BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 11 2013, pages 1382–1389
doi:10.1093/bioinformatics/btt148

Gene expression

Advance Access publication April 4, 2013

Detecting regulatory gene–environment interactions with unmeasured environmental factors

Nicoló Fusi[1,*], Christoph Lippert[2], Karsten Borgwardt[3,4], Neil D. Lawrence[1] and Oliver Stegle[3,5,*]

[1]Department of Computer Science, University of Sheffield, Sheffield S10 2HQ, UK, [2]Microsoft Research, Los Angeles, CA 90024, USA, [3]Machine Learning and Computational Biology Research Group, Max Planck Institutes, 72076 Tübingen, Germany, [4]Zentrum für Bioinformatik, Eberhard Karls Universität, 72074 Tübingen, Germany and [5]EMBL-European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, UK

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Genomic studies have revealed a substantial heritable component of the transcriptional state of the cell. To fully understand the genetic regulation of gene expression variability, it is important to study the effect of genotype in the context of external factors such as alternative environmental conditions. In model systems, explicit environmental perturbations have been considered for this purpose, allowing to directly test for environment-specific genetic effects. However, such experiments are limited to species that can be profiled in controlled environments, hampering their use in important systems such as human. Moreover, even in seemingly tightly regulated experimental conditions, subtle environmental perturbations cannot be ruled out, and hence unknown environmental influences are frequent. Here, we propose 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.

Results: In extensive simulation studies, we show that our method is able to accurately reconstruct environmental factors and their interactions with genotype in a variety of settings. We further illustrate the use of our model in a real-world dataset in which one environmental factor has been explicitly experimentally controlled. Our method is able to accurately reconstruct the true underlying environmental factor even if it is not given as an input, allowing to detect genuine genotype–environment interactions. In addition to the known environmental factor, we find unmeasured factors involved in novel genotype–environment interactions. Our results suggest that interactions with both known and unknown environmental factors significantly contribute to gene expression variability.

Availability: and implementation: Software available at http://pmbio.github.io/envGPLVM/.

Contact: oliver.stegle@ebi.ac.uk or nicolo.fusi@sheffield.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 16, 2012; revised on February 28, 2013; accepted on March 21, 2013

*To whom correspondence should be addressed.

1 INTRODUCTION

Large-scale genotyping and expression profiling initiatives have fostered expression quantitative trait loci (eQTL) analyses, investigating the genetic component of gene expression variability. eQTL studies in human (Montgomery *et al.*, 2010; Pickrell *et al.*, 2010; Stranger *et al.*, 2007, 2012), segregating yeast strains (Brem *et al.*, 2002; Smith and Kruglyak, 2008), mouse (Schadt *et al.*, 2005) and *Arabidopsis thaliana* (Gan *et al.*, 2011; West *et al.*, 2007) have revealed an abundance of associations between genetic polymoprhisms and the expression levels of individual genes.

While building this compendium of 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 toward 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, also 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 are not directly measurable; 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.

Here, we present an integrated probabilistic model, LInear Mixed Model Interaction (*LIMMI*), which allows for recovering 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 (Fig. 1). At the same time, the model enforces that the estimated factors are truly environmental and not themselves under genetic control.

We evaluate LIMMI on synthetic data, where we assess the ability of LIMMI to (i) recover the true simulated environmental state, (ii) better detect direct genetic effects and, in particular, (iii) identify genotype–environment interactions with unmeasured environmental factors. We then 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.

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).

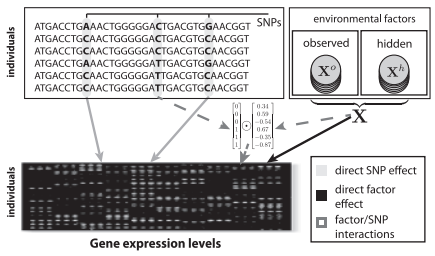

Fig. 1. Illustration of regulatory effects on gene expression modeled by LIMMI. First, non-genetic environmental factors can either be measured (observed) or hidden. Their effect on gene expression is typically dominated by direct effects (blue). In addition, some factors may act in a genotype-specific 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)

Expression variability is modeled as the sum of effects from single nucleotide polymorphisms (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

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

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, 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).

