Accurate modeling of confounding variation in eQTL studies leads to a great increase in power to detect *trans*-regulatory effects

Nicoló Fusi[1,†,*], Oliver Stegle[2,†,*], Neil D. Lawrence[3,*]
1 Sheffield Institute for Translational Neuroscience, University of Sheffield, UK
2 Machine Learning & Computational Biology Research Group, Max Planck Institute for Developmental Biology Tübingen, Germany
∗ E-mail: nicolo.fusi@sheffield.ac.uk, oliver.stegle@tuebingen.mpg.de, N.Lawrence@sheffield.ac.uk
† these authors contributed equally

Abstract

Expression quantitative trait loci (eQTL) studies are an integral tool to investigate the genetic component of gene expression variation. A major challenge in the analysis of such studies are hidden confounding factors, such as unobserved covariates or unknown environmental influences. These factors can induce a pronounced artifactual correlation structure in the expression profiles, which may create spurious false associations or mask real genetic association signals. Here, we report PANAMA (Probabilistic ANAlysis of genoMic dAta), a novel probabilistic model to account for confounding factors within an eQTL analysis. In contrast to previous methods, PANAMA learns hidden factors jointly with the effect of prominent genetic regulators. As a result, PANAMA can more accurately distinguish between true genetic association signals and confounding variation. We applied our model and compared it to existing methods on a variety of datasets and biological systems. PANAMA consistently performs better than alternative methods, and finds in particular substantially more *trans* regulators. Importantly, PANAMA not only identified a greater number of associations, but also yields hits that are biologically more plausible and can be better reproduced between independent studies.

Introduction

Genome-wide analysis of the regulatory potential of polymorphic loci on gene expression has been carried out in a range of different study designs and biological systems. For example in human, association mapping has uncovered an abundance of *cis* associations that are responsible for the variation of a third of all human genes [1,2]. In segregating yeast strains, linkage studies have provided evidence for extensive *trans* regulation, with a few regulatory hotspots controlling the expression profiles of tens or hundreds of genes [3,4].

Despite the success of expresion quantitative trait loci (eQTL) studies, it also has become clear that their statistical analysis comes along with statistical challenges [5]. External confounding factors, such as environmental influences or technical variation, can substantially alter the outcome of an eQTL study. Unobserved confounders can both obscure true association signals and create new spurious associations that are false [6,7].

Suitable data preprocessing, or careful design of randomized studies are helpful measures to avoid confounders in the first place [8], however they rarely rule out confounding influences entirely. It is also relatively straight-forward to account for those factors that are known and measured. For example, it is standard procedure to include covariates such as age and gender in the analysis [9,10]. Similarly, the effect of populational relatedness between samples, a confounding effect that is observed or can be reliably estimated form the genotypes [11,12], is widely included in the model. However, other factors, including subtle environmental or technical influences, often remain unknown to the experimenter, but still need to be accounted for. The impact of such typically hidden effects has been investigated in multiple studies; for example [13, 14] showed that virtually any aspect of sample handling can complicate the analysis.

Several techniques have been developed to account for unknown confounding variation within the eQTL analysis [2, 6, 7, 15]. A common assumption these methods build upon is that confounders are prone to exhibit broad influences, affecting large fractions of all measured gene expression levels. This characteristics has been exploited to learn hidden confounders using PCA-like models. For example [6] and [2] employed factor analysis models, a variant of PCA, to recover the hidden confounders. Once learnt, these factors can then be included in the analysis analogously to known covariates. Another branch of methods avoids recovering the hidden factors explicitly, instead correcting for the correlation structure induced by them in the samples [7,15]. Here, the inter-sample correlation is estimated from the expression profiles first, to then account for its effect in an association scan using mixed linear models. Both types of methods have been applied in a number of studies. Advantages versus naive analysis include better calibrated test statistics [15], and improved reproducibility of hits between independent studies [7]. Perhaps most strikingly, statistical methods to correct for hidden confounders have also been shown to substantially boost the power to detect eQTLs, increasing the number of significant *cis* associations by up to 3-fold [2,16].

While improved sensitivity to detect *cis*-acting eQTLs is an important and necessary step, we expect that even more valuable insights can be gained from those loci that regulate multiple target genes in *trans*. The interest in these regulatory hotspots has been tremendous in recent years, however it has also been shown that their reproducibility between studies is limited (see for example the discussion in [17]). Accurate correction for confounding factors is key to improve the reliability of these regulatory associations, however statistical overlap between confounding factors and true association signals from downstream effects can hamper the identification and fitting of confounders. For example, methodology that merely accounts for broad variance components, such as PCA, is doomed to fail. If the effect size of *trans* regulatory hotspots is large enough, they induce a correlation structure that is very similar to the one caused by confounding factors. As a result, true *trans* regulators tend to be mistaken for confounders and are erroneously explained away.

