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc Natl Acad Sci U S A*, 76(9):4350–4, 1979. 3

Tsai, I.J., Burt, A., and Koufopanou, V. Conservation of recombination hotspots in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 107(17):7847–7852, 2010. 93

Valdar, W. et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, 38:879–887, 2006. 91

van Nas, A., Ingram-Drake, L., Sinsheimer, J.S., Wang, S.S., et al. Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. *Genetics*, 185(3):1059–68, 2010. 19

Visscher, P.M., Hill, W.G., and Wray, N.R. Heritability in the genomics era–concepts and misconceptions. *Nat Rev Genet*, 9(4):255–66, 2008. 13

Vogel, C., de Sousa Abreu, R., Ko, D., Le, S.Y., et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology*, 6, 2010. 21

Voigt, B., Kuramoto, T., Mashimo, T., Tsurumi, T., et al. Evaluation of LEXF/FXLE rat recombinant inbred strains for genetic dissection of complex traits. *Physiol Genomics*, 32(3):335–42, 2008. 6

Watson, J.D. and Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953. 3

Weedon, M.N., Lango, H., Lindgren, C.M., Wallace, C., et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet*, 40(5):575–83, 2008. 8

Wei, W., McCusker, J.H., Hyman, R.W., Jones, T., et al. Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. *Proceedings of the National Academy of Sciences of the United States of America*, 104(31):12825–12830, 2007. 6

Wenger, J.W., Schwartz, K., and Sherlock, G. Bulk Segregant Analysis by High-Throughput Sequencing Reveals a Novel Xylose Utilization Gene from Saccharomyces cerevisiae. *PLoS Genet*, 6(5):e1000942+, 2010. 91

Williams, R.B., Chan, E.K., Cowley, M.J., and Little, P.F. The influence of genetic variation on gene expression. *Genome research*, 17(12):1707–1716, 2007. 53

Winn, J. and Bishop, C. Variational Message Passing. *Journal of Machine Learning Research*, 6(1):661, 2006. 41

Wykoff, D., Rizvi, A., Raser, J., Margolin, B., et al. Positive feedback regulates switching of phosphate transporters in S. cerevisiae. *Molecular cell*, 27(6):1005–1013, 2007. 78

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–9, 2010. 13

Yuste, R. Fluorescence microscopy today. *Nat Methods*, 2(12):902–4, 2005. 15

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature genetics*, 35(1):57–64, 2003. 18, 80, 82, 85

Zhang, W., Zhu, J., Schadt, E.E., and Liu, J.S. A Bayesian Partition Method for Detecting Pleiotropic and Epistatic eQTL Modules. *PLoS Comput Biol*, 6(1):e1000642+, 2010. 27, 88

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M., et al. Genetic analysis of variation in transcription factor binding in yeast. *Nature*, 464(7292):1187–1191, 2010. 6, 14, 114

Zhu, J., Zhang, B., Smith, E.N., Drees, B., et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics*, 40(7):854–861, 2008. 27, 88

Appendix A

Variational inference in PEER

Supplementary Methods

Implementation of non-Bayesian models

Standard expression QTL model

To ensure a common ground when comparing different methods, we used a well established linear regression approach introduced by Lander and Botstein (1989) to detect associations. For each tested SNP n with genotype $s_{n,j}$ and gene g with expression level $y_{g,j}$, we evaluated the log-odds (LOD) score

$$L_{n,g} = \log \left\{ \prod_j \frac{P(y_{g,j} \,|\, s_{n,j}, \boldsymbol{\theta}_1)}{P(y_{g,j} \,|\, \boldsymbol{\theta}_0)} \right\} = \log \left\{ \prod_j \frac{\mathcal{N}(y_{g,j}; u_{n,j} s_{n,j} + \mu_{g,1}, \sigma^2_{g,1})}{\mathcal{N}(y_{g,j}; \mu_{g,0}, \sigma^2_{g,0})} \right\} \tag{A.1}$$

which assess how well a particular gene expression level is modelled when the observed genetic state $s_{n,j}$ is taken into account, compared to how well it is model-led by a background model ignoring the genetic effect. The probe expression levels $y_{g,j}$ can either be the raw measurements, residuals after subtracting the estimated effect of hidden and known factors, or ranks for a non-parametric statistic.

Significance of an association was evaluated in three different ways:

1. **2-tailed t test on expression values** uses the Student's t distribution with $N-2$ degrees of freedom to assess the significance of the statistic $t = (N-2)^{0.5}\rho(1-\rho^2)^{-0.5}$ based on the correlation coefficient $\rho^2 = 1 - \exp(-2L_{n,g}N^{-1})$ between the genotype and the expression levels. We called an association significant if $|t|$ was greater than the $\frac{10^{-3}}{2S}$ tail of the t_{N-2} distribution, which corresponds to a 10^{-3} Bonferroni-corrected per-gene false positive rate when performing tests for S SNPs.

2. **Rank correlation** uses the same test, but on the ranks of expression values.

3. **Permutation testing** (Lynch and Walsh, 1998) repeats the analysis in Equation (A.1) with permuted expression levels with respect to the genetic state, calculating the distribution of null log-odds scores. An eQTL was called significant if $L_{n,g}$ was greater than $\hat{L}_{n,g}$, the δ tail of the null distribution for a given false positive rate (FPR) δ. The same set of permutations was used for all methods. To account for multiple testing, we estimated a single significance threshold $\hat{L}_g$ per gene for all tested SNPs. This was done by taking the maximum LOD score over SNPs for a given permutation and using this score distribution when estimating the δ tail (Stranger et al., 2007).

The posterior of the switch variable for the probabilistic genetic model is not used for the final tests to put all methods on equal footing.

PEER framework

VBQTL and the alternative compared methods are implemented within the PEER (Probabilistic Estimation of Expression Residuals) framework. Here, we give a full self-contained treatment of the framework and the implemented inference algorithms.

Likelihood models

The likelihood model of PEER for observed expression levels $\mathbf{Y}$ is

$$P(\mathbf{Y}\,|\,\mathbf{Y}^{(1)},\ldots,\mathbf{Y}^{(M)},\boldsymbol{\tau}) = \mathcal{N}(\mathbf{Y}\,|\,\mathbf{Y}^{(1)}+\cdots+\mathbf{Y}^{(M)},\boldsymbol{\Sigma}), \tag{A.2}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}\{\frac{1}{\tau_g}\}$ is the diagonal matrix constructed from noise precisions $\{\tau_g\}$ and $\{\mathbf{Y}^{(m)}\}$ are the contributions of expression variability for each of M models. The noise model is per gene, similar to a factor analysis model, where gamma priors are put on the noise precisions,

$$P(\tau_g) = \Gamma(\tau_g\,|\,a_\tau, b_\tau). \tag{A.3}$$

In experiments we used vague gamma prior parameters, $a_\tau = 1, b_\tau = 100$. Each of the M models itself depends on parameters $\boldsymbol{\theta}^{(m)}$ and possibly other data $\mathcal{D}^{(m)}$

$$P(\mathbf{Y}^{(m)}\,|\,\boldsymbol{\theta}^{(m)},\mathcal{D}^{(m)}). \tag{A.4}$$

