Accounting for Non-Genetic Factors Improves the Power of eQTL Studies

Oliver Stegle[1], Anitha Kannan[2], Richard Durbin[3], and John Winn[2]

[1] University of Cambridge, UK
os252@cam.ac.uk

[2] Microsoft Research, Cambridge, UK
{ankannan,jwinn}@microsoft.com

[3] Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
rd@sanger.ac.uk

Abstract. The recent availability of large scale data sets profiling single nucleotide polymorphisms (SNPs) and gene expression across different human populations, has directed much attention towards discovering patterns of genetic variation and their association with gene regulation. The influence of environmental, developmental and other factors on gene expression can obscure such associations. We present a model that explicitly accounts for non-genetic factors so as to improve significantly the power of an expression Quantitative Trait Loci (eQTL) study. Our method also exploits the inherent block structure of haplotype data to further enhance its sensitivity. On data from the HapMap project, we find more than three times as many significant associations than a standard eQTL method.

1 Introduction

Discovering patterns of genetic variation that influence gene regulation has the potential to impact a broad range of biological endeavours, such as improving our understanding of genetic diseases. Recent advances in microarray and genotyping methods have made it feasible to investigate complex multi-gene associations on a genome-wide level, through expression Quantitative Trait Loci (eQTL) studies (see (1) and references therein). The vast number of potential associations and relatively small numbers of individuals in current data sets makes it challenging to discover statistically significant associations between genome and transcript. Methods for improving the sensitivity and accuracy of such studies are therefore of considerable interest.

In this paper, we describe a method to improve substantially the number of significant associations found in an eQTL study. The main insight is that much of the variation in gene expression is due to non-genetic factors, such as differing environmental conditions or developmental stages (2). By explicitly accounting for non-genetic variation, we can greatly improve the statistical power of eQTL methods as most of the non-genetic variation is removed and real associations stand out to a greater extent.
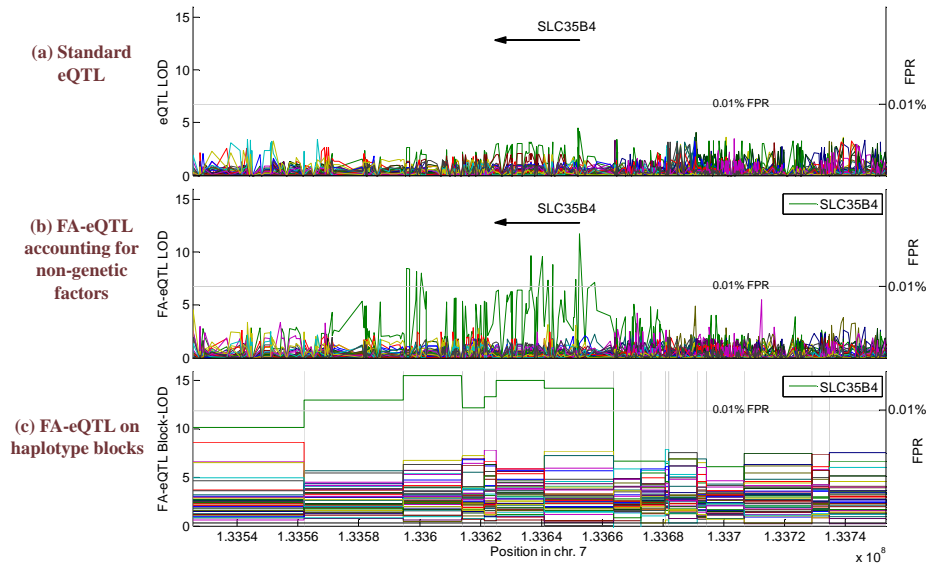
Fig. 1. Example results of **(a)** standard eQTL, **(b)** our proposed method FA-eQTL which accounts for non-genetic factors and **(c)** the same method applied to haplotype blocks. In this region, standard eQTL does not find any significant associations but our proposed methods finds a *cis* association for the gene *SLC35B4*. The significance of the association is improved when haplotype blocks are considered instead of individual SNPs.