Here, we report an integrated probabilistic model PANAMA (Probabilistic ANAlysis of genoMic dAta) to adress these important shortcoming of established approaches. PANAMA learns a dictionary of confounding factors from the observed expression profiles. The key novelty of our approach is to jointly learn these factors while accounting for the effect of loci with a pronounced *trans* regulatory effect, thereby avoiding overlaps between true genetic association signals and the covariance structure induced by the learnt confounders. The statistical model underlying our algorithm is simple and computationally tractable for large eQTL datasets. PANAMA is based on the framework of mixed linear models, and combines the advantages of factor-based methods, such as PCA, SVA [6] or PEER [2] with methods that estimate the implicit covariance structure induced by confounding variation [12, 15]. The model is fully automated and can be easily adapted to include additional observed confounding sources of variation, such as population structure or known covariates.

We applied PANAMA to a range of eQTL datasets, including synthetic data and studies from yeast, mouse and human. Across datasets, PANAMA performed better than previous methods, identifying more statistically significant eQTLs and in particular additional *trans* regulators. We provide multiple sources of evidence that the associations recovered by PANAMA are indeed likely to be real. Most strikingly on yeast, the findings by PANAMA can be better reproduced between independent studies and are more consistent with prior knowledge about the underlying regulatory network. Finally, we also give insights into the limitations of current methods to account for confounders that help to understand the relationship between confounding variation, *cis* regulation and *trans* effects.

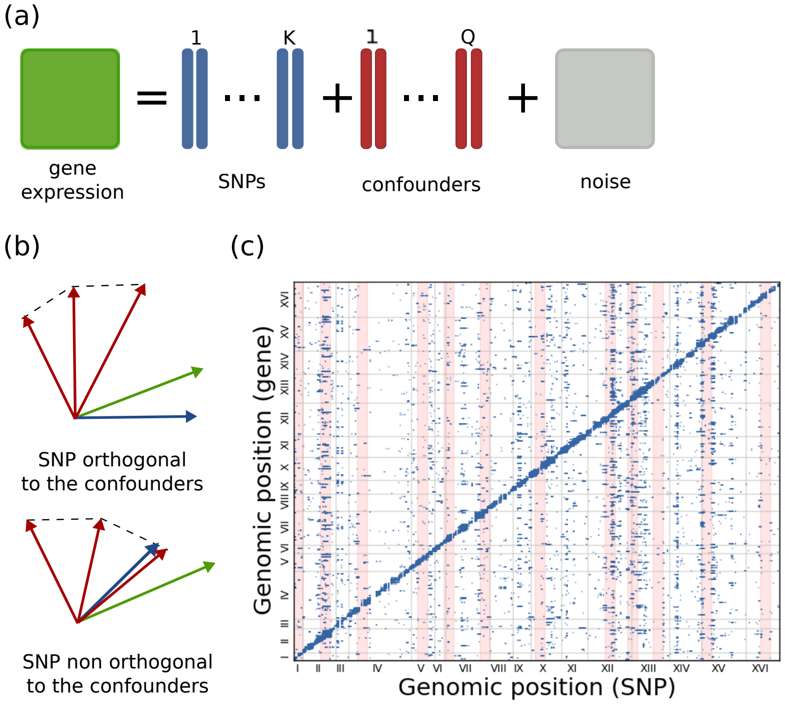


Figure 1. (a) Effects of causal factors on gene expression variation that are accounted for by PANAMA. **(b)** PANAMA applied to the yeast eQTL dataset. Jointly learned *trans* regulators identified by PANAMA are highlighted in red. **(c)** Illustration of the difference between conventional approaches that assume orthogonality of confounding factors and genetic signals (lower figure) and PANAMA, allowing to disentangle causal signals from confounders despite overlaps.

Results

Learning of confounding factors in the presence of *trans* regulators

The statistical model underlying PANAMA assumes additive contributions from true genetic effects and hidden confounding factors. Briefly, this linear model expresses the gene expression of gene g in measured in N individuals as the sum of weighted contributions from a set of K SNPs $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_K\}$ as well as Q confounders $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_Q\}$ and a noise term $\epsilon_{\mathbf{g}}$ (See Figure 1a).

