Chapter 2

Association mapping with high-dimensional traits

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

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

2.1 Expression QTLs

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

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

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

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

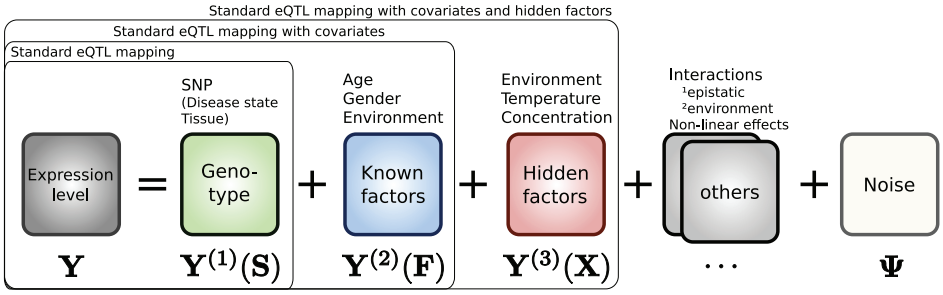


Figure 2.1: General additive model for sources of gene expression variability. The $G \times J$ matrix $\mathbf{Y}$ of measured gene expression levels of G genes from J individuals is modelled by additive contributions from components $\{\mathbf{Y}^{(m)}\}$ and observation noise $\mathbf{\Psi}$. Here,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

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

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

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

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

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

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

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

2.2 The PEER framework

Here, we present PEER, a general framework for modelling diverse sources of gene expression variability. The model underlying this framework assumes that

gene expression levels are influenced by additive effects from independent sources, e.g. in the case of VBQTL these are contributions from genotype, known factors, and hidden factors (Figures 2.1, 2.2a). We cast the full model in a probabilistic setting, treating its parameters as random variables.

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

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

2.2.1 Model

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

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

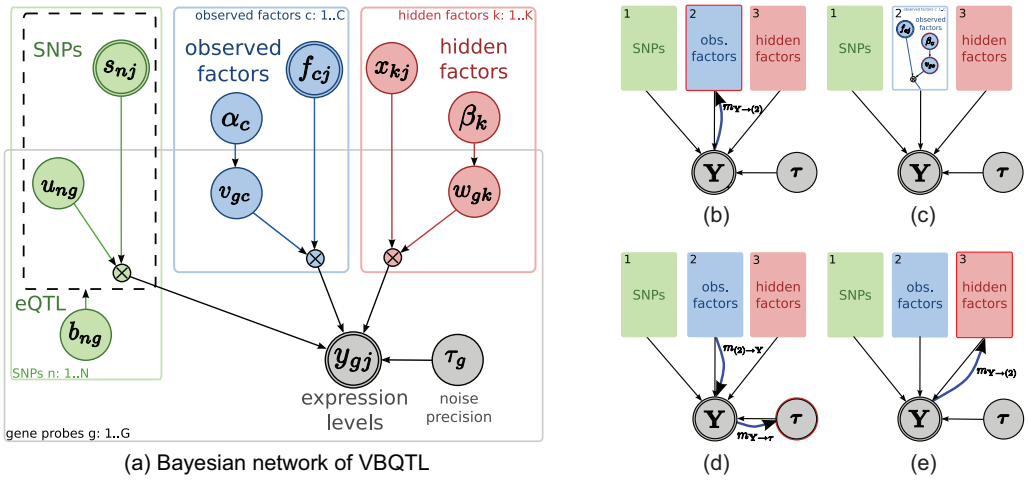


Figure 2.2: Bayesian network and outline of the inference schedule for VBQTL. **(a)** The Bayesian network for the model of gene expression variation used in VBQTL. The full model combines genetic (green), known factor (blue) and hidden factor (red) models to explain the observed gene expression levels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

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

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

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

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

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

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

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

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

blue plate):

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

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

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

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

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

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

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

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

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

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

2.2.2 Inference

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

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

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

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

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

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

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

In experiments, we compare two alternative inference schedules of VBQTL. In iterative VBQTL (iVBQTL), the parameters are learned using several iterations through all model components, first updating the genetic model, then known

and hidden factors. An important property of iVBQTL is that hidden factors are estimated jointly with the genetic state and known factors. This choice of schedule and the iterative learning help to ensure that variability that is due to genetic associations is not explained away by other parts of the model (Chapter 2.3).