Following (3), we also improve the accuracy of eQTL by exploiting the inherent block structure present in haplotype data. By jointly considering all SNPs in a haplotype block, it is possible to detect weaker associations than can be found using single SNPs. For example, if the relevant SNP lies between the measured marker SNPs, a haplotype block model can effectively perform imputation of this missing SNP value leading to a stronger detected association.

The contributions of this paper are best illustrated by the plots of Fig. 1 showing the results of different eQTL methods over the same region of chromosome 7. The top plot demonstrates that no associations have been found using a standard eQTL method, whilst the second plot shows a significant *cis* association which only becomes visible when non-genetic factors are accounted for. The bottom plot shows that, when haplotype blocks are used instead of individual SNPs, the significance of this association is further increased to well above the 0.01% False Positive Rate (FPR) level.

The structure of this paper is as follows. In Section 2, we compare several models of how non-genetic factors influence gene expression. The best of these models is incorporated into an eQTL method in Section 3 and their power demonstrated on data from the HapMap project (4). Section 4 describes how this eQTL approach can be extended to exploit the block structure of haplotype data. Section 5 concludes with a discussion.

2 Modelling Non-genetic Factors

In addition to variation due to genomic differences, human gene expression levels vary because of differing developmental stages, environmental influences and other physiological and biological factors. In principle, when collecting gene expression data sets for eQTL, non-genetic factors should be controlled to be con-
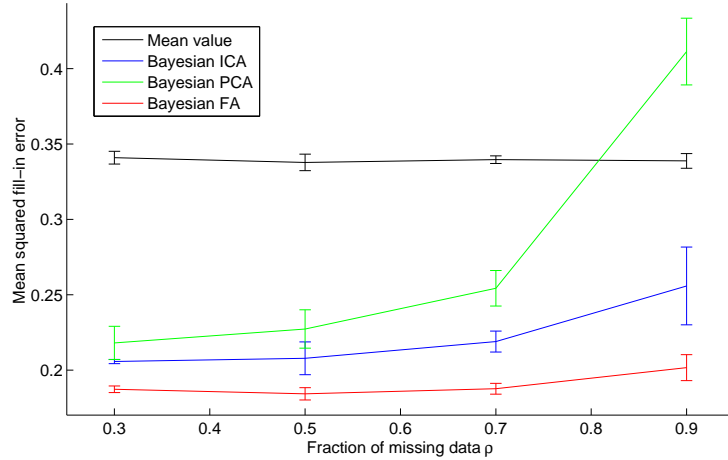
Fig. 2. Comparative performance of various linear Gaussian models for filling-in missing gene expression values for X chromosome genes. The plot shows the mean squared error in the fill-in predictions against the fraction of missing data ρ, averaged over four runs with different training/test splits. Error bars show one standard deviation. The factor analysis model gives the lowest fill-in error over a range of missing data rates.

For each method, we use an Automatic Relevance Determination (ARD) prior on the variance of each column of $\mathbf{W}$, so that the number of latent non-genetic factors is learned automatically (11).

2.1 Investigation on HapMap expression data

We investigated these three models on gene expression measurements of individuals from the HapMap project, consisting of the expression profiles for 47,294 gene probes profiled in EBV-transformed lymphoblastoid cell lines (12). The parameters of each model were learned from the expression levels of 512 X chromosome gene probes from a randomly-selected 75% of the HapMap individuals, with the maximum number of non-genetic factors set to 40 (during learning several of these factors were switched off by ARD). Bayesian learning was achieved with a fully-factorised variational approximation using the VIBES software package (13). For the 25% of individuals not used for training, we removed a fraction ρ of the expression measurements and applied each learned model to fill-in these missing values. The idea behind this experiment is that models which better capture the latent causes of the observed gene expression levels, will better predict missing expression levels from partial observations. The accuracy of the fill-in predictions for each model was assessed in terms of mean squared error. The results are shown in Figure 2, along with a baseline prediction given by the mean expression across the training individuals. These results show that the factor analysis model gives the best fill-in performance, even when large fractions of the data are missing. Hence, we use factor analysis to model non-genetic effects.