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

Neither the regression weights $w_{g,q}$ nor the profiles of the confounding factors $\mathbf{x}_q$ are known *a priori* and hence need to be learnt from the expression data. The parameters of PANAMA are learnt in the framework of mixed linear models [12]. In this hierarchical model, the regression weights of the hidden factors are marginalized out, yielding a covariance structure in a Gaussian model that captures the covariance structure induced by the confounders. Intuitively, the objective during learning in PANAMA is to find a configuration of the hidden factors such that the empirical correlation structure between samples shared across genes is explained by the state of the hidden factors. In the presence of extensive *trans* regulation this approach leads to over-correction, running the risk of explaining away true genetic association signals. To circumvent this effect, PANAMA also accounts for a subset of all SNPs in the mixed model framework, resulting in a full covariance structure that satisfies an appropriate balance between explaining confounding variation and preserving true genetic signals (See Figure 1b,c). The contribution of these signal SNPs and the state of the hidden factors are estimated in a joint fashion. Furthermore, an appropriate number of hidden factors that are needed is determined automatically during learning. As a result, PANAMA is statistically robust and inference of hidden factors is feasible without setting of any tuning parameters. If existing, additional observed covariates can also be included within the model; see Methods and the text S1 for full details.

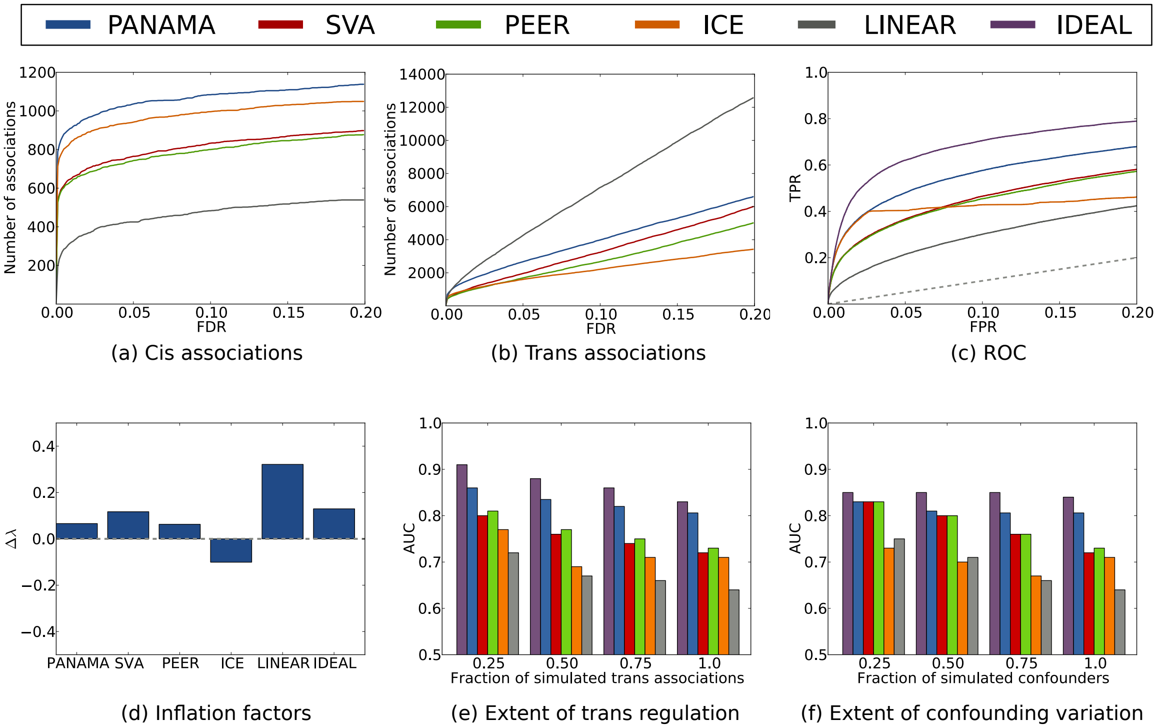
Simulation study

The evaluation of methods to call eQTLs is challenging, as reliable ground truth information is not available. Following previous work [2, 18, 19], we hence used synthetic data to assess and compare PANAMA with alternative approaches to correct for confounding factors. To minimize any assumptions we need to impose on the simulation procedure, we created an artificial dataset that shares key characteristics with a real eQTL dataset from yeast [4]. In the simulation procedure, we first fitted PANAMA to the original yeast eQTL data described below (Application to segregating yeast strains), estimating the number of *cis* and *trans* associations, an empirical distribution of association strengths and the characteristics of confounding variation. Based on these estimates, we then simulated in silico associations from a standard linear model. To ensure that the simulated dataset was not biased towards our method, we also considered a simulation setting based on alternative method to estimate the statistics of empirical associations on the yeast dataset (see below).