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

2.2.3 Alternatives

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

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

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

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

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

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

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

learned and subtracted off the hidden factor effects on the residuals. All these alternative methods are also implemented in the general framework.

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

2.3 Method comparison

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

2.3.1 Comparison on simulated data

Simulation setup

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

by all the considered approaches, and it gives intuition for the results on real datasets discussed later.

Complexity control determines the accuracy of the hidden factor model.

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

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

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

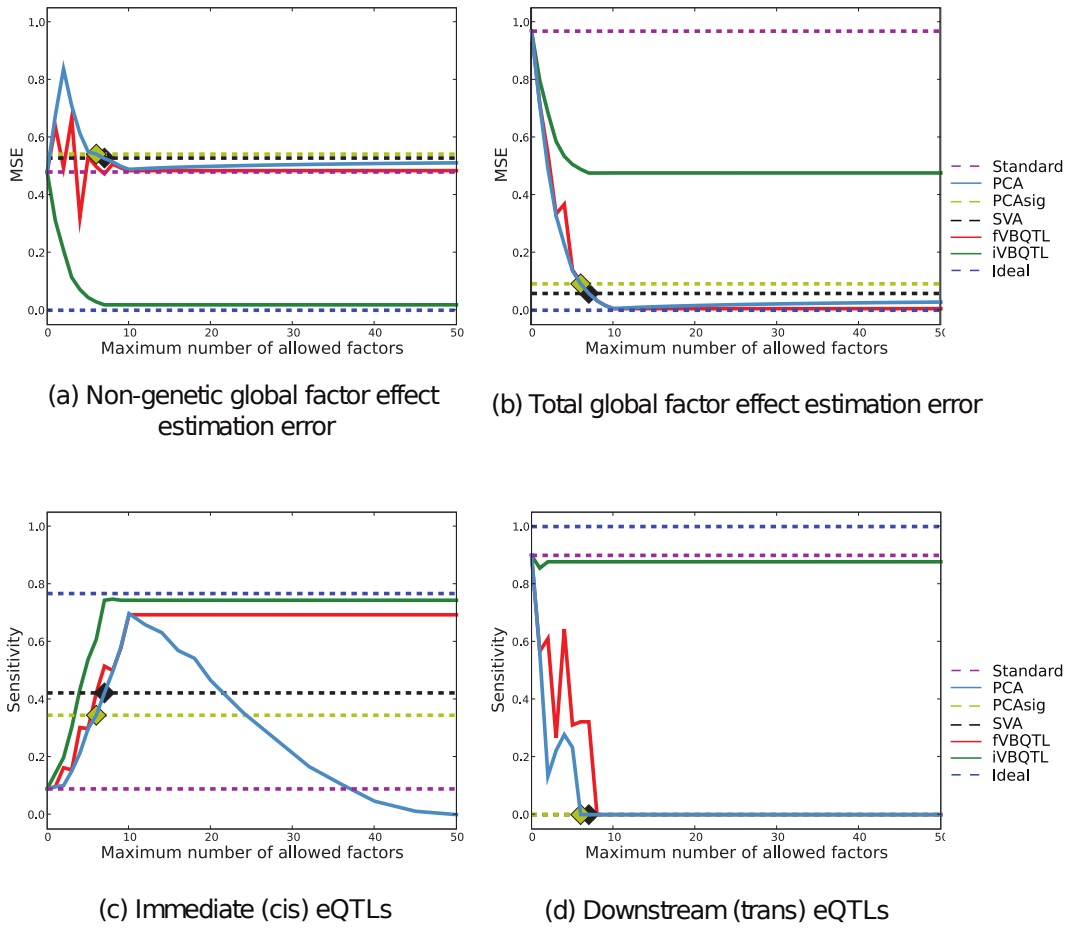


Figure 2.3: Sensitivity of recovering simulated hidden factor effects and eQTLs for Bayesian and non-Bayesian methods. **(a)** Mean-squared error in estimating only the hidden factor contribution. Methods that do not explicitly retain the genetic factors explain them away as hidden global factors, resulting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

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

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

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

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

All methods except iVBQTL and standard method explained away simulated *trans* eQTL hotspots (Figure 2.3d). This is due to the global factor effect estimation accuracy, where iVBQTL alone refrained from explaining the hotspots away as a global factor. The standard method found nearly all the original

trans associations, actually outperforming methods that explain away confounding variability. Thus, in cases where there is true genetic signal with widespread downstream effects, its contribution needs to be taken into account to retain its relation to genotype, and avoid attributing it to a confounding global cause. This is straightforward in our framework, and is demonstrated by the good performance of iVBQTL in this scenario. iVBQTL retained the original associations, while explaining away non-genetic causes of expression variability, thus adding power to detect the weaker, masked eQTLs. This effect is also observed in the study of crosses of inbred strains below.