Genotype effect model. The expression level $y_{g,j}^{(1)}$ of the gth gene probe in the jth individual is explained by linear effects of genotypes of N SNPs $\mathbf{s}_j = (s_{1,j},\ldots,s_{N,j})$:

$$P(y_{g,j}^{(1)}\,|\,\mathbf{s}_j,\mathbf{b}_g,\mathbf{u}_g,\tau_g) = \mathcal{N}(y_{g,j}^{(1)}\,|\,\sum_{n=1}^{N} b_{n,g}\cdot(u_{n,g}s_{n,j})\,,\frac{1}{\tau_g}) \tag{A.5}$$

$$P(b_{n,g}) = \mathrm{Bernoulli}(b_{n,g}\,|\,p_{\mathrm{ass}}) \tag{A.6}$$

$$P(u_{n,g}) = \mathcal{N}(u_{n,g}\,|\,0,1). \tag{A.7}$$

The weight $\mathbf{u}_g = (u_{1,g},\ldots,u_{N,g})$ indicates the magnitude of the effect, and the binary variables $\mathbf{b}_g = (b_{1,g},\ldots,b_{N,g})$ determine whether it is significant (true) or not (false), taking the Bernoulli prior on the switch variable $P(b_{n,g}) = \mathrm{Bernoulli}(b_{n,g}\,|\,p_{\mathrm{ass}})$ into account. When the switch variable is on, the expression level is linearly influenced by the SNP, and unaffected otherwise. The LOD score of the association model (Section Standard expression QTL model) is closely related to the switch variable $b_{n,g}$. For a particular parameter setting, the posterior probability over the switch state $b_{n,g}$ is a monotonically increasing function of the LOD score. The exact relation is $P(b_{n,g} = 1 \,|\, y_{g,j}, s_{j,n}) = \sigma(\text{LOD score})$ where $\sigma()$ is the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$.

2) Known factor model. The effect of the measured C covariates in the jth individual, $\mathbf{f}_j = (f_{1,j}, \ldots, f_{C,j})$, where the weights of their effect on a gene g is $\mathbf{v}_g = (v_{g,1}, \ldots, v_{g,C})$ is modelled as:

$$P(y^{(2)}_{g,j} \,|\, \mathbf{f}_j, \mathbf{v}_g, \tau_g) = \mathcal{N}(y^{(2)}_{g,j} \,|\, \sum_{c=1}^{C} v_{g,c}\, f_{c,j}, \frac{1}{\tau_g}) \tag{A.8}$$

$$P(v_{g,c} \,|\, \alpha_c) = \mathcal{N}(v_{g,c} \,|\, 0, \frac{1}{\alpha_c}) \tag{A.9}$$

$$P(\alpha_c) = \Gamma(\alpha_c \,|\, a_\alpha, b_\alpha). \tag{A.10}$$

The gamma prior on the inverse covariances for each factor introduces automatic relevance detection (ARD) Mackay (1995); Neal (1996), driving the weights of unused factors to 0 and thereby switching them off. This is explained in more detail below.

3) Hidden factor model. Analogously to known factors, expression variability is modelled by linear effects from K hidden factors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$:

$$P(y^{(3)}_{g,j} \,|\, \mathbf{x}_j, \mathbf{w}_g, \tau_g) = \mathcal{N}(y^{(3)}_{g,j} \,|\, \sum_{k=1}^{K} w_{g,k}\, x_{k,j}, \frac{1}{\tau_g}) \tag{A.11}$$

$$P(w_k, \beta_k) = \prod_{g=1}^{G} \mathcal{N}(w_{g,k} \,|\, 0, \frac{1}{\beta_k}) \tag{A.12}$$

$$P(x_{k,j}) = \mathcal{N}(x_{k,j} \,|\, 0, 1) \tag{A.13}$$

$$P(\beta_k) = \Gamma(\beta_k \,|\, a_\beta, b_\beta). \tag{A.14}$$

The factor activations $\mathbf{X}$ are random variables that are not observed, but instead inferred from the expression levels. Again, the ARD prior allows unused factors to be switched off. This forces the model to learn factors which have a broad effect on many expression levels. In experiments we used values $a_\alpha = 10^{-7}G$ and $b_\alpha = 10^{-1}G$, where G is the total number of gene probes. Similar prior settings were used for the weights of the known factors $\mathbf{v}_c$. We put a standard normal prior on the hidden factors, $x_{k,j} \sim \mathcal{N}(x_{k,j} \,|\, 0, 1)$.

Variational inference

As outlined in Methods we use variational Bayesian inference Jordan et al. (1999) for parameter learning in the framework. The basic principle of variational methods is to approximate the exact joint posterior distribution over all parameters by a factorised Q-distribution. Individual factors of the Q-distribution are refined by minimisation of the KL-divergence between the exact and the approximate distributions with respect to the parameters of a single factor. This leads to an iterative algorithm, updating individual factors of the approximate distribution given the state of all others. Here, we give the factorisations and update rules for the general framework and the individual models.

PEER framework. We approximate the exact joint posterior distribution over all parameters

$$P(\{\mathbf{Y}^{(m)}\}_{m=1}^{M}, \{\boldsymbol{\theta}^{(m)}\}_{m=1}^{M}, \,|\, \mathcal{D}) \tag{A.15}$$

by a factorised approximation over parameters for individual models

$$Q(\boldsymbol{\Theta}) = \prod_{m=1}^{M} Q(\boldsymbol{\theta}^{(m)}) Q(\mathbf{Y}^{(m)}). \tag{A.16}$$

Here we defined the abbreviation $\mathcal{D} = \{\mathbf{Y}, \{\mathcal{D}^{(m)}\}_{m=1}^{M}\}$, summarising all observed data; expression levels $\mathbf{Y}$ as well as model-specific data $\{\mathcal{D}^{(m)}\}_{m=1}^{M}$. Note that as the expression contributions $\mathbf{Y}^{(m)}$ are not observed they also resemble parameters that need to be inferred. Strictly speaking these are not treated as random variables of the model, but Gaussian messages that comprise the first and second moments of the expression variability contribution of a respective model. The distributions of parameters $\boldsymbol{\theta}^{(m)}$ for individual models are in turn factorised. The set $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(M)}\}$ denotes the set of all parameters from all models.

The approximate Q-distributions are updated iteratively, taking the current state of all others into account. Update equations for a particular Q_i can be derived by functional minimisation of the KL-divergence between P and Q with respect to Q_i which leads to

$$\tilde{Q}(\Theta_i) \propto \exp\{\langle \log P(\mathcal{D}, \boldsymbol{\Theta})\rangle_{Q(\Theta_j), i \neq j}\}. \qquad \text{(A.17)}$$

The term in the exponent is the expectation of the model log-likelihood under all other Q-distributions. Together with the expression data likelihood

$$P(\mathbf{Y} \mid \boldsymbol{\Theta}) = \mathcal{N}(\mathbf{Y} \mid \mathbf{Y}^{(1)} + \cdots + \mathbf{Y}^{(M)}, \boldsymbol{\Sigma}) \prod_{m=1}^{M} P(\mathbf{Y}^{(m)} \mid \boldsymbol{\theta}^{(m)}, \mathcal{D}^{(m)}) \qquad \text{(A.18)}$$

this allows generic update rules for all model parameters to be derived. Substituting in Equation (A.16) for each $Q(\cdot)$, we obtain the following approximate distributions:

(Approximate distributions)

$$Q(\boldsymbol{\tau}) = \prod_{g=1}^{G} \Gamma(\tau_g \mid \tilde{a}_{\tau_g}, \tilde{b}_{\tau_g}) \qquad \text{(A.19)}$$

$$Q(\mathbf{Y}^{(m)}) = \prod_{g=1}^{G} \prod_{j=1}^{J} \mathcal{N}(y_{g,j}^{(m)} \mid \tilde{m}_{Y_{g,j}^{(m)}}, \frac{1}{\tilde{\tau}_{Y_{g,j}^{(m)}}}), \qquad \text{(A.20)}$$

and similar factorisations for each of the models (given below). The parameter update equations for the framework parameters follow as:

(Update rules)

$$\tilde{a}_{\tau_g} = a_\tau + \frac{1}{2}\sum_{j=1}^{J}\left\langle\left(y_{g,j} - \sum_{m=1}^{M} y_{g,j}^{(m)}\right)^2\right\rangle \tag{A.21}$$

$$\tilde{b}_{\tau_g} = b_\tau + \frac{J}{2}. \tag{A.22}$$

Genotype effect model The update equations for the models introduced in the main text (Inference) follow similarly. For the models, we give the approximate factorisations employed, and the resulting update equations that are derived in identical manner to the treatment above.

(Approximate distributions)

$$Q(\mathbf{B}) = \prod_{n=1}^{N}\prod_{g=1}^{G}\text{Bernoulli}(b_{n,g}\,|\,\tilde{p}_{b_{n,g}}) \tag{A.23}$$

$$Q(\mathbf{U}) = \prod_{n=1}^{N}\prod_{g=1}^{G}\mathcal{N}(\mathbf{u}_{n,g}\,|\,\tilde{\mathbf{m}}_{\mathbf{u}_{n,g}}, \tilde{\mathbf{\Sigma}}_{\mathbf{u}_{n,g}}) \tag{A.24}$$

(Update rules)

$$\tilde{\Sigma}_{\mathbf{u}_{n,g}} = \mathbf{I} + \langle \tau_g \rangle \left\langle b_{n,g}^2 \right\rangle \sum_{j=1}^{J} \mathbf{s}_{n,j}^{\mathrm{T}} \mathbf{s}_{n,j} \tag{A.25}$$

$$\tilde{\mathbf{m}}_{\mathbf{u}_{n,g}} = \tilde{\Sigma}_{\mathbf{u}_{n,g}}^{-1} \left(\langle \tau_g \rangle \langle b_{n,g} \rangle \sum_{j=1}^{J} \mathbf{s}_{n,j} \left\langle z_{g,j}^{(1)^{\setminus n}} \right\rangle \right) \tag{A.26}$$

$$\tilde{m}_{y_{g,j}^{(1)}} = \sum_{n=1}^{N} \langle b_{n,g} \rangle \langle \mathbf{u}_{n,g} \rangle \mathbf{s}_{n,j} \tag{A.27}$$

$$\tilde{\tau}_{y_{g,j}^{(1)}} = \left[\sum_{n=1}^{N} \left\langle b_{n,g}^2 \right\rangle \left\langle \mathbf{u}_{n,g}^2 \right\rangle \mathbf{s}_{n,j}^2 \right], \tag{A.28}$$

where we define

$$\left\langle z_{g,j}^{(1)^{\setminus n}} \right\rangle = z_{g,j}^{(1)} - \sum_{m \neq n} \langle b_{m,g} \rangle \langle \mathbf{u}_{m,g} \rangle \mathbf{s}_{m,j} \tag{A.29}$$

and the residual expression dataset for the mth model

$$z_{g,j}^{(m)} = y_{g,j} - \sum_{l \neq m}^{M} y_{g,j}^{(l)}. \tag{A.30}$$

$$\tag{A.31}$$

The approximate posterior over the indicator variables can be obtained from

$$\begin{aligned} \tilde{p}_{b_{n,g}} &\propto p_{ass} \cdot \exp \left\{ -\frac{1}{2} \sum_{j=1}^{J} \left\langle \left(z_{g,j}^{(1)^{\setminus n}} - b_{n,g} \mathbf{u}_{n,g} \mathbf{s}_{n,j} \right)^2 \right\rangle \right\} \\ (1 - \tilde{p}_{b_{n,g}}) &\propto (1 - p_{ass}) \cdot \exp \left\{ -\frac{1}{2} \sum_{j=1}^{J} \left\langle \left(z_{g,j}^{(1)^{\setminus n}} \right)^2 \right\rangle \right\}, \end{aligned} \tag{A.32}$$

which after normalisation gives rise to $\tilde{p}_{b_{n,g}}$.

$$\tag{A.33}$$

Known factor model is identical in treatment to the **hidden factor model**, without the need for updates of the factor activations. Thus, we only present the hidden factor model here.

$$\tag{A.34}$$

(Approximate distributions)

$$Q(\mathbf{X}) = \prod_{j=1}^{J} \mathcal{N}(\mathbf{x}_j \mid \tilde{\mathbf{m}}_{\mathbf{x}_j}, \tilde{\Sigma}_{\mathbf{x}_j}) \tag{A.35}$$

$$Q(\mathbf{W}) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{w}_g \mid \tilde{\mathbf{m}}_{\mathbf{w}_g}, \tilde{\Sigma}_{\mathbf{w}_g}) \tag{A.36}$$

$$Q(\boldsymbol{\beta}) = \prod_{k=1}^{K} \Gamma(\beta_k \mid \tilde{a}_{\beta_k}, \tilde{b}_{\beta_k}) \tag{A.37}$$

(Update rules)

$$\tilde{\Sigma}_{\mathbf{x}_j} = \Sigma_{\mathbf{x}_j} + \left\langle \mathbf{W}^T \mathrm{diag}\left(\boldsymbol{\tau}\right) \mathbf{W} \right\rangle \tag{A.38}$$

$$\tilde{\mathbf{m}}_{\mathbf{x}_j} = \tilde{\Sigma}_{\mathbf{x}_j}^{-1} \left\langle \mathbf{W}^{\mathrm{T}} \right\rangle \mathrm{diag} \left\langle \boldsymbol{\tau} \right\rangle \left(\left\langle \mathbf{z}_j^{(3)} \right\rangle \right) \tag{A.39}$$

$$\tilde{\Sigma}_{\mathbf{w}_g} = \mathrm{diag}\left\langle \boldsymbol{\beta} \right\rangle + \left\langle \tau_g \right\rangle \sum_{j=1}^{J} \left\langle \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}} \right\rangle \tag{A.40}$$

$$\tilde{\mathbf{m}}_{\mathbf{w}_g} = \tilde{\Sigma}^{-1}_{\mathbf{w}_g} \left(\left\langle \tau_g \right\rangle \sum_{j=1}^{J} \left\langle \mathbf{x}_j \right\rangle \left(\left\langle \mathbf{z}_j^{(3)} \right\rangle \right) \right) \tag{A.41}$$

$$\tilde{m}_{y^{(3)}_{g,j}} = \sum_{k=1}^{K} \left\langle w_{g,k} \right\rangle \left\langle x_{j,k} \right\rangle \tag{A.42}$$

$$\tilde{\tau}_{y^{(3)}_{g,j}} = \left[\sum_{n=1}^{N} \left\langle b^2_{n,g} \right\rangle \left\langle \mathbf{u}^2_{n,g} \right\rangle \mathbf{s}^2_{n,j} \right] \tag{A.43}$$

$$\tag{A.44}$$

Initialisation. The initial states of hidden factor model weights $Q(\mathbf{w}_g)$ and levels $Q(\mathbf{x}_j)$ are determined from a PCA solution, and the weights for known factors $Q(\mathbf{v}_g)$ are initialised to the maximum likelihood estimate. The parameters for remaining Q distributions for all models are deterministically initialised to corresponding prior means. A random initialisation is possible as well, however, additional computation time is required for multiple restarts, and the inference becomes non-deterministic. We have not explored the implications of this alternative here as the maximum likelihood initialisation performs robustly well in practise.