3 Accounting for Non-genetic Factors in eQTL

Standard eQTL methods assess how well a particular gene expression level is modelled when genetic factors are taken into account, compared to how well it is modelled by a background model that ignores genetic factors (14). The relevant quantity is the log-odds (LOD) score,

$$\log_{10} \left\{ \prod_j \frac{P(y_{gj} \,|\, \mathbf{s}_j, \theta_g)}{P(y_{gj} \,|\, \theta_{\mathrm{bck}})} \right\} \tag{2}$$

where $\mathbf{s}_j$ is a SNP measurement and y_{gj} the gene expression level of probe g, for the jth individual. The terms θ_g, θ_{bck} are parameters for probe g of the genetic and background models respectively. The LOD scores can then be plotted against the location of the SNP over a large genomic region to give an eQTL scan for each gene expression g.

To account for non-genetic factors, we modify this approach to use the factor analysis model of the previous section, denoting the new method FA-eQTL. In FA-eQTL, the LOD score compares a *full model* of both genetic and non-genetic factors to a *background model* which only includes non-genetic factors,

$$\log_{10} \left\{ \prod_j \frac{P(y_{gj} \,|\, \mathbf{s}_j, \theta_g, \mathbf{w}_g, \mathbf{x}_j)}{P(y_{gj} \,|\, \mathbf{w}_g, \mathbf{x}_j)} \right\} \tag{3}$$

where $\mathbf{x}_j$ are the latent non-genetic causes for the jth individual and $\mathbf{w}_g$ is the gth row of the weight matrix $\mathbf{W}$ described in the previous section. For the full model, which incorporates both genetic and non-genetic factors, we model the expression value of gene g for the jth individual by

$$P(y_{gj} \,|\, \mathbf{s}_j, \theta_g, \mathbf{w}_g, \mathbf{x}_j) = \mathcal{N}\Big(\overbrace{\mathbf{s}_j.\theta_g}^{\text{genetic}} + \overbrace{\mathbf{w}_g\mathbf{x}_j}^{\text{non-genetic}} \,,\, \psi_g \Big), \tag{4}$$

where $\mathcal{N}(m, \tau)$ represents a Gaussian distribution with mean m and variance τ. The variable $\mathbf{s}_j$ encodes the state of a particular SNP whose relevance we want to assess, and θ_g captures the change in gene expression caused by this SNP. The SNP state $\mathbf{s}_j$ is the sum of two indicator vectors encoding the two alleles measured for this SNP. Each indicator vector has a one at the location corresponding to the measured allele and zeroes elsewhere. The noise is Gaussian with learned variance ψ_g. The full model is shown graphically in the Bayesian network of Figure 3.

For the background model, we use exactly the factor analysis model of the previous section. Hence, $P(y_{gj} \,|\, \mathbf{w}_g, \mathbf{x}_j)$ also has the Gaussian form of Eqn. 4, but where the mean consists of only the non-genetic term. For completeness, we also tested ICA and PCA as alternatives to FA but found that these led to weaker associations, showing that the results of the fill-in experiments of Section 2 do indeed indicate how well each model accounts for non-genetic effects.
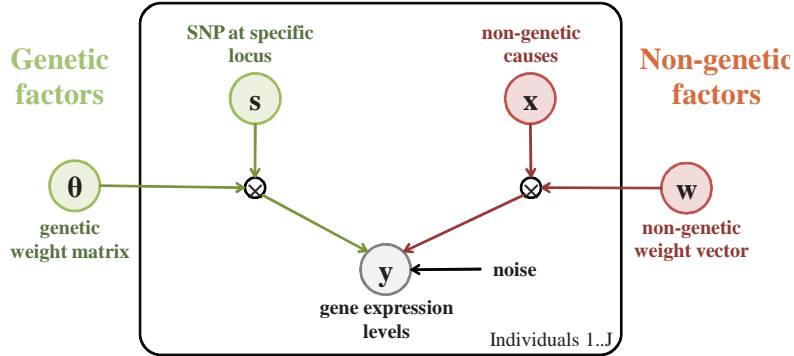
Fig. 3. The Bayesian network for the full model that includes both genetic (green) and non-genetic factors (red) when explaining gene expression levels. The rectangle indicates that contained variables are duplicated for each individual. See the text for a detailed explanation of this model.

