Probabilistic Latent Variable Models in Statistical Genomics

Nicoló Fusi

Department of Computer Science

University of Sheffield

A thesis submitted for the degree of

Doctor of Philosophy

August 2014

Abstract

In this thesis, we propose different probabilistic latent variable models to identify and capture the hidden structure present in commonly studied genomics datasets. We start by investigating how to correct for unwanted correlations due to hidden confounding factors in gene expression data. This is particularly important in expression quantitative trait loci (eQTL) studies, where the goal is to identify associations between genetic variants and gene expression levels. We start with a naïve approach, which estimates the latent factors from the gene expression data alone, ignoring the genetics, and we show that it leads to a loss of signal in the data. We then highlight how, thanks to the formulation of our model as a probabilistic model, it is straightforward to modify it in order to take into account the specific properties of the data. In particular, we show that in the naïve approach the latent variables "explain away" the genetic signal, and that this problem can be avoided by jointly inferring these latent variables while taking into account the genetic information. We then extend this, so far additive, model to additionally detect interactions between the latent variables and the genetic markers. We show that this leads to a better reconstruction of the latent space and that it helps dissecting latent variables capturing general confounding factors (such as batch effects) from those capturing environmental factors involved in genotype-by-environment interactions. Finally, we investigate the effects of misspecifications of the noise model in genetic studies, showing how the probabilistic framework presented so far can be easily extended to automatically infer non-linear monotonic transformations of the data such that the common assumption of Gaussian distributed residuals is respected.

Declaration of Authorship

I, Nicoló Fusi, declare that this thesis titled "Probabilistic Latent Variable Models in Statistical Genomics" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Acknowledgements

First of all, I would like to thank Neil Lawrence for his guidance, inspiration and constant support throughout my PhD. His ideas and insights have shaped the way I think about research and will surely accompany me into the future. This thesis would have not been possible without him. Also thanks to Marta Milo for being, together with Neil, part of my extended family in Sheffield. She has been instrumental in supporting my first steps in computational biology and in learning how to prepare truly Neapolitan pizza.

I would also like to thank Oliver Stegle for extremely productive and inspiring collaborations. Most of the material presented in this thesis is the result of our work together.

I would like to thank James Hensman for all the interesting discussions and pair programming sessions we had over the years. I learned a great deal from him and it was always fantastic to arrive in the lab every morning to hear about the exciting problems he was working on. Thanks to Andreas Damianou for all the insights on latent variable models and on the correct procedure to prepare Greek cafe frappe. Also thanks to the current and past members of the groups in Sheffield and Manchester: Richard Allmendinger, Mauricio Alvarez, Ricardo Andrade, Arjun Chandra, Teo De Campos, Nicholas Durrande, Peter Glaus, Antti Honkela, Ciira Maina, Jens Nielsen, Alfredo Kalaitzis, Jaakko Peltonen, Adam Pocock, Arif Rahman, Jon Roberts, Kevin Sharp, Michalis Titsias, Alessandra Tosi, Manuela Zanda and Max Zwiessele. Also thanks to Magnus Rattray for all his helpful ideas and suggestions over the years.

During my internship at Microsoft Research I had the pleasure to work with Christoph Lippert, Jennifer Listgarten, Jonathan Carlson and David Heckerman. I would like to thank all of them.

Finally I would like to thank my parents Marina and Roberto, and my sister Viola for their never ending support. This thesis is dedicated to them and to my grandmother Ada, who would have loved to see this thesis finally submitted.

Contents

List of Figures

Chapter 1

Introduction

In recent years, technological advantages in high-throughput genotyping have allowed researchers to measure with increasing precision thousands to millions of common and rare genetic variants. At the same time, advances in high-throughput sequencing of molecular traits and the digitization of clinical charts have greatly increased the number of phenotypes that can be investigated.

Despite the wealth of measurements available, transforming all the available data into useful biological knowledge is still challenging and there's a significant demand for advanced methods that can successfully incorporate all the information available. This is particularly true for genetic association studies, which are the main focus of this thesis.

The objective of genome-wide association studies (GWAS) is to find a link between changes in the genotype (single nucleotide polymorphisms, SNPs) and changes in the phenotype of a set of individuals. This apparently simple task is complicated by the vast number of potential associations, the underlying sparsity of the set of causal associations, and by the relatively small sample sizes of many current studies. For this reason, in recent years there has been interest in gathering larger datasets with thousands of individuals, with the aim of eventually analyzing millions of individuals. While this has helped greatly in increasing statistical power, it has also generated new modelling challenges due to the introduction of additional structured (e.g. non i.i.d) noise in the data.

Two of the main main sources of structured noise in GWAS are population structure and environmental factors. Both of these confounders introduce cor-

relation between individuals, violating the assumption of independence across samples in GWAS and producing a loss of power and an increase in the number of false positives. In the case of population structure, this correlation is due to the shared genetic background between two individuals belonging to the same family or population. In the case of environmental factors, this correlation is due to exposure to similar environments, such as cigarette smoke, pollution or diet.

Another common assumption in GWAS is that the noise is Gaussian distributed. This is not always true on real world datasets, so it's common practice to apply transformations to phenotypes to make them as Gaussian as possible. For instance, if the scale of the phenotype spans several orders of magnitude, it is common to apply a log-transformation as a preprocessing step and perform genetic analyses on this new scale. Log transformations can also be appropriate when the phenotypic measurement is defined as the ratio between a foreground and a background signal, such as in gene expression measurements from microarrays or when analyzing composite phenotypes (e.g. the ratio between total cholesterol and high density lipoprotein). Nonetheless, the set of transformations that are being used in genetic studies goes far beyond just log transformations and no single transformation can be considered a universal solution.

In the rest of this thesis we are going we are going to propose novel methods to tackle these problems using a specific family of probabilistic models called latent variable models.

1.1 Probabilistic latent variable models

Latent variable models are a popular class of mathematical models that aims to extract the hidden structure present in a data set [Bishop, 1998]. The idea behind these models is that there are some latent factors, either continuous or discrete, that influence the observable variables, thus introducing correlations betweeen them.

Given some observed data $\mathbf{Y} \in \mathcal{R}^{\mathrm{N}\times\mathrm{D}}$, the goal of a latent variable model is to embed these observations in a lower dimensional space $\mathbf{X} \in \mathcal{R}^{\mathrm{N}\times\mathrm{Q}}$ where $\mathrm{Q} < \mathrm{D}$. From a probabilistic standpoint, this is equivalent to expressing the distribution over the observed variables $p(\mathbf{Y})$ using a smaller number of latent

variables $\mathbf{X}$. Following the presentation in Bishop [1998], we start from the joint distribution $p(\mathbf{Y}, \mathbf{X})$ and refactor it in terms of the marginal distribution over the latent variables $p(\mathbf{X})$ and and the conditional distribution $p(\mathbf{Y} \mid \mathbf{X})$. Assuming that the conditional distribution factorizes across dimensions, we have

$$p(\mathbf{Y}, \mathbf{X}) = p(\mathbf{X}) \prod_{j=1}^{\mathrm{D}} p(\mathbf{y}_{:,j} \mid \mathbf{X}). \tag{1.1}$$

This factorization property is really an assumption of conditional independence, and it's equivalent to saying that the observed variables $\mathbf{y}_{:,1}, \ldots, \mathbf{y}_{:,j}$ are independent given $\mathbf{X}$. Sometimes this assumption is not true and is necessary to adjust the model accordingly, for instance by conditioning on other relevant variables (see for example Chapters 2 and 3, where we condition both on latent factors and on observed genetic data). Next, we express $p(\mathbf{Y} \mid \mathbf{X})$ as a noisy mapping from the latent space to the observed space, or equivalently

$$\mathbf{Y} = f(\mathbf{X}; \mathbf{W}) + \boldsymbol{\epsilon}, \tag{1.2}$$

where $f(\mathbf{X}; \mathbf{W})$ is a function of the latent variables with parameters $\mathbf{W}$ and $\boldsymbol{\epsilon}$ is a random noise term. The definition of the model can then be completed by specifying the prior distribution over the noise term $p(\boldsymbol{\epsilon})$, the latent variables $p(\mathbf{X})$ and the mapping function $f(\mathbf{X}; \mathbf{W})$.

Interestingly, many popular dimensionality reduction techniques can be cast under this framework by simply choosing different probabilities distributions and mapping functions. For instance, principal component analysis (PCA), a dimensionality reduction technique which seeks a lower dimensionally embedding where the projected variance of the data is maximized [Bishop, 1998; Hotelling, 1933] can be interpreted as a probabilistic latent variable model (probabilistic PCA, PPCA). In PPCA [Roweis and Ghahramani, 1999; Tipping and Bishop, 1999] the mapping $f(\mathbf{X}; \mathbf{W})$ is chosen to be linear so that

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \boldsymbol{\epsilon}. \tag{1.3}$$

The noise model is drawn from $\mathrm{N}(\mathbf{0}, \sigma^2\mathrm{I})$ and the prior over the latent variables is chosen to be a standard multivariate Gaussian $\mathrm{N}(\mathbf{0}, \mathrm{I})$. Factor analysis [Basilevsky, 2009; Knott and Bartholomew, 1999] can also be presented in a similar way by allowing the noise distribution to be non-isotropic ($\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a diagonal matrix).

In this thesis, we are going to focus mainly on two methods for latent variable modelling: Gaussian process latent variable models (GP-LVMs) [Lawrence, 2005] and warped Gaussian processes [Snelson et al., 2004]. In GP-LVMs, a Gaussian process prior is placed over the function $f(\mathbf{X}; \mathbf{W})$, resulting in Gaussian process mappings from a latent space $\mathbf{X}$ to an observed data space $\mathbf{Y}$. If the GP prior is chosen to be linear, the resulting model is equivalent to probabilistic PCA; if it's not linear, the model can be used to perform non-linear dimensionality reduction. Similarly to GP-LVMs, warped Gaussian processes also allow non-linear functions $f(\mathbf{X}; \mathbf{W})$, but instead of choosing a GP prior, they assume a specific parametric form for the mapping function.

1.2 Genome-wide association studies

Throughout this thesis, we focus our attention on genome-wide association studies (GWAS). In this type of study, the strength of a potential relationship between a single nucleotide polymorphism and a phenotype is quantified using a statistical model.

Given a phenotype $\mathbf{y} \in \mathcal{R}^{\mathrm{N}\times 1}$ and genotypes $\mathbf{S} \in \mathcal{R}^{\mathrm{N}\times K}$, the simplest approach that can be used to assess this relationship is linear regression. In this model, the phenotype is seen a linear function of the genotype corrupted by noise. For an individual n and a single nucleotide polymorphism k, the phenotype y_n is given by

$$y_n = \mu + s_{n,k} v_k + \epsilon_n, \tag{1.4}$$

where μ is a bias term shared across samples, v_k is a regression weight and ϵ_n is noise independently sampled from a Gaussian distribution with variance σ^2. The likelihood of this model can be written as

$$P(\mathbf{y} \mid \mathbf{S}) = \prod_{n=1}^{N} \mathrm{N}(y_n \mid \mu + s_{n,k} v_k, \sigma^2). \tag{1.5}$$

To assess the strength of the association between each SNP and the phenotype, the model just described is compared to a model that assumes that the SNP has no effect on $\mathbf{y}$ ($v_k = 0$):

$$P(\mathbf{y}) = \prod_{n=1}^{N} \mathrm{N}(y_n \mid \mu, \sigma^2). \tag{1.6}$$

Sections 2.2.4 and 3.2.3 provide more details on how the model comparison and hypothesis testing are performed.

1.2.1 Expression Quantitative Trait Loci Studies

Expression quantitative trait loci (eQTL) studies are a particular type of GWAS where the phenotype consists of gene expression levels. The aim in this case is to identify which genetic variants lead to changes in expression levels between different individuals. In the simplest case, it's possible to use the same linear model with Gaussian noise presented in the last section, with the only difference being the fact that the target variable is not a vector of size N but rather a matrix of size $N \times D$, where D is the number of genes.

$$P(\mathbf{Y} \mid \mathbf{S}) = \prod_{d=1}^{D} \prod_{n=1}^{N} \mathrm{N}(y_{n,d} \mid \mu + s_{n,k} v_{k,d}, \sigma^2). \tag{1.7}$$

This simple model makes two independence assumptions. First, the SNPs are typically treated as independent, even if in reality they are correlated for instance because of linkage disequilibrium and population structure. This assumption is often reasonable in practice [Kang et al., 2010; Lippert et al., 2011], especially if the task is to simply identify *associated* variants, rather than identifying *causal* variants, predicting risk or performing heritability estimation. The second assumption, the independence of the noise across individuals, is often not valid in practice. This is mostly due to the fact that gene expression levels are easily influenced by a multitude of non-genetic factors such as environmental effects (diet, lifestyle, etc.) [Balding et al., 2008; Johnson et al., 2007] and techical effects (lab conditions, type of reagents, etc.) [Locke et al., 2003; Plagnol et al., 2008]. These factors cause the gene expression levels of groups of genes to be (or appear to be) jointly upregulated or downregulated. In turn, these causes different samples to be correlated through the, often unknown, factor that caused such a change in the gene expression levels. To better understand this point, imagine to have a cohort of 10 patients, 5 of which are vegetarians and 5 of which are not. If we analyzed their gene expression levels from peripheral blood and computed the correlation between each pair of individuals based on their gene expression levels, it's likely that we would find that pairs of individual that are both vegeterians are more correlated than pairs composed of vegeterians and non-vegetarians. If we don't have any information about the diet of the patients we are analyzing (i.e. their diet is a *latent variable*), this will act as a source of structured (i.e. non-diagonal) noise that violates the assumption of indipendence across samples.

One way to account for the confounding influence of these unobserved factors is to exploit the fact that they affect multiple gene expression levels at once. Indeed, the approaches proposed so far to correct for confounding factors in eQTL studies can be broadly grouped in two categories: approaches that are based on linear mixed models and condition on all the measured expression levels [Kang

et al., 2008a; Listgarten et al., 2010], and approaches that are based on latent variable models and estimate these latent variables from the gene expression levels [Fusi et al., 2012, 2013; Leek and Storey, 2007; Stegle et al., 2010, 2012].

Two prominent examples of models belonging to the first category are ICE [Kang et al., 2008a] and eLMM [Listgarten et al., 2010]. They are both based on linear mixed models and the basic idea is to go from the model with just a fixed effect (the effect of the SNP) and diagonal noise:

$$P(\mathbf{Y} \mid \mathbf{S}) = \prod_{d=1}^{D} \mathrm{N}(\mathbf{y}_d \mid \mu + \mathbf{s}_k v_{k,d}, \sigma^2 \mathbf{I}), \tag{1.8}$$

to a mixed model with the same fixed effect and a random effect $\mathbf{K} = \mathbf{Y}\mathbf{Y}^\top$ that is obtained by conditioning on all the genes

$$P(\mathbf{Y} \mid \mathbf{S}) = \prod_{d=1}^{D} \mathrm{N}(\mathbf{y}_d \mid \mu + \mathbf{s}_k v_{k,d}, \mathbf{K} + \sigma^2 \mathbf{I}). \tag{1.9}$$

One drawback of these models, examined in more detail in Chapter 2, is that in the case of extensive genetic co-regulation of groups of genes, the choice of conditioning on all the gene expression levels can result in explaining away most of the genetic signal present in the data.

An alternative modelling approach consists in trying to explicitly reconstruct the unobserved confounding factors using latent variable models. Examples of methods belonging to this category include SVA [Leek and Storey, 2007], PEER [Stegle et al., 2010, 2012], PANAMA (Chapter 2) and LIMMI (Chapter 3).

The simplest of these models, SVA, is based on a principal component analysis of the gene expression levels. In probabilistic terms, SVA can be summarized as

$$P(\mathbf{Y} \mid \mathbf{S}) = \prod_{d=1}^{D} \mathrm{N}(\mathbf{y}_d \mid \mu + \mathbf{s}_k v_{k,d} + \mathbf{X}\mathbf{w}_d, \sigma^2 \mathbf{I}), \tag{1.10}$$

where $\mathbf{X} \in \mathcal{R}^{\mathrm{N} \times \mathrm{Q}}$ is a matrix of latent variables and $\mathbf{W} \in \mathcal{R}^{\mathrm{Q} \times \mathrm{D}}$ is a matrix of

regression weights. PEER is very similar to SVA, but rather than being based on PCA, it's based on factor analysis. One problem with these two models is that in estimating the latent variables, they only use the gene expression levels. Again, in the case of extensive genetic co-regulation, these models are likely to mistake genetic signal for confounding noise. This happens because they ignore all the genetic information while estimating the latent variables. PANAMA and LIMMI (Chapters 2 and 3) solve this problem by conditioning on the genetic information while estimating the latent variables. The difference between PANAMA and LIMMI is that while the first is an additive linear model, the second one additionally accounts for multiplicative interactions between the estimated environmental effects and the genetic variables.

1.3 Outline

The focus of chapters 2 and 3 is on eQTL studies. In chapter 2 we propose an approach for estimating and correcting for hidden confounders, leading to a remarkable increase in power to detect associations. Importantly, we propose joint model that takes into account prominent genetic regulators while estimating the latent variables, and thus avoids "explaining away" genetic signal using the latent variables.

In Chapter 3 we consider the problem of identifying interactions between the genotype and the phenotype that have a regulatory effect on gene expression levels. While this can be done with existing methods, these approaches require a complete control of the environment and careful experimental design. Given that it's extremely difficult to completely control the environmental factors of human subjects, these requirements that can really be fully respected only when considering model organisms. For this reason, we use the insights gained in Chapter 2 to estimate unmeasured or unknown environmental factors from the gene expression alone. While in Chapter 2 the emphasis was on *correcting* for the effect of both the environment and batch effects, in Chapter 3 we focus on *estimating* the environmental component and identifying interactions between these hidden factors and the genotype. As shown in the experiments, our method is able to accurately reconstruct environmental factors and their interactions with

genotype in a variety of settings. In particular, in real data from yeast, our results suggest that interactions with both known and unknown environmental factors significantly contribute to gene expression variability.

In Chapter 4 we focus our attention on genome-wide association studies on univariate phenotypes. One of the fundamental assumptions of all the models typically used in association studies is that the residuals are Gaussian distributed. Here, we show that this leads to significant losses of power in genome-wide association studies and biases in parameter estimation, leading to wrong heritability estimates. Typical approaches to mitigate this problem consisted in performing a pre-processing transformation of the phenotypic data (e.g. applying a log-transform). However, choosing a "good" transformation is challenging because of the need to manually define a set of transformations, and then try each one out, without any objective way of selecting one over the other. In Chapter 4 we comprehensively address this important problem by introducing a principled statistical model to infer these transformations from the data itself. In extensive synthetic and real experiments, we find up to twofold increases in GWAS power, reduced bias in heritability estimation of up to 30%, and significantly increased accuracy in phenotype prediction.

1.4 Software

Scientific publications are only part of the expected output of a research project, in particular when the aim is to produce novel methods to be used by other scientists. For this reason, we made all the software and related resources available to other researchers. In particular, during the development of the methods described in this thesis we have contributed to the development of a general purpose Gaussian process library (GPy) which is freely available online at https://github.com/SheffieldML/GPy/. Implementations of the methods described in Chapters 2 and 3 are available in online source control repositories (https://github.com/PMBio/envGPLVM) and on the python package index (https://pypi.python.org/pypi/panama), where they have been downloaded on average more than 700 times every month. An implementation of the method described in Chapter 4 and all of the analysis scripts are available online

(https://github.com/PMBio/warpedLMM).

Chapter 2

Joint modelling of confounding factors and genetic regulators

The material presented in this chapter is joint work with Oliver Stegle and Neil Lawrence, and has been published in *"Joint Modelling of Confounding Factors and Prominent Genetic Regulators Provides Increased Accuracy in Genetical Genomics Studies"* [Fusi et al., 2012].

2.1 Overview

Genome-wide analysis of the regulatory role of polymorphic loci on gene expression has been carried out in a range of different study designs and biological systems. For example, association mapping in human has uncovered an abundance of associations between a gene and neighboring SNPs (also known as *cis* associations) that contribute to the variation of a third of all human genes [Stegle et al., 2010; Stranger et al., 2007]. In segregating yeast strains, linkage studies have revealed extensive genetic regulation controlled by SNPs far away from the gene being regulated (also known as *trans* associations), with a few regulatory hotspots controlling the expression profiles of tens or hundreds of genes [Brem et al., 2002; Smith and Kruglyak, 2008].

Despite the success of such expression quantitative trait loci (eQTL) studies, it has also become clear that the analysis of these data comes along with

non-trivial statistical hurdles [McCarthy et al., 2008]. Different types of external confounding factors, including environment or technical influences, can substantially alter the outcome of an eQTL scan. Unobserved confounders can both obscure true association signals and create new spurious associations that are false [Kang et al., 2008a; Leek and Storey, 2007].

Suitable data preprocessing, or careful design of randomized studies are helpful measures to avoid confounders in the first place [Churchill, 2002], however they rarely rule out confounding influences entirely. It is also relatively straightforward to account for those factors that are known and measured. For example, it is standard procedure to include covariates such as age and gender in the analysis [Balding et al., 2008; Johnson et al., 2007]. Similarly, the effect of populational relatedness between samples, a confounding effect that is observed or can be reliably estimated form the genotype data [Kang et al., 2008b, 2010], is usually included in the model. However, other factors, including subtle environmental or technical influences, often remain unknown to the experimenter, but still need to be accounted for. Their potential impact has previously been characterized in multiple studies; for example Plagnol et al. [Plagnol et al., 2008] and Locke et al. [Locke et al., 2003] showed that virtually any aspect of sample handling can impact the analysis.

The goal of this chapter is to present an integrated probabilistic model, PANAMA, to address these shortcoming of established approaches. PANAMA learns a dictionary of confounding factors from the observed expression profiles while accounting for the effect of loci with a pronounced *trans* regulatory effect, thereby avoiding overlaps between true genetic association signals and the covariance structure induced by the learnt confounders. As shown in sections 2.4 and 2.3, this results in a remarkable improvement in accuracy in the detection of both *cis* and *trans* effects.

The rest of the chapter is organized as follows. In section 2.2, the statistical model underlying PANAMA is presented. In section 2.3, the proposed model is compared to existing approaches on a realistic simulated dataset, while section 2.4 contains extensive experimental validation on several real-world datasets. Section 2.5 gives insight into the limitations of current methods to account for confounders that help to understand the relationship between confounding variation,

cis regulation and *trans* effects.

2.1.1 Related Work

Several computational methods have been developed to account for unknown confounding variation within eQTL analyses [Kang et al., 2008a; Leek and Storey, 2007; Listgarten et al., 2010; Stegle et al., 2010, 2012]. A common assumption these methods built on is that confounders are prone to exhibit broad effects, influencing large fractions of the measured gene expression levels. This characteristic has been exploited to learn the profile of hidden confounders using models that are related to PCA [Leek and Storey, 2007; Stegle et al., 2010, 2012]. Once learnt, these factors can then be included in the analysis analogously to known covariates. Another branch of methods avoids recovering the hidden factors explicitly, instead correcting for the correlation structure they induce between the samples [Kang et al., 2008a; Listgarten et al., 2010]. Here, the inter-sample correlation is estimated from the expression profiles first, to then account for its influence in an association scan using mixed linear models. Both types of methods have been applied in a number of studies. Advantages versus naive analysis include better-calibrated test statistics [Listgarten et al., 2010] and improved reproducibility of hits between independent studies [Kang et al., 2008a]. Perhaps most strikingly, statistical methods to correct for hidden confounders have also been shown to substantially increase the power to detect eQTLs, increasing the number of significant *cis* associations by up to 3-fold [Nica et al., 2011; Stegle et al., 2010].

2.2 Methods

While improved sensitivity to detect *cis*-acting eQTLs is an important and necessary step, we expect that even more valuable insights can be gained from those loci that regulate multiple target genes in *trans*. The interest in these regulatory hotspots has been tremendous in recent years, but limited reproducibility between studies has been a concern (see for example the discussion in Breitling et al. [2008]). While accurately accounting for confounding factors is necessary

for an accurate and reproducible identification of regulatory associations, statistical overlap between confounding factors and true association signals from downstream effects can hamper the identification and fitting of confounders. For example, methods that merely accounts for broad variance components, such as PCA, are doomed to fail. If the effect size of *trans* regulatory hotspots is large enough, they induce a correlation structure that is similar to the one caused by confounding factors. Both in the case of a confounding factor and a regulatory hotspot, multiple gene expression levels co-vary jointly. Techniques, such as PCA, that are designed to simply extract the latent variables that explain the most variance in the data, cannot discriminate between a latent factor and a true genetic regulator.As a result, true *trans* regulators tend to be mistaken for confounders and are erroneously explained away.

The statistical model underlying our algorithm is simple and computationally tractable for large eQTL datasets. PANAMA is based on the framework of mixed linear models, and combines the advantages of factor-based methods, such as PCA, SVA [Leek and Storey, 2007] or PEER [Stegle et al., 2010, 2012] with methods that estimate the implicit covariance structure induced by confounding variation [Kang et al., 2010; Listgarten et al., 2010]. The model is fully automated and can be easily adapted to include additional observed confounding sources of variation, such as population structure or known covariates.

The statistical model underlying PANAMA assumes additive contributions from true genetic effects and hidden confounding factors. Briefly, this linear model expresses the gene expression of gene d measured in N individuals as the sum of weighted contributions from a set of K SNPs $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_K\}$, where each $\mathbf{s}_K$ is an N dimensional vector. There are also Q latent confounders $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_Q\}$, where again each $\mathbf{x}_Q$ is a N dimensional vector, as well as a mean term μ_d and a noise term $\epsilon_\mathbf{d}$ (See Figure 2.1a)

$$\mathbf{y}_d = \mu_d + \sum_{k=1}^{K} v_{k,d}\mathbf{s}_k + \sum_{q=1}^{Q} w_{d,q}\mathbf{x}_q + \boldsymbol{\epsilon}_d.$$

Neither the regression weights $w_{d,q}$ nor the profiles of the confounding factors $\mathbf{x}_q$ are known *a priori* and hence need to be learnt from the expression data. Param-

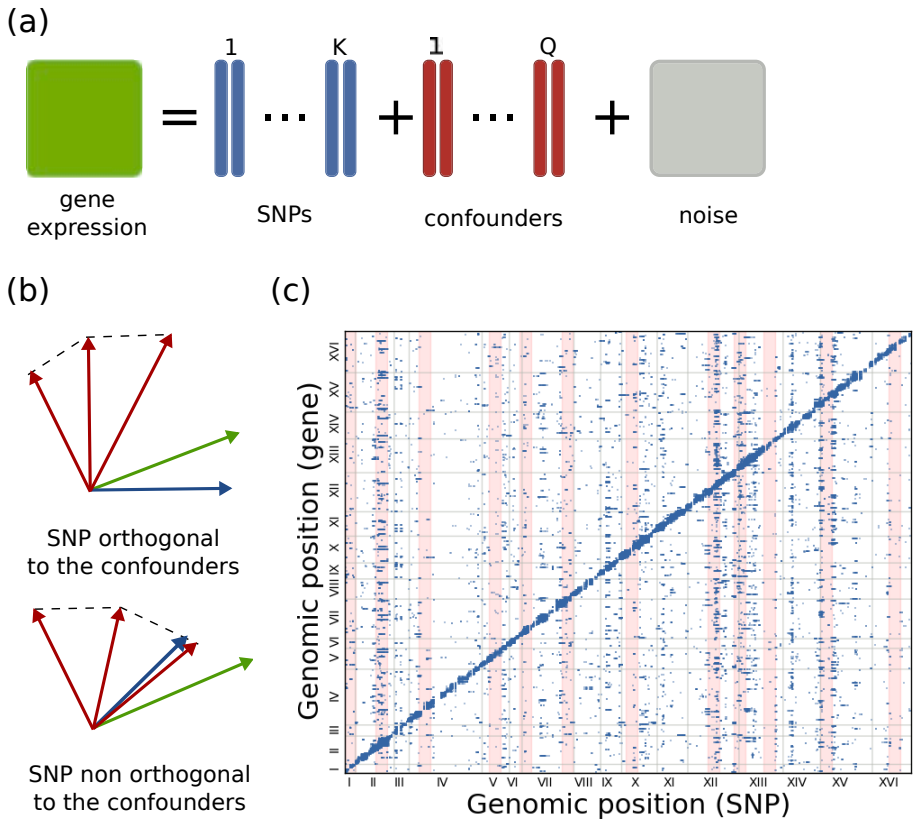


Figure 2.1: **(a)** Effects of causal factors on gene expression variation that are accounted for by PANAMA. **(b)** PANAMA applied to the yeast eQTL dataset. Jointly learned *trans* regulators identified by PANAMA are highlighted in red. **(c)** Illustration of the difference between conventional approaches that assume orthogonality of confounding factors and genetic signals (lower figure) and PANAMA, allowing to disentangle causal signals from confounders despite overlaps.

eter inference in PANAMA is done in the mixed model framework [Kang et al., 2010; Lippert et al., 2011]. In this hierarchical model, the regression weights of the hidden factors are marginalized out, yielding a covariance structure in a multivariate Gaussian model to capture the effect of confounders. Intuitively, the objective during learning in PANAMA is to explain the empirical correlation structure between samples shared across genes by the state of the hidden factors. In the presence of extensive *trans* regulation this approach leads to overcorrection, running the risk of explaining away true genetic association signals. To circumvent this side effect, PANAMA also includes a subset of all SNPs in

the model, resulting in a more complete covariance structure that satisfies an appropriate balance between explaining confounding variation and preserving true genetic signals (Figure 2.1b,c). In this approach, the variance contribution of few major signal SNPs and the state of the hidden factors are then jointly estimated. Moreover, an appropriate number of hidden factors is determined automatically during learning. As a result, PANAMA is statistically robust and inference of hidden factors is feasible without manual setting of any tuning parameters.