Bottleneck approximation. The genetic association model accounts for additive association signals from all considered SNPs. The corresponding variational updates of the indicator variables in Equation (A.32) can be unstable in practise. In particular, if multiple correlated SNPs are in association to a single gene, variational learning is prone to being trapped in local optima, attributing the effect to only one of them. Hence, the inferred state of the indicator variables $\mathbf{B}$ depends on the order in which these updates are carried out. To obtain meaningful results, the update sequence needs to be randomised and typically large numbers of restarts are required. This procedure implies prohibitive computational cost, particularly for large datasets. To avoid this additional computation, these updates are instead implemented greedily. For each gene g only a single non-zero entry in the indicator matrix is permitted, corresponding to the SNP with the greatest evidence for an association. This leads to a sparse association matrix $\mathbf{B}$.

VBQTL

Both the iterative (iVBQTL) and the fast variant (fVBQTL) of the studied algorithms use these update equations presented above. iVBQTL uses the full variational approximation with a specific update order of the $Q(\boldsymbol{\theta}_i)$ distributions. In experiments, we used 3 iterations of the full model. Within each full iteration, the genetic model was iterated 3, known factor model 30 and hidden factor model 30 times.

To compare the eQTL detection performance of VBQTL with standard methods and previous studies, we do not directly evaluate the linkage probabilities $P(b_{n,g})$ which are obtained during learning. Instead, we apply the standard association model (Section Standard expression QTL model) on the residuals of the known and unknown factor models after convergence similarly to the traditional methods.

fVBQTL is a faster approximate variant of iVBQTL. Rather than performing full inference in the model, the genetic part of the model is ignored when inferring the parameters for the factor models, which can be cast as a specific update schedule.

Simulation dataset

We simulated 80 diploid individuals with 100 SNPs and 400 probe expression measurements. The simulated minor allele frequency was 0.4 for each SNP, and the allele configuration $s_{n,j}$ of SNP n was encoded as $(1, 0)$, $(1, 1)$, or $(1, 2)$, including a column for the mean. We independently simulated effects of known and hidden factors, as well as genetic associations, noise, and downstream effects. Noise level ψ_g of probe g was drawn from a normal distribution with mean 0 and inverse variance τ_g drawn from $\Gamma(3, 1)$, $\psi_g \sim \mathcal{N}(0, \tau_g^{-1})$. We simulated associations between SNP genotypes and gene expression levels for 1% of the SNP-gene pairs. The genetic weight $\theta_{g,n}$ for an association between probe g and SNP n was drawn from $\mathcal{N}(0,4)$. A total of 10 global factors affecting all gene expression levels were simulated. Individual factor levels $x_{j,k}$ for factor k were drawn from $\mathcal{N}(0,0.6)$. Weights $w_{k,g}$ of factor k for probe g were drawn from $N(0,\sigma_k^2)$, where $\sigma_k^2 \sim 0.8(\Gamma(2.5,0.6))^2$ for a heavy-tailed weight distribution. Three of the 10 simulated global factors were designated as known covariates $f_{c,j}$. Further three probes that had a simulated SNP association were designated to have downstream effects on 30 other probes. The effect of probe g on probe h in individual j was simulated as additive factor of $w'_{g,h}y_{g,j}$, where $w'_{g,h} \sim \mathcal{N}(8,0.8)$ for strong downstream effects, and $y_{g,j}$ is the expression level of probe g in individual j.

Appendix B

Supplementary Tables

Chr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y	**Total**	**FDR**
Probes	1009	644	540	384	449	571	468	338	387	380	545	520	189	330	348	426	549	154	618	266	120	238	328	15	9816	-
CEU																										
Standard	23	21	12	24	14	26	18	12	3	17	21	24	5	16	15	21	35	9	29	8	8	14	7	0	382	2.57 %
fVBQTL	61	69	53	57	45	83	44	36	12	48	61	68	16	41	32	55	82	20	69	29	17	30	23	0	1051	0.93 %
YRI																										
Standard	37	32	23	19	21	42	27	17	9	27	31	30	9	24	16	24	38	12	30	18	8	26	9	0	529	1.86 %
fVBQTL	79	94	75	48	56	91	66	38	17	58	79	65	26	48	48	59	94	22	77	40	19	43	27	0	1269	0.77 %
ASI																										
Standard	36	37	19	28	19	48	30	15	9	24	33	36	10	19	12	24	43	16	42	16	10	19	9	0	554	1.77 %
fVBQTL	91	105	88	55	58	111	73	55	19	59	87	78	31	56	52	61	109	30	96	43	22	37	28	0	1444	0.68 %
pooled																										
Standard	68	77	56	48	42	79	52	32	14	46	48	66	21	39	34	43	82	21	71	31	19	37	19	0	1045	0.94 %
fVBQTL	159	191	158	115	120	202	138	101	36	120	168	159	54	104	96	113	181	51	170	78	33	85	60	4	2696	0.36 %

Table B.1: Number of probes with a *cis* association for individual chromosomes and per-probe false discovery rate for the considered populations (per-probe FPR= 0.100%, Bonferroni corrected for testing multiple SNPs per probe, 2-tailed t test) on raw expression data (Standard) and after accounting for hidden factors (fVBQTL).

		Standard eQTLs			
		CEU (382)	YRI (529)	CHB+JPT (554)	Pooled (1045)
Standard	CEU (382)	382 (100%)	194 (50%)	236 (61%)	356 (93%)
	YRI (529)	194 (36%)	529 (100%)	228 (43%)	409 (77%)
	CHB+JPT (554)	236 (42%)	228 (41%)	554 (100%)	490 (88%)
	Pooled (1045)	356 (34%)	409 (39%)	490 (46%)	1045 (100%)
fVBQTL	CEU (1051)	365 (34%)	282 (26%)	358 (34%)	662 (62%)
	YRI (1269)	276 (21%)	510 (40%)	356 (28%)	675 (53%)
	CHB+JPT (1444)	305 (21%)	322 (22%)	531 (36%)	788 (54%)
	Pooled (2696)	370 (13%)	486 (18%)	527 (19%)	1028 (38%)
		fVBQTL eQTLs			
		CEU (1051)	YRI (1269)	CHB+JPT (1444)	Pooled (2696)
Standard	CEU (382)	365 (95%)	276 (72%)	305 (79%)	370 (96%)
	YRI (529)	282 (53%)	510 (96%)	322 (60%)	486 (91%)
	CHB+JPT (554)	358 (64%)	356 (64%)	531 (95%)	527 (95%)
	Pooled (1045)	662 (63%)	675 (64%)	788 (75%)	1028 (98%)
fVBQTL	CEU (1051)	1051 (100%)	591 (56%)	717 (68%)	1007 (95%)
	YRI (1269)	591 (46%)	1269 (100%)	697 (54%)	1120 (88%)
	CHB+JPT (1444)	717 (49%)	697 (48%)	1444 (100%)	1350 (93%)
	Pooled (2696)	1007 (37%)	1120 (41%)	1350 (50%)	2696 (100%)

Table B.2: Magnitude and fraction of overlap between probes with a **Standard** of **fVBQTL***cis* eQTL respectively, for different populations and methods. Total numbers for each population and method are given in parenthesis after the population. 955 probes had a standard eQTL in some population, and 148 in every population. 2236 probes had a fVBQTL eQTL in some population, and 477 in every population.