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. Furthermore, let $\mathbf{X} = [\mathbf{X}^o, \mathbf{X}^h]$ 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 manner (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:

$$p(\mathbf{Y}|\mathbf{S}, \mathbf{X}, \Theta_K) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{y}_g | \mu_g \mathbf{1}, \underbrace{\sum_{k=1}^{K} \beta_k^2 \mathbf{s}_k \mathbf{s}_k^\top}_{\mathbf{K_S}} + \underbrace{\sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^\top}_{\mathbf{K_X}} + \underbrace{\sum_{k=1}^{K} \sum_{q=1}^{Q} \gamma_{k,q}^2 (\mathbf{s}_k \odot \mathbf{x}_q)(\mathbf{s}_k \odot \mathbf{x}_q)^\mathrm{T}}_{\mathbf{K_I}} + \sigma_p^2 \mathbf{K_P} + \sigma_e^2 \mathbf{I}) \tag{2}$$

The set $\Theta_K = \{\alpha^2, \beta^2, \gamma^2, \sigma_p^2, \sigma_e^2\}$ 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 α_q^2, β_k^2 and $\gamma_{k,q}^2$, respectively. The covariance model in Equation (2) can also be derived from a linear model equivalence (see Supplementary Methods, Section 1), assuming a hierarchical prior on the effect sizes for interactions and direct effects across all genes.

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 (2) infeasible, both computationally and

statistically (see also Section 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 Θ_K are learnt for given active sets $\mathcal{S}$ and $\mathcal{I}$. Second, for fixed state of $\mathbf{X}, \Theta_K$, 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 2.2, we describe this EM-like iterative training scheme. The technical building blocks of the individual training steps are presented in Section 2.2.1, describing the gradient-based optimization of model parameters, and in Section 2.2.2, addressing the selection of SNPs to be included in the model. More details about the implementation and the initialization of the model are available in the Supplementary Methods.

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 Θ_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 2.2.1). As previously shown (Fusi *et al.*, 2012), 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 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 with any of the learnt factors $\mathbf{X}$. As a result of this process, genotype and the learnt hidden factors are quasi-orthogonal (see also Fusi *et al.*, 2012 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 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 2.3.

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 (2) reduces to

$$p(\mathbf{Y} \mid \mathbf{S}, \mathbf{X}, \Theta_K, \mathcal{I}, \mathcal{S}) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{y}_g \mid \mathbf{0}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} = \underbrace{\sum_{\forall k \in \mathcal{S}} \beta_k^2 \mathbf{s}_k \mathbf{s}_k^\top}_{\mathbf{K_S}} + \underbrace{\sum_{q=1}^{Q} \alpha_q^2 \mathbf{x}_q \mathbf{x}_q^\top}_{\mathbf{K_X}} + \underbrace{\sum_{\forall (k,q) \in \mathcal{I}} \gamma_{k,q}^2 (\mathbf{s}_k \odot \mathbf{x}_q)(\mathbf{s}_k \odot \mathbf{x}_q)^\mathrm{T}}_{\mathbf{K_I}} + \sigma_p^2 \mathbf{K_p} + \sigma_e^2 \mathbf{I} \tag{3}$$

where $\boldsymbol{\Sigma}$ is the overall covariance, which in turn is parametrized by $\mathbf{X}$, Θ_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) can be written as

$$\begin{aligned} \ln p(\mathbf{Y} \mid \mathbf{S}, \mathbf{X}, \Theta_K, \mathcal{I}, \mathcal{S}) &= \ln \prod_{g=1}^{G} \mathcal{N}(\mathbf{y}_g \mid \mathbf{0}, \boldsymbol{\Sigma}) \\ &= -\frac{GN}{2} 2\pi - \frac{G}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \mathrm{Tr}(\boldsymbol{\Sigma}^{-1} \mathbf{Y}\mathbf{Y}^\mathrm{T}) \end{aligned} \tag{4}$$

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

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

and combing it with the covariance derivative with respect to the ith kernel parameter, $\frac{\mathrm{d}}{\mathrm{d}\Theta_{K_i}} \boldsymbol{\Sigma}$, 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 Θ_K

$$\widehat{\boldsymbol{\Theta}_K}, \widehat{\mathbf{X}} = \underset{\boldsymbol{\Theta}_K, \mathbf{X}}{\mathrm{argmax}} \ln p(\mathbf{Y} \mid \mathbf{S}, \mathbf{X}, \Theta_K, \mathcal{I}, \mathcal{S}). \tag{5}$$

A standard gradient-based optimizer, such as L-BFGS-B (Zhu *et al.*, 1997), can be used to take advantage of the availability of closed-form gradients with respect to the elements of $\mathbf{X}$ and Θ_K. A discussion on how the latent dimensionality Q is chosen and the implications on the model fitting are provided in Supplementary Methods, Section 2.