When there is extra information about each individual, it can also be incorporated into the full and background models. For example, the HapMap individuals are divided into three distinct populations with African ancestry (YRI), European ancestry (CEU) and Asian ancestry (CHB,JPT). Certain SNP and probe measurements have differing statistics in each population, which can lead to false associations. To avoid such false associations, we introduce an additional three-valued 'virtual' SNP measurement encoding the population each individual belongs to, and extend the sum in the mean of Eqn. 4 to include a linear relation to this measurement. For similar reasons, we also include a binary measurement encoding each individual's gender. If desired, we can investigate the association of a probe to multiple SNPs jointly by extending Eqn. 4 to a sum over all SNPs in a region (or in multiple disjoint regions) as described in (3).

3.1 FA-eQTL on HapMap Data

We applied both the standard eQTL method and the FA-eQTL method to the HapMap Phase II genotype data (4) and corresponding gene expression measurements (12). Both methods were applied to chromosomes 2, 7, 11 and X. For each chromosome, only probes for genes within that chromosome were included, so that only within-chromosome associations were tested.

An issue with using the FA-eQTL model for chromosome-wide scans is the very high computational cost of re-learning the factor analysis model at each locus. To avoid this, we learned each $\mathbf{w}_g$ and $\mathbf{x}_j$ once for the background model and kept them fixed when learning the full model. This approximation is accurate only if the genetic and non-genetic models are nearly orthogonal. To test this assumption, we estimated the contribution to the gene expression levels due to the non-genetic factors alone, given by $\mathbf{Wx}$, and treated it as expression data in standard eQTL. If the genetic and non-genetic models are nearly orthogonal, then we would expect that no significant association would be found between any SNP and these reconstructed expression levels. This was indeed the case,
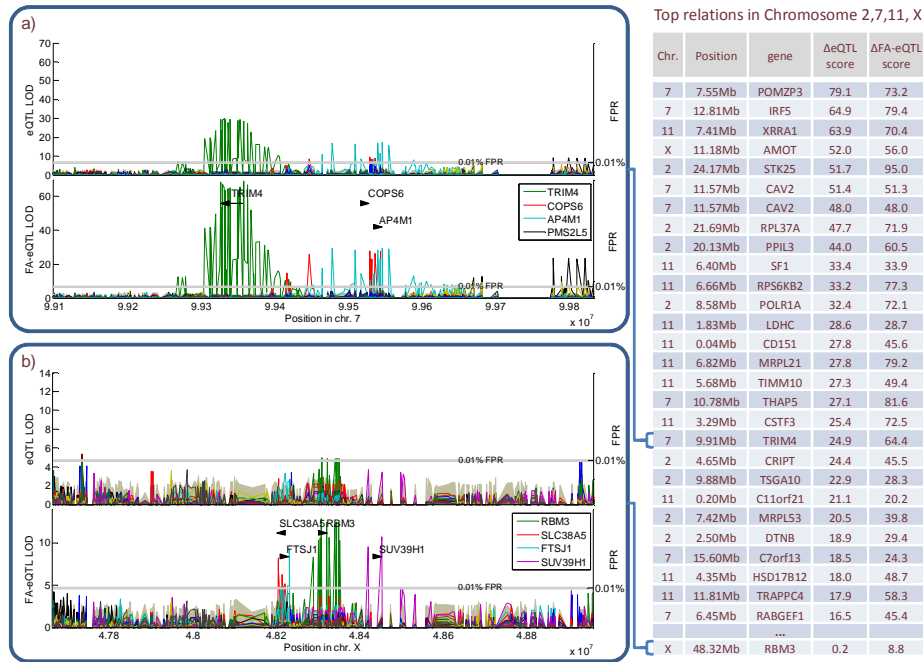
Top relations in Chromosome 2,7,11, X