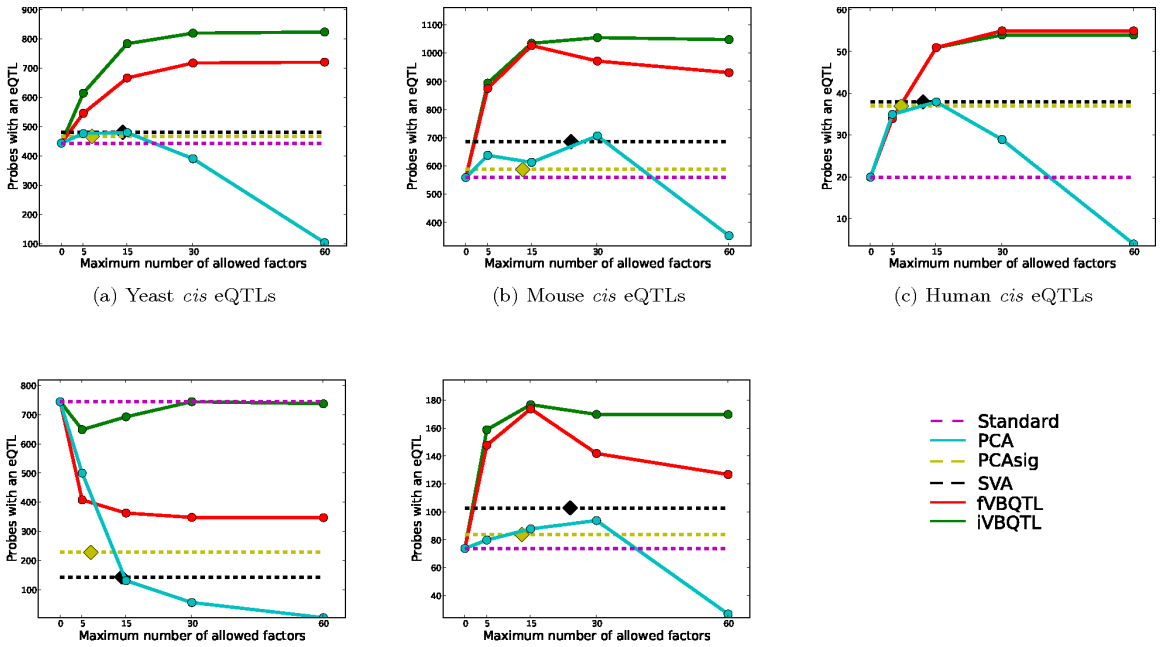
Taken together these results suggest that it is important to account for the confounding sources of variation in expression levels, while keeping the signal of the genetic state. Correct complexity control is required to avoid over- and underfitting in order to achieve optimal sensitivity for detecting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

2.3.2 Comparison on real data

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

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

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



(a) Yeast *cis* eQTLs (b) Mouse *cis* eQTLs (c) Human *cis* eQTLs

(d) Yeast *trans* eQTLs (e) Mouse *trans* eQTLs

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

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

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

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

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

fVBQTL and iVBQTL found more probes with an association (55 and 54) than all other methods, representing an almost threefold increase in the number of genes with a *cis* eQTL. Complexity control assured that the performance saturated for large enough number of factors and did not degrade as for PCA. None of the estimated hidden factors was significantly correlated to a SNP genotype,

suggesting that individual genetic variants do not have global effects on many gene expression levels in this dataset.

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

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

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

The relative performance of different methods was similar to their ability to detect *cis* eQTLs in the outbred population dataset (Figures 2.4a, 2.4b). The absolute performance gain was significantly lower for all methods, however. This finding suggests that the genetic signal is stronger compared to confounding sources of variation, which is not unexpected from the study design. All factor methods identified additional associations compared to the standard method. PCA overfitted for larger numbers of principal components used, explaining away genetic association signal. SVA and PCAsig found the same number of significant hidden factors for a range of significance cutoffs considered, exhibiting little

flexibility. Again, their performance was similar to extrapolation of PCA results with matching numbers of effective components. fVBQTL and iVBQTL found additional genetic associations in *cis* compared to the standard model and other methods for accounting for confounding variance, as observed in simulations and human dataset. Summary statistics for the method performance can be found in tables B.6 to B.8.