2.2.2 Inclusion of genetic effects Individual SNPs are selected for inclusion in $\mathcal{S}$. We follow the approach taken in Fusi *et al.*, (2012), 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 this significance cutoff has previously been demonstrated (Fusi *et al.*, 2012).

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 2.3.1). Significant interaction terms ($qv \leq \alpha_{\mathrm{GxE}}$) 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.

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, in contrast, 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 ($\alpha_{\mathrm{SNP}} \leq 0.1$ in the case of the yeast dataset presented in Section 3.2) and interaction terms ($\alpha_{\mathrm{GxE}} \leq 0.05$ 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 previous work (Fusi *et al.*, 2012).

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; 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 (see Fig. 2c and d and Supplementary Fig. S3).

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 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, ~50 000 gene expression levels and ~100 000 SNPs.

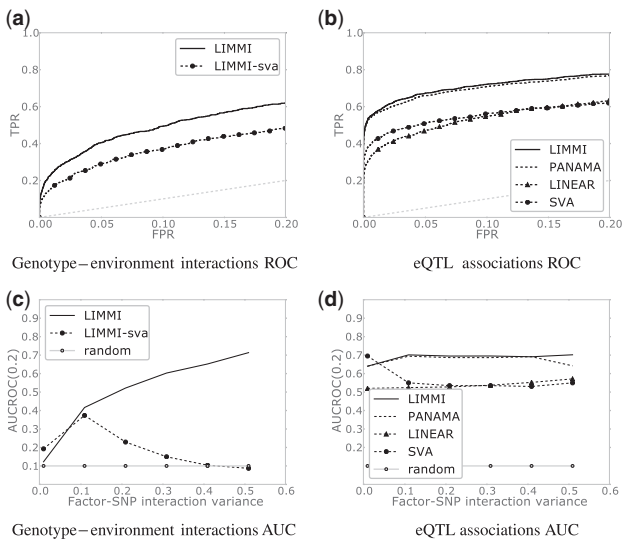

Fig. 2. Comparative evaluation of LIMMI and alternative methods on simulated datasets. (**a**) ROC for recovering simulated interactions between hidden factors and genotype. Linear regression has been omitted because it is not applicable to test for hidden environment interactions. The light grey line 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. AUC 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

For example, on the yeast dataset analyzed in Section 3.2, LIMMI converged within only 50 min [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® Xeon® X7542 12C CPU and 64 gigabytes of RAM. The python scientific libraries (Numpy and Scipy) were compiled against the Intel® Math Kernel Library]. These datasets contained 218 samples, 2956 SNPs and 5493 gene expression levels.

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 2.1).

For testing, we use 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 assesses the relevance of the effect of interest

$$p(\mathbf{y}_g \mid \sigma_g^2, \delta_g, \mathbf{\Sigma}) = \mathcal{N}(\mathbf{y}_g \mid \underbrace{\mathbf{f}()}_{\text{fixed effect}}, \underbrace{\sigma_g^2(\mathbf{\Sigma}_g + \delta_g \mathbf{I})}_{\text{random effect}}). \quad (6)$$

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 use the approximation proposed in Kang *et al.* (2008b) and 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 2.1). A mechanistic understanding of the individual covariance components can be gained when deriving the individual covariance parameters from a linear model equivalence (Supplementary Methods, Section 1).

In Section 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 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 (FDR) estimates from the q-value package (Storey and Tibshirani, 2003).