2.2.1 Model overview

PANAMA is based on an additive linear model, accounting for effects from K observed SNPs $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_K)$ and contributions from a dictionary of Q hidden factors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_Q)$. The resulting generative model for D gene expression levels $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_D)$ can then be cast as

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{SV} + \mathbf{XW} + \boldsymbol{\epsilon}. \tag{2.1}$$

We assume that expression levels and SNPs are observed in each of $n = 1, \ldots, N$ individuals, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)$ is a vector of gene-specific mean effects and $\boldsymbol{\epsilon}$ denotes Gaussian distributed observation noise, $\epsilon_{n,d} \sim \mathcal{N}(0, \sigma_e^2)$. The matrices $\mathbf{V}$ and $\mathbf{W}$ represent the weights for the SNP effects and hidden factor effects respectively. To improve parameter estimation, we introduce a hierarchy on the weights of genetic influences and hidden factors in Equation (2.1). We marginalize out the effect of the latent factors, $\mathbf{X}$ and a subset of the SNPs with a strong regulatory role (see Section 2.2.3 for more details), resulting in a mixed linear model. We choose independent Gaussian priors for the factors weights $\mathbf{w}_q$ and the weights of respective SNPs $\mathbf{v}_k$

$$p(\mathbf{W}) = \prod_{q=1}^{Q} \mathcal{N}\left(\mathbf{w}_q \,\middle|\, \mathbf{0}, \alpha_q^2 \mathbf{I}\right),$$
$$p(\mathbf{V}) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{v}_k \,\middle|\, \mathbf{0}, \beta_k^2 \mathbf{I}\right),$$

The variance parameters for each factor α_q^2 and each SNP β_k^2 modulate the relevance of the corresponding regulatory variables.

Integrating over the weights $\mathbf{W}$ and $\mathbf{V}$ yields the marginal likelihood that factorizes across genes

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta}) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{y}_d \,\middle|\, \mathbf{0}, \sum_{k=1}^{K} \beta_k^2 \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}} + \sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\right). \tag{2.2}$$

For notational convenience we dropped the mean term $\boldsymbol{\mu}$, since it's always possible to renormalize the data such that each gene has mean 0, and we have defined $\boldsymbol{\Theta} = \{\{\beta_k^2\}, \{\alpha_q^2\}, \sigma_e^2\}$ as the set of all hyperparameters of the model.

In addition to marginalising out the factors weights $\mathbf{w}_q$, it could also be desirable to marginalise out the latent variables $\mathbf{X}$ themselves. Unfortunately, this leads to an intractable marginal likelihood. Titsias and Lawrence [2010] (see also [Hensman et al., 2013] for a different derivation) have proposed a variational approach in which the likelihood has the form of a reduced rank Gaussian process.

Known covariates If available, additional covariates can directly be included in the background covariance structure from Equation (2.2)

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta}) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{y}_d \,\middle|\, \mathbf{0}, \sum_{k=1}^{K} \beta_k^2 \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}} + \sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^{\mathrm{T}} + \gamma^2 \mathbf{K}_0 + \sigma_e^2 \mathbf{I}\right), \tag{2.3}$$

where $\mathbf{K}_0$ denotes the covariance induced by these additional covariates and γ^2 the corresponding scaling parameter. Examples for possible choices of this covariance include the covariance induced by a fixed covariate vectors, i.e. $\mathbf{K}_0 = \mathbf{c}\mathbf{c}^{\mathrm{T}}$ or a kinship matrix that accounts for the genetic relatedness (see for example Kang et al. [2010] and Listgarten et al. [2010]).

2.2.2 Model fitting

Parameter learning, i.e. determining the most probable state of the hyperparameters $\boldsymbol{\Theta}$ and the latent factors $\mathbf{X}$, can be carried out using a straightforward maximum likelihood approach (Equation (2.2))

$$\{\hat{\boldsymbol{\Theta}}, \hat{\mathbf{X}}\} = \underset{\boldsymbol{\Theta},\mathbf{X}}{\operatorname{argmax}} \underbrace{\ln p(\mathbf{Y} \mid \mathbf{S}, \mathbf{X}, \boldsymbol{\Theta})}_{\mathcal{L}}$$

$$\mathcal{L} = -\frac{ND}{2} \ln 2\pi - \frac{D}{N} \ln|\boldsymbol{\Sigma}| - \frac{1}{2} \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \mathbf{Y}\mathbf{Y}^{\top}), \tag{2.4}$$

where the covariance $\boldsymbol{\Sigma}$ implicitly depends on the model parameters $\mathbf{X}$ and $\boldsymbol{\Theta}$. Analytical expression for the gradients of the objective function with respect to particular a particular element of the parameter set θ_i can be determined in closed form

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\Sigma}} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i} = \left(\boldsymbol{\Sigma}^{-1} \mathbf{Y}\mathbf{Y}^{\top} \boldsymbol{\Sigma}^{-1} - G\boldsymbol{\Sigma}^{-1} \right) \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}, \tag{2.5}$$

where $\frac{\partial \boldsymbol{\Sigma}}{\partial \theta_i}$ is the matrix derivative of the covariance with respect to a particular parameter. The objective function and gradients can be used in combination with a gradient-based optimizer such as the limited memory BFGS algorithm (L-BFGS, see [Byrd et al., 1995]). Complete details on parameter inference in Gaussian process models can be found elsewhere [Lawrence, 2005; Rasmussen and Williams, 2006].

In practical applications of PANAMA, this model fitting (Equation (2.4)) is not carried out with the set of all genome-wide SNPs included in Equation (2.1), because the number of weight parameters β_k^2 for each SNP would be prohibitive. Only those genetic regulators with strong effects on multiple genes do play a role during the estimation of hidden factors and thus need to be accounted for. Our inference scheme determines the set of relevant regulators in an iterative procedure.

The number of hidden factors to be learnt, Q is not set *a priori* and instead Q is set to a sufficiently large value. During the optimization, the individual variance parameters for each factors, α_q^2, automatically determine an appropriate number of effective factors, switching off unused ones. See Section 2.2.5 for a discussion.

2.2.3 Iterative learning of the complete model

The presentation so far neglects a strategy to identify regulatory SNPs to be accounted for in the covariance structure (Equation (2.2)). Accounting for the complete set in the covariance is computationally infeasible and difficult to identify statistically, because the number of relevance parameters α_k typically exceeds the number of samples. Here, we suggest an iterative procedure, where only key regulators that are essential to accurately estimate the hidden factors are included during learning. In each iteration we add the SNPs that are most overlapping with the span of the current latent dimensionality, as defined by a linear association test between all latent factors and SNPs. As a convergence criterion we use a q-value [Storey and Tibshirani, 2003] cutoff for statistical significance of the association scan between factors and SNPs. In the following, we refer to this cutoff as FDR addition cutoff. While there is no guarantee that this algoritm (also outlined in Algorithm 1) will converge after selecting a subset of SNPs, in the worst case the algorithm will select all the SNPs for inclusion into the model, simply increasing the time needed to train it. In practice, we found that this procedure always terminates after selecting a small subset of SNPs and that the number of SNPs selected depends on the FDR cutoff. The empirical stability of this procedure for different FDR cutoffs is evaluated on simulated data in section 2.3.

2.2.4 Mixed model testing approaches

Once the confounding-correcting covariance structure is determined from the maximum likelihood solution of Equation (2.4), significance testing can be carried out in the framework of mixed linear models. In an LMM, the trained covariance structure effectively acts as a random effect background model to account for non-genetic confounding variation. Given the covariance structure, it's possible to perform a likelihood ratio test to determine the strength of an association between a SNP and a gene. This type of test is potentially expensive because it requires an inversion of the covariance matrix for each test. Fortunately, several efficient approaches that avoid this problem have been proposed before [Kang et al., 2008b, 2010; Lippert et al., 2011]. The association between a SNP k and

Input: Matrix $\mathbf{Y}$ of individuals $\times$ genes, matrix $\mathbf{S}$ of individuals $\times$ SNPs
Output: Final covariance structure $\mathbf{\Sigma}$

initialize $\mathcal{I} = \emptyset$;
Estimate initial latent dimensionality from PCA $Q = \text{PCA}(\mathbf{Y}, 95\%)$;
$\mathbf{X} = \text{PCA}(\mathbf{Y}, Q) + \mathcal{N}(0, 1)$;
$t = 1$;
Initialise genetic regulators empty $\mathcal{I}_t = \{\}$;
repeatupdate $\{\boldsymbol{\theta}_\text{K}, \mathbf{X}\}$:
  $(\boldsymbol{\theta}^*_\text{K}, \mathbf{X}^*) = \text{argmax}_{\mathbf{X}, \boldsymbol{\theta}_\text{K}}\, p(\mathbf{Y} \,|\, \mathbf{S}, \mathbf{X}, \boldsymbol{\theta}_\text{K}, \mathcal{I}_t)$; `/* optimise covariance */`
  $k^*, q^* = \text{argmax}_{k,q}\, \text{LOD}_{k,q}(\mathbf{s}_k, \mathbf{x}_q)$; `/* scan factor-SNP associations */`
  if *LOD_{k^*,q^*} significant (qv <FDR addition cutoff)* **then**
    $\mathcal{I}_{t+1} = \mathcal{I}_t \cup \{k^*\}$; `/* add overlapping SNP to covariance */`
  end
  $t = t + 1$
until $\mathcal{I}_t = \mathcal{I}_{t+1}$;

Algorithm 1: Algorithm summary of the iterative learning in performed in PANAMA. SNPs that overlap with current estimate of the hidden factors ($\mathbf{X}$) are greedily included in the covariance structure until convergence is reached.

gene d to be tested is treated as fixed effect, allowing to construct a likelihood ratio statistics of the form

$$\text{LOD}_{d,k} = \log \frac{\mathcal{N}\left(\mathbf{y}_d \,|\, \theta \mathbf{s}_k, \sigma_k^2 \mathbf{K} + \sigma_e^2 \mathbf{I}\right)}{\mathcal{N}\left(\mathbf{y}_d \,|\, \mathbf{0}, \;\; \sigma_k^2 \mathbf{K} + \sigma_e^2 \mathbf{I}\right)}. \tag{2.6}$$

where σ_k^2 and σ_e^2 weight the respective distribution of the confounding covariance $\mathbf{K}$ and additive noise contributions, which are refitted for every test. The confounding covariance matrix $\mathbf{K}$ is derived from components of the complete covariance $\mathbf{\Sigma}$ of the fitted PANAMA model (Equation (2.2)), with different choices corresponding to alternative correction strategies. Computationally, the likelihood ratio tests (Equation (2.6)) can be efficiently implemented using recently proposed computational tricks [Lippert et al., 2011], allowing for application to large-scale genomic data.

In PANAMA, this correction covariance structure $\mathbf{K}$ only accounts for the

confounding factors, excluding the genetic regulators (See Equation (2.2))

$$\mathbf{K} = \sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^{\mathrm{T}}.$$

Alternatively, in PANAMA$_{\text{trans}}$, also correcting for the *trans* factors, the covariance also includes *trans* regulators

$$\mathbf{K}_{\text{trans}} = \sum_{k=1}^{K} \beta_k^2 \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}} + \sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^{\mathrm{T}}.$$

PANAMA$_{\text{trans}}$ accounts for the putative confounding influence of broad variance components that do have a genetic basis. While these are not confounding *per se*, accounting for their effect may increase the power for identifying smaller effects that are otherwise overshadowed.

Efficient mixed model implementations Several computational advances have been presented to efficiently carry out the mixed model tests for all SNP/gene pairs (Equation (2.6)) [Kang et al., 2008b, 2010; Lippert et al., 2011]. In the software implementation that accompanies PANAMA, we follow the route taken in most recent development, allowing for exact inference while retaining linear-time complexity in the number of samples per test [Lippert et al., 2011]. Similar to what done in EMMAX [Kang et al., 2010], we carry out a single cubical decomposition of the full-rank matrix $\mathbf{K}$ upfront. Briefly, the underlying idea is to decompose the testing covariance once, which allows to efficiently adapting the weights σ_e^2 and σ_k^2 for each individual test. These measures allow PANAMA to be applicable to genome-scale datasets (See Section 2.2.6).

Significance testing and multiple testing correction In experiments, all considered methods were applied to carry out independent association tests between individual SNPs and genes. We assessed genome-wide significance of individual associations using the q-value method [Storey, 2003; Storey and Tibshirani, 2003].

PANAMA residuals for alternative downstream models For applications other than eQTL testing, it maybe desirable to account for the confounding factors explicitly, subtracting their contribution from the expression data. Such an approach is useful when using the expression levels in combination with other analyses such as clustering or network reconstruction.

In PANAMA, a residual dataset can be obtained by considering the joint Gaussian distribution on the observed data and the test dataset. Completing the square yields a closed form mean-prediction of this Gaussian covariance model

$$\hat{\mathbf{y}}_d = \mathbf{K}\left(\sigma_k^2\mathbf{K} + \sigma_e^2\mathbf{I}\right)^{-1}\mathbf{y}_d. \tag{2.7}$$

Similar as for mixed model testing, the relative weights of the correction and the noise component σ_k^2 and σ_e^2 are refit for every gene. See also [Rasmussen and Williams, 2006] for further details on the usage of Gaussian models as predictors.

2.2.5 Determining the latent dimensionality

In addition to the hyperparameters, the dimensionality of the latent space, Q, is an important implicit parameter of factor-models such as PANAMA. Choosing Q too large results in over-correction, with the model explaining away true genetic associations. In contrast, choosing too few hidden factors, leads to under-correction, where the full hidden variation is not accounted for, ultimately leading to reduced sensitivity.

In related work, several of approaches have been proposed to select an appropriate latent dimensionality. One approach is to consider the explained variance, choosing a user-defined cutoff that determines the fraction of variance explained away by factor components [Stegle et al., 2010]. Alternatively, in [Leek and Storey, 2007], the authors estimate the number of factors using a permutation procedure alongside with additional heuristics that yield the expected number of target genes of a true confounding factor. Also in [Minka, 2001], Minka suggests to employ Bayesian model comparison, evaluating the marginal likelihood of the observed data in the light of alternative models that correspond to different choices of the latent dimensionality.

Here, we follow the approaches presented in Bishop [1999]; MacKay [1995];

Neal [1995] and employ automatic relevance determination (ARD). The principle underlying ARD is to allow each latent dimension to be controlled by a relevance parameter that has a non-zero value only if it is supported by the data. This means that it's possible to avoid choosing a cutoff value for the number of factors explicitly and instead determine the dimensionality of the latent space while training the model. Another advantage of ARD is that it results in a linear combination of different dimensionalities (due to the fact that the relevance parameters are continuous), rather than selecting a specific one. In PANAMA, the variance explained by each hidden factor is controlled by the values of α_q^2, with small values corresponding to irrelevant factors and larger values to factors that explain significant amounts of variation.

In practice, we first obtain a coarse estimate of the latent dimension by using PCA, choosing a cutoff point Q for the number of latent factors when 95% of the total variance is explained. This approach yields an upper bound of the latent dimensionality, which we use a starting point in PANAMA. The learning procedure of PANAMA then determines the number of factors with non-zero relevances α_q^2 automatically while optimizing the marginal likelihood (Equation (2.2)). This approach is both computational efficient and avoids the need of user specified tuning parameters.

The state of the latent factors is initialized by using a perturbed PCA solution (as suggested in [Lawrence, 2005]). Empirically, this approach yields similar results than initialising the factor randomly, however greatly decreases the time for convergence of the optimization.

2.2.6 Software implementation and scalability

Due to the continuous increase in the size of genomics studies, the computational efficiency of the current approaches for eQTL testing is of crucial importance. The Python implementation exploits several properties of the model, in order to allow for applicability to larger datasets. First, the marginal likelihood for parameter inference (Equation (2.4)) has a low-rank structure and hence allows for efficient evaluation of the matrix inverses, speeding up parameter learning. Second, the association tests given the trained PANAMA model build on recent

advances for mixed models that scale linearly with the number of samples and tests [Lippert et al., 2011].

Efficient testing and parallelization Typically, in large scale data the bottleneck lies in the association testing, thus demanding for particular attention of this step. PANAMA builds on recent advances for fast mixed model testing [Lippert et al., 2011], which accompany the PANAMA software package in form of an integrated C++ library. While good performance on a single process/thread is needed, scientific software also requires to be easily parallelized for computing on clusters and clouds. To this end, PANAMA natively allows for jobs to be distributed across multiple processes, multiple machines on the local network, on a cluster and on the most popular cloud computing platforms (provided they have a working Python/numpy/scipy installation).

Empirical computational cost and runtime To compare the computational demands of PANAMA and alternative methods, we carried out a timing experiment on a benchmark dataset consisting of 193 samples, 8,598 genes and 8,311 SNPs (based on the cortical gene expression dataset, chromosome 17, as described in section 2.4.2). The size of this problem was chosen as to ensure that the slowest approach converges within an acceptable time interval. Table 2.1 shows the cpu-time required for various methods used to correct for confounding factors in eQTL studies[1]. All tests were performed on a GNU/Linux machine with an Intel(R) Xeon(R) X7542 CPU and 64 gigabytes of RAM, the python scientific libraries (Numpy and Scipy) were compiled against the Intel(R) Math Kernel Library.

We also extrapolated the computational runtime for current human-scale data, assuming 193 samples, 40,000 genes and 10 million SNPs. These estimates are based on the assumption that the final testing step dominates the computational cost in all methods. This is especially true for the methods that use a low-rank representation of the confounding factors (PANAMA, SVA, PEER), since

[1]The computationally dominating testing step in LINEAR, SVA, PEER has been identically implemented in python; testing of PANAMA in C++ and ICE is fully based on R scripts from the authors. Such difference in the implementation may have implications for the exact runtime estimates provided.

their computational cost for learning of confounders scales with respect to the number of individuals, not with respect of the number of genes. PANAMA, carrying out iterative learning to derive the confounding covariance (Section 2.2.3) requires additional tests between the learnt factors and all SNPs (Algorithm 1). Importantly, because the typical number of confounders is much smaller than genes, this cost can be neglected in practice. Even with 10 million SNPs and 40 factors (more than the typical number of factors in human), this association scan only takes 3 hours compared to 137 days of computation that are needed for genome-wide application of mixed model tests between all SNPs and genes.

Model	CPU-time (in **minutes**)	projected CPU-time (in **days**)
LINEAR	35	136
SVA	39	150
PEER	45	152
PANAMA	62	159
ICE	8,540	33,197

Table 2.1: Empirical computation time for experiments on parts of the human cortical dataset (chromosome 17) and extrapolations for a full-genome dataset with 10 million SNPs and 40,000 probes.

2.3 Simulation study

The evaluation of methods to call eQTLs is difficult as reliable ground truth information is not available. Following previous work [Price et al., 2006; Stegle et al., 2010; Yu et al., 2005], we have used synthetic data to assess and compare PANAMA with alternative approaches. To minimize assumptions we need to impose on the simulation procedure we created an artificial dataset that borrows key characteristics from a real eQTL study in yeast [Smith and Kruglyak, 2008] (See also Application to segregating yeast strains). In this approach, we first fit PANAMA to the real eQTL data, estimating the confounding variation and *cis* and *trans* associations. Given the fitted model of independent tests, we reduced the association matrix between all SNPs and genes to at most one association

per chromosome and gene, avoiding inflated association counts due to linkage disequilibrium. To also include weak associations, we considered association with a q-value of at most 0.3. On the residual dataset, after removing the effect of the estimated confounders, we then fitted a linear model of all significant associations for each gene. Next, we estimated final residuals by removing the confounders and the fitted associations to estimate a distribution of noise levels across genes.

Finally, we used the fitted model parameters from the real dataset to create a synthetic eQTL dataset with known ground truth associations. We considered the same number of simulated *cis* and *trans* associations as found on the real data as well as the empirical distribution of associations weights and noise estimates obtained form the empirical fit. Using the real genotypes we randomly chose associations between SNPs and genes, simulating effects drawing from the empirical distribution of weights. Finally, we added confounding variation by drawing a sample from the fitted confounding covariance structure and added simulated noise from the fitted distribution of noise levels.

Variation of fitted simulation parameters Comparative evaluation of methods on the simulated data were repeated for variations of the fitted simulation parameters. To create datasets of variable levels of difficulty, we considered different numbers of true simulated *trans* regulators (Figure 2.2e) and different numbers of simulated confounders (Figure 2.2f). In both cases, we ran the same simulation approach as previously described, however removing random fractions of the simulated *trans* regulators or confounders respectively.

Given the synthetic eQTL study, we employed alternative methods to recover the underlying simulated associations. We compared PANAMA to standard linear regression (LINEAR), ignoring the presence of confounders entirely, as well as SVA [Leek and Storey, 2007], ICE [Kang et al., 2008a] and PEER [Stegle et al., 2010, 2012], established and widely used approaches to correct for hidden confounders. For reference, we also compared to an idealized model with the simulated confounders perfectly removed (IDEAL). First, Figure 2.2a and 2.2b show the respective number of significant *cis* and *trans* associations as a function of the false discovery rate (FDR) cutoff. To avoid overly optimistic association counts due to linkage disequilibrium, we considered at most a single *cis* association per

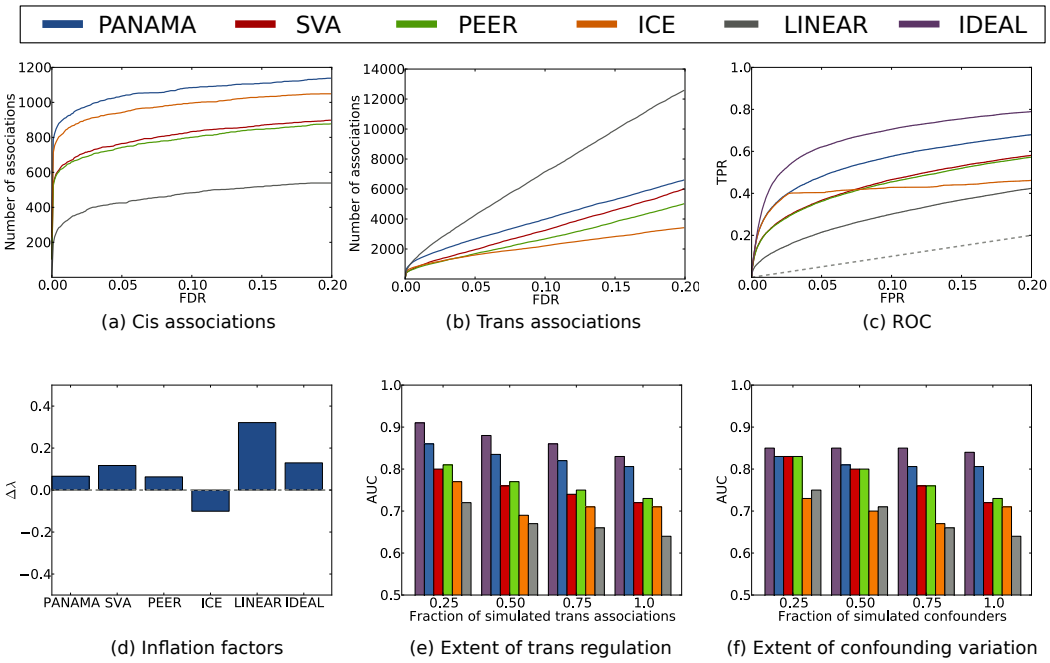


Figure 2.2: Accuracy of alternative methods in recovering simulated *cis* or *trans* associations. **(a,b)** number of recovered *cis* and *trans* associations as a function of the false discovery rate cutoff. At most one association per chromosome and gene was counted. The x-axis is truncated at an FDR of 0.2 in order to highlight the region of most interest for practical purposes. **(c)** Receiver Operating Characteristics (ROC) for recovering true simulated associations, showing the true positive rate (TPR) as a function of the permitted false positive rate (FPR), evaluated on the simulated ground truth. **(d)** inflation factors, defined as $\Delta\lambda = \lambda - 1$, indicate either inflated p-value distributions ($\Delta\lambda > 0$) or deflation ($\Delta\lambda < 0$) of the p-value statistics of different methods. **(e)** Area under the ROC curve for alternative methods as a function of the extent of *trans* regulation. **(f)** Area under the ROC curve for alternative methods for varying extent of confounding variation.

gene and at most one *trans* association per chromosome for each gene. PANAMA found more *cis* associations than any other approach and retrieved the greatest number of *trans* associations among methods that correct for hidden confounders. Notably, the linear model appeared to find even more *trans* associations, however the majority of these calls were inconsistent with the simulated ground truth and were spurious false positives. The extent of false associations called by the linear

model is also reflected in Figure 2.2c, which shows the receiver operating characteristics for each method. All approaches that correct for confounders performed strikingly better than the linear model. Among these, PANAMA was most accurate, achieving greater sensitivity than any other method for a large range of false positive rates (FPR), approaching the performance of an ideal model (IDEAL).

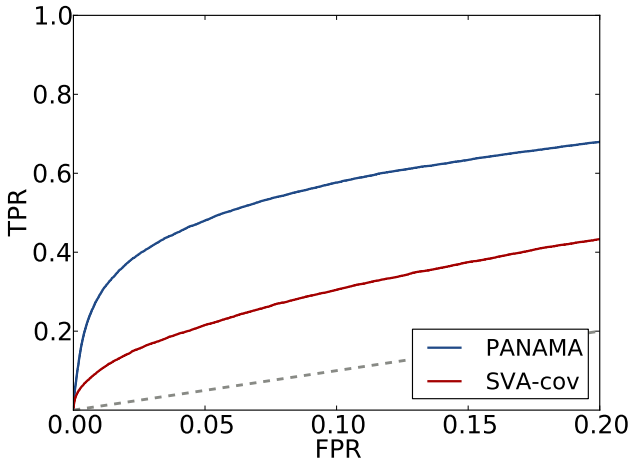
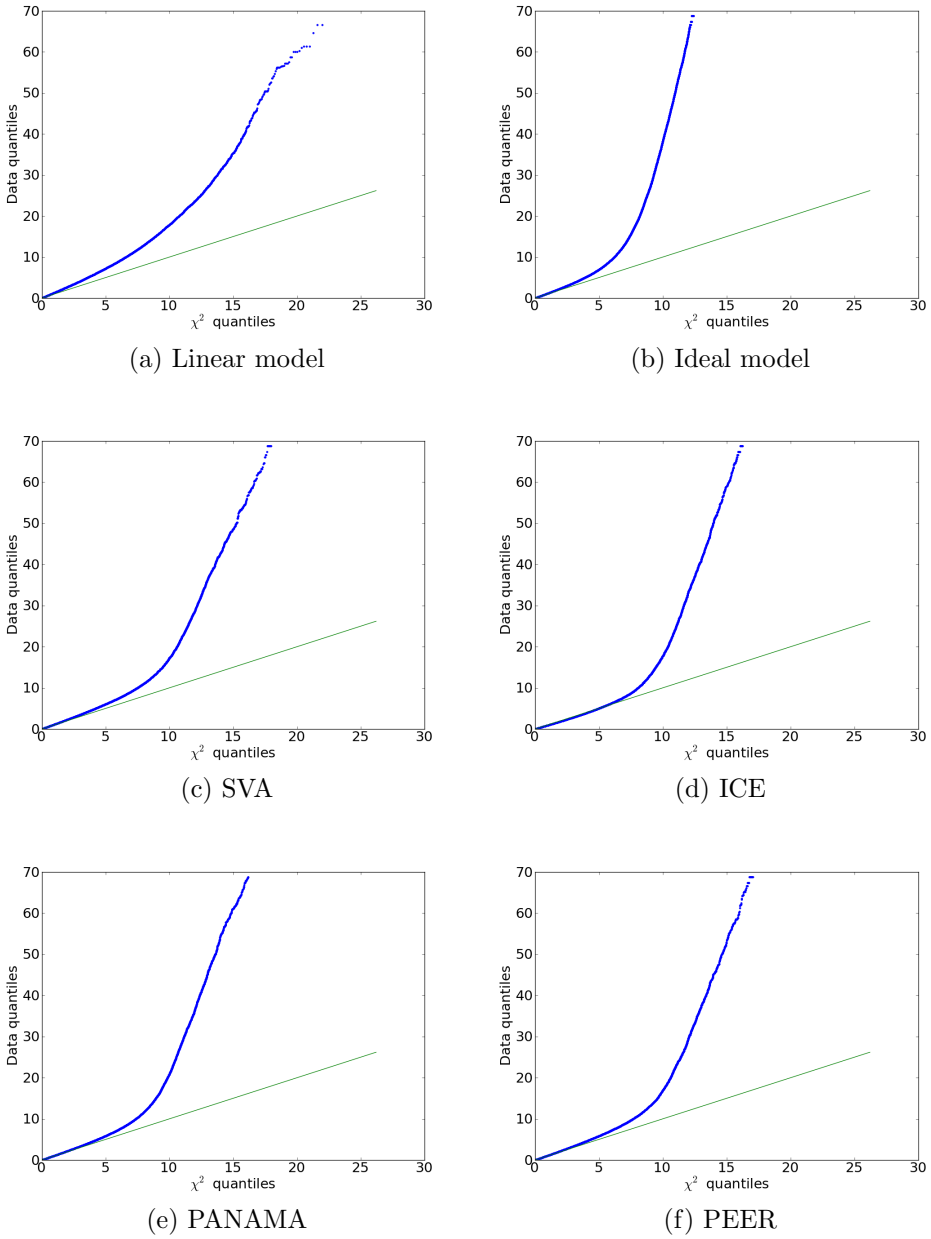


Figure 2.3: Receiver operating characteristics (ROC) curve comparing PANAMA to a modified version of SVA that models the most prominent genetic regulators as covariates.