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

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

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

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

2.4 Expression QTL mapping in large human populations

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

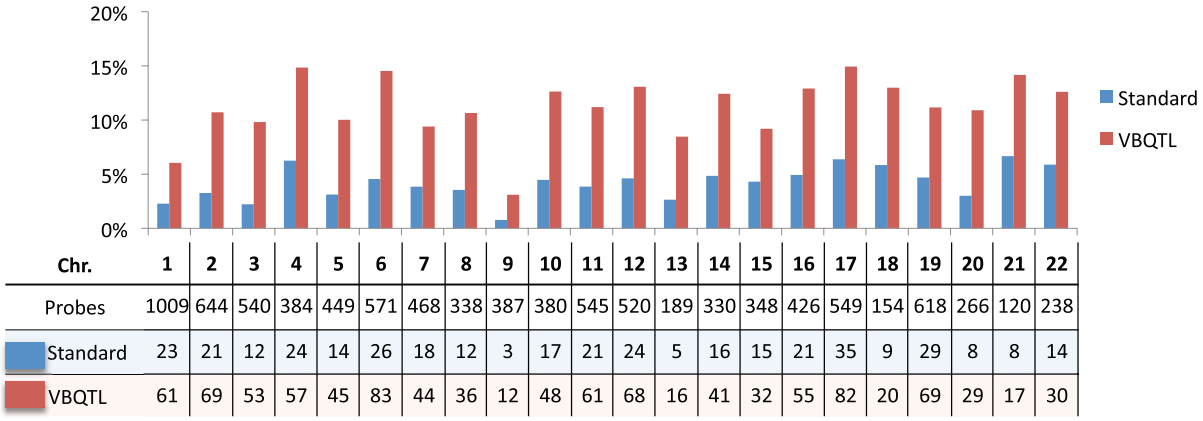
2.4.1 HapMap phase 2 dataset.

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

VBQTL increases power threefold

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

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



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

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

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

Additional associations are due to increased sensitivity.