Population	1. eQTLs	2. fVBQTLs	3. Pooled eQTLs	2. & 3.	2. - 1.	3. - 1.	(2. - 1.) &(3. - 1.)
CEU	382	1051	871	485	686	582	204
YRI	529	1269	796	476	759	507	188
CHB+JPT	554	1444	709	501	913	378	170

Table B.3: Overlap of VBQTLs in one population (2.) with standard eQTLs found when pooling the other two populations (3.). Overlaps are given both for all QTLs (2. & 3.) and only for additional ones (2. - 1. & 3. - 1.) compared to standard eQTLs in the population. Per-probe eQTL FPR=0.1%, Bonferroni corrected for testing multiple SNPs per probe, 2-tailed t test.

Standard			
Population	CEU (47)	YRI (78)	CHB+JPT (46)
CEU (47)	47 (100%)	18 (38%)	22 (47%)
YRI (78)	18 (23%)	78 (100%)	18 (23%)
CHB+JPT (46)	22 (48%)	18 (39%)	46 (100%)
All populations	13		
> 1 populations	32		
Any population	126		

fVBQTL			
Population	CEU (72)	YRI (87)	CHB+JPT (76)
CEU (72)	72 (100%)	26 (36%)	41 (57%)
YRI (87)	26 (30%)	87 (100%)	31 (36%)
CHB+JPT (76)	41 (54%)	31 (41%)	76 (100%)
All populations	25		
> 1 populations	48		
Any population	162		

Table B.4: Count and percent overlap between probes in *trans* associations on different populations using standard method and after using fVBQTL.

Factor	1	2	3	4	5	6
Gender	0.12	0.16	**-0.81**	0.19	0.08	-0.00
CEU	**0.68**	-0.47	-0.21	-0.04	-0.27	0.04
CHB+JPT	-0.43	0.28	-0.24	**-0.64**	-0.08	0.03
YRI	-0.25	0.19	0.46	**0.69**	0.35	-0.08

Table B.5: Pearson correlation coefficient between top 6 factors learned on the pooled HapMap data, and 4 indicator variables relating to the background of the individual. Correlations with absolute value above 0.6 are highlighted.

Method	K	α	Factors found	Variance explained	*cis* probes	*cis* spec.	*cis* sens.	*trans* probes	*trans* spec.	*trans* sens.
Standard	–	–	0	0.00	20	1.00	1.00	0	0.00	0.00
PCA	5	–	5	0.52	35	0.54	0.95	0	0.00	0.00
PCA	15	–	15	0.70	38	0.45	0.85	0	0.00	0.00
PCA	30	–	30	0.82	29	0.45	0.65	0	0.00	0.00
PCA	60	–	60	0.94	4	0.75	0.15	0	0.00	0.00
PCAsig	–	0.01	7	0.56	37	0.51	0.95	0	0.00	0.00
PCAsig	–	0.1	7	0.56	37	0.51	0.95	0	0.00	0.00
PCAsig	–	0.3	7	0.56	37	0.51	0.95	0	0.00	0.00
SVA	–	0.01	12	0.65	38	0.50	0.95	0	0.00	0.00
SVA	–	0.1	12	0.65	38	0.50	0.95	0	0.00	0.00
SVA	–	0.3	12	0.65	38	0.50	0.95	0	0.00	0.00
fVBQTL	5	–	5	0.52	34	0.59	1.00	0	0.00	0.00
fVBQTL	15	–	15	0.69	51	0.39	1.00	0	0.00	0.00
fVBQTL	30	–	30	0.70	55	0.36	1.00	0	0.00	0.00
fVBQTL	60	–	60	0.70	55	0.36	1.00	0	0.00	0.00
iVBQTL	5	–	5	0.52	34	0.59	1.00	0	0.00	0.00
iVBQTL	15	–	15	0.69	51	0.39	1.00	0	0.00	0.00
iVBQTL	30	–	30	0.70	54	0.37	1.00	0	0.00	0.00
iVBQTL	60	–	60	0.70	54	0.37	1.00	0	0.00	0.00

Table B.6: Summary statistics for method performances on the human chromosome 19 dataset presented in the main text. The parameters for different methods are varied by the number of allowed factors K (PCA, VBQTL) or by the significance cutoff α (PCAsig, SVA). Hidden factor summary is given by the number of factors found and the variance explained by the hidden factor effects. The number of probes with a *cis* and *trans* eQTL, as well as the sensitivity and specificity of recovering probes with a standard eQTL are given. Per-probe eQTL FPR = 0.001, Bonferroni corrected for testing multiple SNPs per probe, 2-tailed t test.

Method	K	α	Factors found	Variance explained	*cis* probes	*cis* spec.	*cis* sens.	*trans* probes	*trans* spec.	*trans* sens.
Standard	–	–	0	0.00	445	1.00	1.00	746	1.00	1.00
PCA	5	–	5	0.28	478	0.77	0.82	501	0.79	0.53
PCA	15	–	15	0.53	481	0.64	0.69	132	0.77	0.14
PCA	30	–	30	0.70	392	0.60	0.53	57	0.75	0.06
PCA	60	–	60	0.86	105	0.66	0.16	5	1.00	0.01
PCAsig	–	0.01	7	0.34	468	0.72	0.76	229	0.80	0.25
PCAsig	–	0.1	7	0.34	468	0.72	0.76	229	0.80	0.25
PCAsig	–	0.3	7	0.34	468	0.72	0.76	229	0.80	0.25
SVA	–	0.01	14	0.52	482	0.65	0.71	144	0.78	0.15
SVA	–	0.1	14	0.52	482	0.65	0.71	144	0.78	0.15
SVA	–	0.3	14	0.52	482	0.65	0.71	144	0.78	0.15
fVBQTL	5	–	5	0.34	547	0.72	0.89	409	0.81	0.45
fVBQTL	15	–	15	0.55	668	0.59	0.88	364	0.80	0.39
fVBQTL	30	–	30	0.62	719	0.54	0.87	349	0.79	0.37
fVBQTL	60	–	60	0.62	722	0.54	0.87	348	0.78	0.37
iVBQTL	5	–	5	0.32	616	0.68	0.95	650	0.76	0.66
iVBQTL	15	–	15	0.50	785	0.54	0.96	694	0.73	0.68
iVBQTL	30	–	30	0.57	821	0.52	0.95	746	0.71	0.71
iVBQTL	60	–	60	0.57	825	0.51	0.95	739	0.71	0.70

Table B.7: Summary statistics for method performances on the yeast dataset presented in the main text. The parameters for different methods are varied by the number of allowed factors K (PCA, VBQTL) or by the significance cutoff α (PCAsig, SVA). Hidden factor summary is given by the number of factors found and the variance explained by the hidden factor effects. The number of probes with a *cis* and *trans* eQTL, as well as the sensitivity and specificity of recovering probes with a standard eQTL are given. Per-probe eQTL FPR = 0.001, Bonferroni corrected for testing multiple SNPs per probe, 2-tailed t test.