Since some models, including SVA and PEER, allow to include additional known covariates, we investigated their performance when adding the strongest genetic regulators as covariates. This procedure is mimicking the central concept of PANAMA using previous methods. As shown in Figure 2.3, the iterative learning procedure of PANAMA still produces a significantly better receiver operatic characteristic (ROC) curve for the recovery of the true simulated associations.

Next, we studied the statistics of obtained p-values, checking for departure from a uniform distribution that either indicates inflation (genomic control $\lambda > 1$) or deflation (genomic control $\lambda < 1$) of the respective methods (Figures 2.2d and 2.4). All methods except for ICE yielded an inflated p-value distribution. Notably, this observation also applies to the ideal model where the effect of confounders had been perfectly removed. Thus, in settings with sufficiently strong *trans* regulation, inflated statistics are not necessarily due to poor calibration



(a) Linear model (b) Ideal model

(c) SVA (d) ICE

(e) PANAMA (f) PEER

Figure 2.4: Comparison of theoretical PV statistics with empirical distribution. Figure shows the quantile-quantile plots for alternative methods evaluated on the simulated dataset.

because of confounders, but instead may occur as a consequence of an excess of true biological signals themselves.

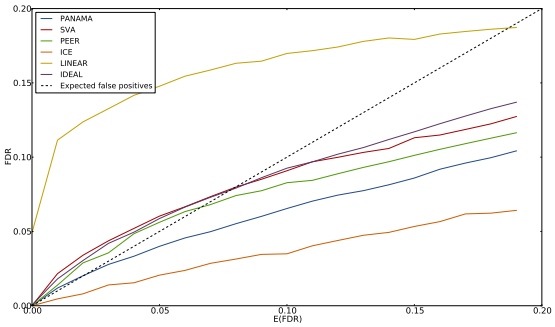


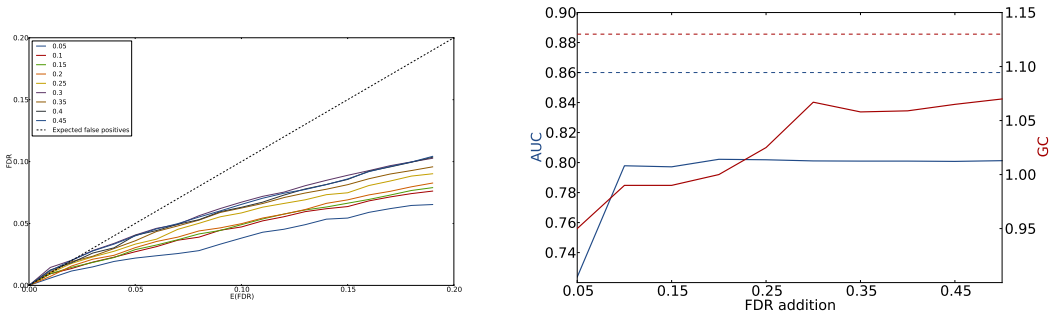
Figure 2.5: Comparison of the calibration accuracy of false discovery estimates for alternative methods. Shown is the estimated false discovery rate (E(FDR)) as a function of the empirical false discovery rate for associations called on the simulated dataset. In summary, PANAMA is better calibrated than any other method, neither underestimating nor overestimating the FDR.

False discovery rate estimates from all methods but the linear model were approximately in line with the empirical rate of errors when taking the ground truth into account (Figure 2.5), with PANAMA being the best calibrated method.

We then repeated the same analysis on a broader range of simulated datasets, varying particular aspects of the simulation procedure around the parameters obtained from the fit to the real yeast data. Figure 2.2e shows the accuracy of alternative methods when reducing the extent of simulated *trans* regulation by subsampling from the set of initial *trans* effects. These results highlight that previous methods only work well in the regime of little *trans* regulation, while PANAMA provides for accurate calls for a wider range of settings. Similarly, Figure 2.2f shows results for strong *trans* regulation, now varying the extent of confounding factors from weaker to stronger influences. Again, PANAMA was found to be more robust than previous approaches, recovering true simulated associations with great accuracy irrespectively of the magnitude of simulated confounding.

Alternative simulation using ICE for real data fitting The simulation procedure described yields eQTL datasets that share key properties with the real dataset used for fitting. For comparison, we repeated the fitting process using

ICE as an alternative method to correct for confounders. All other details on the exact simulation procedure remained identical.



(a) E(FDR) vs FDR

(b) AUC/GC

Figure 2.6: Impact of choosing more stringent (0.05) to less stringent (0.5) cutoff parameters for adding *trans* associations into PANAMA while learning hidden confounders. **(a)** Estimated false discovery rate (E(FDR)) versus the empirical false discovery rate of called associations on the simulated dataset. **(b)** Area under the Receiver Operating Characteristics and inflation of the test statistics, λ. For comparison this figures includes AUC and λ of an ideal model, with the confounders being removed. The results show that PANAMA is not sensitive to the choice of the stringency parameter for including *trans* factors and generally achieves better performance for higher values.

Sensitivity to FDR addition cutoff While most of the model parameters are automatically inferred from the data, the FDR addition cutoff value needs to be set manually. As discussed in section 2.2.3, this parameter is a q-value cutoff that controls the inclusion of individual genetic regulators in the model. If after the association test between all latent factors and SNPs, no SNP-factor pair has a q-value lower than the FDR addition cutoff, the iterative training procedure stops. Given the importance of this parameter for the convergence of the model, we checked that the performance of PANAMA is not sensitive to the exact setting of the FDR addition cutoff value. Figure 2.6a shows the impact on the performance of PANAMA (as measured by the area under the receiver operating characteristic curve) when using alternative cutoff values that regulate the extend of *trans* regulators to be included in the model covariance structure.

Reassuringly, PANAMA approached the performance of the ideal model for less stringent cutoffs corresponding to a greater number of regulators that were included during the learning process. We also checked the calibratedness of the test statistics of PANAMA. In general, less stringent cutoffs that lead to larger numbers of regulators to be included in the model did not impact the calibration of resulting q-value estimates (See Figure 2.6b). Hence, in practical applications the increased computational cost of determining the genetic weight parameters β_k^2 is the limiting factor when choosing less stringent FDR addition cutoffs values.

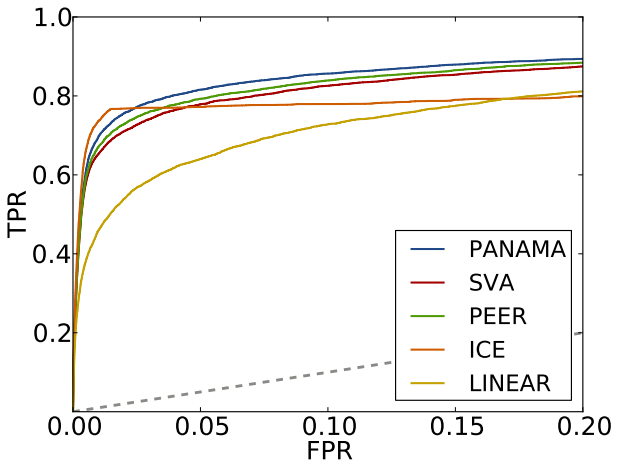


Figure 2.7: Receiver operating characteristics for an alternative simulated dataset based on a fit of ICE to the original yeast dataset. While the general performance differences are smaller, the general trends remain. The kink in ICE is due to deflation of the model.

Figure 2.7 shows summary results for a second synthetic dataset fitted using ICE. As ICE tends to be the most conservative approach among the considered methods, the extent of *trans* regulation on this simulated data was severely reduced. As a consequence, the differences between methods were considerably smaller, however confirming the previously observed trends.

2.4 Experiments on real data

2.4.1 Application to segregating yeast strains

Having established the accuracy of PANAMA in recovering hidden confounders in a simulation study, we applied PANAMA and the alternative methods to the primary eQTL dataset from segregating yeast strains [Smith and Kruglyak, 2008]. These data cover a set of 108 genetically diverse strains that have been expression profiled in two environmental conditions, glucose and ethanol. First, we focused on the glucose condition, which has previously been expression profiled [Brem et al., 2002], providing an independent study for the purpose of comparison.

Figure 2.10a and 2.10b show the number of *cis* and *trans* associations for different methods as a function of the FDR cutoff. Again, we considered at most one association per chromosome to avoid confounding the size of associations with their number. In line with previously reported results [Kang et al., 2008a; Stegle et al., 2010] and our own simulations, the standard linear model identified fewer *cis* associations than methods that correct for confounding variation.

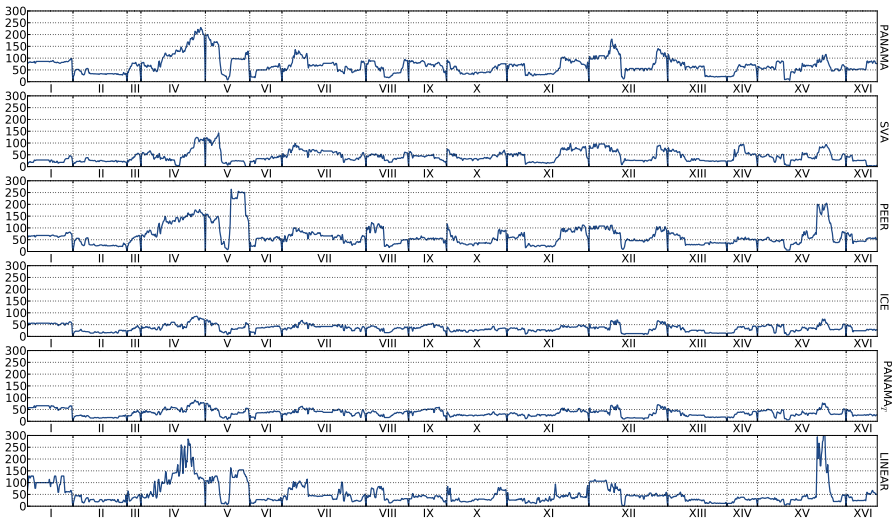
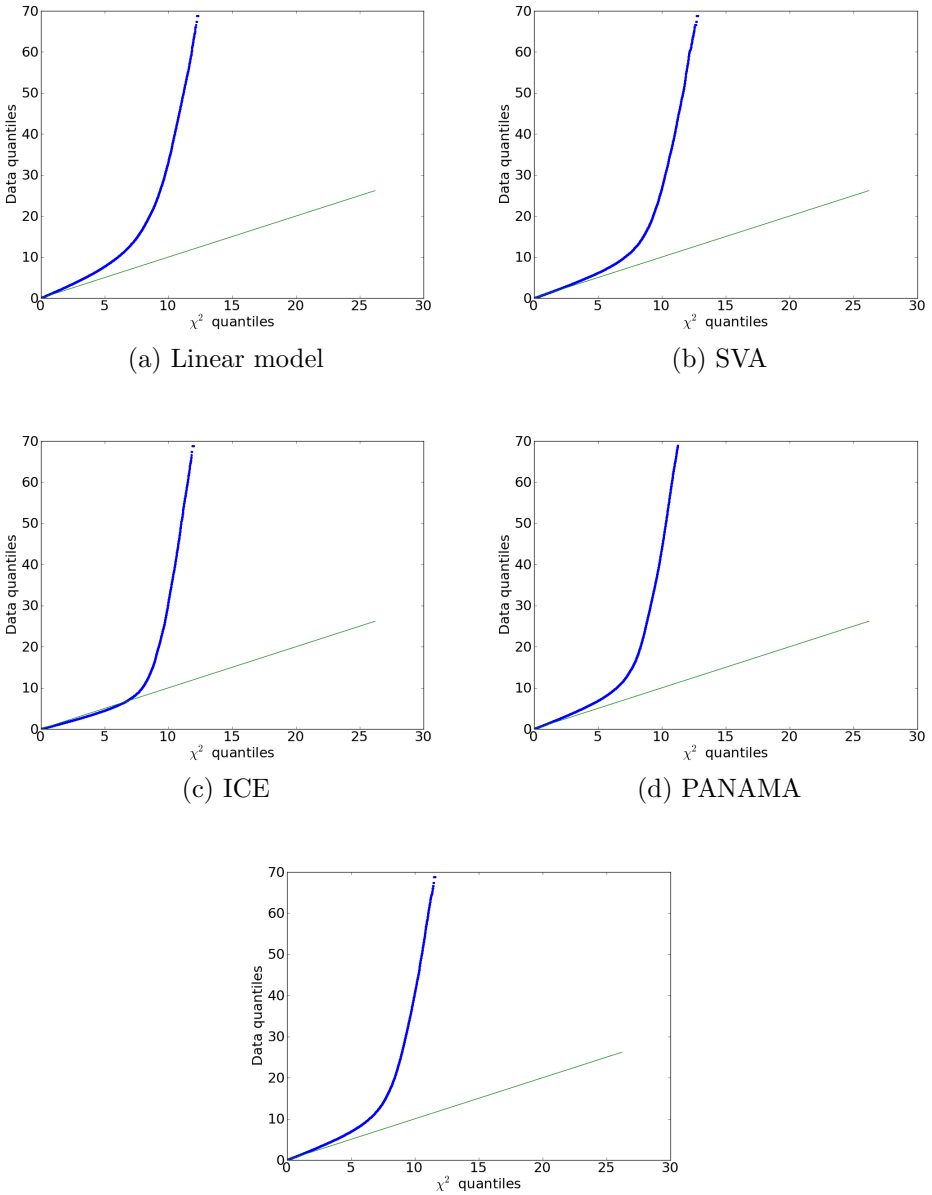


Figure 2.8: Number of associations called as a function of the genomic position for alternative methods on the eQTL dataset from segregating yeast strains (glucose condition).

The trends from the simulated dataset also carried over for *trans* associations, where the linear model called many more associations than methods that account for confounders, yielding an excess of regulatory hotspots (See Figure 2.8). It has previously been suggested that many of these are likely to be false; see for example the discussion in Kang et al. [2008a]. Among the methods that correct for confounding variation, PANAMA identified the greatest number of associations. Among the alternative methods, ICE appeared to be more sensitive in recovering *cis* associations while PEER and SVA retrieved a greater number of *trans* associations.

It should be noted that models that account for confounding factors yielded slightly inflated p-value distributions (Figure 2.10c, Figure 2.9), supporting that also in real settings, a certain degree of inflation may be caused by extensive *trans* regulation. Finally, Figure 2.8 shows the number of associations called by different methods as a function of the genomic position. This summary of genome-wide eQTLs confirms that ICE is most conservative in detecting hotspots, whereas all other methods do find multiple *trans* bands. For comparison, we also included a version of PANAMA that corrects for the *trans* regulators that are accounted for while learning ($\text{PANAMA}_{\text{trans}}$). The resulting model, named $\text{PANAMA}_{\text{trans}}$, shows that explicitly overcorrecting for confounders can lead to explaining away all the regulatory hotspots, both spurious and non-spurious, found by the other models. Interestingly, $\text{PANAMA}_{\text{trans}}$ yields near-identical results to ICE, suggesting that the difference in performance between the two models can be explained, at least in part, by the fact that ICE does not explicitely model pronounced regulators.

Reproducibility of eQTLs between studies To objectively shed light on the validity of the associations called, we considered the consistency of calls between two independent studies. The glucose environment from Smith et al. [Smith and Kruglyak, 2008] has previously been studied [Brem et al., 2002], sharing a common set of segregants. We checked the consistency in calling genes with a *cis* association for increasing FDR cutoffs (Figure 2.10d). Alternatively, focusing on the consistency of regulatory hotspots, Figure 2.10e shows the ranking consistency of polymorphisms ordered by their regulatory potential on multiple genes. Re-



(a) Linear model

(b) SVA

(c) ICE

(d) PANAMA

(e) PEER

Figure 2.9: Comparison of theoretical PV statistics with empirical distribution. Figure shows the quantile-quantile plots for alternative methods evaluated on the yeast dataset.

assuringly, for both *cis* effects and *trans* regulatory hotspots, PANAMA yielded results with far greater consistency than any other currently available method.

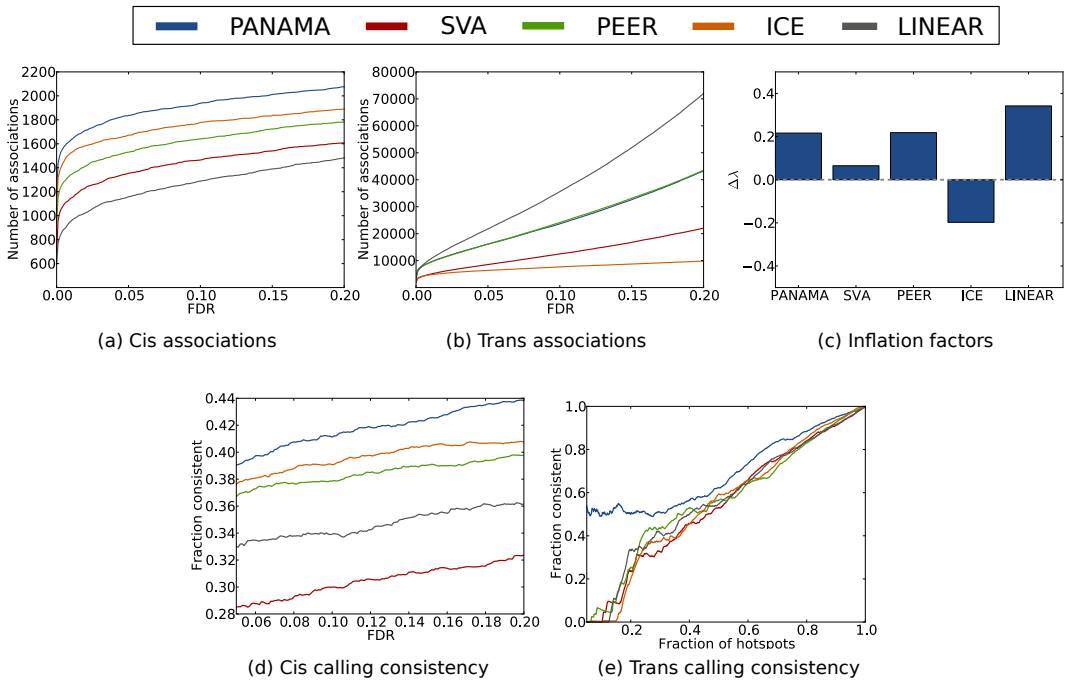


Figure 2.10: Evaluation of alternative methods on the eQTL dataset from segregating yeast strains (glucose condition). **(a,b)**: number of *cis* and *trans* associations found by alternative methods as a function of the FDR cutoff. **(c)** Inflation factors of alternative methods, defined as $\Delta\lambda = \lambda - 1$. **(d)** Consistency of calling *cis* associations between two independent glucose yeast eQTL datasets. **(e)** Consistency of calling eQTL hotspots between two independent glucose yeast datasets, where SNPs are ordered by extent of *trans* regulation as determined by $-\log_{10}(pv)$.

In particular the consistency of *trans* hotspots suggest that PANAMA achieved an appropriate balance between explaining away spurious signals as confounding variation and identifying hotspots that are likely to have a true genetic underpinning.

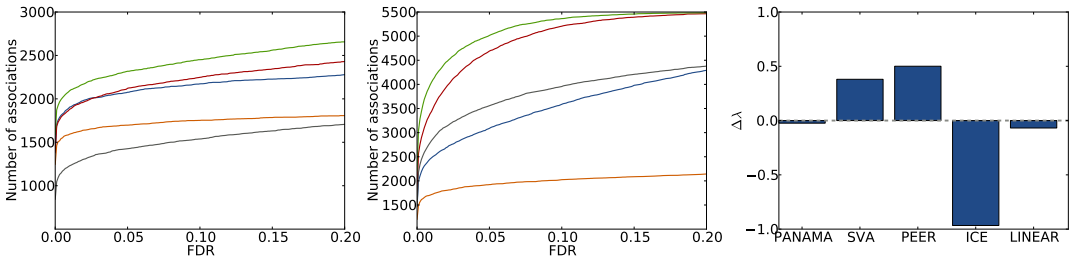
Consistency of *trans* regulatory hotspots with respect to known regulatory mechanisms in yeast As a second means of validating *trans* eQTLs, we investigated to what extent polymorphisms that regulate multiple genes in *trans* can be interpreted as indirect effects that are mediated by known transcriptional regulators. For this analysis we considered an established regulatory

network of transcription factors extracted from Yeastract [Teixeira et al., 2006]. Although we do not expect *trans* associations to be exclusively mediated by direct transcriptional regulation, the degree of associations that are consistent with this regulatory structure is nevertheless an informative indicator for the validity of eQTL calls from different models.

For each transcription factor, we considered polymorphisms in the vicinity of the coding region of the transcription factor ($\pm$ 10kb around the coding region), and tested the fraction of associations with genes that are known targets of the transcription factor versus other associations with genes that are no direct targets. For half of the 129 TFs, PANAMA yielded a higher F-score than any of the other methods considered. Interestingly, the standard linear models performed second best under this metric, achieving the greatest F-score in 36% of all cases, followed by PEER (28%), SVA (15%) and ICE (6%). Among the methods that correct for confounders, PANAMA consistently yielded the highest F-score.

Detecting eQTLs that are shared across environments Finally, we considered the full expression dataset from Smith et al. [Smith and Kruglyak, 2008], combining expression measurement in an ethanol and glucose background. Because each yeast strain was profiled twice, the set of samples was not independent, but instead had a replicate population structure. Similarly to what has been done in previous work [Listgarten et al., 2010], we accounted for this genetic relatedness in PANAMA by adding a population covariance term (Material and Methods).

Figure 2.11 shows the number of associations retrieved by PANAMA and alternative methods on this joint yeast dataset. Because PANAMA accounted for the replicate structure of the dataset, the increase in the number of associations compared to the analysis of the single-condition analysis was modest. Other methods, not accounting for the replicate structure of the genotypes, yielded severely inflated test statistics, identifying a *trans* effect for the great majority of all genes. To check the impact of the population structure covariance, we also applied PANAMA without the correction for artificial genetic relatedness, yielding similarly inflated results (data not shown).



(a) Cis associations (b) Trans associations (c) Inflation factors

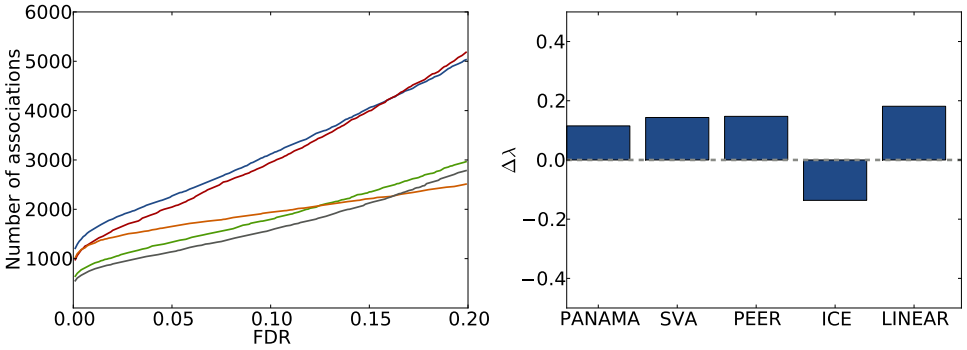
Figure 2.11: Evaluation of alternative methods on the eQTL dataset from segregating yeast strains (glucose and ethanol jointly). **(a,b)** number of recovered *cis* and *trans* associations as a function of the false discovery rate cutoff. At most one association per chromosome and gene was counted. **(b)** inflation factors, defined as $\Delta\lambda = \lambda - 1$. Note that PANAMA included a covariance term that accounts for the genetic relatedness of identical individuals profiled in two conditions. As a result, PANAMA yielded better calibrated results, calling fewer associations than other methods.

2.4.2 Application to further eQTL studies

We have also successfully applied PANAMA to additional ongoing and retrospective studies. For example, on a dataset from inbred mouse crosses [Schadt et al., 2005], PANAMA identified a greater number of associations than other methods (Figure 2.12). In contrast to the yeast dataset, the distribution of p-values on this dataset was almost uniform, suggesting that the extent of true *trans* regulation is lower.

We also investigated parts of a dataset of the genetics of human cortical gene expression [Myers et al., 2007]. On chromosome 17, methods that account for confounders identified more genes in associations than a linear model, with SVA and PANAMA retrieving the greatest number (Figure 2.13). Results on other four other chromosomes were similar (data not shown).

Finally, results of PANAMA applied to an RNA-Seq eQTL study on *Arabidopsis* [Gan et al., 2011] indicate that expression heterogeneity as accounted for by PANAMA is also present on expression estimates from short read technologies, which is consistent with previous reports in human RNA-Seq studies [Pickrell et al., 2010]. This suggests that statistical challenges due to confounding varia-



(a) Associations (b) Inflation factors

Figure 2.12: Evaluation of alternative methods on the eQTL dataset from mouse. **(a)** Number of *cis* and *trans* associations found by alternative methods as a function of the FDR cutoff. **(b)** Inflation factors of alternative methods, defined as $\Delta\lambda = \lambda - 1$.

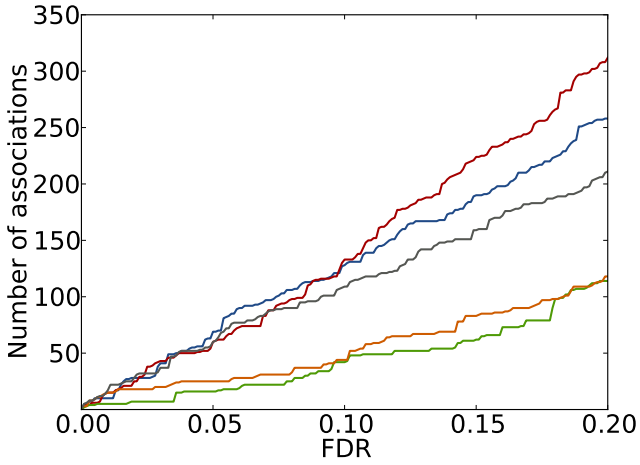


Figure 2.13: Number of associations as a function of the false discovery rate cutoff on the human dataset.

tion are not specific to a particular platform for measuring gene expression.

2.5 Discussion

We have reported the development of PANAMA, an advanced statistical model to correct for confounding influences while preserving genuine genetic association signals. We have shown that this approach is of substantial practical use in a range of real settings and studies. The correction approach of PANAMA, for the first time, is able to not only find more *cis* eQTLs, but also greatly improves the statistical power to uncover true *trans* regulators. PANAMA finds a greater number of associations, and calls eQTLs that are more likely to be real, as validated by means of realistic simulated settings and an analysis of eQTL consistency between independent studies. Most notably, PANAMA identified several strong *trans* hotspots on yeast, out of which at least 40% could be reproduced on a replication dataset.

There are several previous approaches to correct for confounding influences in eQTL studies. These methods can be broadly grouped into factor-based models like PCA, SVA [Leek and Storey, 2007] and PEER [Stegle et al., 2010, 2012], and approaches that employ a mixed linear model [Kang et al., 2008a; Listgarten et al., 2010], estimating a covariance structure that captures the confounding variation. An important reason why PANAMA performs well is the intermediate approach taken here, which consists in learning a covariance structure within a linear mixed model (LMM), but at the same time retaining the low-rank constraint that yields an explicit representation of factors. Moreover, PANAMA systematically exploits the flexibility provided by the representation in terms of covariance structures, jointly accounting for genetic regulators while estimating the confounding factors. Our approach is stable and robust, avoiding the need to first subtract off the genetic contribution greedily, as for example suggested and implemented in SVA [Leek and Storey, 2007] and PEER [Stegle et al., 2010, 2012]. Although this is not the focus of this work, we have shown how our approach can be combined with additional measures to correct for observed sources of confounding variation, such as known covariates or populational relatedness. The utility of such measures has been illustrated in the joint analysis on data from two environmental conditions. A more specialized approach that is aimed at the combined correction for expression confounders and population structure has recently been

	Low rank	LMM	Preserves genetic signal
SVA	✓		✓(partially)
PEER	✓		✓(partially)
ICE		✓	
LMM-EH		✓	
PANAMA	✓	✓	✓
LINEAR			✓

Table 2.2: Comparison of the different models that account for confounders (SVA,PEER, ICE, LMM-EH, PANAMA) and LINEAR. A mark indicates that the model exhibits that property. The properties are: *Low rank*: is the model using a low-rank representation of the confounders? *LMM*: is it a linear mixed model? *Preserve genetic signal*: is the model explicitly preserving the genetic signal or is it greedily subtracting the confounding effects? PANAMA is the only model that spans all the different properties, since it imposes a low-rank structure for the confounders, but is efficiently implemented as a linear mixed model. Moreover, the latent confounders are learned in conjuction with the genetics, thereby preserving true genetic signals.

proposed by Listgarten et al. [Listgarten et al., 2010]. This LMM-EH approach is methodologically related to what is done here, as the contribution from multiple sources of variation are combined within a single covariance structure. Importantly, the main contribution in PANAMA is an integrated model that does not include additional confounders but true genetic regulators. Unique to PANAMA, these regulators are jointly identified and accounted for during learning of the confounding factors. Our analysis shows, that this approach yields a significant improvement in the sensitivity of recovering *trans* associations and plausible regulatory hotspots.