Chr.	Position	gene	ΔeQTL score	ΔFA-eQTL score
7	7.55Mb	POMZP3	79.1	73.2
7	12.81Mb	IRF5	64.9	79.4
11	7.41Mb	XRRA1	63.9	70.4
X	11.18Mb	AMOT	52.0	56.0
2	24.17Mb	STK25	51.7	95.0
7	11.57Mb	CAV2	51.4	51.3
7	11.57Mb	CAV2	48.0	48.0
2	21.69Mb	RPL37A	47.7	71.9
2	20.13Mb	PPIL3	44.0	60.5
11	6.40Mb	SF1	33.4	33.9
11	6.66Mb	RPS6KB2	33.2	77.3
2	8.58Mb	POLR1A	32.4	72.1
11	1.83Mb	LDHC	28.6	28.7
11	0.04Mb	CD151	27.8	45.6
11	6.82Mb	MRPL21	27.8	79.2
11	5.68Mb	TIMM10	27.3	49.4
7	10.78Mb	THAP5	27.1	81.6
11	3.29Mb	CSTF3	25.4	72.5
7	9.91Mb	TRIM4	24.9	64.4
2	4.65Mb	CRIPT	24.4	45.5
2	9.88Mb	TSGA10	22.9	28.3
11	0.20Mb	C11orf21	21.1	20.2
2	7.42Mb	MRPL53	20.5	39.8
2	2.50Mb	DTNB	18.9	29.4
7	15.60Mb	C7orf13	18.5	24.3
11	4.35Mb	HSD17B12	18.0	48.7
11	11.81Mb	TRAPPC4	17.9	58.3
7	6.45Mb	RABGEF1	16.5	45.4
		...		
X	48.32Mb	RBM3	0.2	8.8

Fig. 4. Examples of the improved association significance for FA-eQTL over standard eQTL. The plots a) and b) give two example regions where FA-eQTL increases the significance of *cis* associations found by standard eQTL and also finds additional *cis* associations. Horizontal lines indicate the threshold value corresponding to 0.01% FPR for relevant genes. Arrows indicate gene coding regions. The table shows associations ranked how much the eQTL score is above the FPR threshold (ΔeQTL) along with the corresponding score above threshold for FA-eQTL (ΔFA-eQTL). For the associations found by both methods, the FA-eQTL score is on average 21.8 higher than the eQTL score, demonstrating the advantage of accounting for non-genetic factors.

for example, the highest LOD score over the entirety of chromosome 2 was just 11.4 which is not statistically significant. Also for computational reasons, we apply maximum likelihood methods to estimate the parameters θ, rather than the variational approach which is Bayesian but much more expensive. Because maximum likelihood methods perform poorly with little data, we remove SNPs where two or fewer individuals have the minor allele.

Fig. 4 shows the results of FA-eQTL and standard eQTL applied to the HapMap data. The table lists associations ranked by the difference between the eQTL score and the 0.01% FPR threshold. For comparison the corresponding FA-eQTL score difference (computed from the highest score within 50 loci of the eQTL peak) is also listed. We consider all associations found in any 100kbp window as a single association, so as not to over count associations due to linkage disequilibrium between SNPs. Across all associations found by eQTL, the FA-eQTL scores are higher in all but five cases, with an average score change of

+21.8. The plot of Fig. 4a gives an example of the improvement gained, where FA-eQTL increases the LOD score of four *cis* associations found by eQTL. The plot of Fig. 4b illustrates that some weaker *cis* associations missed by eQTL are picked up by FA-eQTL, for the genes *SLC38A5*, *FTSJ1* and *SUV39H1*.

To quantify the improvement in power given by the FA-eQTL model, we counted the number of associations found at a 0.01% FPR for each model (Table 1). Using the factor analysis model to explain away non-genetic effects more than doubles the number of significant associations found (from 81 to 222).

	Chr 2	Chr 7	Chr11	Chr X	Total	FDR
eQTL	24	13	39	5	**81**	2%
FA-eQTL	82	44	84	12	**222**	2%

Table 1. Number of associations found at a 0.01% FPR and corresponding FDR

False Positive Rates (FPRs) were estimated empirically for each chromosome using 30,000 permutations across randomly selected regions of length 500 SNPs. For each gene the threshold score was set to give a FPR of 0.01%. False Discovery Rates were calculated from this fixed FPR, the number of conducted tests and the number of associations found for a specific gene. Since almost all of the associations we find are *cis* each gene generally has either exactly one or no

-eQTth3374923)283362(9e0i)D.5((2)1.332758486(la0.4B3T40(60)0.62(a4769.32(s)731(P82)7.5J762(r)-0um