Method	K	α	Factors found	Variance explained	*cis* probes	*cis* spec.	*cis* sens.	*trans* probes	*trans* spec.	*trans* sens.
Standard	–	–	0	0.00	560	1.00	1.00	369	1.00	1.00
PCA	5	–	5	0.25	639	0.84	0.96	418	0.76	0.86
PCA	15	–	15	0.48	614	0.82	0.90	409	0.72	0.80
PCA	30	–	30	0.74	708	0.70	0.88	488	0.59	0.78
PCA	60	–	60	0.91	354	0.82	0.52	178	0.76	0.37
PCAsig	–	0.01	12	0.39	601	0.84	0.91	376	0.76	0.77
PCAsig	–	0.1	13	0.41	589	0.85	0.90	371	0.75	0.76
PCAsig	–	0.3	13	0.41	589	0.85	0.90	371	0.75	0.76
SVA	–	0.01	24	0.67	687	0.74	0.91	501	0.58	0.79
SVA	–	0.1	24	0.67	687	0.74	0.91	501	0.58	0.79
SVA	–	0.3	24	0.67	687	0.74	0.91	501	0.58	0.79
fVBQTL	5	–	5	0.32	876	0.63	0.98	590	0.56	0.90
fVBQTL	15	–	15	0.51	1028	0.54	0.99	716	0.46	0.89
fVBQTL	30	–	30	0.67	973	0.56	0.98	657	0.49	0.88
fVBQTL	60	–	60	0.70	932	0.59	0.98	626	0.51	0.87
iVBQTL	5	–	5	0.32	895	0.62	0.99	613	0.55	0.91
iVBQTL	15	–	15	0.51	1036	0.53	0.99	723	0.46	0.90
iVBQTL	30	–	30	0.55	1056	0.52	0.99	729	0.46	0.90
iVBQTL	60	–	60	0.55	1049	0.53	0.99	728	0.45	0.90

Table B.8: Summary statistics for method performances on the mouse dataset presented in the main text. The parameters for different methods are varied by the number of allowed factors K (PCA, VBQTL) or by the significance cutoff α (PCAsig, SVA). Hidden factor summary is given by the number of factors found and the variance explained by the hidden factor effects. The number of probes with a *cis* and *trans* eQTL, as well as the sensitivity and specificity of recovering probes with a standard eQTL are given. Per-probe eQTL FPR = 0.001, Bonferroni corrected for testing multiple SNPs per probe, 2-tailed t test.

Factor	Q-value	mean(LOD$_s$)	Covariate
Oaf1p	5.54E-03	42.9 (r^2=0.30)	Probe
Pdr3p	2.09E-02	14.6	SNP XV 132423
Rtg3p	3.01E-02	21.4	SNP XIV 449639
Reb1p	3.70E-02	41.5	Env
Reb1p	0.00E+00	78.1 (r^2=0.51)	Probe
Thi2p	0.00E+00	52.2	SNP VI 5852
Kar4p	0.00E+00	45.7	SNP V 183958
Hcm1p	0.00E+00	38.9 (r^2=0.29)	Probe
Rpn4p	2.25E-02	56.1	Env
Rpn4p	2.44E-02	35.4 (r^2=0.24)	Probe
Pdc2p	1.84E-02	16.4	SNP XII 611967
Gis1p	4.18E-02	11.9	SNP XV 193911
Ino2p	1.48E-02	11.9	SNP II 603790
Upc2p	2.90E-02	11.7	SNP I 55215
Adr1p	4.98E-02	41.7	Env
Met32p	1.90E-02	15.8	SNP IX 277908
Met32p	1.04E-03	23.4 (r^2=0.19)	Probe
Sum1p	0.00E+00	115.2	SNP XV 838599
Stp1p	1.36E-02	23.6 (r^2=0.19)	Probe
Gcn4p	2.28E-02	66.7	Env
Gcn4p	3.00E-02	72.4 (r^2=0.42)	Probe
Swi4p	6.09E-03	39.7	Env
Spt2p	8.70E-05	34.1	SNP XV 10337
Gat1p	2.44E-02	23.5 (r^2=0.19)	Probe
Hac1p	4.56E-02	20.5	Env
Cdc14p	0.00E+00	42.3	SNP X 307178
Pho4p	2.90E-02	15.5	SNP XIII 28694
Mig1p	5.77E-04	151.3	Env
Mig1p	3.30E-02	51.1 (r^2=0.35)	Probe
Aft1p	3.83E-02	10.9	SNP XV 180210
Hsf1p	2.60E-02	64.3	Env
Hsf1p	3.79E-04	31.1 (r^2=0.24)	Probe
Tos8p	5.79E-03	60.0	Env
Tos8p	1.92E-02	14.7 (r^2=0.12)	Probe
Gts1p	7.33E-03	43.1	SNP V 17399
Yap3p	1.53E-03	21.6	SNP VII 73452
Opi1p	3.24E-02	22.5	SNP V 15817
Stp2p	1.63E-02	70.4	Env
Stp2p	3.41E-02	61.7 (r^2=0.39)	Probe
Rsc30p	1.00E-03	29.7	SNP VIII 221933

Factor	Q-value	mean(LOD$_s$)	Covariate
Rsc30p	4.97E-02	60.7 (r^2=0.41)	Probe
Ste12p	2.22E-02	156.7	Env
Ste12p	4.31E-05	85.1 (r^2=0.51)	Probe
Zap1p	3.67E-02	35.1	Env
Gzf3p	3.63E-02	110.2	SNP III 210748
YJL206C	8.80E-04	46.0	SNP VIII 92978
Cbf1p	3.70E-02	34.7	Env
Put3p	1.47E-02	10.3	Env
Put3p	2.26E-02	7.0 (r^2=0.06)	Probe
Phd1p	2.51E-02	12.9	SNP XIII 46084
Phd1p	6.45E-04	24.5 (r^2=0.19)	Probe
Hap4p	4.84E-02	79.0 (r^2=0.41)	Probe
Abf1p	0.00E+00	52.4	Env
Bas1p	3.46E-02	72.9	SNP IV 289639
Rfx1p	4.78E-02	29.7	Env
Ifh1p	4.61E-02	15.7	Env
Hap1p	0.00E+00	38.7	SNP XII 607076
Hap1p	0.00E+00	96.4 (r^2=0.59)	Probe
Pdr8p	5.93E-03	14.2	SNP XII 27765
Sfp1p	0.00E+00	104.6	Env
Yap1p	0.00E+00	225.2	Env
Yap1p	0.00E+00	84.9 (r^2=0.52)	Probe
Yox1p	0.00E+00	93.6	Env
War1p	8.89E-03	36.5	SNP III 301446
Msn2p	3.35E-02	21.0	SNP XV 154309
Mcm1p	8.37E-03	76.7	Env
Mcm1p	3.28E-02	21.5 (r^2=0.17)	Probe
Fkh2p	4.90E-02	17.7	Env
Fkh2p	4.42E-02	10.5 (r^2=0.09)	Probe
Met4p	2.21E-04	79.0	Env
Met4p	4.77E-02	32.9 (r^2=0.24)	Probe
Sko1p	1.76E-02	36.3	SNP XV 180222
Gcr2p	6.25E-04	22.7	SNP XIV 486861
Gcr2p	4.36E-02	8.2 (r^2=0.07)	Probe
Gis2p	3.79E-02	12.6	SNP XIV 582954
Cin5p	2.35E-02	45.6	Env
Hms1p	3.21E-02	27.3	Env
Sfl1p	0.00E+00	39.1	SNP I 186488
Pip2p	4.34E-02	35.4 (r^2=0.25)	Probe
Usv1p	9.62E-04	41.3	SNP XI 98330