A tabular overview of the relation between alternative methods is shown in Table 2.2.

In conclusion, PANAMA is an important step towards exhaustively addressing common types of confounding variation in eQTL studies. The number of datasets that benefit from careful dissection of true genetic signals and confounders, as done here, is expected to rise quickly. Growing sample sizes and expression profiling in more than one environment allow for the estimation of more subtle confounding influences and at the same time provide the statistical power to

detect many more *trans* effects than possible as of today.

2.6 Conclusions

In this chapter we have presented and studied probabilistic latent variable models to correct for gene expression heterogeneity while accounting for the effect of strong genetic regulators. Across several different datasets, the approach presented here has been shown to perform better than previous methods, identifying a greater number of significant eQTLs and in particular additional *trans* regulators. Multiple sources of evidence support that these additional associations are likely to be real. Most strikingly in yeast, the findings by PANAMA can be better reproduced between independent studies and are more consistent with prior knowledge about the underlying regulatory network.

While the focus of this chapter has been on correcting for the influence of unobserved batch effects and environmental factors, in the next chapter we are going to leverage these inferred environmental factors to identify genotype-by-environment interactions with a regulatory effect on gene expression.

Chapter 3

Modelling GxE interactions with unmeasured environments

The material presented in this chapter is joint work with Christoph Lippert, Karsten Borgwardt, Oliver Stegle and Neil Lawrence, and has been published in *"Detecting regulatory gene-environment interactions with unmeasured environmental factors"* [Fusi et al., 2013].

3.1 Overview

In Chapter 2, we have proposed a latent variable model to capture the effect of environmental confounders and batch effects on gene expression. Accounting for these factors while performing eQTL studies resulted in an overall increase in power (ability to indentify true SNP-gene associations), detecting both more *cis* and *trans* associations across a wide range of datasets. In this chapter, we still use a largely similar latent variable model, but with a slightly different aim. Instead of estimating latent factors affecting the gene expression levels and simply explaining them away, we want to use them as surrogate estimates of hidden environmental factors in genotype-by-environment studies.

Indeed, while analyzing eQTLs in different genetic systems and species, it has become clear that the cellular and environmental context needs to be taken

into account to fully understand the genetic architecture of gene expression [McCarthy et al., 2008]. One route towards investigating such context dependency is explicit experimental stratification. In human, expression profiling in different tissue types, both in unrelated individuals [Fu et al., 2012; Nica et al., 2011] and families [Grundberg et al., 2012], has shown that eQTLs frequently have tissue-specific effect sizes, and in some cases exhibit opposite effects. Analogously, different environmental backgrounds and cellular contexts may modulate the genetic control of molecular traits [Smith and Kruglyak, 2008; Vinuela et al., 2010], suggesting that environment-specific genetic effects, also called genotype-environment interactions, are the rule rather than the exception.

Despite their relevance, molecular studies with explicit environmental perturbations are difficult to carry out in population-scale studies. Precise control of the environmental state cannot be achieved for many important organisms. For example in human, the relevant environment could be of climatological or social nature and hence is either completely unknown [Gibson, 2008] or can only be indirectly influenced via targeted sample selection [Nath et al., 2012]. Furthermore, the most relevant factors for molecular regulation may not be a global external condition but rather cellular factors, which are in turn driven by genetic or external factors [Litvin et al., 2009]. In all of these settings, the most relevant context and environment is not directly measurable and hence statistical inference of these factors is needed to study their implications on the transcriptional state.

Recently, several methods have been proposed to account for unknown confounding in eQTL studies, a substantial proportion of which can be attributed to subtle environmental effects [Fusi et al., 2012; Leek and Storey, 2007; Listgarten et al., 2010; Stegle et al., 2010]. While these methods have been shown to substantially increase power in detecting true eQTLs, the potential of using such recovered factors to identify genotype-environment interactions has largely been overlooked.

In this chapter, we present an integrated probabilistic model, **LI**near **M**ixed **M**odel **I**nteraction (*LIMMI*), which allows to recover unknown environmental or cellular factors from gene expression profiles and detecting genotype-environment interactions. LIMMI allows for a flexible class of environmental and genetic effects

that act on gene expression, including direct effects and interactions between them (see Figure 3.1). At the same time, the model enforces that the estimated factors are truly environmental and not themselves under genetic control.

In section 3.3.1 we evaluate LIMMI on synthetic data where we assess the ability of LIMMI to (i) recover the true simulated environmental state, to (ii) better detect direct genetic effects and in particular to (iii) identify genotype-environment interactions with unmeasured environmental factors. In section 3.3.2 revisit an eQTL study on yeast [Smith and Kruglyak, 2008], where we compare the inference of LIMMI with a measured environmental variable. Beyond accurately recovering this known environmental effect, LIMMI retrieves an additional 14 factors that are orthogonal to the genetic state. When using these factors to test for environment-specific genetic effects, we find hotspots of genotype-environment interactions, some of which are enriched for known response processes to environmental stimuli. Finally, we demonstrate that including interactions between genotype and learnt factors in a mixed model improves both detection power as well as calibration of test statistics for direct genetic effects in an eQTL scan.

3.2 Methods

LIMMI is based on a linear additive model that explains phenotype variability as the sum of genetic and non-genetic factors. Formally, assume we are given an eQTL dataset comprising a gene expression matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_G]$ of G gene expression levels. Each expression profile $\mathbf{y}_g$ is observed in N individuals, i.e. $\mathbf{y}_g = [y_{(g,1)}, \ldots, y_{(g,N)}]$. We assume that the expression estimates $\mathbf{Y}$ are variance stabilized, i.e. the measurement error is independent of the expression level. Suitable variance stabilizing transformations have previously been proposed for both data from microarray technologies [Lin et al., 2008] and RNA-Seq data [Anders and Huber, 2010].

Similarly to what was done in Chapter 2, expression variability is modelled as the sum of effects from SNPs $\mathbf{S}$ and non-genetic (environmental) factors $\mathbf{X}$. The generative model underlying LIMMI allows for direct effects on the phenotype, as well as interaction effects between SNPs and environmental factors. Using the framework of linear mixed models, the joint contribution to the expression

variability of a single gene g can be written as the sum of individual covariance matrices for each of these respective effect types

$$\mathbf{y}_g \sim \mathcal{N}\Big(\underbrace{\mu_g \mathbf{1}}_{\text{mean}}, \underbrace{\mathbf{K_S}}_{\substack{\text{SNP} \\ \text{effects}}} + \underbrace{\mathbf{K_X}}_{\substack{\text{direct factor} \\ \text{effects}}} + \underbrace{\mathbf{K_I}}_{\substack{\text{SNP-factor} \\ \text{interactions}}} + \underbrace{\sigma_p^2 \mathbf{K_P}}_{\substack{\text{population} \\ \text{structure}}} + \underbrace{\sigma_e^2 \mathbf{I}}_{\text{noise}} \Big). \tag{3.1}$$

Here, the individual $N \times N$ covariance matrices explain the joint covariation across genes due to genetic effects ($\mathbf{K_S}$) and environmental factors ($\mathbf{K_X}$), while $\mathbf{K_I}$ explains the joint covariation due to genotype-environment interactions. Additionally, we include a genetic relatedness matrix $\mathbf{K_P}$ as a variance component, in order to account for confounding due to population structure, which can be estimated from the genotype data itself [Kang et al., 2008b, 2010; Lippert et al., 2011].

In order to determine suitable expressions for the individual covariance matrices, let the matrix of genotypes for the same N individuals be $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_K]$ of K SNPs. We use a binary $(0, 1)$ encoding for homozygous and a $(0, 1, 2)$ encoding for heterozygous organisms, however other encodings can be considered as well. Further, let $\mathbf{X} = \left[\mathbf{X}^o, \mathbf{X}^h\right]$ denote the set of non-genetic factors that influence the gene expression levels, where $\mathbf{X}^o \in \mathbb{R}^{N \times C}$ are *a priori* observed (measured) environmental covariates and $\mathbf{X}^h \in \mathbb{R}^{N \times L}$ denote *unobserved* factors we would like to infer from the expression profiles.

Let the symbol $\odot$ denote the element-wise product. An interacting pair of a SNP $\mathbf{s}_k$ and a factor $\mathbf{x}_q$ can then be represented by the vector $(\mathbf{s}_k \odot \mathbf{x}_q)$. In this form, the factor effect is masked for all samples where the genetic state is zero, here the major allele. Other interaction models can be implemented in an analogous fashion [Hallgrímsdóttir and Yuster, 2008].

Assuming only linear additive effects of single SNPs, environmental factors and their interactions, we write all variance components in the form of linear kernels:

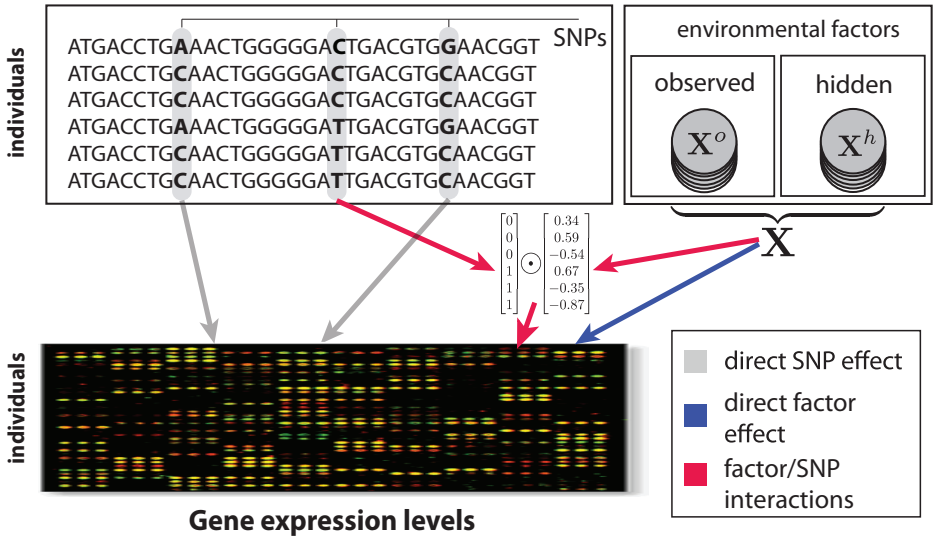


Figure 3.1: **Illustration of regulatory effects on gene expression modelled by LIMMI.** First, non-genetic environmental factors can either be measured (observed) or hidden. Their effect on gene expression is typically dominated by direct effects (blue). In addition, some factors may act in a genotype-specific manner, for example with effects only standing out in a particular genetic background (red). Finally, there are standard genetic expression QTLs with individual genetic loci regulating gene expression levels (black).

$$p(\mathbf{Y} \,|\, \mathbf{S}, \mathbf{X}, \boldsymbol{\theta}_{\mathrm{K}}) = \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{y}_g \,|\, \mu_g \mathbf{1}, \boldsymbol{\Sigma}\right) \tag{3.2}$$

$$\Sigma = \underbrace{\sum_{k=1}^{K} \beta_k^2 \mathbf{s}_k \mathbf{s}_k^\top}_{\mathbf{K_S}} + \underbrace{\sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^\top}_{\mathbf{K_X}} + \underbrace{\sum_{k=1}^{K} \sum_{q=1}^{Q} \gamma_{k,q}^2 (\mathbf{s}_k \odot \mathbf{x}_q)(\mathbf{s}_k \odot \mathbf{x}_q)^{\mathrm{T}}}_{\mathbf{K_I}} + \sigma_p^2 \mathbf{K_p} + \sigma_e^2 \mathbf{I}.$$

The set $\boldsymbol{\theta}_{\mathrm{K}} = \{\boldsymbol{\alpha}^2, \boldsymbol{\beta}^2, \boldsymbol{\gamma}^2, \sigma_p^2, \sigma_e^2\}$ denotes all kernel parameters. The relevance (variance) of individual direct factor effects, direct SNP effects and factor-SNP interactions is controlled by the relevance parameters parameters $\alpha_q^2, \beta_k^2, \gamma_{k,q}^2$ respectively.

3.2.1 Inference

The large number of SNPs in real-world datasets renders learning the relevance parameters for all K genetic effects (β_k^2) and $K \times Q$ interaction terms ($\gamma_{k,q}^2$) in Equation (3.2) infeasible, both computationally and statistically (see also Section 3.2.2.4). However, it is safe to assume sparsity where only a small fraction of all genome-wide SNPs have a non-zero SNP effect or SNP-factor interaction effect [Smith and Kruglyak, 2008; Stranger et al., 2007]. In the following, we call SNPs with a non-zero main effect or interaction effect *active*; the relevance parameters (β_k^2 and $\gamma_{k,q}^2$) of all remaining SNPs are implicitly assumed to be zero, which is equivalent to them being dropped from the model. We exploit this assumption to construct an algorithm similar, in principle, to expectation maximization (EM). Let us denote the set of active direct effect SNPs ($\beta_k^2 > 0$) as $\mathcal{S}$. Analogously, the set of active SNP-factor pairs with non-zero relevances ($\gamma_{k,q}^2 > 0$) will be denoted $\mathcal{I}$. Inference in the full model is then achieved by alternating between two operations. First, the factors $\mathbf{X}$ and model parameters $\boldsymbol{\theta}_{\mathrm{K}}$ are learnt for given active sets $\mathcal{S}$ and $\mathcal{I}$. Second, for fixed state of $\mathbf{X}, \boldsymbol{\theta}_{\mathrm{K}}$, additional SNPs are added to the active sets $\mathcal{S}$ and $\mathcal{I}$ using a greedy forward selection strategy. A specific schedule of these updates is used to ensure convergence to accurate solutions.

In Section 3.2.2, we describe this EM-like iterative training scheme. The technical building blocks of the individual training steps are presented in Section 3.2.2.1, describing the gradient-based optimization of model parameters and in Section 3.2.2.2, addressing the selection of SNPs to be included in the model.

3.2.2 Iterative training of LIMMI

Training is achieved in three steps. First, the state of the environmental factors $\mathbf{X}$ and the model parameters $\boldsymbol{\theta}_{\mathrm{K}}$ is inferred for empty active sets, where both the set of SNPs with a direct effect ($\mathcal{S}$) and the interactions ($\mathcal{I}$) have no elements. The necessary parameter inference for given active sets is achieved using a gradient-based optimization approach (see Section 3.2.2.1). As shown in the previous chapter, this simplistic inference that ignores the effect of genotype, may result in learnt hidden factors that are correlated with genotype and hence

have a genetic component. To rule out genetic control of the latent factors, SNPs that are correlated with these hidden variables are included in the set $\mathcal{S}$ (see Section 3.2.2.2), and the model parameters and factors are retrained. This process is iterated until no additional SNPs reach genome-wide significance for association to any of the learnt factors $\mathbf{X}$. As a result of this process, genotype and the learnt hidden factors are orthogonal (see Chapter 2 for further details).

Once the environmental factors have been determined, genotype-environment interactions are detected and SNP-factor pairs that participate in a significant interaction are included in the set $\mathcal{I}$ (Section 3.2.2.3). The model parameters are once again updated. This step completes the training. Individual components of the final covariance can then be used to test for specific hypotheses; see Section 3.2.3.

3.2.2.1 Gradient-based inference of covariance parameters

If the SNP effects and interactions are only present for a defined active set of direct SNP effects ($\mathcal{S}$) and interactions between pairs of SNPs and factors ($\mathcal{I}$) the full likelihood in Equation (3.2) reduces to

$$p(\mathbf{Y} \mid \mathbf{S}, \mathbf{X}, \boldsymbol{\theta}_{\mathrm{K}}, \mathcal{I}, \mathcal{S}) = \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{y}_g \mid \mathbf{0}, \boldsymbol{\Sigma}\right), \tag{3.3}$$

$$\boldsymbol{\Sigma} = \underbrace{\sum_{\forall k \in \mathcal{S}} \beta_k^2 \mathbf{s}_k \mathbf{s}_k^\top}_{\mathbf{K_S}} + \underbrace{\sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^\top}_{\mathbf{K_X}} + \underbrace{\sum_{\forall (k,q) \in \mathcal{I}} \gamma_{k,q}^2 (\mathbf{s}_k \odot \mathbf{x}_q)(\mathbf{s}_k \odot \mathbf{x}_q)^{\mathrm{T}}}_{\mathbf{K_I}} + \sigma_p^2 \mathbf{K_P} + \sigma_e^2 \mathbf{I},$$

where $\boldsymbol{\Sigma}$ is the overall covariance, which in turn is parametrized by $\mathbf{X}$,$\boldsymbol{\theta}_{\mathrm{K}}$ and the active sets $\mathcal{S}$ and $\mathcal{I}$. Here, we have dropped the mean effect to unclutter the notation and the summation is restricted to the elements in the respective active sets. The log of the marginal likelihood from Equation (3.3) can be written as

$$\begin{aligned} \ln p(\mathbf{Y} \mid \mathbf{S}, \mathbf{X}, \boldsymbol{\theta}_{\mathrm{K}}, \mathcal{I}, \mathcal{S}) &= \ln \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{y}_g \mid \mathbf{0}, \boldsymbol{\Sigma}\right) \\ &= -\frac{GN}{2} 2\pi - \frac{G}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \mathrm{Tr}(\boldsymbol{\Sigma}^{-1} \mathbf{Y} \mathbf{Y}^{\mathrm{T}}). \end{aligned} \tag{3.4}$$

Gradients of the marginal likelihood with respect to individual elements of $\mathbf{X}$ and hyperparameters $\boldsymbol{\theta}_{\mathrm{K}}$ can be calculated in closed form using the matrix derivative

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Sigma}} \ln \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{y}_g \,|\, \mathbf{0}, \boldsymbol{\Sigma}\right) = \boldsymbol{\Sigma}^{-1}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} - G\boldsymbol{\Sigma}^{-1}$$

and combing it with the covariance derivative with respect to the ith kernel parameter, $\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\Theta}_{K_i}}\boldsymbol{\Sigma}$, using the chain rule [Lawrence, 2005].

Parameter learning can then be done using a maximum likelihood approach, jointly determining the most probable state of the hidden environmental states $\mathbf{X}$ and model parameters $\boldsymbol{\theta}_{\mathrm{K}}$

$$\hat{\boldsymbol{\theta}_{\mathrm{K}}}, \hat{\mathbf{X}} = \underset{\boldsymbol{\theta}_{\mathrm{K}}, \mathbf{X}}{\operatorname{argmax}} \ln p(\mathbf{Y} \,|\, \mathbf{S}, \mathbf{X}, \boldsymbol{\theta}_{\mathrm{K}}, \mathcal{I}, \mathcal{S}). \tag{3.5}$$

A standard gradient-based optimizer, such as L-BFGS-B [Zhu et al., 1997], can be employed to take advantage of the availability of closed-form gradients with respect to the elements of $\mathbf{X}$ and $\boldsymbol{\theta}_{\mathrm{K}}$. A discussion on how the latent dimensionality Q is chosen and the implications on the model fitting is provided in Section 2.2.5.

3.2.2.2 Inclusion of genetic effects

Individual SNPs are selected for inclusion in $\mathcal{S}$. We follow the approach taken in Chapter 2, and test for correlation between individual factors $\mathbf{x}_1, \ldots, \mathbf{x}_Q$ and all genome-wide SNPs $\mathbf{s}_1, \ldots, \mathbf{s}_K$. In each iteration, SNPs that are in significant association (assessed using q-values [Storey and Tibshirani, 2003] $qv \leq \alpha_{\mathrm{SNP}}$) are added to the active set $\mathcal{S}$. The exact cutoff α_{SNP} is not critical as it merely alters the number of SNPs in the model, thereby affecting computational speed. Robustness with respect to to this significance cutoff has previously been discussed in section 2.2.3.

3.2.2.3 Inclusion of interaction effects

After the iterative procedure to determine the state of the environmental factors has converged, it is possible to test for interactions between factors and individual SNPs. We do so by exhaustively testing for interactions between SNPs $k \in \{1, \ldots, K\}$ and factors $q \in \{1, \ldots, Q\}$ (Section 3.2.3.1). Significant interaction terms ($qv \leq \alpha_{\text{GxE}}$) are then added to the active set $\mathcal{I}$. Finally, LIMMI relearns all the model parameters while taking into account the newly-added interactions, which allow the model to explain non-linear dependencies due to genotype-environment interactions.

3.2.2.4 Identifiability and robustness

Naive inclusion of all possible effects is both computationally intractable and statistically not identifiable as this would result in $K + (Q \cdot K)$ relevance parameters. Greedy step-wise strategies, on the other hand suffer from convergence to local optima. To reduce such side effects, we enforce sparsity in a two-step procedure. First, a cutoff is used for the inclusion of genetic markers ($\alpha_{\text{SNP}} \leq 0.1$ in the case of the yeast dataset presented in section 3.3.2) and interaction terms ($\alpha_{\text{GxE}} \leq 0.05$ again in the case of the yeast dataset) into the model. Then, irrelevant variance parameters (β_k^2, $\gamma_{k,q}^2$) are set to zero during inference by means of automatic relevance determination [MacKay, 1995]. The empirical stability of this approach has been explored in the previous chapter. In particular, in Section 2.3 we have investigated the robustness of the model while varying the cutoff for inclusion of genetics effects ((α_{SNP}).

Although we have taken measures to ensure that the learnt factors are likely environmental, there are fundamental limitations on statistical identifiability. The correct identification of factors that exhibit genotype-specific interactions affecting large numbers of target genes is particularly challenging. The variance explained by such an interaction hotspot can be similar to the variance of a direct factor effect, such that a single factor may mistakingly be learnt as two separate factors. When testing for interactions with the main effect factor, the second one can explain away the interaction signals and hence the interaction hotspot may not be detected. Thus, our approach depends on the assumption that the

direct contribution of environmental factors dominates genotype-specific effects. This assumption is reasonable in practice and we found LIMMI to be robust with respect to deviations from it (Figure 3.2(c-d)).

3.2.2.5 Computational efficiency

There are two components of the LIMMI model that determine the computational complexity. First, the Gaussian process latent variable model (Section 3.2.2.1), estimating the covariance parameters and the environmental factors, has a complexity that is independent of the number of genes. Instead, its runtime is dominated by inversions of the covariance matrix, which scale cubically with the number of samples. Thanks to modern linear algebra implementations, these computations are tractable even for thousands of samples. Second, given the latent variables, LIMMI carries out mixed model interaction and association tests relating inferred factors, genes and SNPs. Here, we build on recent advances in mixed models [Lippert et al., 2011], reducing the computational complexity of these statistical tests to a cost that is linear in the number of samples and tested hypotheses. Moreover, this second step can easily be parallelized across hypotheses, which is supported in our software implementation.

As a result, LIMMI can be applied to human-scale datasets with hundreds of samples, $\sim 50,000$ gene expression levels and $\sim 100,000$ SNPs. For example, on the yeast dataset analyzed in Section 3.3.2, LIMMI converged within 50 minutes[1]. This datasets contained 218 samples, $2,956$ SNPs and $5,493$ gene expression levels.

3.2.3 Statistical association and interaction testing

The ultimate goal is to use the covariance models described above to carry out tests for genetic associations (eQTLS) as well as tests for genotype-environment interactions. Statistical testing is also used to iteratively expand the LIMMI covariance model (Section 3.2.1).

[1]Implementation based on a Gaussian Processes framework in python, while association and interaction scans are implemented in C++. Runtime estimates are given for a GNU/Linux machine with an Intel(R) Xeon(R) X7542 12C CPU and 64 gigabytes of RAM. The python scientific libraries (Numpy and Scipy) were compiled against the Intel(R) Math Kernel Library.

For testing, we employ a strategy based on linear mixed models, where a fitted covariance structure $\mathbf{\Sigma}$ accounts for confounding and other factors that cause expression variability, whereas the fixed effect assess the relevance of the effect of interest

$$p(\mathbf{y}_g \,|\, \sigma_g^2, \delta_g, \mathbf{\Sigma}) = \mathcal{N}\left(\mathbf{y}_g \,\middle|\, \underbrace{\mathbf{f}()}_{\text{fixed effect}}, \underbrace{\sigma_g^2(\mathbf{\Sigma}_g + \delta_g \mathbf{I})}_{\text{random effect}}\right). \tag{3.6}$$

The overall variance of the trait σ_g^2 can be efficiently determined in closed form for each test (SNP-gene pair), whereas δ_g requires a grid-based optimization. We employ the approximation proposed in Kang et al. [2008b]; Lippert et al. [2011], and determine δ_g once on the null model, and keep this variance ratio fixed for all genome-wide tests.

When testing for associations and interactions with LIMMI, the covariance $\mathbf{\Sigma}$ is intended to capture the effects from other SNPs, confounding factors and interactions. The covariance is derived from the components fitted on the null model (Section 3.2.1).

In Section 3.2.3.1, we describe the covariance used for genotype-factor interaction tests (genotype-environment interactions). Association tests between genotype and expression traits (eQTL) are described in Section 3.2.3.2. Both, for interaction and association scans we obtain p-values by applying a likelihood ratio test. For genome-wide significance estimates we used false discovery rate estimates from the q-value package [Storey and Tibshirani, 2003].

3.2.3.1 Interaction test

The likelihood ratio corresponding to the test for a particular SNP k and factor q affecting gene g can be expressed as

$$\text{LOD}^{\text{inter}}_{k,q,g} = \log \frac{\mathcal{N}\left(\mathbf{y}_g \,\middle|\, \theta_{i,g}(\mathbf{s}_k \odot \mathbf{x}_q) + \theta_{k,g}\mathbf{s}_k + \theta_{q,g}\mathbf{x}_q, \sigma_g^2(\mathbf{\Sigma}_a + \delta_g \mathbf{I})\right)}{\mathcal{N}\left(\mathbf{y}_g \,\middle|\, \theta_{k,g}\mathbf{s}_k + \theta_{q,g}\mathbf{x}_q, \sigma_g^2(\mathbf{\Sigma}_a + \delta_g \mathbf{I})\right)}, \tag{3.7}$$

where $\theta_{i,g}, \theta_{k,g}$ and $\theta_{q,g}$ correspond to the fitted fixed effect weight of the interaction term, the SNP effect and the factor effect respectively. We have dropped the

mean effect μ_g to unclutter the notation. The background covariance includes all other additive effects and is defined as $\mathbf{\Sigma}_a = \sigma_p^2 \mathbf{K_P} + \sum_{q' \neq q} \alpha_{q'}^2 \mathbf{x}_{q'} \mathbf{x}_{q'}^\top$, accounting for known covariates and the direct effects of all factors but factor q which is tested.

3.2.3.2 Association test

Analogous likelihood ratio tests can be derived for the hypothesis that SNP k is in association with gene g

$$\text{LOD}_{k,g}^{\text{asso}} = \log \frac{\mathcal{N}\left(\mathbf{y}_g \,\middle|\, \theta_{k,g}\mathbf{s}_k, \sigma_g^2(\mathbf{\Sigma}_i + \delta_g \mathbf{I})\right)}{\mathcal{N}\left(\mathbf{y}_g \,\middle|\, \mathbf{0} \quad\quad , \sigma_g^2(\mathbf{\Sigma}_i + \delta_g \mathbf{I})\right)}. \tag{3.8}$$

Here, the fixed-effect term includes the direct effect of the SNP and the confounding covariance accounts for direct effects of the learnt environmental factors ($\mathbf{K_X}$) as well as the detected interactions ($\mathbf{K_I}$) i.e. $\mathbf{\Sigma}_i = \mathbf{K_X} + \mathbf{K_I} + \mathbf{K_P}$. Again, we have dropped the mean effect term from equation 3.8.

3.2.4 LIMMI-sva

In principle, in the first step of the procedure outlined in Section 3.2.2.1, any latent variable model could be used to infer environmental factors. For comparison, we have implemented and compared to a variant of LIMMI called LIMMI-sva. LIMMI-sva uses surrogate variable analysis (SVA) [Leek and Storey, 2007], which does not encourage orthogonality of learnt factors and genotype and does not rely on the iterative model refinement described in Section 3.2.2. The implementation of LIMMI-sva is straightforward and relies on just two steps. First, an estimate of the latent factors is obtained using SVA Leek and Storey [2007]. The resulting $\hat{\mathbf{X}}$ can then be directly used in associations and interaction tests

We also considered a variant of LIMMI-sva, called LIMMI-sva-cov that in addition to known covariates, also accounts for the direct effect of all the factors not being tested ($\mathbf{\Sigma}_a = \mathbf{K_P} + \sum_{q' \neq q} \alpha_{q'}^2 \mathbf{x}_{q'} \mathbf{x}_{q'}^\top$).

3.3 Results

We evaluated the ability of LIMMI to retrieve genuine genotype-environment interactions. In particular, we studied the relative performance of two approaches, LIMMI and LIMMI-sva, that share the same testing procedure but infer the unknown environment in different ways (Section 3.2.4). We also considered a standard linear association test as a baseline method.

3.3.1 Simulation study

First, we tested LIMMI on simulated data, where the underlying true associations and genotype-environment interactions are known. The simulation procedure largely follows previous studies to assess the performance of eQTL methods [Fusi et al., 2012; Listgarten et al., 2010]. Each simulated dataset consisted of 800 SNPs simulated as from an F2 cross and 1,000 gene expression levels. We simulated 5 environmental factors that have both direct effects on gene expression and interactions with genotype. In addition, we also considered 5 simulated technical factors that affect gene expression directly but are independent of genotype.