Given the synthetic eQTL dataset, we then employed alternative methods to recover the simulated associations. We compared PANAMA to standard linear regression (LINEAR), ignoring the presence of confounders entirely, as well as SVA [6], ICE [7] and PEER [2], established and widely used approaches to correct for hidden confounders. For reference, we also compared to an idealized model with the simulated simulated confounders perfectly removed (IDEAL). First, Figure 2a and 2b show the number of significant *cis* and *trans* associations as a function of the false discovery rate (FDR) cutoff, for each considered method. To avoid inflated association counts due to linkage disequilibrium, we considered at most a single *cis* association per gene and at most one *trans* association per chromosome for each gene. PANAMA found more *cis* associations than any other approach and retrieved the greatest number of *trans* associations among the methods that correct for hidden confounders. Notably, the linear model appeared to find even more *trans* associations. However although statistically significant, the majority of these calls from the linear model were not consistent with the simulated associations, but instead spurious artifacts due to the confounding variation. The extent of false associations called by the linear model is also reflected in Figure 2c, which shows the receiver operating characteristics for each method. All approaches that correct for confounders performed strikingly better than the linear model. Among these, PANAMA was most powerful approach, achieving greater sensitivity than any other approach for a large range of false positive rates (FPR), approaching the performance of an ideal model.

Next, we studied the statistics of obtained p-values, checking for departure from a uniform distribution that either indicates inflation (genomic control $\lambda > 1$) or deflation (genomic control $\lambda < 1$) of the respective methods (Figure 2d). All methods except for ICE showed an inflated p-value distribution. Notably, this also applied to the ideal model where the confounders had been perfectly removed. This observation shows that in settings with sufficiently strong *trans* regulation, inflated statistics are not necessarily due to poor calibration because of confounders, but instead may be caused by an excess of true biological effects themselves. We also checked that calls by the various methods were not overly optimistic and artificially inflated. Indeed, false discovery rates estimates from all methods but the linear model were approximately in line with the empirical rate of errors when taking the ground truth into account, with PANAMA being the best calibrated approach.

We then repeated the same analysis on a broader range of simulated datasets, varying particular



(a) Cis associations

(b) Trans associations

(c) ROC

(d) Inflation factors

(e) Extent of trans regulation

(f) Extent of confounding variation

Figure 2. Accuracy of alternative methods in recovering simulated *cis* or *trans* associations. **(a,b)** number of recovered *cis* and *trans* associations as a function of the false discovery rate cutoff. At most one association per chromosome and gene was counted. **(c)** Receiver Operating Characteristics (ROC) for recovering true simulated associations, showing the true positive rate (TPR) as a function of the permitted false positive rate (FPR), evaluated on the simulated ground truth. **(d)** inflation factors, defined as $\Delta\lambda = \lambda - 1$, indicate either inflated p-value distributions ($\Delta\lambda > 0$) or deflation ($\Delta\lambda < 0$) of the p-value statistics of different methods. **(e)** Area under the ROC curve for alternative methods as a function of the extent of *trans* regulation. **(f)** Area under the ROC curve for alternative methods for varying extent of confounding variation.

aspects of the simulation procedure. Figure 2e shows the accuracy of different methods in recovering true simulated associations when varying the extent of *trans* regulation, reducing the number of simulated *trans* effects to a certain fraction. These results show, that existing methods only work well in the regime of little *trans* regulation, while PANAMA achieves accurate estimates for a wider range of settings, outperforming alternative approaches. Similarly, Figure 2f shows results for strong trans regulation, now varying the extent of confounding variation from weak to strong confounding influences. Again, PANAMA was the most robust approach, recovering true simulated associations with great accuracy irrespectively of the strength of confounders, whereas the performance of other methods degraded quickly towards more difficult settings.

Finally, we considered the impact of the type of model used to fit the association characteristics to the real yeast dataset. As ICE tent to be the most conservative approach among the considered methods, the extent of *trans* regulation on the simulated data was severely reduced. As a result the differences between methods was considerably smaller, however confirming the superior sensitivity and specificity of PANAMA as observed in the primary simulation setting.