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

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

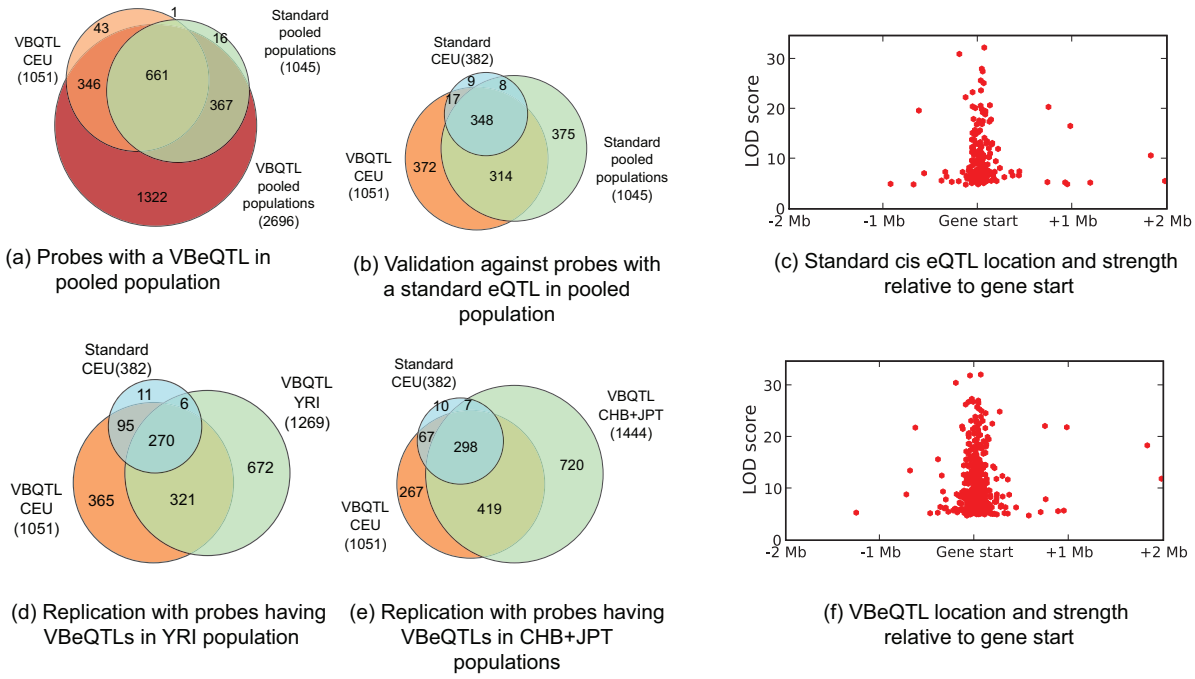


Figure 2.6: Validation of VBQTLs by comparison to standard eQTLs. **(a,b,d,e)** Venn diagrams depicting overlap of probes with a standard eQTL or VBeQTL in the CEU population and probes with an eQTL in other populations.

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

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

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

2.4.2 The MuTHER study

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

Fitting hyperparameters to maximise consistency

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

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

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

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

Many more QTLs are found

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

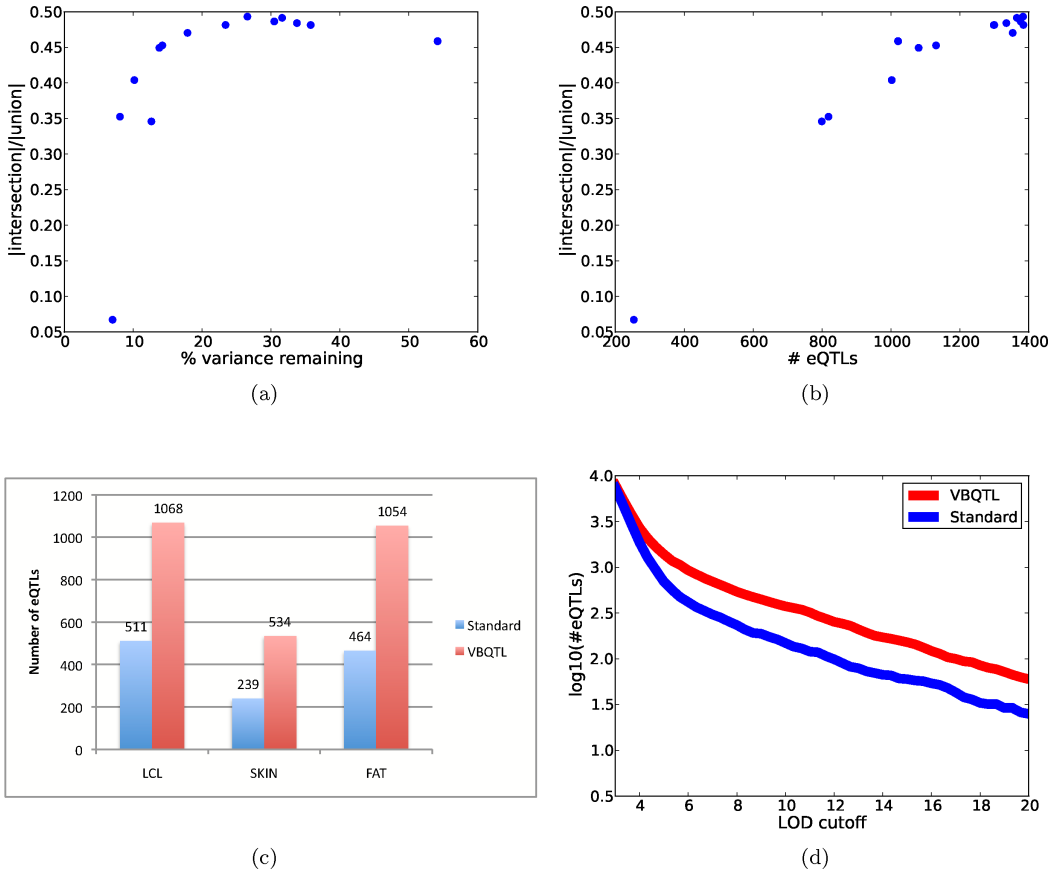


Figure 2.7: eQTL finding results on the MuTHER dataset. **(a)** Overlap of co-twin eQTLs as a function of variance explained by the factor analysis model **(b)** Correlation of co-twin eQTL overlap and total number of discoveries **(c)** Total number of discovered eQTLs in the three tissues with standard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