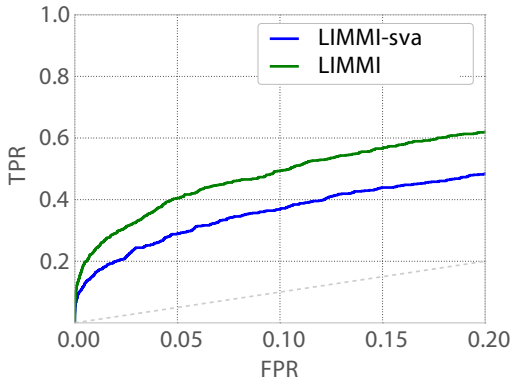
The factor profiles were independently drawn from $\mathcal{N}(0, 1)$ and the effect sizes of factors q on genes g was sampled from $w_{g,q} \sim \mathcal{N}(0, 0.45)$, which is similar to empirical estimates from the yeast dataset [Smith and Kruglyak, 2008]. We added 800 simulated associations with effect sizes sampled from $w_{g,k} \sim \mathcal{N}(0, 0.05)$ as well as 5 interactions between randomly chosen pairs of genetic loci and environmental factors, each affecting 15% of the genes and with an effect size sampled from $\mathcal{N}(0, 0.15)$. Broad genetic effects, such as *trans*-acting genetic variants, can complicate the recovery of the confouding factors ([Fusi et al., 2012], Chapter 2). If the genetics and the environment are not modelled jointly, part of the genetic signal will be captured by the estimated confouding factors, making the discovery of genotype-environment interactions even harder. In order to further investigate this hypothesis, we simulated 5 broad *trans*-acting genetic variants each affecting 20% of the genes and with an effect size sampled from $\mathcal{N}(0, 0.2)$. Finally, we added independent measurement noise to each gene $\psi_g \sim \mathcal{N}(0, 0.15)$. The simulation framework employed here does not favor any of the considered methods, since they all share the assumption that the environmental state is characterized

by few environmental factors, i.e. is low rank.

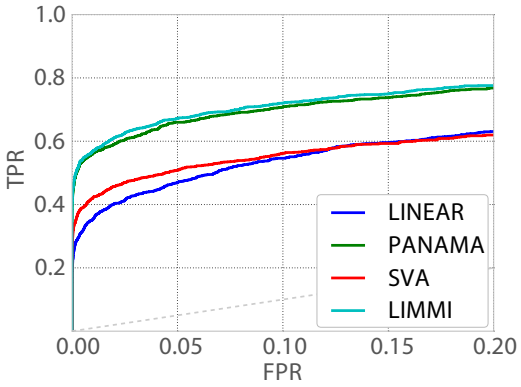
First, we checked that factor models like LIMMI are able to recover environmental variables and gene-environment regulatory interactions. Figure 3.2a depicts the receiver operating characteristics (ROC), assessing the true positive rate of alternative methods as a function of the permitted false positive rate (FPR). For practical applications, the regime of few false positives is most relevant and hence we consider the ROC analysis on the range of FPR between 0 and 0.2. Determining an explicit mapping between the learnt environmental factors and the simulated ones is difficult and may introduce biases. Thus, we assessed the accuracy of recovering SNP-gene pairs with a detected interaction for any of the learnt environmental factors. Both LIMMI-sva and LIMMI detected many of the simulated genotype-environment interactions, where LIMMI significantly outperformed LIMMI-sva.

Next, we evaluated alternative methods for detecting eQTLs, i.e. direct associations between polymorphic loci and gene expression levels that are not environment specific (Figure 3.2b). Standard linear regression (LINEAR) ignores the presence of unknown environmental factors, which resulted in a poor recovery of true associations. SVA and PANAMA account for the direct effect of learnt environmental factors, resulting in a considerable improvement compared to the linear model (see also discussion in Chapter 2 and in Fusi et al. [2012]; Listgarten et al. [2010]; Stegle et al. [2010]). Finally, LIMMI also accounts for both the learnt environmental factors and their interactions with the genetic state, resulting in a marginal but consistent improvement over PANAMA.

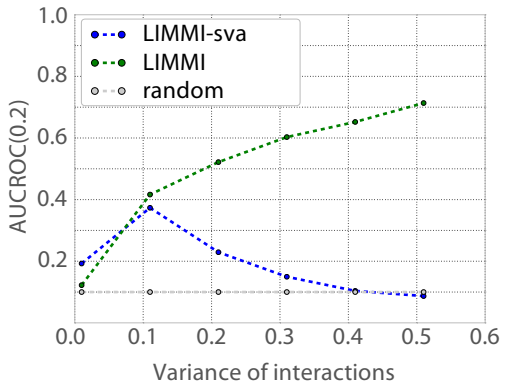
Finally, we investigated the impact when changing the relative magnitude of direct environmental effects and genotype-environment interactions. Figure 3.2c and Figure 3.2d show the respective area under the ROC (AUC) when varying the relative fractions of variance explained by genotype-environment interactions and direct environmental effects. In each plot, the leftmost point corresponds to a setting with very small (0.01) relative proportion of variance explained by interactions whereas the rightmost point corresponds to an equal proportion (0.50) of variance explained by direct effects and interactions. As expected, the ability of LIMMI to detect genotype-environment interactions improved with larger relative effect sizes of the interactions (Figure 3.2c), whereas the performance



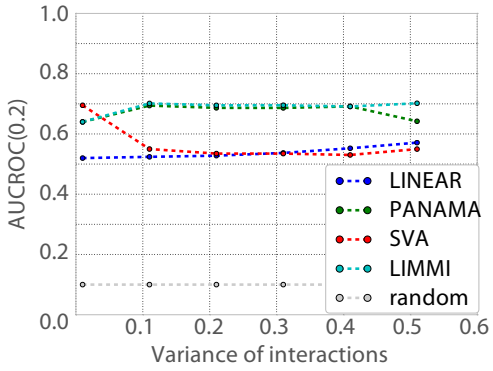
(a) Genotype-environment interactions ROC



(b) eQTL Associations ROC



(c) Genotype-environment interactions AUC



(d) eQTL associations AUC

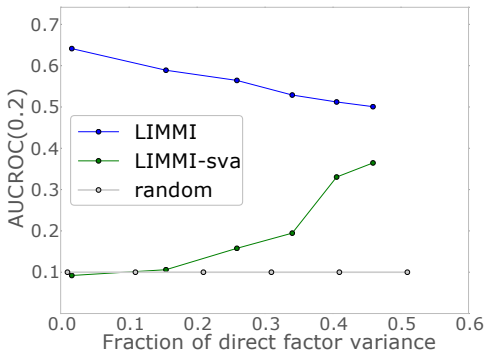
Figure 3.2: **Comparative evaluation of LIMMI and alternative methods on simulated datasets.** **(a)** Receiver operating characteristics (ROC) for recovering simulated interactions between hidden factors and genotype. Linear regression has been omitted because it is not applicable to test for hidden environment interactions. The light grey line indicates the expected performance of a random predictor. **(b)** ROC for recovering simulated associations between genotype and expression. SVA, PANAMA and LIMMI account for the learnt environmental factors during testing, thus outperforming the linear model. LIMMI yields a slightly better ROC than PANAMA, indicating that accounting for interaction effects improves the ability to detect true associations. Area under the ROC for detection of simulated interactions **(c)** and associations **(d)** as a function of the relative variance explained by genotype-environment interactions versus direct factor effects.

of LIMMI-sva degraded when the relative variance explained by interactions exceeded 10%. This observation exemplifies the model misfit of approaches like SVA that ignore genotype-environment interactions during inference. Analogous conclusions hold when considering the performance of the considered methods to detect direct associations or eQTLs (Figure 3.2d). Here, PANAMA came close second and again SVA degraded in performance for increasing relevance of the interaction terms. Remarkably, starting from 30% of the variance explained by genotype-environment interactions, a standard linear association test that ignores unknown environments entirely yielded more accurate results than SVA. A possible explanation for this result is that SVA recovers progressively worse estimates of the latent confounders as the importance of the simulated non-additive confounding component (i.e. due to interaction effects) increases.

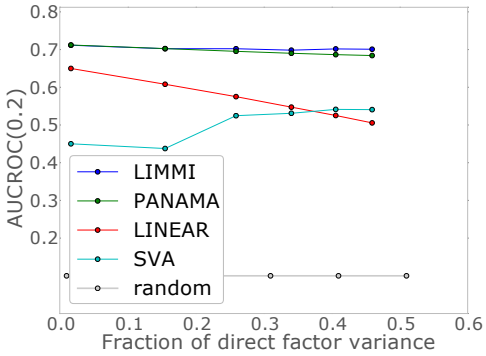
In addition to varying the relative proportion of interactions and direct environmental effects, we also considered varying the variance of each effect type in isolation. Figure 3.3 shows analogous AUC performances when varying the variance explained by direct factor effects (Figure 3.3a-b) or the variance from genotype-factor interactions (Figure 3.3c-d), keeping the other term constant. In contrast to alternative methods, LIMMI was able to detect genotype-environment interactions even for weak interaction effects ($< 10\%$, Figure 3.3c), suggesting that the method is suitable in studies where genotype-environment interactions have a subtle effect.

LIMMI is related to previous approaches, such as SVA [Leek and Storey, 2007] and PANAMA (Chapter 2), that have predominantly been intended to identify and account for the effect of technical factors. To assess the effect of technical factors versus environmental effects, we considered a series of simulated settings, changing the relative proportions of environmental and technical factors. In principle, LIMMI will retrieve both types of factors on equal footing, however only environmental influences are expected to yield interactions with the genetic state.

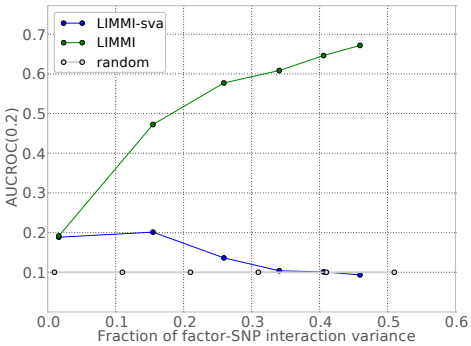
Indeed, the results presented in Figure 3.4 support that even when almost all factors are technical and do not interact with genotype, LIMMI is still able to recover the small number of genuine genotype-environment interactions.



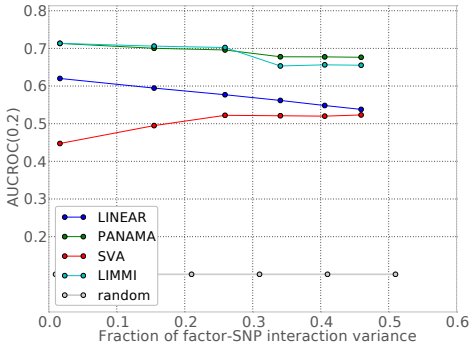
(a) AUC for recovery of genotype-environment interactions for different direct factor effect sizes



(b) AUC for recovery of eQTL associations for different direct factor effect sizes

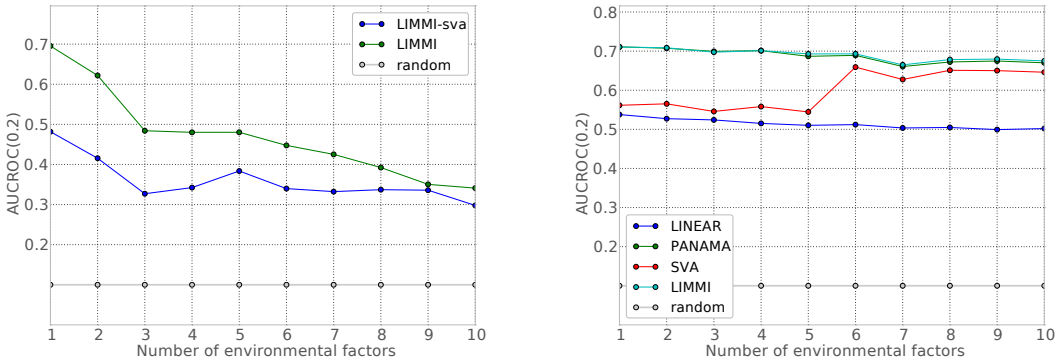


(c) AUC for recovery of genotype-environment interactions for different factor-SNP interaction effect sizes



(d) AUC for recovery of eQTL associations for different factor-SNP interaction effect sizes

Figure 3.3: Performance comparison of alternative methods for recovering genotype-environment interactions (**a,c**) and direct eQTLs (**b,d**). **a,b**: area under the receiver operating curve in the FPR interval 0..0.2 (AUC0.2) for different effect sizes of direct contribution of environmental factors, keeping all other effect sizes fixed. For larger effect sizes, estimation of the hidden environmental state is easier and hence PANAMA and LIMMI-sva approach the same performance (**a**). At the same time, the difference between PANAMA and LIMMI for discovering eQTL increases (**b**). **c,d**: AUC for increasing variance explained by factor-SNP interactions, while keeping all other variance components fixed. LIMMI is able to make useful predictions starting from 10% relative variance explained. The performance difference compared to LIMMI-sva is most pronounced for strong interactions.



(a) AUC for recovery of genotype-environment interactions

(b) AUC for recovery of genotype-environment associations

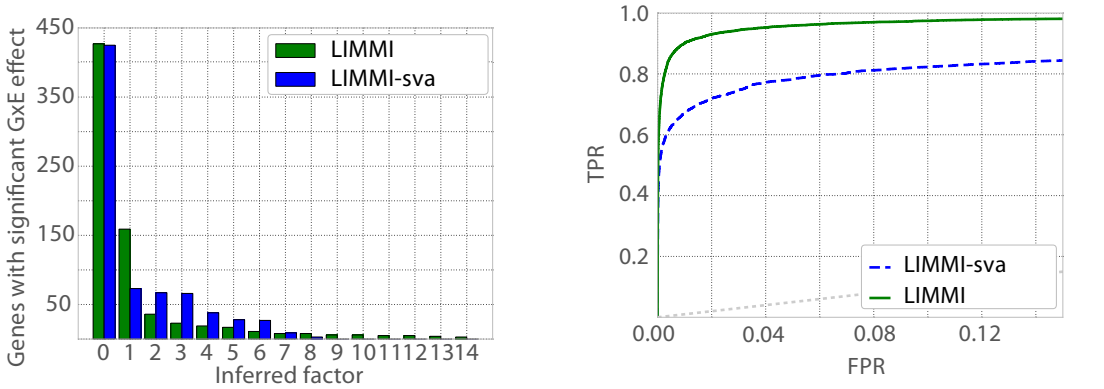
Figure 3.4: Analysis of the sensitivity against batch effects on a simulated dataset. The leftmost point in both plots corresponds to a setting where there's only 1 true environmental factor interacting with the genotype and 9 batch effects not interacting with the genotype. The rightmost point corresponds to a setting where there are 10 environmental factors and 0 batch effects. **(a)** measures the ability to correctly detect genotype-environment interactions, whereas **(b)** measures the ability to detect eQTL associations.

3.3.2 Applications in yeast genetics of gene expression

We revisited the yeast study from Smith and Kruglyak [2008], studying genetic regulation of gene expression as a function of environmental background. In this study, an F2 population of yeast strains has been expression profiled in two contrasting growth media: glucose and ethanol. Thus, the growth medium is a strong and likely dominant environmental factor. In the primary analysis, both major direct genetic effects (associations) as well as prevalent genotype-environment interactions have been reported [Smith and Kruglyak, 2008].

LIMMI accurately recovers the genotype-environment interactions with a measured environmental factor

We applied LIMMI and LIMMI-sva to the yeast dataset without providing knowledge about the measured environmental factor that corresponds to the growth medium as an input. SVA identified 9 latent factors and LIMMI found 15 factors.



(a) Number of interacting genes per factor (b) ROC for recovering ethanol/glucose GxE

Figure 3.5: **Recovery of known and novel gene-environment interactions**. **(a)** The number of genes with at least one significant genotype-environment interaction (FDR ≤ 0.01) as identified by LIMMI and SVA. The first factor was most correlated with the measured ethanol/glucose contrast, capturing this experimental conditions. **(b)** ROC curves for LIMMI-sva and LIMMI, assessing the accuracy of recovering pairs of genetic loci and genes in statistical interactions with the first factor. Ground truth information was derived from genotype-environment tests with the measured environment (FDR ≤ 0.01). The dashed line indicates the accuracy of a random predictor.

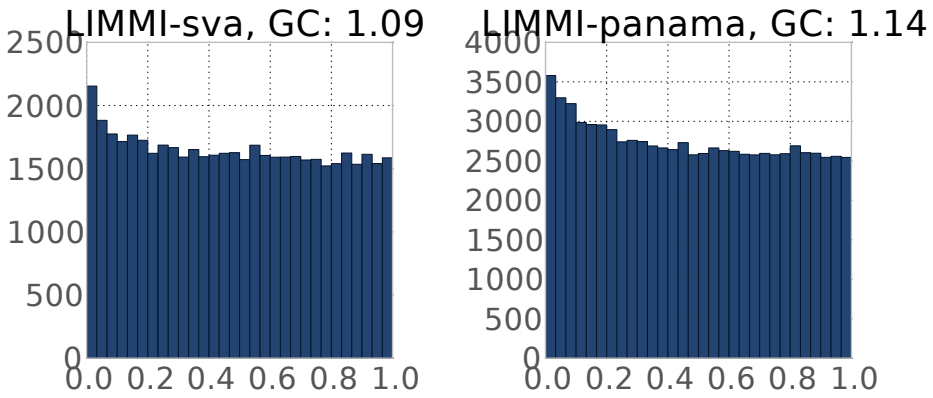


Figure 3.6: P-value histograms and inflation factors for interaction tests on the smith datasets.

When considering each learnt factor to test for genotype-environment interactions with individual gene expression levels, LIMMI-sva retrieved a larger number of genes with significant effects than LIMMI (Figure 3.5a, at comparable statistical

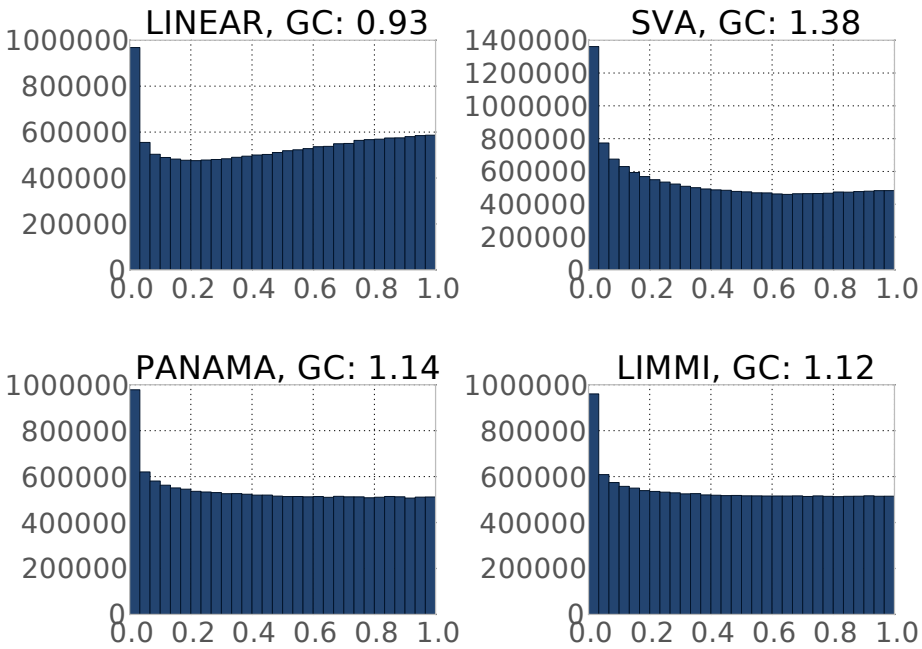


Figure 3.7: P-value histograms and inflation factors for association test on the yeast dataset.

calibration; see Figures 3.6 and 3.7). For both methods, the factor with the greatest number of genotype/factor interactions was strikingly correlated ($r \geq 0.99$) with the known environmental state that corresponds to the ethanol/glucose condition. Other factors were largely uncorrelated with this known environmental variable (Figure 3.8), suggesting that the first factor indeed captures most of the effect due to the ethanol/glucose condition.

First, we focused on the recovered factor that is a likely proxy for the true environmental state. Figure 3.5b depicts the ROC curve, assessing the accuracy of genotype-environment interactions recovered by LIMMI and LIMMI-sva when using genotype-environment effects with the known environment as ground truth (as done in Smith and Kruglyak [2008]). LIMMI outperformed LIMMI-sva, which is likely due to a combination of two important differences between these methods. First, LIMMI incorporates a constraint such that recovered factors are uncorrelated with genotype, whereas many of the factors retrieved by

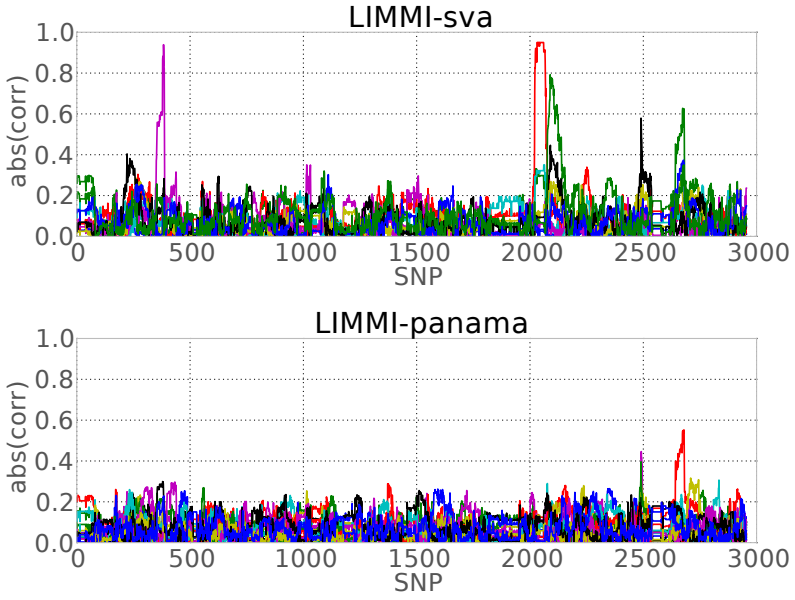


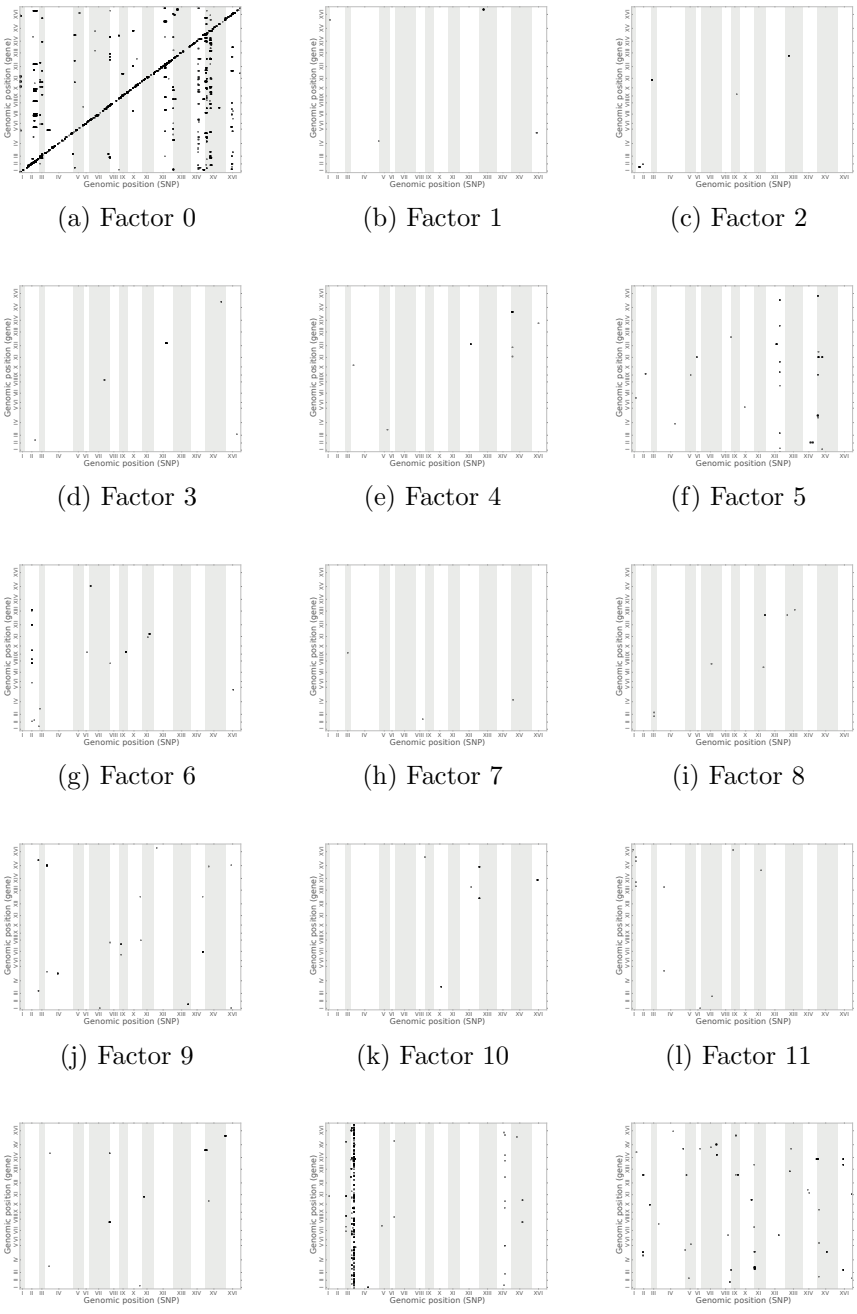
Figure 3.8: Correlation between genome-wide SNPs and learnt factors for LIMMI-sva and LIMMI. With few exceptions, LIMMI retrieved factors that are not genetically driven and hence environmental.

SVA are themselves under genetic control (Figure 3.8). Second, the statistical test for interactions employed in LIMMI accounts for direct effects of all other learnt factors, explaining away nuisance variation due to other environmental axes (Section 3.2.3.1).

Novel genotype-environment interactions with unknown environmental effects

In addition to interactions that correspond to the known environmental factor of the glucose/ethanol contrast, both LIMMI-sva and LIMMI retrieved additional factors, which were considered for possible GxE interactions (Figure 3.5a). The factors recovered by LIMMI-sva tended to be in strong association with genotype, suggesting that they capture genetic signals instead of environmnental effects. The factors retrieved by LIMMI, on the other hand, were found to be orthogonal to the genetic signal (Figure 3.8).

A map of the genetic loci and regulated genes for interactions with all factors



(a) Factor 0 (b) Factor 1 (c) Factor 2

(d) Factor 3 (e) Factor 4 (f) Factor 5

(g) Factor 6 (h) Factor 7 (i) Factor 8

(j) Factor 9 (k) Factor 10 (l) Factor 11

(m) Factor 12 (n) Factor 13 (o) Factor 14

Figure 3.9: Map of genotype-environment interactions recovered when applying LIMMI to the yeast dataset.

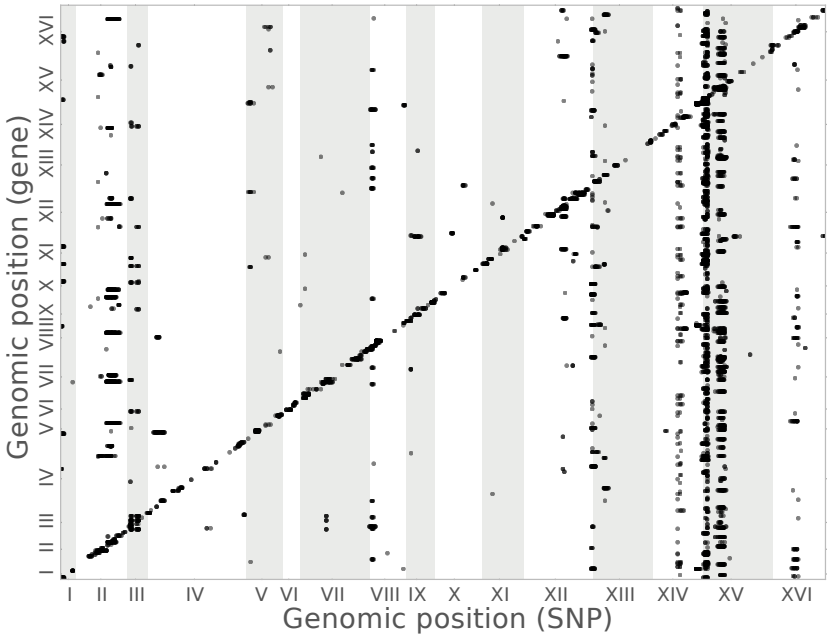


Figure 3.10: Map of genotype-environment interactions recovered when using the known environmental state.

detected by LIMMI is shown in Figure 3.11 (Interaction results for each individual factor are given in Figure 3.9). Notably, genotype-environment interactions with the factor that recapitulates the ethanol/glucose effect (Factor 0) were enriched in the proximity of the regulated genes, suggesting a *cis* mechanism. Other factors yielded interactions that involve distal loci and hence have a putative *trans* mechanism. A particularly prominent hotspot appeared for factor 13 in chromosome 4, where LIMMI detected genotype-environment interactions involving 10 distinct SNPs in that region. In the direct vicinity of these SNPs ($\pm$ 10kb), there were 6 annotated genes, four of which have previously been reported as implicated with temperature response (YDL143W, YDL139C, YDL135C, YDL132W) [Auesukaree et al., 2009; Patton et al., 1998; Shimon et al., 2008; Stoler et al., 2007; Tiedje et al., 2008]. This enrichment suggests that factor 13 may explain subtle temperature variation in the experiment.