We applied a method for discovering haplotype blocks that explains each individual's haplotype using pieces of a small number of learned 'ancestor' haplotypes, where the pieces break at block boundaries. The particular model we used is the recently proposed piSNP model (15) which accounts for population structure to give increased accuracy when learning the ancestral haplotypes. The learning process discovers the haplotype block boundaries and indicates which ancestral haplotype is used for each block. We define Block FA-eQTL to be the FA-eQTL method previously described but where the measurements $\mathbf{s}_j$ are over ancestors rather than alleles and are measured for each block rather than each locus. Also, as there are fewer blocks than SNPs, it is computationally affordable to enhance the noise model for Block FA-eQTL. To do this, we model v as a mixture of a Gaussian distribution and a uniform distribution. This heavy tailed noise model is more robust to outliers and considerably reduces the false positive rate, at the cost of around a fifty-fold increase in computation time.

4.1 Block FA-eQTL on HapMap Data

Due to the computational expense of Block FA-eQTL, it was only applied to 500 SNP regions within each chromosome, with the number of ancestors set to ten. A region was analysed if it was in the top 100 regions ranked by a soft-max criterion, $S = \frac{1}{N^{1/p}} \left(\sum_{n=1}^{N} (\text{LOD}_n)^p \right)^{1/p}$ with $p = 3$. This criterion identifies both regions with high single-locus peaks and regions where the LOD score is high over an extended area. The selected regions contained all the associations found with FA-eQTL in the previous section. Figure 5 illustrates the benefits of learning associations using haplotype blocks. The plot of Fig. 5a shows the LOD scores for FA-eQTL and Block FA-eQTL for a region containing a strong association. The form of this association is shown in Fig. 5b where the expression level of the *AMOT* gene plotted against the SNP allele and haplotype block at the locations marked with red arrows. The block model is able to pull out population-specific associations which are not apparent in the single SNP plot.

This difference is shown more clearly in the second example of Fig. 5 c,d where blocking causes a *cis* association to move from just above to well above the 0.01% FPR significance threshold. Table 2 shows that for each of the four chromosomes tested we identify more significant associations using haplotype blocks compared to the single SNP approach. Out of the total of 274 associations at the 0.01% FPR level, only five are *trans*, all at sufficiently low significance levels to suggest that they are false associations.

	Chr 2	Chr 7	Chr11	Chr X	Total	FDR
FA-eQTL	82	44	84	12	**222**	2%
Block FA-eQTL	117	57	86	14	**274**	2%

Table 2. Number of associations found at a 0.01% FPR and corresponding FDR

Transcription factor study We selected a subset of 960 gene probes listed as transcription factors in the BDB database ver 2.0 (16). FA-eQTL was then applied genome-wide to search for *trans* associations to these probes. Only one
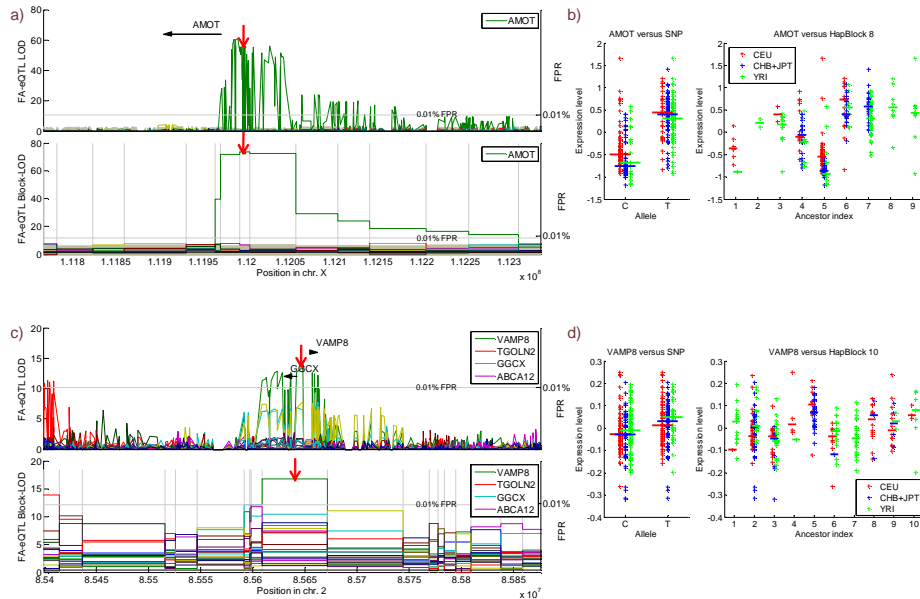
Fig. 5. Left: FA-eQTL scores for SNPs and haplotype blocks, showing the 0.01% FPR threshold for the gene corresponding to the strongest association. Blocking leads to improved association significance. **Right:** Scatter plots of mRNA levels against the SNP allele and block ancestor at the locations marked with red arrows in the left plots.