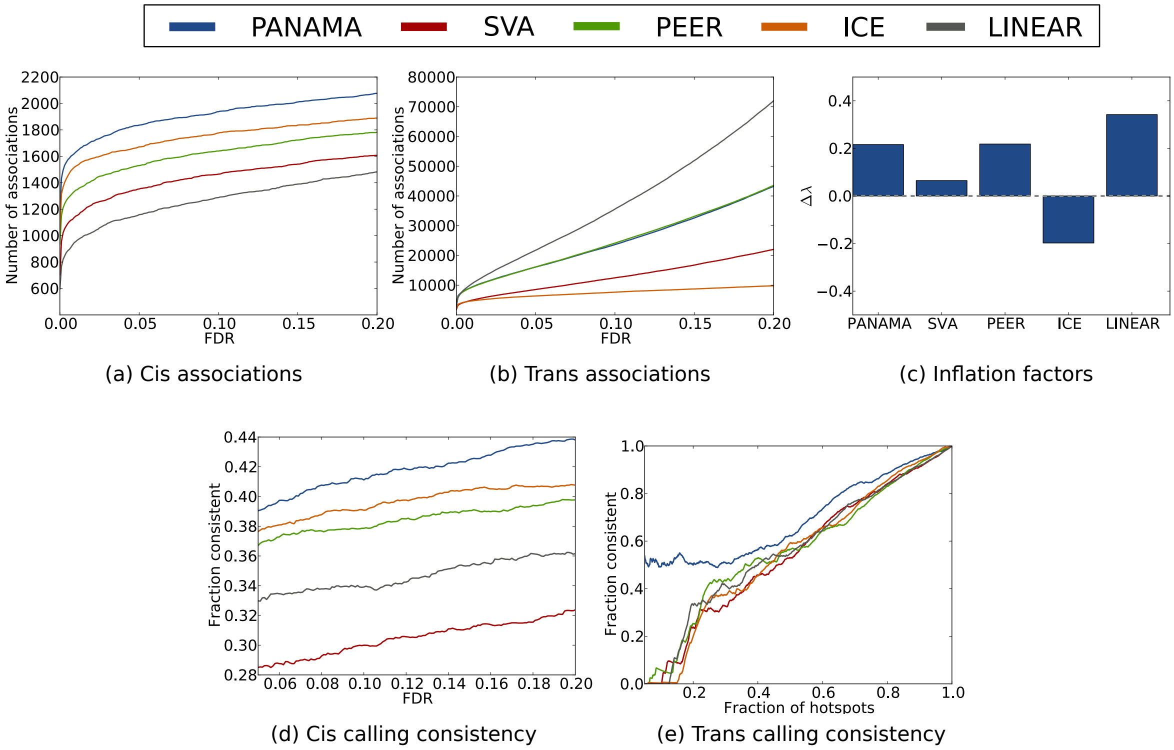
Application to segregating yeast strains

Having established the accuracy of PANAMA in recovering hidden confounders, we applied PANAMA and the alternative methods to the primary eQTL dataset from segregating yeast strains. A total of 108 strains have been expression profiled in two environmental conditions, glucose and ethanol [4]. First, we focused on the glucose condition, which has also previously been examined in [3], providing an independent study for the purpose of comparison.

Figure 3a and 3b show the number of *cis* and *trans* associations for different methods as a function of the FDR cutoff. Again, we considered at most one association per chromosome to avoid confounding the size of associations with their number. In line with previously reported results [2, 7] and the simulated setting (Simulation Study), the standard linear model identified fewer *cis* associations than methods that correct for confounding variation. The trends from the simulated dataset also carried over for *trans* associations, where the linear model called many more associations than methods that account for confounders, yielding an excess of regulatory hotspots. It has previously been suggested, that many of these are likely to be false; see for example the discussion in [7]. Among the methods that correct for confounding variation, PANAMA identified the greatest number of associations. Among the alternative methods, ICE appeared to be more sensitive in recovering *cis* associations while PEER and SVA retrieved a greater number of *trans* associations. Also note that models that account for confounding factors yielded slightly inflated p-value distributions (Figure 3c), supporting that also on real data a certain degree of inflation may be caused by extensive *trans* regulation. This summary of genome-wide eQTLs confirms that ICE is most conservative in detecting hotspots, whereas all other methods do find multiple *trans* bands. For comparison we also included a version of PANAMA that also corrects for the *trans* regulators that are accounted for while learning PANAMA$_{\text{trans}}$yields near-identical results to ICE, which explains the differences and similarities between the two approaches, where PANAMA can be regarded as generalization of ICE. By accounting for pronounced regulators PANAMA circumvents the over-conservative correction of the ICE model.