Figure 3.10 depicts the interaction map when using the known environmental

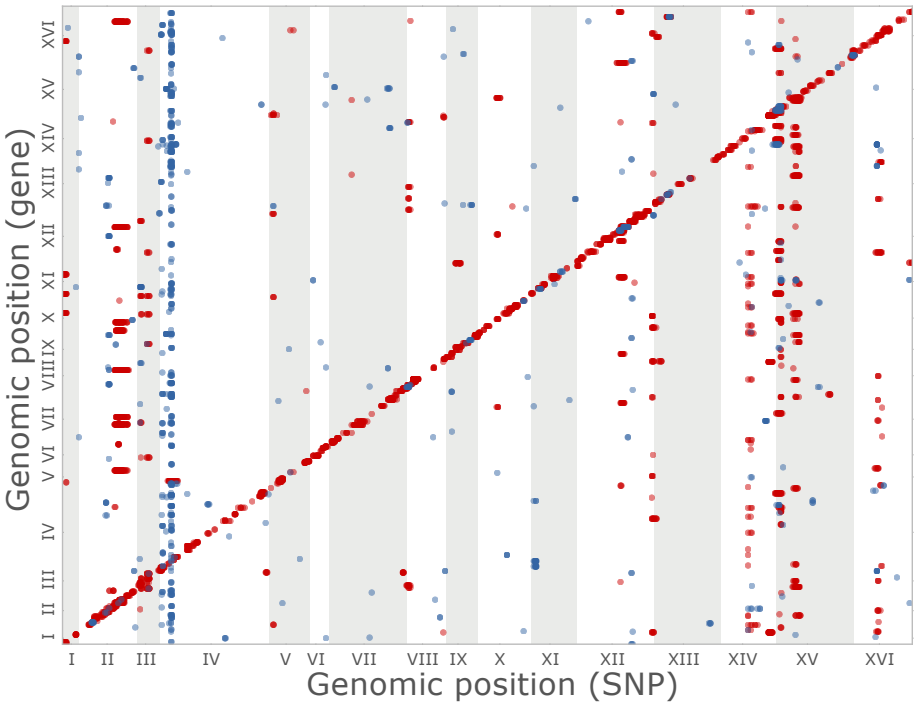
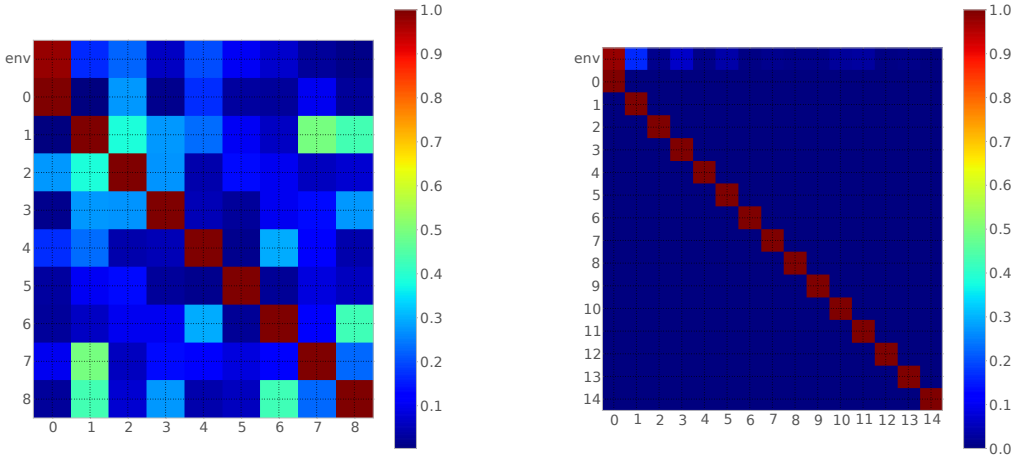


Figure 3.11: **Genomic map of the genotype-environment interactions retrieved by LIMMI** (FDR ≤ 0.01). Shown are the position of the SNP (x-axis) and the gene (y-axis) that participate in each significant genotype-environment interaction. Red circles correspond to interactions with the first latent factor that captures the known ethanol/glucose contrast. Blue interactions correspond to all other 14 factors.

condition (glucose/ethanol) to test for genotype-environment interactions. The results obtained in the latter are remarkably similar to the ones obtained to LIMMI interactions on Factor 0, which is in line with the ROC analyses discussed earlier (Figure 3.5b). Overall, LIMMI identified more *trans* bands for genotype-environment effects than LIMMI-sva.



(a) LIMMI-sva

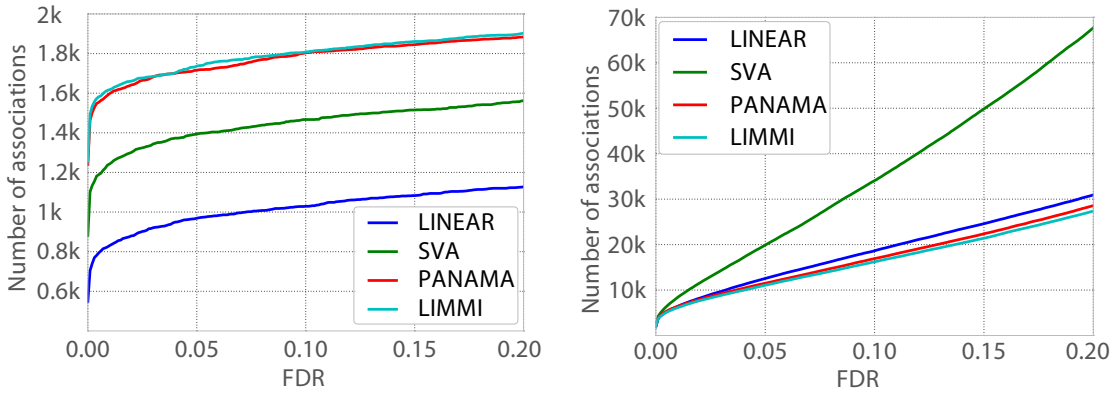
(b) LIMMI

Figure 3.12: Correlation coefficients between the known environmental factor (glucose/ethanol) and the factors retrieved by **(a)** LIMMI-sva and **(b)** LIMMI. Both methods recover one factor that appears to be strikingly correlated with the true environmental state (labelled as "env" in the plot).

Genotype-environment interaction hotspot may confound genetic association analyses

Finally, we considered the ability of different models to call direct eQTL associations between genetic loci and individual gene expression levels.

Figure 3.13 shows the number of associations retrieved by alternative methods as a function of the false discovery rate cutoff. As in the simulated settings (Figure 3.2), LIMMI accounts for the interaction effects found, which controls for nuisance variation due to these effects. As a result, LIMMI identified additional *cis* eQTLs, while the number of *trans* eQTLs decreased when compared to PANAMA. At the same time, the p-values statistics of LIMMI was slightly more uniform than PANAMA, suggesting that better control for confounding has been achieved (Figures 3.6 and 3.7). While more uniform p-values support an improved calibration [Listgarten et al., 2010] of the methods presented here, some inflation of the test statistics was retained, which is an expected consequence of the presence of extensive *trans* hotspots (see Section 2.3 for a discussion). These



(a) *cis* association count

(b) *trans* association count

Figure 3.13: **Number of direct genetic associations (eQTLs) called by different methods as a function of the FDR cutoff. (a)** *cis* associations. **(b)** *trans* associations. We considered at most one association per chromosome in order to avoid confounding the size of associations with their number.

results suggest that including interaction terms into the model can also be beneficial to identify direct genetic effects in real studies. On one hand, this finding supports the conjecture that the interactions retrieved by LIMMI are indeed genuine, since they explain variance that cannot be captured by a model that relies on fully additive effects. Conversely, it is clear that genotype-environment effects contribute to gene expression variability and accounting for their effect in genetic analyses has similar benefits than accounting for hidden confounding [Fusi et al., 2012; Listgarten et al., 2010; Stegle et al., 2010, 2012] or correcting for population structure [Price et al., 2006, 2010]. eQTLs retrieved by LIMMI have a slightly better *cis* enrichment compared to PANAMA, a criterion previously suggested to judge the plausibility of eQTL results [Fusi et al., 2012; Kang et al., 2008a; Listgarten et al., 2010].

3.4 Discussion

Here, we have presented a novel approach to detect genotype-environment interactions with unmeasured environmental factors. LIMMI is able to recover the unmeasured environmental state solely from gene expression data. Once learnt,

these variables can be used in genetic analyses to investigate interactions between environmental factors and genotype with a regulatory effect on gene expression traits.

Approaches like LIMMI are relevant for virtually any genetic study of high-dimensional molecular traits, in particular if the environmental state is only partially measured or remains entirely unknown [Gibson, 2008]. Here, we illustrated and assessed LIMMI in simulated examples and in retrospective analyses of data from yeast genetics. We compared genotype-environment interactions with learnt environments to interactions found when using explicit environmental measurements. First, LIMMI was able to accurately detect previously known interactions. Second, we found novel genotype-environment interactions beyond what can be detected when relying on the measured environmental state. These additional effects were predominantly *trans*-acting, with some loci having widespread effects on large fractions of the expression traits. In the case of the largest hotspot, the interacting locus overlapped with a group of genes involved in temperature sensitivity, providing a plausible explanation of the mechanistic underpinning of this finding. Finally, we have shown how the recovered interactions can be used to refine statistical testing procedures. Accounting for the effect of genotype-environment interactions within a LIMMI eQTL scan resulted in increased power to detect true associations in simulations and yielded improved test statistics on real data.

LIMMI is related to a range of existing factor models, in particular techniques that model hidden expression determinants to correct for their confounding effect. These methods can be broadly grouped in two classes: models that are aimed at retrieving a set of confounding factors explicitly (see Chapter 2 and Leek and Storey [2007]; Stegle et al. [2010]) and models that account for the variance introduced by confounding factors [Kang et al., 2008a; Listgarten et al., 2010]. In principle, any of the models that retrieves and explicit representation of factors can be used for interaction analyses like the one presented here. Specifically, in this paper we compared to a version of our method that was using SVA for this purpose. LIMMI is most closely related and builds on PANAMA (Chapter 2), however we propose a new route towards understanding the role of the environment in a genetic context rather than merely "correcting it away". For this

purpose, we extend PANAMA in several ways. First, we introduce a systematic approach to use inferred environments to test for genotype-environment interactions while accounting for the effect of unknown environments. Second, we show how the detected genotype-environment interactions can be used to further refine the statistical testing of eQTLs. Other methods like SVA [Leek and Storey, 2007], PEER [Stegle et al., 2012] and the method by Listgarten et al. [Listgarten et al., 2010] do not focus on recovering interactions per se, although we have created a modified variant of SVA for the purpose of comparison. The main shortcoming of these techniques is the lack of an effective mechanisms to ensure that the learnt factors are not driven by genotype, which leads to the inferior performance of LIMMI-sva in our experiments.

In conclusion, LIMMI is a methodological advance that allows for refined inference of environmental factors from molecular profiling data. When used in genetic analyses, these learnt variables help to improve the mechanistic understanding of molecular traits, thereby increasing the fraction of phenotype variability that can be explained. Approaches as the one presented here will become even more useful when dataset sizes increase further, providing sufficient power to estimate even more complex models and effect types between the genetic state, known and hidden environments and the transcriptional state of the cell.

3.5 Conclusions

In this chapter, we have proposed and described a model-based approach to simultaneously infer unmeasured environmental factors from gene expression profiles and use them in genetic analyses, identifying environment-specific associations between polymorphic loci and individual gene expression traits. As shown in the experiments, our method is able to accurately reconstruct environmental factors and their interactions with genotype in a variety of settings. In particular, in real data from yeast, our results suggest that interactions with both known and unknown environmental factors significantly contribute to gene expression variability.

So far, we have assumed that the noise distribution of the phenotype being studied (gene expression levels, in the case of Chapters 2 and 3) was Gaussian.

In some cases, this is a reasonable assumption, since the phenotype can be normalized or transformed so that this assumption is respected. Unfortunately, it's not always clear which transformation should be applied and selecting a suitable transformation for a specific dataset is still an open problem. In the next chapter we are going to present an approach to infer the optimal transformation given the data, showing that it leads to an increase of power in GWAS, more accurate heritability estimates and higher phenotype prediction accuracy.

Chapter 4

Warped Linear Mixed Models

The material presented in this chapter is joint work with Christoph Lippert, Neil Lawrence, and Oliver Stegle and has been published in *"Genetic Analysis of Transformed Phenotypes"* Fusi et al. [2014]

4.1 Overview

In the previous two chapters, we have used latent variable models to correct (Chapter 2) and find interactions (Chapter 3) with unobserved experimental factors. The phenotype being analyzed in both cases consisted of gene expression levels, and the latent variables were capturing inter-sample correlations caused by hidden factors. In this chapter we consider univariate phenotypes (even though extensions to the multivariate case are possible), and use latent variable models to alleviate the problem of misspecification of the noise model in genome-wide association studies. The standard linear mixed model is based on the assumption of Gaussian distributed residuals and deviations from it can result in model misspecification [McCulloch and Neuhaus, 2001]. In some special cases, such as binary case/control phenotypes, the true distribution of the phenotype and its residuals are defined *a priori*, motivating use of generalized linear mixed models with specific link functions such as the probit or the logit [McCulloch and Neuhaus, 2001]. However, the vast majority of phenotypes are quantitative and

their precise distribution is unknown [Valdar et al., 2006]. To address possible non-Gaussian residuals, it may be desirable to apply non-linear transformations to the phenotype data as a pre-processing step prior to genetic analysis. Manual assessment of different transformations within a predefined range of alternatives (e.g., log, root, inverse, etc) is common practice [Baranzini et al., 2009; Himes et al., 2009; Kathiresan et al., 2007; Wallace et al., 2008]. However, such an approach can be error-prone, introduces a multiple testing problem (due to repetition of the same analysis under multiple transformations) and can produce biases because the family of transformations that can be manually explored is limited. Moreover, different traits, even if related, may require different transformations [Baranzini et al., 2009; Valdar et al., 2006], and hence the selection of phenotype transformations has to be repeated for every phenotype. To avoid the rigidity of predefined transformations, adaptive procedures such as the Box-Cox transformation [Ahn et al., 2010; Box and Cox, 1964; Chiu et al., 2005; Huang et al., 2007; McCauley et al., 2005; Tian et al., 2011] or non-parametric transformations using rank statistics [Goh and Yap, 2009; Servin and Stephens, 2007; Stephens, 2013; Zhou and Stephens, 2013] have been used with some success. However, the problem of selecting a suitable transformation remains a major challenge, in particular as there is no objective measure of comparison to assess alternatives. This is because the goal is not to obtain Gaussian distributed phenotypes but instead Gaussian distributed residuals of an unknown genetic model. A second concern applies in particular to non-parametric rank transformations, which cannot be directly inverted. As a consequence, the output of a genetic model fitted on the transformed phenotypes cannot be related back to the original phenotype scale, which hinders phenotype prediction.

Here, we address both of shortcomings. First, we show how to assess alternative transformations in the light of the observed genotype and phenotype data. Building on this insight we propose the warped linear mixed model (WarpedLMM), an extension of the linear mixed model that adaptively learns a suitable transformation from a flexible class of permitted functions. In simulations we find that this approach is able to recover complex phenotype transformations solely from genotype and phenotype data, greatly reducing biases when estimating narrow-sense heritability. At the same time, the transformations recovered by

WarpedLMM can be non-ambiguously inverted, allowing to map genetic effects estimated on the transformed scale back to the original phenotype scale, thus enabling phenotype prediction.

In experiments on data from mouse, yeast and human, we find that warpedLMM is widely applicable to a wide range of genetic analyses, reducing bias in narrow-sense heritability estimation, improving out-of-sample prediction and increasing power in GWAS.

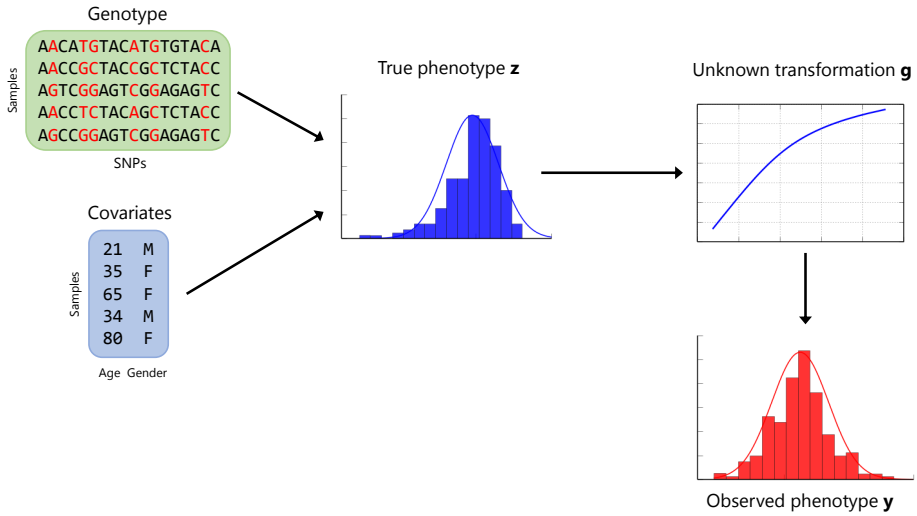


Figure 4.1: The genetic model of interest determines the latent phenotype profiles $\mathbf{z}$ (blue histogram), the measured phenotype data $\mathbf{y}$ (red histogram) are then derived from $\mathbf{z}$ via an unknown transformation $g(\mathbf{z})$. WarpedLMM is then able to reconstruct the original phenotype $\mathbf{z}$ by estimating the inverse transformation function $f(\mathbf{y}) = g^{-1}(\mathbf{y})$ from the observed phenotype, genetic markers and covariates.

4.2 Methods

WarpedLMM assumes that the genetics don't affect the observed phenotype $\mathbf{y}$, but rather have an effect on a Gaussian-distributed latent phenotype $\mathbf{z}$. Intuitively, the observed phenotype data $\mathbf{y}$ results from a non-linear distortion function g to the latent phenotype $\mathbf{z}$. Thus, in order to recover the true genetic model that gives rise to $\mathbf{z}$, an estimate of the inverse transformation g^{-1} is

needed. In some instances, expert knowledge may help to guide the choice of suitable functions to approximate the true inverse g^{-1}, which is the ideal phenotype transformation. However, such knowledge may be subjective and misleading, or may be missing entirely. As an alternative, we propose the Warped Linear Mixed Model (WarpedLMM), which generalizes a number of previous approaches (Section 4.2.1). This model extends the standard linear mixed model and allows for assessing the fit of alternative candidate transformations using the likelihood principle. The most probable transformation is then obtained by maximizing the sum of the log-likelihood and a regularization term that penalizes the complexity of the fitted monotonic function f. The fitted function can then be used to obtain latent phenotypes $\mathbf{z}$, which are then amenable to analysis using standard methods.

4.2.1 WarpedLMM

We model the observed non-normal distributed phenotype y_n of each individual indexed by n by an unobserved normal distributed phenotype z_n that results from transforming y_n by the monotonic function f, parameterized by $\boldsymbol{\psi}$.

$$z_n = f(y_n; \boldsymbol{\psi}). \tag{4.1}$$

On the normal distributed scale, the representation z_n of the phenotype is given by the following linear mixed model:

$$z_n = \mathbf{x}_n \boldsymbol{\beta} + \mathbf{g}_n^{\star} \boldsymbol{\alpha} + \epsilon_n, \tag{4.2}$$

where $\mathbf{x}_n$ holds the covariates for individual n, $\boldsymbol{\beta}$ are fixed effects, $\mathbf{g}_n^{\star}$ contains the genotype of the individual at $S^{\star}$ causal genetic loci, $\boldsymbol{\alpha}$ are normal distributed random genetic effects, and ϵ_n is independent normal distributed noise.

Given this linear mixed model, the likelihood for the N-by-1 vector $\mathbf{z} = f(\mathbf{y}; \boldsymbol{\psi})$ of transformed phenotypes for a sample of N individuals is

$$\mathbf{z} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}\ ,\ \sigma_{\mathrm{g}}^2 \mathbf{C}_N + \sigma_{\mathrm{e}}^2 \mathbf{I}), \tag{4.3}$$

where $\mathbf{C}_N$ is the relationship matrix at the causal loci, σ_g^2 is the total amount of genetic variance and σ_e^2 is the error noise variance.

In practice, we use a genomic relatedness matrix [Lynch and Ritland, 1999] computed from all S genotyped common SNPs, that are pre-processed to have zero-mean and unit-variance, stored in the N-by-S matrix $\mathbf{G}$.

$$\mathbf{C}_N = \frac{1}{S}\mathbf{G}\mathbf{G}^{\mathrm{T}}. \tag{4.4}$$

Choosing a monotonic warping function

Instead of specifying a fixed transformation, we find the optimal transformation $f_{\hat{\boldsymbol{\psi}}}$ for a given data set by maximizing the likelihood (4.3) of the transformed phenotype over a flexible class of monotonic functions parameterized by $\boldsymbol{\psi}$.

In the following, we consider a particular family of functions initially proposed by Snelson et al. [2004] in the context of Gaussian process regression. For the phenotype y_n of each sample, the transformation is chosen as

$$f(y_n; \boldsymbol{\psi}) = d \cdot y_n + \sum_{i=1}^{I} a_i \cdot \tanh\left(b_i \cdot (y_n + c_i)\right) \qquad a_i \geq 0, \quad b_i \geq 0 \quad d \geq 0, \quad \forall i \tag{4.5}$$

where $\boldsymbol{\psi} = (d, a_1, b_1, c_1, \ldots, a_I, b_I, c_I)$.

In this equation, f is a sum over I non-linear step functions, where for each step function with an index i, a_i controls the step size, b_i controls the steepness and c_i controls the location. Additionally, the parameter d is a coefficient for the linear part (in y_n) of the function.

The only parameter requiring manual setting is the number of step functions I. We followed the recommendation in Snelson et al. [2004] and used $I = 3$ step function in all of our experiments, yielding a good empirical performance.

In principle, any parametric monotonic function can be used in place of the function suggested above. For instance, a warping function based on the popular

Box-Cox [Box and Cox, 1964] transformation could be used as an alternative:

$$f_{Box-Cox}(y_n;\psi) = \begin{cases} \frac{y_n^{\psi}-1}{\psi} & \text{if } \psi \neq 0 \\ \ln(y_n) & \text{if } \psi = 0 \end{cases} \tag{4.6}$$

This classical warping function is controlled by a single parameter, and thus can be useful when the large number of parameters of the function proposed above is a concern.

Parameter estimation

The model parameters are estimated by maximizing a penalized form of the linear mixed model likelihood. By taking the logarithm of (4.3), the negative log likelihood $\mathcal{L}$ for the hidden normal distributed phenotype $\mathbf{z}$ is obtained as

$$\begin{aligned} \mathcal{L} &= -\log P(\mathbf{z} \,|\, \mathbf{X}, \mathbf{G}) = \\ &= \frac{1}{2}\log \det \mathbf{C}_N + \frac{1}{2}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{C}_N^{-1}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) + \frac{N}{2}\log 2\pi. \end{aligned} \tag{4.7}$$

Equation 4.7 is not accounting for the fact that $\mathbf{z}$ is really a transformation of the observed phenotype $\mathbf{y}$. This transformation can be taken into account with a change of variable, yielding the negative log likelihood for $\mathbf{y}$ as

$$\begin{aligned} \mathcal{L} = \frac{1}{2}\log \det \mathbf{C}_N + \frac{1}{2}(f(\mathbf{y};\boldsymbol{\psi}) - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{C}_N^{-1}(f(\mathbf{y};\boldsymbol{\psi}) - \mathbf{X}\boldsymbol{\beta}) \\ - \sum_{n=1}^{N} \log \frac{\partial f(\mathbf{y};\boldsymbol{\psi})}{\partial \mathbf{y}} + \frac{N}{2}\log 2\pi. \end{aligned} \tag{4.8}$$

It's then possible to fit the model by minimizing (4.8) with respect to the parameters of the model and the transformation.

Incorporating strong genetic effects

While the realized relationship matrix $\mathbf{K}$ can accurately capture the relatedness between individuals in the presence of many causal variants with small effect sizes, it doesn't necessarily do so when the genetic signal is mostly due to a small number of causal variants. For this reason, several approaches (Chapters 2 and

3, citepSegura2012Efficient) have been proposed to select strong genetic effects for inclusion in the model. Here we follow the approach presented in Chapters 2 and 3 and we perform a forward selection by iteratively adding a new variance component representing strongest effect to the random effects term.

At iteration t the model is defined as

$$\mathbf{z} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}\ ,\ \sigma_{\mathrm{g}}^2\mathbf{K} + \sum_{i=1}^{t} \sigma_i^2 \mathbf{G}_i\mathbf{G}_i^\top + \sigma_{\mathrm{e}}^2\mathbf{I}), \tag{4.9}$$

where the parameters Ψ, $\boldsymbol{\beta}$, σ_{g}^2, σ_i^2, σ_{e}^2 are re-estimated each iteration.

In each iteration t, the next genotype $\mathbf{G}_{t+1}$ with the strongest individual effect is determined by fixed effects testing [Lippert et al., 2011] of all genetic markers against current transformed phenotype $\mathbf{z}_t$ using the current set of variance components $\sigma_{\mathrm{g}}^2\mathbf{K} + \sum_{i=1}^{t} \sigma_i^2 \mathbf{G}_i\mathbf{G}_i^\top + \sigma_{\mathrm{e}}^2\mathbf{I}$ as the relatedness matrix. A marker is selected if its q-value [Storey, 2003; Storey and Tibshirani, 2003] is $\leq \alpha_{\mathrm{FDR}}$. The algorithm converges when no marker achieves genome-wide significance at the FDR specified. We used $\alpha_{\mathrm{FDR}} = 0.05$ for all our experiments.

The genetic effects incorporated in the model at the end of this procedure can in general be beneficial in certain tasks such as phenotype prediction. Here, we use them only to better reconstruct the transformation function f and we don't take them into account while doing prediction or heritability estimation. Finally, it's important to notice that alternatives to the forward-selection technique can be used to perform feature selection.

Phenotype prediction

Under this model we can predict the unobserved phenotype of a new individual indexed by $\star$ given its genotype alone. Given a fully observed sample of N individuals, we can use the parameter estimates under model (4.3) to compute the best linear unbiased predictor (BLUP) $\hat{z}_\star$ of the new individual's phenotype on the normal distributed scale.

$$\hat{z}_\star = \mathbf{x}_\star\boldsymbol{\beta} + \hat{\sigma}_{\mathrm{g}}^2\mathbf{k}_\star \left(\hat{\sigma}_{\mathrm{g}}^2\mathbf{K} + \hat{\sigma}_{\mathrm{e}}^2\mathbf{I}\right)^{-1} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}), \tag{4.10}$$

where $\mathbf{x}_\star$ is the vector of covariates for the new individual, the 1-by-N vector $\mathbf{k}_\star$ is the genomic relatedness between the new individual and all individuals in the original sample. To get an estimate $\hat{y}_\star$ of the phenotype on the original scale, we then apply the reverse transformation f^{-1} on the BLUP.

$$\hat{y}_\star = f^{-1}(\hat{z}_\star \,;\, \hat{\boldsymbol{\psi}}). \tag{4.11}$$

The reverse transformation f^{-1} is obtained by numerically inverting f by applying Newton-Raphson.

Estimating heritability

We obtain an estimate of the narrow-sense heritability h^2 on the normal distributed scale by computing a chip heritability $\hat{h}^2$ from common genotyped markers in the linear mixed model (4.3).

$$\hat{h}^2 = \frac{\hat{\sigma}_\mathrm{g}^2}{\hat{\sigma}_\mathrm{e}^2 + \hat{\sigma}_\mathrm{g}^2}, \tag{4.12}$$

where $\hat{\sigma}_\mathrm{g}^2$ and $\hat{\sigma}_\mathrm{e}^2$ are restricted maximum likelihood estimates of σ_e^2 and σ_g^2.

4.3 Results

In this section, we investigate the practical relevance of phenotype transformations in the context of key applications of LMMs in genetics. In particular, we consider both extensive simulation studies, as well as real data from human, mouse and yeast, comparing WarpedLMM to established preprocessing approaches for phenotypes, such as Box-Cox transformations or rank transformations, in combination with a standard LMM, demonstrating that WarpedLMM more accurately recovers the true underlying warping functions.