Factor	Q-value	mean(LOD$_s$)	Covariate
Rox1p	4.72E-02	35.5	SNP XIV 449639
Fhl1p	3.76E-02	31.7 (r^2=0.25)	Probe
Arr1p	3.50E-02	111.9	Env

Table B.9: Properties of inferred yeastract factor activations. Q-value and average LOD score of association with SNPs (with best locus) or environment indicator is given for associations with combined Q-value < 0.050

Factor	Q-value	mean(LOD$_s$)	Covariate
Glycolysis / Gluconeogenesis (00010)	4.63E-02	19.9	SNP XIV 486861
Nitrogen metabolism (00910)	0.00E+00	119.9	SNP XII 433955
Lysine biosynthesis (00300)	4.00E-05	25.6	SNP II 479166
Tryptophan metabolism (00380)	0.00E+00	29.2	SNP XV 779974
Arginine and proline metabolism (00330)	0.00E+00	46.7	SNP XV 59733
Aminoacyl-tRNA biosynthesis (00970)	4.50E-02	21.7	SNP XIV 486861
Metabolic pathways (01100)	0.00E+00	393.2	Env
Fatty acid metabolism (00071)	7.66E-03	67.1	SNP I 55329

Table B.10: Properties of inferred kegg factor activations. Q-value and average LOD score of association with SNPs (with best locus) or environment indicator is given for associations with combined Q-value < 0.050

Factor	Q-value	mean(LOD$_s$)	Covariate
Factor 1	0.00E+00	289.5	Env
Factor 4	0.00E+00	19.9	SNP XV 89211
Factor 5	0.00E+00	61.4	SNP XIV 449639
Factor 7	1.96E-02	10.9	SNP XV 446514
Factor 8	2.34E-04	16.2	SNP XII 681096
Factor 9	1.11E-03	15.6	SNP XII 659357
Factor 10	1.32E-03	15.1	SNP XII 672779
Factor 11	0.00E+00	19.2	SNP XII 634225
Factor 12	0.00E+00	17.7	SNP II 506661
Factor 14	6.23E-03	12.6	SNP XI 180221
Factor 15	1.99E-03	14.2	SNP III 76127
Factor 16	1.65E-02	11.2	SNP XIII 404546
Factor 17	2.54E-02	10.1	SNP XV 838599
Factor 18	3.12E-02	9.8	SNP XIII 216022
Factor 19	3.15E-02	9.7	SNP XV 619862
Factor 20	0.00E+00	21.3	SNP II 506661
Factor 21	2.25E-03	13.8	SNP XV 842027
Factor 22	0.00E+00	24.1	SNP V 395442
Factor 23	2.36E-03	14.1	SNP XIII 78655
Factor 24	0.00E+00	18.5	SNP III 75021
Factor 25	1.08E-02	11.5	SNP XV 496730
Factor 26	9.58E-03	11.6	SNP IX 195965
Factor 27	1.98E-02	10.9	SNP II 486640
Factor 28	3.32E-02	9.7	SNP XVI 454307

Table B.11: Properties of inferred freeform factor activations. Q-value and average LOD score of association with SNPs (with best locus) or environment indicator is given for associations with combined Q-value < 0.050

Locus	Factor	Q-value	mean(LOD$_s$)
III 79091	War1p	4.53E-02	26.3
III 79091	Thi2p	9.65E-03	12.3
III 79091	Gzf3p	3.63E-02	110.2
IV 106892	Bas1p	3.46E-02	72.9
IV 106892	Gzf3p	3.90E-02	9.8
IV 106892	Yap3p	1.73E-03	35.2
V 6335	Gts1p	7.33E-03	43.1
V 6335	Opi1p	3.24E-02	22.5
V 6335	Kar4p	0.00E+00	45.7
V 420595	Rsc30p	3.60E-02	10.5
V 420595	Kar4p	0.00E+00	40.5
V 420595	Hap1p	1.89E-02	10.4
V 420595	Sfl1p	3.78E-04	36.4
VII 55458	Gts1p	1.79E-02	13.0
VII 55458	Yap3p	1.53E-03	21.6
VII 449898	Gzf3p	4.24E-02	12.2
VII 449898	Pdr8p	4.52E-02	14.7
XII 611810	Hap1p	0.00E+00	38.7
XII 611810	Pdc2p	1.84E-02	16.4
XII 611810	Pdr8p	2.21E-02	13.2
XIII 46084	Pho4p	2.90E-02	15.5
XIII 46084	Phd1p	2.51E-02	12.9
XIII 46084	Ino2p	4.69E-02	11.3
XIV 449639	Rox1p	4.72E-02	35.5
XIV 449639	Gcr2p	6.25E-04	22.7
XIV 449639	Rtg3p	3.01E-02	21.4
XIV 449639	Gis2p	3.79E-02	12.6
XV 174364	Pdr3p	2.09E-02	14.6
XV 174364	Sko1p	1.76E-02	36.3
XV 174364	Spt2p	8.70E-05	34.1
XV 174364	Aft1p	3.83E-02	10.9
XV 174364	Gis1p	4.18E-02	11.9
XV 174364	Msn2p	3.35E-02	21.0
XV 380725	Gis1p	4.79E-02	9.5
XV 380725	Sum1p	6.18E-03	13.2
XVI 932310	Rsc30p	4.52E-02	14.2
XVI 932310	Sfl1p	2.76E-02	14.8

Table B.12: Associations to loci with more than one yeastract factor association. Q-value and average LOD score are given for all factors associated to each locus.

Factor	Q-value	mean(LOD$_s$)	Covariate
Locus	Factor	Q-value	mean(LOD$_s$)
XIV 486861	Aminoacyl-tRNA biosynthesis (00970)	4.50E-02	21.7
XIV 486861	Glycolysis / Gluconeogenesis (00010)	4.63E-02	19.9

Table B.13: Associations to loci with more than one kegg factor association. Q-value and average LOD score are given for all factors associated to each locus.