Reproducibility of eQTLs between studies To objectively shed light on the correctness of the associations called, we considered the consistency of calls between two independent studies. The glucose environment from [4] has previously been studied in [3], sharing a common set of segregants. We checked the consistency in calling genes with a *cis* association for increasing FDR cutoffs (Figure 3d). Alternatively, focusing on the consistency of regulatory hotspots, Figure 3e shows the ranking consistency of polymorphisms ordered by their regulatory potential on multiple genes. Reassuringly, for both *cis* effects and *trans* regulatory hotspots, PANAMA yielded results with far greater consistency than any other currently available method. In particular the consistency of *trans* hotspots suggest that PANAMA achieved an appropriate balance between explaining away spurious signals as confounding variation and identifying hotspots that are likely to hav a true genetic basis.

Consistency of *trans* regulatory hotspots with respect to known regulatory mechanisms in yeast As a second means of validating *trans* eQTLs, we investigated to what extent polymorphisms that regulate multiple genes in *trans* can be interpreted as indirect effects that are mediated by known transcriptional regulators. For this analysis we considered an established regulatory network of transcription factors extracted from Yeastract [20]. Although we do not expect *trans* associations to be exclusively mediated by direct transcriptional regulation, the degree of associations that are consistent with this regulatory structure is nevertheless an informative indicator for the validity of eQTLs calls from different models. For each transcription factor, we considered polymorphisms in the vicinity of the coding region of the TF (± 1.5Mb around the coding region), and tested the fraction of associations with genes that are known targets of the TF versus other associations with genes that are no direct targets. Table S1 shows the F-score (harmonic mean between precision and recall) for each of 129 transcription factors which had



(a) Cis associations

(b) Trans associations

(c) Inflation factors

(d) Cis calling consistency

(e) Trans calling consistency

Figure 3. Evaluation of alternative methods on the eQTL dataset from segregating yeast strains (glucose condition). **(a,b)**: number of *cis* and *trans* associations found by alternative methods as a function of the FDR cutoff. **(c)** Inflation factors of alternative methods, defined as $\Delta\lambda = \lambda - 1$. **(d)** Consistency of calling *cis* associations between two independent glucose yeast eQTL datasets. **(e)** Consistency of calling eQTL hotspots between two independent glucose yeast datasets, where SNPs are ordered by extent of *trans* regulation as determined by $< -\log(pv) >$.

at least one SNP in the local *cis* window. For half of the 129 TFs, PANAMA yielded a higher F-score than any of the other methods considered. Interestingly, the standard linear models performed second best under this metric, achieving the greatest F-score in 36% of all cases, followed by PEER (28%), SVA (15%) and ICE (6%). Among the methods that correct for confounders, PANAMA consistently yielded the highest F-score.

Detecting eQTLs that are shared across environments Finally, we considered the full expression dataset from [4], combining expression measurement in an ethanol and glucose background. Because each yeast strain was profiled twice, the set of samples was not independent, but instead had a replicate population structure. Similarly as done in [15], we added this inherent genetic relatedness to the PANAMA covariance structure (Material and Methods). Because PANAMA accounted for the replicate structure of the dataset, the increase of the number of associations compared to the analysis of the single-condition analysis was modest. Other methods, not accounting for the replicate structure of the genotypes, yielded severely inflated test statistics, identifying a *trans* effect for almost every gene. To check this hypothesis we also applied PANAMA without the correction for artificial genetic relatedness, yielding similarly inflated results (data not shown).

Application to further eQTL studies

We successfully applied PANAMA to additional ongoing and retrospective studies. For example, on a dataset from inbred mouse crosses [21], PANAMA identified a greater number of associations than other methods. In contrast to the yeast dataset, the distribution of p-values on this dataset was almost uniform, suggesting that the extent of true *trans* regulation was lower. We also investigated parts of a dataset of the genetics of human cortical gene expression [22]. On chromosome 17, methods that account for confounders identified more genes in associations than a linear model, with SVA and PANAMA retrieving the greatest number. Results on other 4 other chromosomes were similar (data not shown). Finally, preliminary results to an RNA-Seq eQTL study on *Arabidopsis* indicate that expression heterogeneity as accounted for by PANAMA is also presented on expression estimates from short read technologies. This results shows that statistical challenges due to confounding variation are not specific to a particular platform for measuring gene expression.