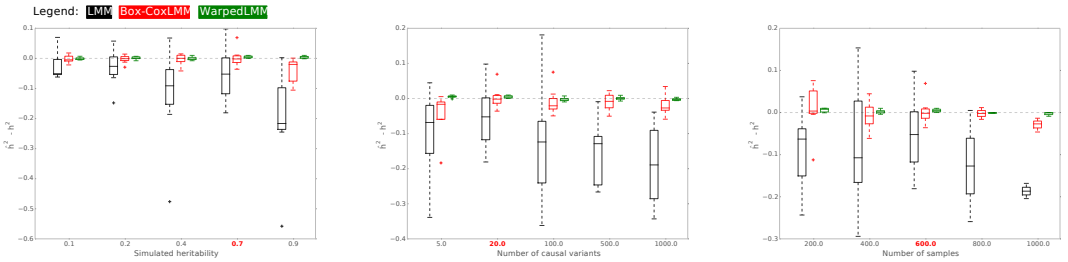
4.3.1 Narrow-sense heritability estimation and out-of-sample phenotype prediction

4.3.1.1 Simulations

First, to verify the the accuracy of the transformations recovered by WarpedLMM, we considered simulated data. In an effort to consider representative settings, we used genotype data from the HapMap project [Gibbs et al., 2003] and simulated phenotypes from a broad range of alternative genetic models. We considered variable proportions of variance explained by genotype, altered the number of causal variants, the observed sample size and the type and magnitude of the phenotype transformation (interpolating between a linear function and an exponential transformation); see also Figure 4.1 for data from a typical simulation experiment. In each simulation, we sample an h^2 from $\{0.1, 0.20, 0.40, 0.70, 0.9\}$ (as done in [Zaitlen and Kraft, 2012]), we sample the number of causal variants from $\{5, 20, 100, 500, 1000\}$, the number of samples from $\{200, 400, 600, 800, 1000\}$, the variance explained by covariates from $\{0.0, 0.25, 0.5, 0.70, 0.9\}$ (we can then recover the noise level conditioned on h2, and the covariates variance). Finally we pick a transformation $f(y)$ from the set of transformations used in Valdar et al. [2006]. We then transform the phenotype as $z = t \cdot y + (1-t)f(y)$. Where t is a parameter that determines the intensity of the transformation and is sampled from $\{0.0, 0.25, 0.50, 0.75, 1.0\}$. We repeated this simulation procedure $50,000$ times in order to have a sufficiently large sample size to investigate all the regimes described above.

WarpedLMM was used to recover the initial untransformed phenotype data, followed by a standard mixed model to estimate narrow-sense heritability. For comparison, we also considered heritability estimates obtained by applying a linear mixed model to the untransformed data (LMM) [Yang et al., 2011; Zaitlen and Kraft, 2012], or to phenotype data that have been preprocessed using the popular Box-Cox transformation (Box-CoxLMM) [Ahn et al., 2010; Chiu et al., 2005; Huang et al., 2007; McCauley et al., 2005; Tian et al., 2011].

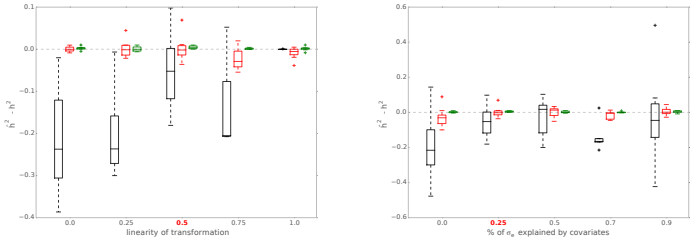
When comparing the estimated heritability to the simulated truth, most methods tended to underestimate heritability in difficult regimes, which is in line with previous findings [Speed et al., 2012]. Poor performance is found for strongly



(a) Varying the simulated heritability

(b) Varying the number of causal variants

(c) Varying the number of samples



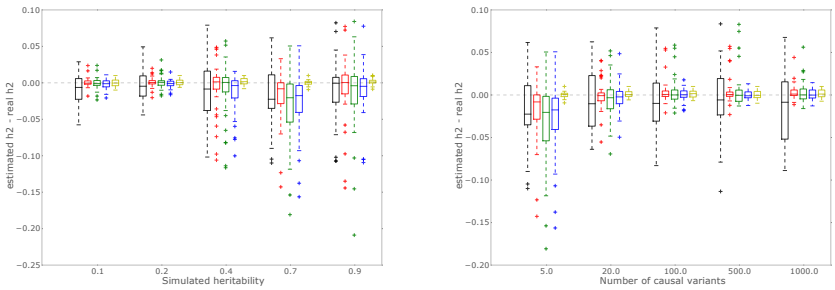
(d) Varying the intensity of the transformation

(e) Varying $\sigma_{covariates}$

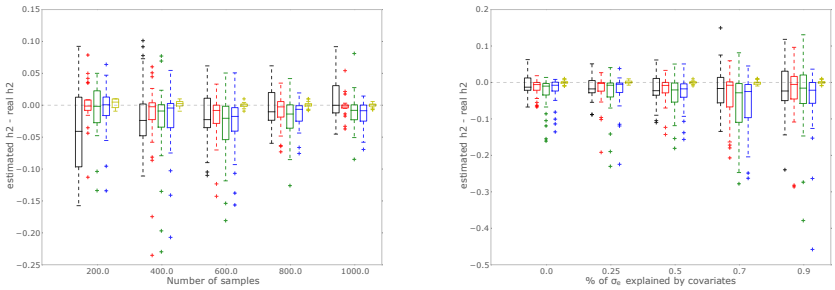
Figure 4.2: Comparison of alternative linear mixed-model approaches for estimating the genetic proportion of phenotype variability (narrow-sense heritability, h^2). Shown is the difference between the estimated and the true genetic proportion of variance for 50,000 simulated experiments, stratified by different simulation settings: **(a)**, variable simulated heritability, **(b)**, considering alternative numbers of causal variants, **(c)**, for variable numbers of samples and **(d)**, different extents of the non-linearity of the true simulated transformation. For each parameter, the remaining simulation settings remained constant with the default parameters being highlighted in red bold face font. Heritability estimates were obtained either using WarpedLMM fitting, Box-Cox preprocessed LMM and a standard linear mixed model.

heritable traits (Figure 4.2a), when the numbers of causal SNPs was small [Ryoo and Lee, 2013] (Figure 4.2b), the dataset had low sample size (Figure 4.2c) or when true phenotype transformations was strongly non-linear (Figure 4.2d).

Across these regimes, Box-Cox as preprocessing approach improved the accuracy of heritability estimates compared to a standard linear mixed model. A further improvement, however, was achieved by WarpedLMM which reduced the variance of heritability estimates compared to the true values. We confirmed the



(a) Varying the simulated heritability (b) Varying the number of causal variants



(c) Varying the number of samples (d) Varying the intensity of the transformation

Figure 4.3: Comparison of alternative linear mixed-model approaches for estimating the genetic contribution to phenotype variability (narrow sense heritability, h^2). In this particular experiment we considered a different transformation ($\mathbf{z} = \sqrt{\mathbf{y^2}}$) and included comparisons to a rank-based transformation and a simpler version of the WarpedLMM model which incorporates genetic information with a full rank kernel only (realized relationship matrix). Legend: LMM, Box-Cox, WarpedLMM, WarpedLMM with full RRM only, Rank transformation

performance of WarpedLMM for a different non-linear phenotype transformation and further compared it to additional alternative methods, including rank-based transformations as preprocessing [Baranzini et al., 2009; Himes et al., 2009; Kathiresan et al., 2007; Wallace et al., 2008] (Figure 4.3).

4.3.1.2 Analysis of data from yeast

Next, we considered a study on a F2 yeast cross [Bloom et al., 2013], to understand the implication of phenotype transformation in a well-powered study with

highly heritable traits. Figure 4.4a shows narrow-sense heritability estimates using a standard linear mixed model versus heritability estimates using transformations fitted by WarpedLMM. These methods results in significantly deviating heritability estimates (paired t-test, $\alpha = 0.05$) for 17 phenotypes (38%), for most of which WarpedLMM estimated a larger fraction of genetic variance than the standard approach (11 of 17, 65%). This suggest that even phenotypes obtained in a controlled lab environment tend to be subject to transformations, leading to both overestimation and underestimation of the narrow-sense heritability. To validate the genetic models derived using WarpedLMM, we performed out-of-sample phenotype prediction using both a WarpedLMM and a standard LMM (Figure 4.4c). Reassuringly, the WarpedLMM model consistently yielded superior prediction accuracy, irrespective of whether the estimated heritability was larger or smaller than those obtained using a standard LMM (Figure 4.6a).

4.3.1.3 Analysis of data from mouse

Next, we revisited an association study in a structured mouse population [Valdar et al., 2006]. In the original analysis, the authors manually defined inverse phenotype transformation for each of the 47 phenotypes considered. While this process was guided by an initial Box-Cox fit, the authors performed further manual tuning of the resulting function for each phenotype independently. Here, we compared a linear mixed model on untransformed phenotypes (LMM) to estimates derived using WarpedLMM. Covariates such as age, gender, body weight, litter number and cage density were included as fixed effects in both models.

As shown in Figure4.4b, the two models considered yielded significantly (t-test, $pv \leq 0.05$) different heritability estimates for 18 of the phenotypes (0.38%), supporting the results we obtained in the simulations and the yeast dataset. For 17 out of 18 phenotypes (0.94%), WarpedLMM found higher heritability than a standard LMM. We further validated these findings by comparing the LMM and the WarpedLMM in an out-of-sample prediction task. Again, WarpedLMM consistently improved out-of-sample prediction accuracy over a standard LMM (Figure 4.4d), even when the estimated heritability was lower (Figure 4.6a), suggesting that appropriate phenotype transformations are needed to void overfitting

in applications of mixed models. Finally, we compared the fitted transformations of WarpedLMM model to those manually derived in Valdar et al. [2006]. The transformations recovered by WarpedLMM were consistently in the same functional class as those reported by Valdar et al. [2006] (linear, logarithmic, etc.), however with slight differences in parametrization (Figure 4.5).

In summary, the results in yeast and mouse studies provide confidence that WarpedLMM model yields a better fit to phenotype data in a broad range of settings, resulting in more reliable parameter estimates and improved prediction.

4.3.2 Phenotype preprocessing for genome-wide association studies

Analogously to narrow-sense heritability estimation and prediction, WarpedLMM can be used to define quantitative traits for analysis in GWAS. To investigate this, we revisited genotype and phenotype data from the Northern Finnish birth cohort [Sabatti et al., 2009]. We considered four related metabolic traits [Ridker et al., 2005] (high density lipoprotein, low density lipoprotein, triglycerides and C-reactive protein) that have previously been considered for pairwise genetic analysis [Korte et al., 2012] and joint analysis of all four traits [Zhou et al., 2013].

Previous investigations using the same data have considered a range of alternative normalization procedure to estimate the hidden phenotype variables. In the initial publication, Sabatti et al. [2009] the authors considered a log transformation of some phenotypes (triglycerides, CRP) while leaving the remaining phenotypes (HDL, LDL) on the original scale. To avoid the need to decide upon an explicit transformation, authors of follow-up studies have considered semi-parametric transformation approaches Zhou and Stephens [2013], employing a three-step procedure which consisted of rank transforming the phenotype, regressing out the covariates and rank transforming the residuals again. This approach assumes that the genotype explains only a small portion of the variance and hence "Gaussianizing" phenotype data on the null model is valid. In the following, we assessed whether this assumption is valid for this particular dataset by comparing the transformations recovered by WarpedLMM and by the semi-parametric approach just described.

Phenotype	WarpedLMM	LMM
High density lipoprotein	0.06 ± 0.02	0.035 ± 0.01
Low density lipoprotein	0.05 ± 0.02	0.04 ± 0.02
Triglycerides	0.14 ± 0.04	0.13 ± 0.03
C-reactive protein	0.08 ± 0.03	0.02 ± 0.02

Table 4.1: Out of sample r^2 computed over 10 random train/test splits on the human dataset. Shown are the average and the standard error computed over different test sets.

Indeed, we observed striking correlations between the p-values when applying a standard mixed model to phenotype data preprocessed using WarpedLMM and the non-parametric approach ($\rho = 0.99 \pm 0.01$, Figure 4.9). In contrast, using non-normalized phenotypes resulted in a substantial loss of power (Figure 4.9). For instance, for LDL the LMM on untransformed phenotypes yielded 7 associations at genome-wide significance level 5×10^{-8}, whereas WarpedLMM preprocessing identified 9 associations. With the exception of the CRP phenotypes (3 associations irrespective of the processing approach), the same trend was observed for the remaining phenotypes (triglycerides: 8 vs 10 associations, HDL: 3 versus 10 associations).

Furthermore, separate application of WarpedLMM to each of the 4 phenotypes increased pairwise correlations structure between phenotypes, which is key for multivariate linear mixed models Korte et al. [2012]; Zhou et al. [2013] (4.8).

Finally, we validated the full genetic model implied by WarpedLMM using out-of-sample phenotype prediction. Importantly, the transformations functions fit by WarpedLMM can be inverted, thus permitting to assess prediction accuracy on the natural scale, which is not possible for rank-based methods. In comparison with a naive mixed model ignoring phenotype transformation, we observed a consistent improvement in out-of-sample prediction when employing WarpedLMM, suggesting that it accurately models the phenotype data (Table 4.1). Overall, these experiments support that WarpedLMM can be used as robust preprocessing approach for GWAS.

4.4 Discussion

Although preprocessing methods are widely used in practice to invert an unknown phenotype transformation [Ahn et al., 2010; Baranzini et al., 2009; Chiu et al., 2005; Goh and Yap, 2009; Himes et al., 2009; Huang et al., 2007; Kathiresan et al., 2007; McCauley et al., 2005; Servin and Stephens, 2007; Stephens, 2013; Tian et al., 2011; Wallace et al., 2008; Zhou and Stephens, 2013], so far there has been no principled approach to assess alternative transformations.

Here, we have shown how the classical linear mixed model can be extended to learn phenotype transformations from the observed data itself. In experiments, we found that the resulting warped linear mixed model significantly improves the accuracy and robustness in several different types of genetic analyses. Although an important application of WarpedLMM is the generation of transformed phenotypes for downstream analysis, we emphasize that the model is much more than just another normalization procedure. The objective function of the model can be derived from first principles, resulting in an extension of the mixed model to balance the data likelihood and the complexity of the fitted transformation (Section 4.2.1). As a result, our approach is ideal for use in combination with major applications of the linear mixed model, including GWAS, heritability estimation and phenotype prediction.

When applied to studies in yeast and mouse, we found that WarpedLMM results in an overall increase of the proportion of variance that could be attributed to genetic factors. Although in a minority of traits the heritability estimates decreased, we note that the model yielded consistent improvements for out-of-sample prediction. This shows that inappropriate phenotype transformations can lead to overoptimistic heritability estimates and overfitting, a fact that has previously been noted by others [Ryoo and Lee, 2013]. Remarkably, although the WarpedLMM model has a larger number of parameters, the model did not overfit even when applied to datasets with smaller sample sizes (Figure 2a).

Although we have focused on a the most established tasks in genetic analysis, WarpedLMM can easily be used in more specialized tasks. For example, the model can be combined with multi locus mixed models [Rakitsch et al., 2013; Segura et al., 2012], mixed models that jointly consider multiple phenotypes [Korte

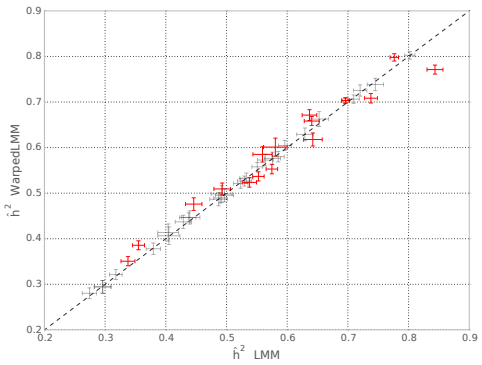
et al., 2012; Zhou and Stephens, 2013] or expression quantitative trait loci studies (Chapters 2 and 3, Kang et al. [2008a]; Listgarten et al. [2010]).

WarpedLMM finds the transformation function while jointly taking into account all the available covariates and the genotype data. This joint approach helps to ensure that the model residuals are Gaussian distributed, rather than the phenotype itself. This has been recognized in previous work by Zhou and Stephens [2013], where the authors employ an ad-hoc but accurate three-step procedure, comprising of rank transforming the phenotype, regressing out the covariates and rank transforming the residuals again. While this approach is similar in spirit to WarpedLMM, it assumes that the genotype explains only a small portion of the variance and hence "Gaussianizing" phenotype data on the null model is valid. While this is reasonable for some analyses, deviations from this assumption remain a concern, as discussed in Stephens [2013]. Our approach is able to overcome these limitations by a principled jointly modeling approach, taking the effect of covariates and genotype data into account.

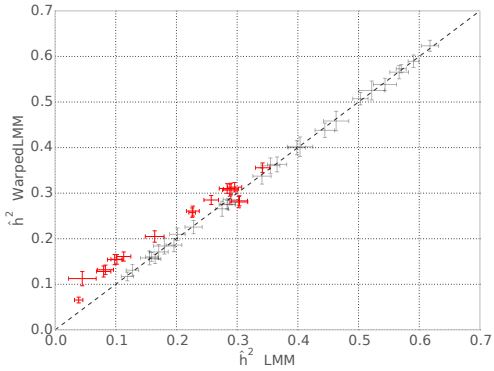
Finally, we note that there may be scenarios where also WarpedLMM does not achieve optimal results. Similar to other existing methods, the model learns a transformation but assumes that that the noise level in the transformed phenotype space is constant. This assumption may be violated for special data types such as count data or binary phenotypes. In such instances, it will remain appropriate to use generalized linear mixed models with non-Gaussian likelihood models that incorporate stronger assumptions about the nature of the data. However, the breadth of phenotype data being generated is increasing at a rapid rate and the majority is quantitative with an unknown underlying scale. In these instances there are clear advantages of the WarpedLMM model: the model allows for the robust and failure safe analysis of a broad spectrum of phenotypes without the need to develop specialized methods or revert to manual processing steps. This will open new opportunities to analyze data from high-throughput phenotyping platforms.

4.5 Conclusions

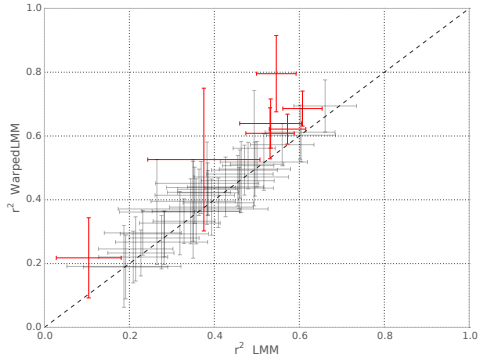
In this chapter, we have presented an extension of the linear mixed model framework that estimates an optimal transformation from the observed data. In extensive simulations and applications to real data from human, mouse and yeast we have shown that using transformations inferred by our model leads to increased power in genome-wide association studies and higher accuracy in heritability estimates and phenotype predictions.



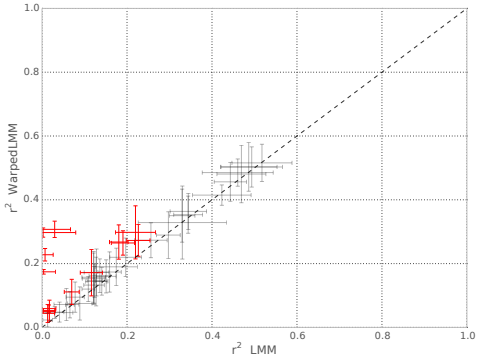
(a) LMM $\hat{h}^2$ vs WarpedLMM $\hat{h}^2$ in the yeast dataset



(b) LMM $\hat{h}^2$ vs WarpedLMM $\hat{h}^2$ in the mouse dataset



(c) LMM out-of-sample r^2 vs WarpedLMM out-of-sample r^2



(d) LMM out-of-sample r^2 vs WarpedLMM out-of-sample r^2

Figure 4.4: Comparative analysis of WarpedLMM and a standard LMM on the yeast and mouse datasets. Panels **(a)** and **(b)** show comparative estimates of the heritability using a linear mixed model on the untransformed phenotype versus the heritability estimates obtained by WarpedLMM. Empirical error bars were obtained from 10 bootstrap replicates, using 90 % of the data in each replicate. Significant differences are colored in red (paired t-test, $\alpha = 0.05$). **(a)** $\hat{h}^2$ estimated by a LMM on the untransformed data and by WarpedLMM for the yeast dataset. **(b)** $\hat{h}^2$ estimated by a LMM on the untransformed data and by WarpedLMM for the mouse dataset. Panels **(c)** and **(d)** show out-of-sample prediction accuracy assessed by the squared correlation coefficient r^2, considering either a linear mixed model on the untransformed data (LMM) and a warped linear mixed model (WarpedLMM), for **(c)** yeast and **(d)** mouse. Prediction accuracies were assessed from 10 random train-test splits. Phenotypes with significant deviations in prediction accuracy of the LMM and the WarpedLMM are highlighted in red (paired t-test, $pv \leq 0.05$).

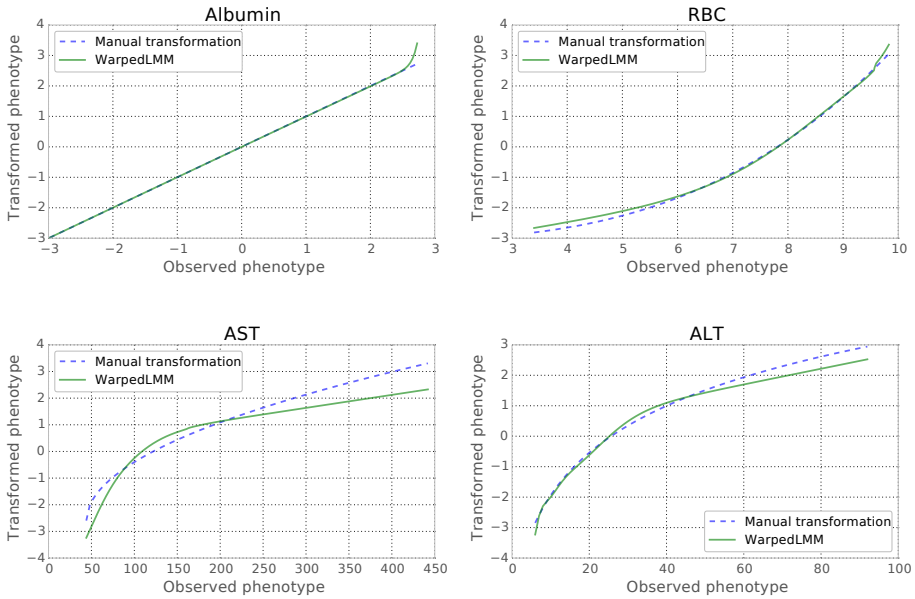
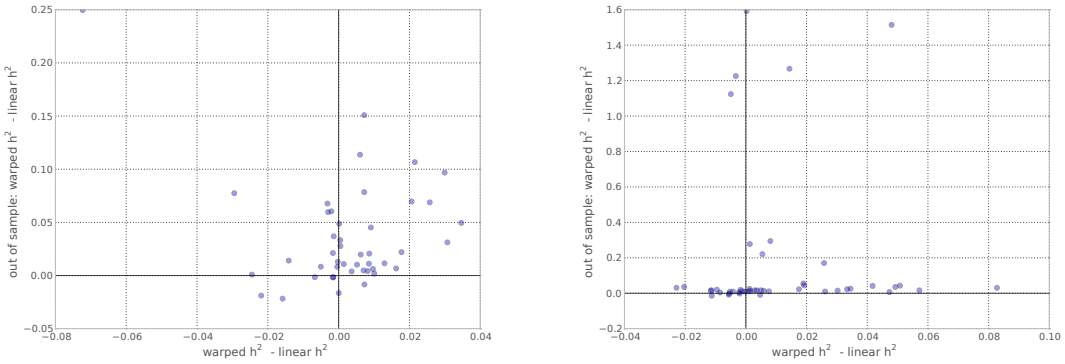


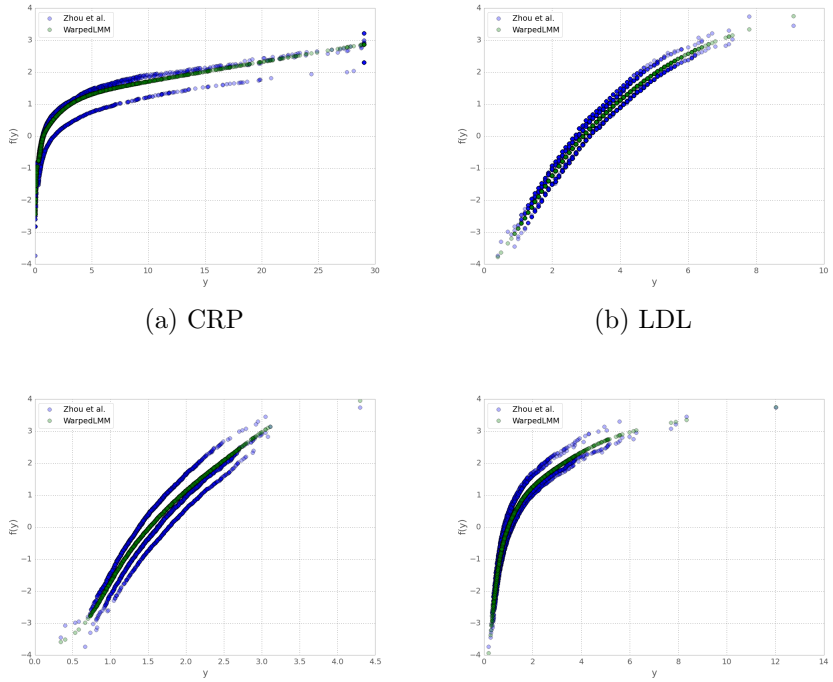
Figure 4.5: Comparison of the transformation recovered by WarpedLMM and the transformation found manually in Valdar et al. [2006]. In the original study on mouse, the authors first applied a Box-Cox transformation then manually tuned the resulting function. In all 4 phenotypes shown here, WarpedLMM and the manual transformations appear to belong to the same class (log, exp, etc.) of functions, with some minor differences in parametrizations and complexity.



(a) Difference in $\hat{h}^2$ vs difference in r^2 in the yeast dataset

(b) Difference in $\hat{h}^2$ vs difference in r^2 in the mouse dataset

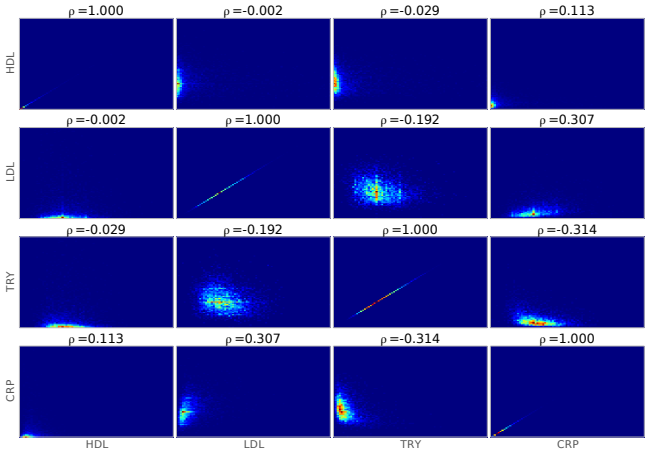
Figure 4.6: Comparison of narrow-sense heritability estimates and out-of-sample r^2 in the yeast and mouse datasets. The x-axis represents the difference in estimated heritability between the WarpedLMM and a LMM. The y-axis represents the difference in out of sample r^2. This means that for every point on the right of the vertical line, the WarpedLMM found more heritability than the LMM. Similarly, for every point above the horizontal line, the WarpedLMM had a better out-of-sample prediction performance than a LMM. Both these plots show that even in cases where the estimated heritability is lower, the out-of-sample prediction performance of the WarpedLMM is better than the LMM's.



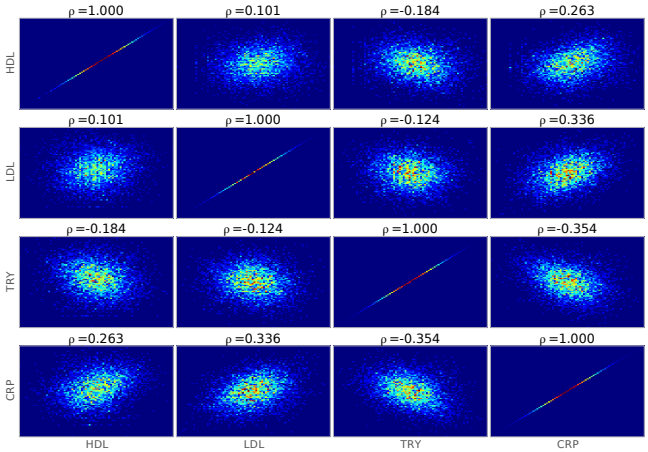
(a) CRP (b) LDL

(c) HDL (d) TRI

Figure 4.7: Comparison of the transformation described in Zhou and Stephens [2013] and the transformation obtained by WarpedLMM. For all the 4 phenotypes considered, the two methods find qualitatively very similar transformations. The main difference between the two functions is that the rank transformation seems to produce multiple functions (multiple blue lines). This is a consequence of the two-step procedure used (rank transform the phenotype, subtract off covariates, rank transform again).