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

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

2.4.3 The 1000 Genomes low coverage pilot

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

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

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

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

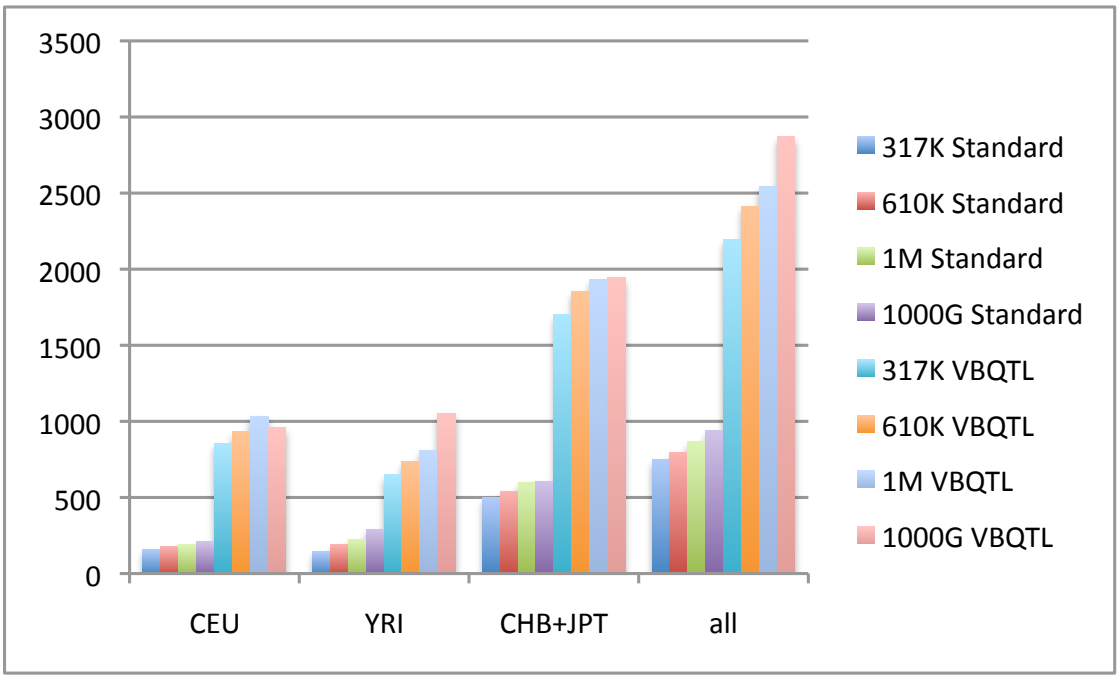


Figure 2.8: eQTL finding results on the HapMap2 dataset with 1000 Genomes genotypes. Number of eQTLs in three different populations as well as combined population significant at 5% FDR using the standard eQTL model (Standard), and residuals from PEER factor analysis model (VBQTL). Spearman's Rank Correlation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

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

2.5 Interpretation of learned hidden factors.

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

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

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

procedures needed to isolate RNA from the resistant skin tissue. Low RNA quality implies degradation of RNA molecules among other effects, which has a broad effect on many gene expression levels, and was captured with an inferred latent factor.

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

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

2.6 Discussion

We have presented VBQTL, a probabilistic model to dissect gene expression variation in the context of genetic association studies. The model is implemented in a Bayesian inference framework that allows uncertainty to be propagated between

different parts of the model, and yields posterior distributions over parameter estimates for more sensitive analysis. In comparative eQTL mapping experiments, VBQTL outperformed alternative methods for eQTL finding on simulated and real data. In the most striking example, VBQTL found up to three times more eQTLs than a standard method, and 45% more compared to the best alternative in the HapMap 2 expression dataset.

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

VBQTL and other methods that account for hidden factors all found additional expression QTLs in the datasets studied compared to the standard method. It is remarkable that, with only 270 samples, and looking in one tissue type, we can find significant genetic associations to 27% of the expressed genes. The replication of the additional associations in different populations suggests that they are genuine. The increase in power is due to the hidden factor model, which explains away unwanted non-genetic variability, thereby allowing the genetic effects to stand out to a greater extent. The high number of additional associations suggests that association finding studies in human have not saturated, and we expect the fraction of genes with an eQTL will increase further as the number

of samples grows. It may be that the expression of the majority of human genes varies as a result of segregating genetic variation. While previous studies have reported only 12% of heritable variation to be due to *cis* variants (Price et al., 2008), this does not contradict the presence of weak *cis* eQTLs for a large fraction of the genes.

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

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