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

$$\mathrm{LOD}_{k,q,g}^{\mathrm{inter}} = \log \frac{\mathcal{N}(\mathbf{y}_g \mid \theta_{i,g}(\mathbf{s}_k \odot \mathbf{x}_q) + \theta_{k,g}\mathbf{s}_k + \theta_{q,g}\mathbf{x}_q, \sigma_g^2(\mathbf{\Sigma}_a + \delta_g \mathbf{I}))}{\mathcal{N}(\mathbf{y}_g \mid \qquad\qquad \theta_{k,g}\mathbf{s}_k + \theta_{q,g}\mathbf{x}_q, \sigma_g^2(\mathbf{\Sigma}_a + \delta_g \mathbf{I}))} \quad (7)$$

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.

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

$$\mathrm{LOD}_{k,g}^{\mathrm{asso}} = \log \frac{\mathcal{N}(\mathbf{y}_g \mid \theta_{k,g}\mathbf{s}_k, \sigma_g^2(\mathbf{\Sigma}_i + \delta_g \mathbf{I}))}{\mathcal{N}(\mathbf{y}_g \mid \mathbf{0} \qquad , \sigma_g^2(\mathbf{\Sigma}_i + \delta_g \mathbf{I}))} \quad (8)$$

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 (8).

2.4 LIMMI-sva

In principle, in the first step of the procedure outlined in Section 2.2.1, any latent variable model could be used to infer environmental factors. For comparison, we have implemented 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 2.2. The details of the testing procedure are described in Supplementary Methods, Section 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 2.4). We also considered a standard linear association test as a baseline method.

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 1000 gene expression levels. We simulated five environmental factors that have both direct effects on gene expression and interactions with genotype. In addition, we also considered five 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 five 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 confounding factors (Fusi *et al.*, 2012). If the genetics and the environment are not modeled jointly, part of the genetic signal will be captured by the estimated confounding factors, making the discovery of genotype–environment interactions even harder. To further investigate this hypothesis, we simulated five 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 used here does not favor any of the considered methods, as 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 2a depicts the receiver operating characteristic (ROC) curve, 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; 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 (Fig. 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 with the linear model (see also discussion 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.

Supplementary Figure S2 presents the analogous results when considering a precision–recall measure, leading to the same conclusions.

Finally, we investigated the impact when changing the relative magnitude of direct environmental effects and genotype–environment interactions. Figure 2c and d show the respective area under the ROC curve (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 (Fig. 2c), whereas the performance 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 (Fig. 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.

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. Supplementary Figure S3 shows analogous AUC performances when varying the variance explained by direct factor effects (Supplementary Fig. S3a and b) or the variance from genotype–factor interactions (Supplementary Fig. S3c and 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%, Supplementary Fig. S3c), 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 (Fusi *et al*., 2012), 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 Supplementary Figure S4 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.

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 nine latent factors, and LIMMI found 15 factors. 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 (Fig. 3a, at comparable statistical calibration; see Supplementary Figs S5 and S6). 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 (Supplementary Fig. S7), suggesting that the first factor indeed captures most of the effect due to the ethanol/glucose condition.

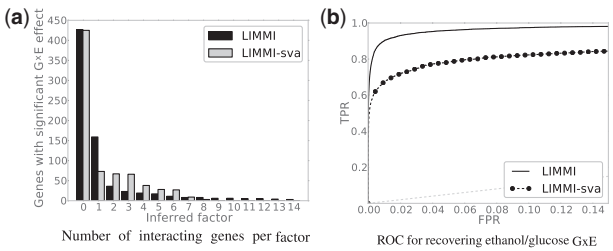

Fig. 3. Recovery of known and novel gene–environment interactions. (**a**) The number of genes with at least one significant genotype–environment interaction (FDR ≤ 0.01) as identified by LIMMI and SVA. The first factor was most correlated with the measured ethanol/glucose contrast, capturing this experimental condition. (**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

First, we focused on the recovered factor that is a likely proxy for the true environmental state. Figure 3b 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 SVA are themselves under genetic control (Supplementary Fig. S8). Second, the statistical test for interactions used in LIMMI accounts for direct effects of all other learnt factors, explaining away nuisance variation due to other environmental axes (Section 2.3.1). To investigate this further, we considered different variants of LIMMI-sva where alternative testing strategies were compared. Despite some improvement, this adjustment alone did not cure the inferior performance of LIMMI-sva (Supplementary Fig. S1), suggesting that both the lower extent of genetic correlation and the refined testing procedure contribute to the performance of LIMMI.