significant *trans* association was found: between *DPF2* and a region in chromosome 12 (54.5Mb). Testing for associations of all expression levels to this genomic region, we identified an additional *cis* association with *RPS26* and a second *trans* association with *FLT1*. The expression profiles of *RPS26* and *FLT1* show strong correlation whilst *RPS26* and *DPF2* are strongly anti-correlated. A plausible biological explanation is that the ribosomal protein *RBS26* is mediating the expression of both *FLT1* (vascular endothelial growth factor) and *DPF2* (apoptosis response zinc finger protein).

5   Conclusion

In (17), Sen and Churchill described two effects that act to obscure QTL associations, "First is the environmental variation inherent in most quantitative phenotypes. Second is the incomplete nature of the genotype information, which can only be observed at the typed markers". By explicitly modelling non-genetic variation and by using a haplotype block model, we have accounted for both of these effects. Our results on HapMap data demonstrate that countering these effects leads to a more than threefold increase in the number of significant associations found. Given this performance of Block FA-eQTL, we now plan to scale it so that it can be applied exhaustively across all probes and SNPs.

Bibliography

[1] Kendziorski, C.M., Chen, M., Yuan, M., Lan, H., Attie, A.D.: Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics **62**(1) (March 2006) 19–27

[2] Brem, R.B., Kruglyak, L.: The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. **102**(5) (2005) 1572–7

[3] Huang, J., Kannan, A., Winn, J.: Bayesian association of haplotypes and non-genetic factors to regulatory and phenotypic variation in human populations. Bioinformatics **23**(13) (2007) i212–221

[4] The International HapMap Consortium: A haplotype map of the human genome. Nature **437** (2005) 1299–1320

[5] Roweis, S.T., Ghahramani, Z.: A unifying review of linear Gaussian models. Neural Computation **11**(2) (1999) 305–345

[6] Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B **21**(3) (1999) 611–622

[7] Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. Bioinformatics **18**(1) (2002) 51–60

[8] Iosifina, P., Lorenz, W.: Factor analysis for gene regulatory networks and transcription factor activity profiles. BMC Bioinformatics

[9] Lan, H., Stoehr, J.P., Nadler, S.T., Schueler, K., Yandell, B., Attie, A.D.: Dimension reduction for mapping mRNA abundance as quantitative traits. Genetics **121** (2003) 1607–1614

[10] Hastie, T., Tibshirani, R., Eisen, A., Levy, R., Staudt, L., Chan, D., Brown, P.: Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology (2000)

[11] Bishop, C.M.: Bayesian PCA. Advances in Neural Information Processing Systems **11** (1999) 382–388

[12] Stranger, B., Forrest, M., Dunning, M., Ingle, C., Beazley, C., et al.: Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science **315** (2007) 848–853

[13] Bishop, C.M., Winn, J., Spiegelhalter, D.: VIBES: A variational inference engine for Bayesian networks. In: Advances in Neural Information Processing Systems. Volume 15. (2002) 793–800

[14] Lander, E., Botstein, D.: Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. Genetics **121**(1) (1989) 185–199

[15] Kamisetty, H., Kannan, A., Winn, J.: A Bayesian model for population-stratified haplotype block inference
http://research.microsoft.com/mlp/bio/piSNP.html

[16] Kummerfeld, S., Teichmann, S.: DBD: a transcription factor prediction database. Nucleic Acids Res **34**(Database issue) (2006) D74–81

[17] Sen, S., Churchill, G.A.: A statistical framework for quantitative trait mapping. Genetics **159** (September 2001) 371–387