Discussion

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

There are several previous approaches to correct for confounding influences in eQTL studies. These methods can be broadly grouped into factor-based models like PCA, SVA and PEER [2], and approaches that employ a mixed linear model [7,15], estimating a covariance structure that captures the confounding variation. An important reason why PANAMA performs well is the intermediate approach taken here, that is, learning a covariance structure within a linear mixed model (LMM), but at the same time retaining the low-rank constraint which yields an explicit representation of factors. Moreover, PANAMA systematically exploits the flexibility provided by the representation in terms of covariance structures, jointly accounting for genetic regulators while estimating the confounding factors. Our approach is stable and robust, avoiding the need to first subtract off the genetic contribution greedily, as for example suggested and implemented in [2,6]. Although this is not the focus of this work, we have shown how our approach can be combined with additional measures to correct for observed sources of confounding variation, such as known covariates or populational relatedness. The utility of such measures has been illustrated in the joint analysis on data from two environmental conditions. A more specialized approach that is aimed at the combined correction for expression confounders and population structure has recently been proposed by [15]. This LMM-EH approach is methodologically related to what is done here, as the contribution from multiple sources of variation are combined within a single covariance structure. Importantly, the main contribution in PANAMA is an integrated model that does not include additional confounders but true genetic regulators. Unique to PANAMA, these regulators are jointly identified and accounted for during learning of the confounding factors. Our analysis shows, that this approach yields a significant improvement in the sensitivity of recovering *trans* associations and plausible regulatory hotspots.

In conclusion, PANAMA is an important step towards exhaustively addressing common types of confounding variation in eQTL studies. The number of datasets that benefit from careful dissection of true genetic signals and confounders, as done here, is expected to rise quickly. Growing sample sizes and expression profiling in more than one environment allow for the estimation of more subtle confounding influences and at the same time provide the statistical power to detect many more *trans* effects than possible as of today. An open source implementation of PANAMA will be made publicly available.

Materials and Methods

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

$$\mathbf{Y} = \mathbf{S}\mathbf{V} + \mathbf{X}\mathbf{W} + \boldsymbol{\epsilon}, \tag{1}$$

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

$$p(\mathbf{W}) = \prod_{q=1}^{Q} \mathcal{N}\left(\mathbf{w}_q \,|\, \mathbf{0}, \alpha_q \mathbf{I}\right),$$

$$p(\mathbf{V}) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{v}_k \,|\, \mathbf{0}, \beta_k \mathbf{I}\right),$$

and integrate them out. The corresponding marginal likelihood, conditioned on the state of the confounding factors $\mathbf{X}$ is now factorized across genes

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

For brevity we have defined $\boldsymbol{\Theta} = \{\{\beta_k\}, \{\alpha_q\}, \sigma^2\}$, the set of all hyperparameters of the model.

Known covariates If available, additional covariates can directly be included in the background covariance structure from Equation (4)

$$p(\mathbf{Y} \,|\, \mathbf{X}, \boldsymbol{\Theta}) = \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{y}_g \,\middle|\, \mathbf{0}, \sum_{k=1}^{K} \beta_k \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}} + \sum_{q=1}^{Q} \alpha_q \mathbf{x}_q \mathbf{x}_q^{\mathrm{T}} + \gamma \mathbf{K}_0 + \sigma_e^2 \mathbf{I}\right), \tag{3}$$

where $\mathbf{K}_0$ denotes the covariance induced by these additional covariates. Examples for possible choices of this covariance include the covariance induced by a fixed covariate vectors, i.e. $\mathbf{K}_0 = \mathbf{c}\mathbf{c}^{\mathrm{T}}$ or a kindship matrix that accounts for the genetic relatedness (see for example [12], [15]).

Model fitting The most probable state of the latent variables $\mathbf{X}$ and the hyperparameters $\boldsymbol{\Theta}$ can be identified via straight-forward maximum likelihood

$$\{\hat{\boldsymbol{\Theta}}, \hat{\mathbf{X}}\} = \underset{\boldsymbol{\Theta}, \mathbf{X}}{\mathrm{argmax}}\, p(\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\Theta}), \tag{4}$$

employing a gradient-based optimizer. In practical applications of PANAMA, this model fitting (Equation (4)) is not carried out with the set of all genome-wide SNPs included in Equation (1), because the number of weight parameters β_k for each SNP would be prohibitive. Only those genetic regulators with strong effects on multiple genes do play a role during the estimation of hidden factors and thus need to be accounted for. Our inference scheme determines the set of relevant regulators in an iterative procedure. The number of hidden factors to be learnt, Q is not set *a prior* and instead Q is set to a sufficiently large value. During the optimization, the individual variance parameters for each factors, α_q^2, automatically determine an appropriate number of effective factors, switching off unused ones.