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 $G \times E$ interactions (Fig. 3a). The factors recovered by LIMMI-sva tended to be in strong association with genotype, suggesting that they capture genetic signals instead of environmental effects. The factors retrieved by LIMMI, in contrast, were found to be orthogonal to the genetic signal (Supplementary Fig. S8).

A map of the genetic loci and regulated genes for interactions with all factors detected by LIMMI is shown in Figure 4 (interaction results for each individual factor are given in Supplementary Fig. S9). 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 (Supplementary Table S1). 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 (±10 kb), there were six 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. Corresponding maps that reflect the interactions retrieved by LIMMI-sva are given in Supplementary

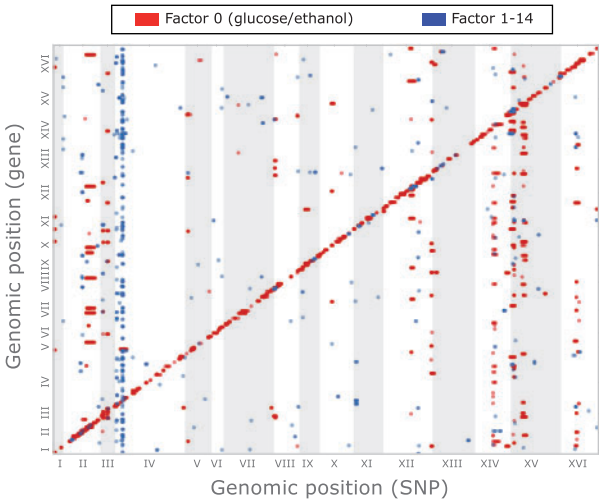

Fig. 4. Genomic map of the genotype–environment interactions retrieved by LIMMI (FDR ≤ 0.01). Shown are the positions of the SNP (x-axis) and the gene (y-axis) that participate in each significant genotype–environment interaction. Red circles correspond to interactions with the first latent factor that captures the known ethanol/glucose contrast. Blue interactions correspond to all other 14 factors

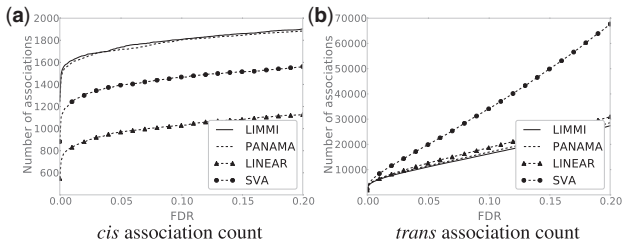

Fig. 5. Number of direct genetic associations (eQTLs) called by different methods as a function of the FDR cutoff. (**a**) *cis* associations. (**b**) *trans* associations. We considered at most one association per chromosome to avoid confounding the size of associations with their number

Figure S10, whereas Supplementary Figure S11 depicts the interaction map when using the known environmental 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 (Fig. 3b). Overall, LIMMI identified more *trans* bands for genotype–environment effects than LIMMI-sva.

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 5 shows the number of associations retrieved by alternative methods as a function of the FDR cutoff. As in the simulated settings (Fig. 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 with PANAMA. At the same time, the *P*-value statistics of LIMMI was slightly more uniform than PANAMA, suggesting that better control for confounding has been achieved (Supplementary Figs S5 and S6). 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 Fusi *et al.*, 2012, for a discussion). These 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, as they explain variance that cannot be captured by a model that relies on fully additive effects. On the other hand, 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.*, 2012, 2010) or correcting for population structure (Price *et al.*, 2006, 2010). eQTLs retrieved by LIMMI have a slightly better *cis* enrichment compared with PANAMA, a criterion previously suggested to judge the plausibility of eQTL results (Fusi *et al.*, 2012; Kang *et al.*, 2008a; Listgarten *et al.*, 2010).

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 with 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 (Fusi *et al.*, 2012;

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 retrieve an explicit representation of factors can be used for interaction analyses like the one presented here. Specifically, in this article, we implemented a version of our method that was using SVA for this purpose. LIMMI is most closely related and builds on PANAMA (Fusi *et al*., 2012); however, we propose a new route toward 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 mechanism 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.

Funding: FP7 PASCAL II Network of Excellence; University of Sheffield doctoral training award (to N.F.); Volkswagen Foundation and a Marie Curie FP7 grant (to O.S.).