Locus	Factor	Q-value	mean(LOD$_s$)
II 486640	Factor 5	1.73E-02	11.1
II 486640	Factor 7	2.00E-02	10.8
II 486640	Factor 8	2.11E-02	10.6
II 486640	Factor 12	0.00E+00	17.7
II 486640	Factor 20	0.00E+00	21.3
II 486640	Factor 27	1.98E-02	10.9
II 697894	Factor 20	3.11E-02	9.8
II 697894	Factor 12	2.01E-03	14.3
III 91287	Factor 8	3.40E-02	9.6
III 91287	Factor 15	1.99E-03	14.2
III 91287	Factor 16	3.51E-02	9.4
III 91287	Factor 17	4.78E-02	8.8
III 91287	Factor 24	0.00E+00	18.5
III 91287	Factor 28	3.49E-02	9.4
V 350744	Factor 14	4.96E-02	8.7
V 350744	Factor 22	0.00E+00	24.1
IX 195965	Factor 25	4.18E-02	9.0
IX 195965	Factor 26	9.58E-03	11.6
IX 195965	Factor 4	3.05E-02	9.8
XII 635380	Factor 4	4.21E-02	9.0
XII 635380	Factor 8	2.34E-04	16.2
XII 635380	Factor 9	1.11E-03	15.6
XII 635380	Factor 10	1.32E-03	15.1
XII 635380	Factor 11	0.00E+00	19.2
XII 635380	Factor 12	1.50E-03	14.9
XII 635380	Factor 23	2.53E-02	10.0
XIII 28622	Factor 18	3.12E-02	9.8
XIII 28622	Factor 23	2.36E-03	14.1
XIII 28622	Factor 7	2.56E-02	10.1
XIV 418269	Factor 5	0.00E+00	61.4
XIV 418269	Factor 30	3.37E-02	9.6
XIV 418269	Factor 8	1.67E-03	14.7
XV 96633	Factor 18	4.94E-02	8.7
XV 96633	Factor 4	0.00E+00	19.9
XV 96633	Factor 5	2.38E-02	10.3
XV 96633	Factor 24	9.55E-03	11.6
XV 838599	Factor 17	2.54E-02	10.1
XV 838599	Factor 21	2.25E-03	13.8

Table B.14: Associations to loci with more than one freeform factor association. Q-value and average LOD score are given for all factors associated to each locus.

Locus	Chr	Pos.	1. Probes with *trans* associations	2. Probes with downstream factor associations	3. (2.) with stronger factor association	1.&2	$\frac{1.\&2.}{1.}$	$\frac{1.\&3.}{1.}$	$\frac{1.\&3.}{1.\&2}$
AMN1	2	555575	51	73	73	3	0.06	0.06	1.00
HAP1	12	644082	66	53	53	31	0.47	0.47	1.00
PHO84	13	46084	31	454	454	11	0.35	0.35	1.00
MKT1	14	449639	218	514	508	21	0.10	0.07	0.71
IRA2	15	174364	271	1443	1438	164	0.61	0.59	0.97

Table B.15: *trans* eQTL peaks with at least 50 associations. For each peak, the number of significant associations to probe expression levels (1.), number of associations for Yeastract factor activations significantly associated with the peak (2.), number of genes more strongly associated with the factor than the peak locus genotype (3.) are given, together with the number and fraction of *trans* eQTLs explained by the factors, fraction of *trans* eQTLs more strongly associated with the factor, and fraction of *trans* eQTLs associated with a factor that are more strongly associated with the factor.

Sample	Generation	Replica	Type	Ploidy	Condition	Timepoint	Coverage
WA-NA_Initial_R1_F6_T0	6	1	Pool	Haploid	Permissive	0	23.8
WA-NA_Initial_R2_F6_T0	6	2	Pool	Haploid	Permissive	0	13.1
WA-NA_Heat_R1_F6_T4	6	1	Pool	Haploid	Heat 40C	2	19.3
WA-NA_Heat_R2_F6_T4	6	2	Pool	Haploid	Heat 40C	2	25.7
WA-NA_Initial_R1_F6_S1	6	1	Segregant	Haploid	Permissive	0	20.3
WA-NA_Initial_R2_F6_S1	6	2	Segregant	Haploid	Permissive	0	27.4
WA-NA_Mock_R1_F12_T4	12	1	Pool	Haploid	Permissive	2	115.4
WA-NA_Heat_R1_F12_T4	12	1	Pool	Haploid	Heat 40C	2	129.3
WA-NA_Mock_R2_F12_T4	12	2	Pool	Haploid	Permissive	2	105.7
WA-NA_Initial_R2_F12_T0	12	2	Pool	Haploid	Permissive	0	107.3
WA-NA_Heat_R2_F12_T2	12	2	Pool	Haploid	Heat 40C	1	54.8
WA-NA_Heat_R2_F12_T4	12	2	Pool	Haploid	Heat 40C	2	83.7
WA-NA_Heat_R2_F12_T6	12	2	Pool	Haploid	Heat 40C	3	65.9
WA-NA_Diploid-heat_R2_F12_T6	12	2	Pool	Diploid	Heat 40C	3	32.6
WA-NA_Diploid-heat_R1_F12_T4	12	1	Pool	Diploid	Heat 40C	2	88.6
WA-NA_Paraquat_R1_F12_T4	12	1	Pool	Haploid	Paraquat	2	150

Table B.16: Average sequencing coverage at segregating sites for different intercross generations, ploidies, conditions, and selection timepoints.

Chromosome	Location	Combined change (R1 + R2)
1	11998	0.38
1	207560	0.29
2	472111	0.33
4	1444248	-0.26
4	373030	0.3
4	430662	0.35
4	474894	0.39
4	572931	0.53
4	700611	-0.35
7	1081499	-0.59
8	261643	0.28
9	77497	0.27
10	420908	0.27
10	450702	0.26
10	492479	0.26
10	613016	0.45
12	388635	-0.38
12	491120	-0.28
12	967942	-0.35
14	49576	0.3
15	184627	0.39
15	580877	-0.28

Table B.17: Regions selected for during intercross rounds between F6 and F12 generations.

Chromosome	Location	Combined allele frequency change (R1 + R2)
1	119382	0.31
2	472031	-0.52
2	517350	-0.68
4	1313885	0.42
4	454021	-0.31
4	496586	-0.3
7	131690	0.3
7	859960	0.83
9	292345	-0.32
10	234117	-0.39
10	420908	-0.42
10	679911	-0.28
12	140165	0.38
12	730764	-0.28
13	743221	-0.27
13	893719	-0.56
14	480623	0.46
15	1032447	-0.76
15	179760	-1.27

Table B.18: Heat QTLs detected with artificial selection. All loci with total allele frequency change of at least 0.3, and at least 0.1 in both replicas are given.

Gene	F12 T4	F12 T0	Change T0 - T4
Q0045	0.4	36.5	-36.1
Q0250	2	37	-35
Q0255	1.7	36.1	-34.4
Q0060	0.3	31.9	-31.6
Q0115	1.2	32	-30.8
Q0275	1.8	32.3	-30.5
Q0105	3.3	32.8	-29.5
Q0050	0.2	27.7	-27.5
Q0120	1.6	28.3	-26.7
Q0070	0.2	26	-25.8
Q0085	1.9	26.1	-24.2
Q0065	0.2	22	-21.8
Q0182	0.7	18.3	-17.6
Q0032	0.9	12.3	-11.4
Q0142	0.3	11.3	-11
YLR162W	44.5	55.4	-10.9
Q0140	3.3	13.2	-9.9
Q0130	2.7	11.4	-8.7
Q0144	2.2	10.8	-8.6
Q0143	0.7	7.9	-7.2
Q0080	0.1	6.2	-6.1
YDR366C	11.5	17.6	-6.1
Q0110	0.7	6	-5.3
Q0010	13.7	18	-4.3
Q0092	0	3.5	-3.5
Q0017	0.1	2.5	-2.4
YEL074W	4.1	5.9	-1.8
YIR044C	1.1	2.9	-1.8
YIL174W	0.7	1.9	-1.2
YJL225C	2.1	3.3	-1.2
YNL337W	1.6	2.8	-1.2
YOL166C	1.6	2.8	-1.2
YHR216W	3.4	4.4	-1
YLR465C	2.6	0.9	1.7
YDR340W	8.3	3.9	4.4

Table B.19: Genes changing in copy number upon selection.