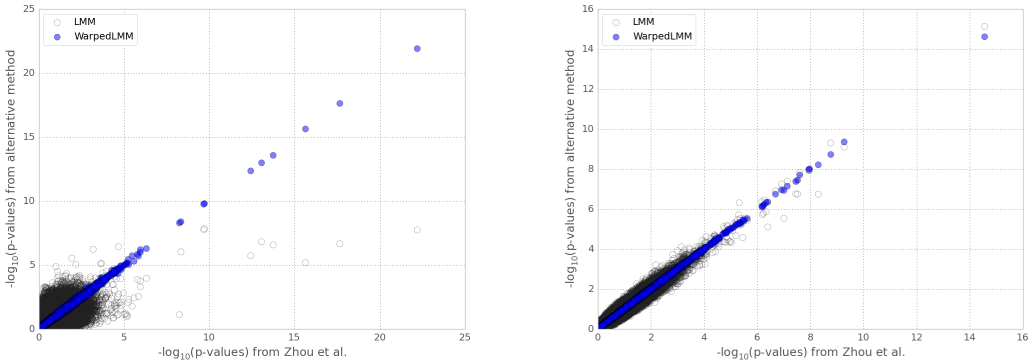


(a) Without transforming the phenotypes



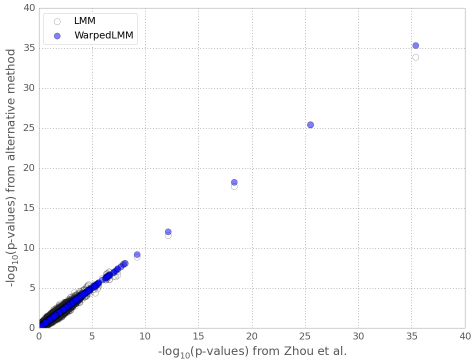
(b) Applying the transformation found by WarpedLMM

Figure 4.8: Correlations between phenotypes in the human dataset. The 4 different phenotypes (High density lipoprotein, low-density lipoprotein, tryglycerides and C-reactive protein) are all biomarkers for cardiovascular diseases and are all known to have some degree of correlation between them [Arena et al., 2006]. While performing our analyses, we noticed that independently transforming the phenotypes with WarpedLMM resulted in a general increase in the inter-phenotype correlations. This is not only more aligned to our prior beliefs, but it also has the potential to uncover new interesting biological findings. For instance, performing a univariate GWAS on the HDL phenotype with WarpedLMM resulted in significant ($pv \leq 5 \times 10^{-8}$) associations (rs1811472 on chr1) found in the CRP cis region . Interestingly, not only these associations were not significant in an analysis with a LMM, but additionally they were not significantly associated to the CRP phenotype itself.

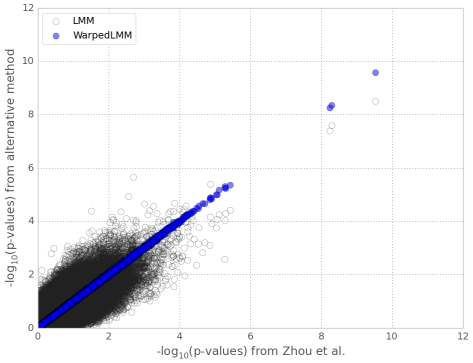


(a) High density lipoprotein

(b) Low density lipoprotein



(c) Triglycerides



(d) C-reactive protein

Figure 4.9: Comparison of p-values obtained from a parametric rank transformation regressing out covariates (Zhou and Stephens [2013]) and WarpedLMM. The plots show the $-\log_{10}$(p-values) of the method described in Zhou et al. [2013] on the x-axis versus the $-\log_{10}$(p-values) obtained when using WarpedLMM (solid blue circles) and a LMM (empty black circles). All the methods considered gave well-calibrated p-values with genomic controls of 1.00 ± 0.01

Chapter 5

Conclusions and future work

5.1 Conclusions

One of the key contributions of this thesis has been the development of probabilistic latent variable models an their application to eQTL studies. In particular, in Chapter 2 we have presented a model that takes into account prominent genetic regulators while determining hidden factors acting on the gene expression profiles. This development is particularly important if considered in the context of previous approaches to correct for confounding influences in eQTL studies. In these methods, the effect of the genotype was largely ignored while estimating the confounders. This led to an improvement in the detection of *cis* associations, but a significant loss of power when detecting *trans* associations. In Chapter 2 and in the related paper, we have identified this problem on a multitude of datasets and proposed a joint probabilistic model to correct for it. While the approach we developed was extremely expressive, we had to face a significant number of modelling challenges. In particular we had to control model complexity and account for the effect of sparse genetic regulators while mantaining computational tractability even for large datasets.

In Chapter 3 we have extended the model of Chapter 2 to identify genotype-by-enviroment interactions. While this is common practice in cases in which the environmental condition is carefully measured and controlled, we have shown that our approach can correctly identify unmeasured enviromental factors from the gene expression data alone. Once learnt, these variables can then be used

in genetic analyses to investigate interactions between environmental factors and genotype with a regulatory effect on gene expression traits. Chapter 3 extends the the work we have presented in Chapter 2 in two different ways. First, we have introduced a systematic approach to test for GxE interactions while accounting for the effect of unknown environments. Second, we have shown how accounting for significant GxE interactions can further refine the statistical testing of eQTLs.

In Chapter 4 we tackled the problem of noise model misspecification in genome-wide association studies. A limiting assumption of linear mixed models is that the model residuals are Gaussian distributed, a requirement that rarely holds in practice. We have shown that violations of this assumption lead to false conclusions and losses in power. To mitigate this problem, it is common practice to pre-process the phenotypic values to make them as Gaussian as possible, for instance by applying logarithmic or other non-linear transformations. Unfortunately, different phenotypes require different transformations, and choosing a "good" transformation is in general challenging and subjective. For this reason we have derived and presented a latent variable model that learns the transformation from the observed data itself. In extensive simulations and applications to real data from human, mouse and yeast we showed that using transformations inferred by our model leads to increased power in genome-wide association studies and higher accuracy in heritability estimates and phenotype predictions.

5.2 Future work

The work presented in this thesis could benefit from a series of extensions.

In Chapter 2, we assumed the mapping function between the latent space and the observed space (the gene expression levels) to be linear. Ideally, one should allow this function to be non-linear, leading to more expressive models. Given that the model presented in Chapter 2 is a Gaussian process latent variable model [Lawrence, 2005] and that this family of models allows to non-linearize the mappings using Gaussian processes, this extension can be very easily obtained. Additionally, in section 2.2 we derived the marginal likelihood by starting from the generative model, placing a spherical Gaussian prior on the weights and integrating them out. Instead of seeking maximum a posteriori solutions for the

latent variables, it would be desirable to also marginalise the latent variables and to optimize with respect to the hyperparameters introduced. Unfortunately this leads to an intractable marginal likelihood. Titsias and Lawrence [2010] (see also [Hensman et al., 2013] for a different derivation) have proposed a variational approach in which the likelihood has the form of a reduced rank Gaussian process. This approximation is derived by introducing M additional input/output pairs to the Gaussian process function. These so called *inducing inputs* are then optimized alongside the other model parameters. This approximate Bayesian approach has been shown [Titsias and Lawrence, 2010] to outperform the MAP derivation presented in [Lawrence, 2005], so it would be of great interest to extend the PANAMA model in this direction. For instance, the approach described here allows a much more robust method for automatic relevance determination than the one described in section 2.2.5, because of the Bayesian treatment of $\mathbf{X}$.

In Chapter 3 we mainly focussed on the detection of GxE interactions in the context of genome-wide association studies. Recently, there has been a lot of interest in models that can estimate the proportion of phenotypic variance explained by the genetics [Speed et al., 2012; Yang et al., 2011] and by environmental factors [Valdar et al., 2006]. GxE interactions have been investigated in this context before [Valdar et al., 2006], but only when the environmental condition was explicitly measured and carefully controlled. Our approach allows to perform this type of study using potentially any type of high-dimensional molecular measurement (for example gene expression levels) to infer the environmental factors even when they are unknown *a priori*.

Finally, in Chapter 4 we presented a warped linear mixed model targeted to the analysis of univariate traits. Given the recent interest in the joint analysis of multiple traits, it would be desirable to extend WarpedLMM to include one warping function for each phenotype. Alternatively, this type of warped model could be investigated in conjuction with linear GP-LVMs [Lawrence, 2005] to perform linear dimensionality reduction with non-Gaussian distributed residuals.

More in general, probabilistic models such as the ones presented in this thesis can be easily extended and augmented to incorporate additional information. Similarly to what we have done in Chapter 3, where we have extended the model presented in Chapter 2 to additionally capture interactions between the genotype

and the environment, it's possible to easily build joint models that account for heterogeneous sources of information. In this thesis, we have considered relatively simple probabilistic latent variable models, consisting of only one set of latent variables and thus having a "shallow" architecture. Thanks to the increase in computational capabilities and to advances in approximate inference techniques [Hensman et al., 2013], it's now possible go beyond these shallow architectures and build hierarchical latent variable models, consisting of many layers of latent variables connected to each other [Damianou and Lawrence, 2013]. These "deep" architectures have proven to be extremely effective in automatically extracting features that explain the structure of the observed data. In this thesis, we have analyzed genotypes, gene expression and, indirectly, environmental factors. In principle, many other types of data, such as epigenomes or even images and clinical charts can be included at different levels of a deep architecture. Similarly to what we have observed in Chapter 3, where the inclusion of genotype-environment interactions in the model resulted in an improved reconstruction of the latent environmental factors, these joint models that incorporate multiple heterogeneous sources of data will allow to develop a much deeper understanding of the factors influencing disease risk. Of course, these expressive models also come with a unique set of challenges. First, they are extremely computationally expensive, and thus require the use of approximate inference techniques [Hensman et al., 2013] that often need to be finetuned for each task or dataset analyzed. Second, the inclusion of several different types of data, each with a potentially different likelihood, can make it difficult to identify a good loss function to minimize during learning. To give an example, if the goal is to find a link between a clinical image (PET scan, X-ray, etc.) and the genotype of an individual, the model would have strike a balance between correctly reconstructing the image (a *generative* task) and correctly classifying which genetic variant is responsible for the presence of a feature in the image (a *discriminative* task). Despite these challenges, the move towards deeper architectures is likely to result in richer, more expressive models that better capture the complex mechanisms underlying diseases.

Appendix A

Datasets

A.1 Yeast datasets.

A.1.1 eQTL studies

We used the yeast expression dataset from Smith et al. Smith and Kruglyak [2008] (GEO accession number GSE9376), which consists of 5,493 probes measured in 109 segregants derived from a cross between BY and RM. The authors provided the genotypes, which consisted of 2,956 genotyped loci. The dataset does not contain any covariates.

An association was defined as *cis* if the location of the SNP and the location of the opening reading frame (ORF) of the gene were within 10kb, and *trans* otherwise. In order to validate the associations found, we also used data from Brem et al. Brem et al. [2002] (GEO accession number GSE1990), which consisted of 7,084 probes and 2,956 genotyped loci in 112 segregants. For the purpose of comparison, we defined *cis* associations in the same way as we did for the previous dataset.

Preprocessing The binary genetic markers have not been preprocessed. Log expression levels of all 5,493 probes were used without any reduction but shifted to zero mean.

A.1.2 Heritability estimation

We used the yeast genotype and phenotype data from Bloom et al. [2013]. This dataset contains genetic information for 1,008 yeast segregants from a BY/RM cross, with a total of 11,623 markers. The phenotypes are fitness traits profiled in 46 different environment conditions.

A.1.3 Yeastract.

We used data from Yeastract [Teixeira et al., 2006], which contains information about the regulatory network between 185 transcription factors and 6,298 genes. Out of these 189 transcription factors, we selected the 129 TFs that had a polymorphism in the vicinity (10kb) of the coding region.

A.2 Mouse datasets

A.2.1 eQTL studies

We used the data described in Schadt Schadt et al. [2005], consisting of 23,698 expression measurements and 137 genotyped loci for 111 F_2 mouse lines.

A.2.2 GWAS and heritability estimation

We used mouse data from Valdar et al. [2006]. This dataset contains between 1700 and 1940 samples (depending on phenotype missingness), 10,132 markers and 47 phenotypes.

A.3 Human datasets

A.3.1 eQTL studies

We used the dataset from Myers et al. [2007] (GEO accession number GSE8919), which consists of 14,078 transcripts and 366,140 SNPs genotyped on 193 human samples.

A.3.2 GWAS and heritability estimation

We used the data from Sabatti et al. [2009] and applied the same filtering criteria described in Zhou et al. [2013]. This resulted in 5,255 individuals and 328,517 SNPs.

References

Jiyoung Ahn, Kai Yu, Rachael Stolzenberg-Solomon, K Claire Simon, Marjorie L McCullough, Lisa Gallicchio, Eric J Jacobs, Alberto Ascherio, Kathy Helzlsouer, Kevin B Jacobs, Qizhai Li, Stephanie J Weinstein, Mark Purdue, Jarmo Virtamo, Ronald Horst, William Wheeler, Stephen Chanock, David J Hunter, Richard B Hayes, Peter Kraft, and Demetrius Albanes. Genome-wide association study of circulating vitamin d levels. *Human molecular genetics*, 19(13): 2739–45, 7 2010. ISSN 1460-2083. 73, 80, 86

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010. 45

Ross Arena, James A Arrowood, Ding-Yu Fei, Shirley Helm, and Kenneth A Kraft. The relationship between c-reactive protein and other cardiovascular risk factors in men and women. *Journal of Cardiopulmonary Rehabilitation and Prevention*, 26(5):323–327, 2006. xvii, 93

C. Auesukaree, A. Damnernsawad, M. Kruatrachue, P. Pokethitiyook, C. Boonchird, Y. Kaneko, and S. Harashima. Genome-wide identification of genes involved in tolerance to various environmental stresses insaccharomyces cerevisiae. *Journal of applied genetics*, 50(3):301–310, 2009. 65

David J Balding, Martin Bishop, and Chris Cannings. *Handbook of statistical genetics*, volume 1. John Wiley & Sons, 2008. 6, 12

Sergio E Baranzini, Joanne Wang, Rachel A Gibson, Nicholas Galwey, Yvonne Naegelin, Frederik Barkhof, Ernst-Wilhelm Radue, Raija LP Lindberg, Bernard MG Uitdehaag, Michael R Johnson, et al. Genome-wide association

analysis of susceptibility and clinical phenotype in multiple sclerosis. *Human molecular genetics*, 18(4):767–778, 2009. 73, 82, 86

Alexander T Basilevsky. *Statistical factor analysis and related methods: theory and applications*, volume 418. John Wiley & Sons, 2009. 4

Christopher M Bishop. Latent variable models. In *Learning in graphical models*, pages 371–403. Springer, 1998. 2, 3

Christopher M Bishop. Bayesian pca. *Advances in neural information processing systems*, pages 382–388, 1999. 22

Joshua S Bloom, Ian M Ehrenreich, Wesley T Loo, Thúy-Lan Võ Lite, and Leonid Kruglyak. Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237, 2013. 82, 100

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26:211–252, 1964. 73, 77

Rainer Breitling, Yang Li, Bruno M Tesson, Jingyuan Fu, Chunlei Wu, Tim Wiltshire, Alice Gerrits, Leonid V Bystrykh, Gerald De Haan, Andrew I Su, et al. Genetical genomics: spotlight on qtl hotspots. *PLoS Genetics*, 4(10): e1000232, 2008. 13

Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568): 752–755, 2002. 11, 33, 34, 99

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. 18

Yen-Feng Chiu, Lee-Ming Chuang, Chin-Fu Hsiao, Yi-Jen Hung, Ming-Wei Lin, Ying-Tsung Chen, John Grove, Eric Jorgenson, Thomas Quertermous, Neil Risch, et al. An autosomal genome-wide scan for loci linked to pre-diabetic

phenotypes in nondiabetic chinese subjects from the stanford asia-pacific program of hypertension and insulin resistance family study. *Diabetes*, 54(4): 1200–1206, 2005. 73, 80, 86

Gary A Churchill. Fundamentals of experimental design for cdna microarrays. *Nature Genetics*, 32:490–495, 2002. 12

Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 207–215, 2013. 98

Jingyuan Fu, Marcel GM Wolfs, Patrick Deelen, Harm-Jan Westra, Rudolf SN Fehrmann, Gerard J Te Meerman, Wim A Buurman, Sander SM Rensen, Harry JM Groen, Rinse K Weersma, et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS genetics*, 8(1):e1002431, 2012. 44

Nicoló Fusi, Oliver Stegle, and Neil D Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS computational biology*, 8(1):e1002330, 1 2012. ISSN 1553-7358. 7, 11, 44, 55, 56, 68

Nicoló Fusi, Christoph Lippert, Karsten Borgwardt, Neil D Lawrence, and Oliver Stegle. Detecting regulatory gene-environment interactions with unmeasured environmental factors. *Bioinformatics (Oxford, England)*, 29(11):1382–9, 6 2013. ISSN 1367-4811. 7, 43

Nicoló Fusi, Christoph Lippert, Neil D Lawrence, and Oliver Stegle. Genetic analysis of transformed phenotypes. *arXiv preprint arXiv:1402.5447*, 2014. 72

Xiangchao Gan, Oliver Stegle, Jonas Behr, Joshua G Steffen, Philipp Drewe, Katie L Hildebrand, Rune Lyngsoe, Sebastian J Schultheiss, Edward J Osborne, Vipin T Sreedharan, et al. Multiple reference genomes and transcriptomes for arabidopsis thaliana. *Nature*, 477(7365):419–423, 2011. 38

Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003. 80

Greg Gibson. The environmental contribution to gene expression profiles. *Nature Reviews Genetics*, 9(8):575–581, 2008. 44, 69

Liang Goh and Von Bing Yap. Effects of normalization on quantitative traits in association test. *BMC bioinformatics*, 10(1):415, 1 2009. ISSN 1471-2105. 73, 86

Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, et al. Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 44(10):1084–1089, 2012. 44

I.B. Hallgrímsdóttir and D.S. Yuster. A complete classification of epistatic two-locus models. *BMC genetics*, 9(1):17, 2008. 46

James Hensman, Nicoló Fusi, and Neil D Lawrence. Gaussian processes for big data. *Uncertainty in Artificial Intelligence*, 2013. 17, 97, 98

Blanca E Himes, Gary M Hunninghake, James W Baurley, Nicholas M Rafaels, Patrick Sleiman, David P Strachan, Jemma B Wilk, Saffron A G Willis-Owen, Barbara Klanderman, Jessica Lasky-Su, Ross Lazarus, Amy J Murphy, Manuel E Soto-Quiros, Lydiana Avila, Terri Beaty, Rasika A Mathias, Ingo Ruczinski, Kathleen C Barnes, Juan C Celedn, William O C Cookson, W James Gauderman, Frank D Gilliland, Hakon Hakonarson, Christoph Lange, Miriam F Moffatt, George T O'Connor, Benjamin A Raby, Edwin K Silverman, and Scott T Weiss. Genome-wide association analysis identifies pde4d as an asthma-susceptibility gene. *American journal of human genetics*, 84(5): 581–93, 5 2009. ISSN 1537-6605. 73, 82, 86

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933. 3

Stephanie R Huang, Shiwei Duan, Wasim K Bleibel, Emily O Kistner, Wei Zhang, Tyson A Clark, Tina X Chen, Anthony C Schweitzer, John E Blume, Nancy J Cox, and M Eileen Dolan. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proceedings of the National Academy of Sciences of the United States of America*, 104(23):9758–63, 6 2007. ISSN 0027-8424. 73, 80, 86

Evan W Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1): 118–127, 2007. 6, 12

Hyun Min Kang, Chun Ye, and Eleazar Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–25, December 2008a. ISSN 0016-6731. doi: 10.1534/genetics.108.094201. 6, 7, 12, 13, 26, 33, 34, 40, 68, 69, 87

Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, March 2008b. 12, 19, 21, 46, 53

Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-Yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42:348–354, 2010. ISSN 1061-4036. 6, 12, 14, 15, 17, 19, 21, 46

Sekar Kathiresan, Alisa K Manning, Serkalem Demissie, Ralph B D'Agostino, Aarti Surti, Candace Guiducci, Lauren Gianniny, Nel P Burtt, Olle Melander, Marju Orho-Melander, Donna K Arnett, Gina M Peloso, Jose M Ordovas, and L Adrienne Cupples. A genome-wide association study for blood lipid phenotypes in the framingham heart study. *BMC medical* , 8 Suppl 1(Suppl 1):S17, 1 2007. ISSN 1471-2350. 73, 82, 86

Martin Knott and David J Bartholomew. *Latent variable models and factor analysis*. Number 7. Edward Arnold, 1999. 4

Arthur Korte, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics*, 44(9):1066–1071, 2012. 84, 85, 86

Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005. 4, 18, 23, 50, 96, 97

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, September 2007. 7, 12, 13, 14, 22, 26, 40, 44, 54, 58, 69, 70

Simon M Lin, Pan Du, Wolfgang Huber, and Warren A Kibbe. Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic acids research*, 36(2):e11–e11, 2008. 45

Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–5, 2011. ISSN 1548-7105. 6, 15, 19, 20, 21, 24, 46, 52, 53, 78

Jennifer Listgarten, Carl Kadie, Eric E Schadt, and David Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107: 16465–16470, 2010. ISSN 0027-8424. 7, 13, 14, 17, 37, 40, 41, 44, 55, 56, 67, 68, 69, 70, 87

Oren Litvin, Helen C Causton, Bo-Juen Chen, and Dana Pe'Er. Modularity and interactions in the genetics of gene expression. *Proceedings of the National Academy of Sciences*, 106(16):6441–6446, 2009. 44

Devin P Locke, Richard Segraves, Lucia Carbone, Nicoletta Archidiacono, Donna G Albertson, Daniel Pinkel, and Evan E Eichler. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome research*, 13(3):347–357, 2003. 6, 12

Michael Lynch and Kermit Ritland. Estimation of pairwise relatedness with molecular markers. *Genetics*, 152(4):1753–1766, 8 1999. 76

David JC MacKay. 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. 22, 51

Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008. 12, 44

Jacob L McCauley, Chun Li, Lan Jiang, Lana M Olson, Genea Crockett, Kimberly Gainer, Susan E Folstein, Jonathan L Haines, and James S Sutcliffe. Genome-wide and ordered-subset linkage analyses provide support for autism loci on 17q and 19p with evidence of phenotypic and interlocus genetic correlates. *BMC medical genetics*, 6:1, 1 2005. ISSN 1471-2350. 73, 80, 86

Charles E McCulloch and John M Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001. 72

Tom P Minka. Automatic choice of dimensionality for PCA. *Advanced in Neural Information Processing Systems*, pages 598–604, 2001. 22

Amanda J Myers, J Raphael Gibbs, Jennifer A Webster, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Doris Leung, Leslie Bryden, Priti Nath, et al. A survey of genetic human cortical gene expression. *Nature genetics*, 39 (12):1494–1499, 2007. 38, 100

Artika Praveeta Nath, Dalia Arafat, and Greg Gibson. Using blood informative transcripts in geographical genomics: impact of lifestyle on gene expression in fijians. *Applied Genetic Epidemiology*, 3:243, 2012. 44

Radford M Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995. 23

Alexandra C Nica, Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena Sekowska, Mary Travers, Simon Potter, Elin Grundberg, Kerrin Small, et al. The architecture of gene regulatory variation across multiple human tissues: the muther study. *PLoS genetics*, 7(2):e1002003, 2011. 13, 44

Elizabeth E Patton, Andrew R Willems, Danne Sa, Laurent Kuras, Dominique Thomas, Karen L Craig, and Mike Tyers. Cdc53 is a scaffold protein for multiple cdc34/skp1/f-box protein complexes that regulate cell division and methionine biosynthesis in yeast. *Genes & development*, 12(5):692–705, 1998. 65

Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010. 38

Vincent Plagnol, Elif Uz, Chris Wallace, Helen Stevens, David Clayton, Tayfun Ozcelik, and John A Todd. Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS One*, 3(8):e2966, 2008. 6, 12

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006. 25, 68

Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010. 68

Barbara Rakitsch, Christoph Lippert, Oliver Stegle, and Karsten Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214, 2013. 86

Carl E Rasmussen and Chris KI Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. 18, 22

Paul M Ridker, Nader Rifai, Nancy R Cook, Gary Bradwin, and Julie E Buring. Non–hdl cholesterol, apolipoproteins ai and b100, standard lipid measures, lipid ratios, and crp as risk factors for cardiovascular disease in women. *Jama*, 294 (3):326–333, 2005. 84

Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. *Neural computation*, 11(2):305–345, 1999. 3

Hyunju Ryoo and Chaeyoung Lee. Underestimation of heritability using a mixed model with a polygenic covariance structure in a genome-wide association study for complex traits. *European Journal of Human Genetics*, 2013. 81, 86

Chiara Sabatti, Susan K Service, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G Jones, Noah A Zaitlen, Teppo Varilo, Marika Kaakinen, Ulla Sovio, Aimo Ruokonen, Jaana Laitinen, Eveliina Jakkula, Lachlan Coin, Clive Hoggart, Andrew Collins, Hannu Turunen, Stacey Gabriel, Paul Elliot, Mark I McCarthy, Mark J Daly, Marjo-Riitta Jrvelin, Nelson B Freimer, and Leena Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41:35–46, 2009. ISSN 1061-4036. 84, 101

Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj GuhaThakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005. 38, 100

Vincent Segura, Bjarni J Vilhjálmsson, Alexander Platt, Arthur Korte, Ümit Seren, Quan Long, and Magnus Nordborg. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, 44(7):825–830, 2012. 86

Bertrand Servin and Matthew Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, 3:e114, 2007. ISSN 1553-7390. 73, 86

Liat Shimon, Gillian M Hynes, Elizabeth A McCormack, Keith R Willison, and Amnon Horovitz. Atp-induced allostery in the eukaryotic chaperonin cct is abolished by the mutation g345d in cct4 that renders yeast temperature-sensitive for growth. *Journal of molecular biology*, 377(2):469–477, 2008. 65

Erin N Smith and Leonid Kruglyak. Gene–environment interaction in yeast gene expression. *PLoS biology*, 6(4):e83, 2008. 11, 25, 33, 34, 37, 44, 45, 48, 55, 60, 62, 99

Edward Snelson, Carl Edward Rasmussen, and Zoubin Ghahramani. Warped gaussian processes. *Advances in neural information processing systems*, 16: 337–344, 2004. 4, 76

Doug Speed, Gibran Hemani, Michael R Johnson, and David J Balding. Improved heritability estimation from genome-wide snps. *American journal of human genetics*, 91(6):1011–21, 12 2012. ISSN 1537-6605. 80, 97

Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. 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. 7, 11, 13, 14, 22, 25, 26, 33, 40, 44, 56, 68, 69

Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3): 500–507, 2012. 7, 13, 14, 26, 40, 68, 70

Matthew Stephens. A unified framework for association analysis with multiple related phenotypes. *PloS one*, 8:e65245, 2013. ISSN 1932-6203. 73, 86, 87

Sam Stoler, Kelly Rogers, Scott Weitze, Lisa Morey, Molly Fitzgerald-Hayes, and Richard E Baker. Scm3, an essential saccharomyces cerevisiae centromere protein required for g2/m progression and cse4 localization. *Proceedings of the National Academy of Sciences*, 104(25):10571–10576, 2007. 65

John D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. ISSN 0090-5364. 21, 78

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003. 19, 21, 50, 53, 78

Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, et al. Population genomics of human gene expression. *Nature genetics*, 39(10):1217–1224, 2007. 11, 48

Miguel C Teixeira, Pedro Monteiro, Pooja Jain, Sandra Tenreiro, Alexandra R Fernandes, Nuno P Mira, Marta Alenquer, Ana T Freitas, Arlindo L Oliveira, and Isabel Sá-Correia. The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucleic acids research*, 34(suppl 1):D446–D451, 2006. 37, 100

Feng Tian, Peter J Bradbury, Patrick J Brown, Hsiaoyi Hung, Qi Sun, Sherry Flint-Garcia, Torbert R Rocheford, Michael D McMullen, James B Holland, and Edward S Buckler. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics*, 43(2): 159–162, 2011. 73, 80, 86

Christopher Tiedje, Imme Sakwa, Ursula Just, and Thomas Höfken. The rho gdi rdi1 regulates rho gtpases by distinct mechanisms. *Molecular biology of the cell*, 19(7):2885–2896, 2008. 65

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. 3

Michalis K. Titsias and Neil D. Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 844–851, 2010. 17, 97

William Valdar, Leah C Solberg, Dominique Gauguier, William O Cookson, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint. Genetic and environmental effects on complex traits in mice. *Genetics*, 174(2):959–84, 10 2006. ISSN 0016-6731. xv, 73, 80, 83, 84, 90, 97, 100

Ana Vinuela, L Basten Snoek, Joost A G Riksen, and Jan E Kammenga. Genome-wide gene expression regulation as a function of genotype and age in *C. elegans*. *Genome Research*, 20(7):929–937, July 2010. 44

Chris Wallace, Stephen J Newhouse, Peter Braund, Feng Zhang, Martin Tobin, Mario Falchi, Kourosh Ahmadi, Richard J Dobson, Ana Carolina B Marano, Cother Hajat, Paul Burton, Panagiotis Deloukas, Morris Brown, John M Connell, Anna Dominiczak, G Mark Lathrop, John Webster, Martin Farrall, Tim Spector, Nilesh J Samani, Mark J Caulfield, and Patricia B Munroe. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *American journal of human genetics*, 82(1):139–49, 1 2008. ISSN 1537-6605. 73, 82, 86

Jian Yang, Beben Benyamin, Brian P Mcevoy, Scott Gordon, Anjali K Henders, R Dale, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common snps explain a large proportion of heritability for human height. *Nature Genetics*, 42:565–569, 2011. ISSN 1546-1718. 80, 97

Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, 2005. 25

Noah Zaitlen and Peter Kraft. Heritability in the genome-wide association era. *Human genetics*, 131(10):1655–64, 10 2012. ISSN 1432-1203. 80

Xiang Zhou and Matthew Stephens. Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *arXiv preprint arXiv:1305.4366*, pages 1–35, 2013. xvi, xvii, 73, 84, 86, 87, 92, 94

Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 1 2013. ISSN 1553-7404. xvii, 84, 85, 94, 101

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997. 50