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*., **11**, R106.

Auesukaree,C. *et al*. (2009) Genome-wide identification of genes involved in tolerance to various environmental stresses in *Saccharomyces cerevisiae*. *J. Appl. Genet*., **50**, 301–310.

Brem,R.B. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Fu,J. *et al*. (2012) Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet*., **8**, e1002431.

Fusi,N. *et al*. (2012) Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol*., **8**, e1002330.

Gan,X. *et al*. (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, **477**, 419–423.

Gibson,G. (2008) The environmental contribution to gene expression profiles. *Nat. Rev. Genet*., **9**, 575–581.

Grundberg,E. *et al*. (2012) Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nat. Genet*., **44**, 1084–1089.

Hallgrímsdóttir,I. and Yuster,D. (2008) A complete classification of epistatic two-locus models. *BMC Genet*., **9**, 17.

Kang,H.M. *et al*. (2008a) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**, 1909–1925.

Kang,H.M. *et al*. (2008b) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

Kang,H.M. *et al*. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet*., **42**, 348–354.

Lawrence,N. (2005) Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J. Mach. Learn. Res*., **6**, 1783–1816.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*., **3**, e161.

Lin,S. *et al*. (2008) Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic Acids Res*., **36**, e11.

Lippert,C. *et al*. (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.

Listgarten,J. *et al*. (2010) Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 16465–16470.

Litvin,O. *et al*. (2009) Modularity and interactions in the genetics of gene expression. *Proc. Natl Acad. Sci. USA*, **106**, 6441–6446.

Mackay,D.J. (1995) Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network*, **6**, 469–505.

McCarthy,M.I. *et al*. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet*., **9**, 356–369.

Montgomery,S. *et al*. (2010) Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, **464**, 773–777.

Nath,A. *et al*. (2012) Using blood informative transcripts in geographical genomics: impact of lifestyle on gene expression in fijians. *Front. Genet*., **3**, 243.

Nica,A. *et al*. (2011) The architecture of gene regulatory variation across multiple human tissues: the muther study. *PLoS Genet*., **7**, e1002003.

Patton,E. *et al*. (1998) Cdc53 is a scaffold protein for multiple cdc34/skp1/f-box protein complexes that regulate cell division and methionine biosynthesis in yeast. *Genes Dev*., **12**, 692–705.

Pickrell,J. *et al*. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

Price,A.L. *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet*., **38**, 904–909.

Price,A.L. *et al*. (2010) New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet*., **11**, 459–463.

Schadt,E. *et al*. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet*., **37**, 710–717.

Shimon,L. *et al*. (2008) ATP-induced allostery in the eukaryotic chaperonin cct is abolished by the mutation g345d in cct4 that renders yeast temperature-sensitive for growth. *J. Mol. Biol*., **377**, 469–477.

Smith,E.N. and Kruglyak,L. (2008) Gene-environment interaction in yeast gene expression. *PLoS Biol*., **6**, e83.

Stegle,O. *et al*. (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol*., **6**, e1000770.

Stegle,O. *et al*. (2012) Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc*., **7**, 500–507.

Stoler,S. *et al*. (2007) Scm3, an essential *Saccharomyces cerevisiae* centromere protein required for g2/m progression and cse4 localization. *Proc. Natl Acad. Sci. USA*, **104**, 10571.

Storey,J. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440.

Stranger,B. *et al*. (2007) Population genomics of human gene expression. *Nat. Genet*., **39**, 1217–1224.

Stranger,B. *et al*. (2012) Patterns of cis regulatory variation in diverse human populations. *PLoS Genet*., **8**, e1002639.

Tiedje,C. *et al*. (2008) The rho gdi rdi1 regulates rho gtpases by distinct mechanisms. *Mol. Biol. Cell*, **19**, 2885–2896.

Vinuela,A. *et al*. (2010) Genome-wide gene expression regulation as a function of genotype and age in *C. elegans*. *Genome Res*., **20**, 929–937.

West,M. *et al*. (2007) Global eqtl mapping reveals the complex genetic architecture of transcript-level variation in arabidopsis. *Genetics*, **175**, 1441–1450.

Zhu,C. *et al*. (1997) Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw*., **23**, 550–560.