Significance testing Once the confounding-correcting covariance structure is determined from the maximum likelihood solution of (4), significance testing can be carried out in the framework of mixed linear models. Such testing can be implemented efficiently, for example using a computational trick recently proposed by [12]. The association between a SNP k and gene g to be tested is treated as fixed effect, yielding the following likelihood ratio

$$\mathrm{LOD}_{g,k} = \log \frac{\mathcal{N}\left(\mathbf{y}_g \mid \theta \mathbf{s}_k, \sigma_k^2 \mathbf{K} + \sigma_e^2 \mathbf{I}\right)}{\mathcal{N}\left(\mathbf{y}_g \mid \ \mathbf{0}, \ \ \sigma_k^2 \mathbf{K} + \sigma_e^2 \mathbf{I}\right)}. \tag{5}$$

Here, the covariance matrix $\mathbf{K}$ denotes the covariance structure explaining confounding variation, which is shared across all genes. In PANAMA, this covariance structure excludes the effect of *trans* regulators:

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

In PANAMA$_{\mathrm{trans}}$, also correcting for the *trans* factors, the covariance equals to

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

For computational efficiency we fix the covariance structure $\mathbf{K}$ that is learnt from the full expression dataset. The relative weighting of the covariance (σ_k^2) and the noise term (σ_e^2) is then adjusted on the null model of every single genes similar as done in EMMAX [12]. For larger datasets we propose to perform independent maximum likelihood tests carried out on the residual datasets of the random effect model of the learnt covariance; see Supporting Material .

Yeast datasets. We used the yeast expression dataset from [4] (GEO accession number GSE9376), which consists of 5493 probes measured in 109 segregants derived from a cross between BY and RM. The authors provided the genotypes, which consisted in 2956 genotyped loci. An association was defined as *cis* if the location of the SNP and the location of the opening reading frame (ORF) of the gene were within 100kb, *trans* otherwise. In order to validate the associations found, we also used data from [3] (GEO accession number GSE1990), which consisted in 7084 probes and 2956 genotyped loci in 112 segregants. For the purpose of comparison, we defined *cis* associations in the same way as we did for the previous dataset.

Mouse dataset. We used the data described in [21], consisting of 23,698 expression measurements and 137 genotyped loci for 111 F$_2$ mouse lines.

Human dataset. We used the dataset from [22] (GEO accession number GSE8919), which consists of 14,078 transcripts and 366,140 SNPs genotyped on 193 human samples.

Yeastract. We used data from Yeastract [20], which contains information about the regulatory network between 185 transcription factors and 6298 genes. Out of these 189 transcription factors, we selected the 129 TFs that had a polymorphism in the vicinity (1.5Mb) of the coding region.

Acknowledgments

This research was supported by the FP7 PASCAL II Network of Excellence. NF was supported by PhD scolarships from the University of Sheffield and the University of Manchester. OS was supported by a fellowship from the Volkswagen Foundation.

References

1. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nature genetics 39: 1217–24.

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

3. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science (New York, NY) 296: 752–5.

4. Smith EN, Kruglyak L (2008) Gene-environment interaction in yeast gene expression. PLoS biology 6: e83.

5. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetic 9: 356–369.

6. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS genetics 3: 1724–35.

7. Kang HM, Ye C, Eskin E (2008) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics 180: 1909–25.

8. Churchill G (2002) Fundamentals of experimental design for cDNA microarrays. Nature genetics 32 Suppl: 490–5.

9. Balding D, Bishop M, Cannings (2003) Handbook of Statistical Genetics. N.Y.: Wiley J. and Sons Ltd., second edition.

10. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England) 8: 118–27.

11. Kang H, Zaitlen N, Wade C, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709.

12. Kang H, Sul J, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nature genetics 42: 348–354.

13. Plagnol V, Uz E, Wallace C, Stevens H, Clayton D, et al. (2008) Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. PLoS One 3: 2966.

14. Locke D, Segraves R, Carbone L, Archidiacono N, Albertson D, et al. (2003) Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome research 13: 347.

15. Listgarten J, Kadie C, Schadt E, Heckerman D (2010) Correction for hidden confounders in the genetic analysis of gene expression. Proceedings of the National Academy of Sciences 107: 16465.

16. Nica A, Parts L, Glass D, Nisbet J, Barrett A, et al. (2011) The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. PLoS genetics 7: e1002003.

17. Breitling R, Li Y, Tesson B, Fu J, Wu C, et al. (2008) Genetical genomics: spotlight on QTL hotspots. PLoS Genet 4: e1000232.

18. Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38: 904–909.

19. Yu J, Pressoir G, Briggs W, Bi I, Yamasaki M, et al. (2005) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38: 203–208.

20. Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, et al. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. Nucleic Acids Research 34: D3–D5.

21. Schadt E, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nature genetics 37: 710–717.

22. Myers A, Gibbs J, Webster J, Rohrer K, Zhao A, et al. (2007) A survey of genetic human cortical gene expression. Nature genetics 39: 1494–1499.