

Bayesian association of haplotypes and non-genetic factors to regulatory and phenotypic variation in human populations

Jim C. Huang^a, Anitha Kannan^b and John Winn^b

^a Probabilistic and Statistical Inference Group, University of Toronto
Toronto, ON, M5S 3G4, Canada

^b Microsoft Research Cambridge
Cambridge, CB3 0FB, UK

ABSTRACT

Motivation: With the recent availability of large scale data sets profiling single nucleotide polymorphisms (SNPs) and quantitative traits data across different human sub-populations, there has been much attention directed towards discovering patterns of genetic variation and their connection to gene regulation and the onset/progression of disease. While previous work has focused primarily on correlating individual SNP markers with gene expression and disease, it has been suggested that using haplotype blocks instead of individual markers can significantly increase statistical power.

Results: We present BlockMapper, a probabilistic generative model for genotype data and quantitative traits data, such as gene expression or phenotype measurements. BlockMapper discovers the block structure of genotype data and associates these inferred blocks to patterns of variation in quantitative traits data, whilst accounting for non-genetic factors. Our model achieves high accuracy for predicting Crohn's disease phenotype in Chromosome 5q31 and reveals novel cis-associations between two haplotype blocks in the ENM006 genomic region and GDI1, a gene implicated in X-linked mental retardation. Our results underscore the importance of accounting for the influence of large sets of SNPs on patterns of regulatory/phenotypic variation and represent a step towards an understanding of human genetic variation.

Contact: jwinn@microsoft.com

Keywords: Bayesian inference, haplotype block, gene expression, phenotypic data, association study

1 INTRODUCTION

An important question in molecular biology and medicine is how patterns of genetic variation influence gene regulation and phenotype in humans. Previous studies for understanding the influence of genetic variation on gene regulation and phenotypic traits involved searching for genetic variants in unrelated individuals and mapping these to patterns of gene expression (Stranger *et al.*, 2007) or phenotypic traits (Cheung *et al.*, 2005; Morley *et al.*, 2004; Wang *et al.*, 2005). These studies are based on one of two approaches. One approach makes use of linkage analysis to find candidate genomic regions which vary between pedigrees, and then perform linkage disequilibrium (LD) mapping to search for genetic markers in these putative disease-correlated regions found by linkage analysis (Cheung *et al.*, 2005; Morley *et al.*, 2004).

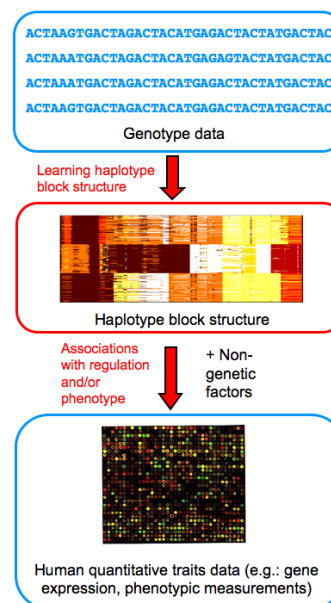


Fig. 1. The joint BlockMapper model for haplotype and quantitative traits data. We first partition genotype data into a set of haplotype blocks and then assign labels to each haplotype block to summarize genetic variation within the block. These labels, along with non-genetic factors such as gender and sub-population, are then related to a set of quantitative traits measurements, such as patterns of gene expression or phenotypic measurements. The result is a set of associations between haplotype and quantitative traits which allows the prediction of new measurements given genotype data.

The problem with this approach is that it may miss weaker associations due to lack of sufficiently large phenotypic differences between individuals in the study (Ardlie *et al.*, 2002). As a way to overcome this, the second approach genotypes a dense set of SNPs across multiple individuals, and then make use of the allele frequencies in these individuals to correlate for disease (Wang *et al.*, 2005). In this genome-wide association approach, high-throughput technologies for profiling quantitative traits (such as microarrays) allow for the detection of more subtle interactions, but at the cost of having to simultaneously test an extremely large number of statistical hypotheses and maintain both high sensitivity and specificity

(Stranger *et al.*, 2007). Instead of testing against individual SNPs, neighboring SNP markers can be grouped together into haplotype blocks (Botstein and Risch, 2003) which can then be correlated with disease. Using haplotype blocks in this way can drastically reduce the number of hypotheses to be tested, whilst providing resistance to genotyping errors. This approach is facilitated by the fact that neighboring SNPs are often in linkage disequilibrium with one another (Gabriel *et al.*, 2002), such that they tend to be inherited jointly in haplotype blocks with relatively little haplotype diversity within a block (Daly *et al.*, 2001). Such blocks can arise due to recombination hotspots in the genome, population bottlenecks or genetic drift, all of which act to preserve low haplotype diversity in distinct regions of the genome.

In this paper, we present a joint probabilistic model for genotype and quantitative traits data which enables associating patterns of genetic variation with patterns of gene expression and other phenotype measurements. We call the model BlockMapper since it learns the relationships between haplotype blocks and the patterns of gene expression and other phenotype measurements. The BlockMapper model has two parts: the first part infers reliable haplotype blocks from the genotype data which are used in the second part to infer relationships to gene expression and other phenotype measurements. While both parts of BlockMapper model can be learned in tandem, for clarity, we show in this paper how to learn them sequentially. A flowchart summarizing the joint BlockMapper model is shown in Fig. 1: the first component accounts for the haplotype block structure of genotype data and allows us to simultaneously infer recombination hotspots and mutations and the phase at each locus in the genotype. The second component then associates the discovered haplotype blocks to a set of quantitative trait measurements across the same set of genotyped individuals. We will show that the BlockMapper model both provides an accurate model for variability in genotype data and infers biologically significant associations between haplotype blocks and patterns of gene expression and disease in both the Chr5q31 and ENm006 genomic regions.

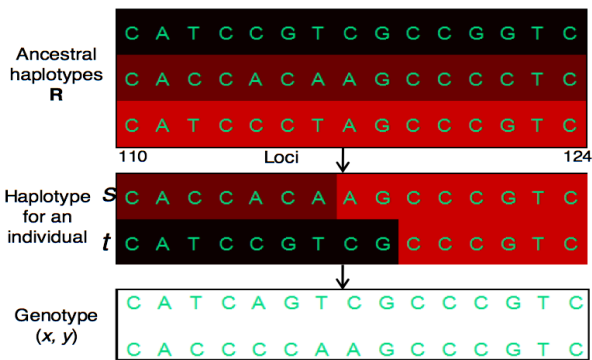


Fig. 2. The generative process used to model the genotype data of each individual (shown for a subset of markers). For each SNP marker, a pair of ancestral indices s and t are generated: alleles are then sampled from the shared ancestral haplotypes \mathbf{R} for ancestors s and t to generate maternal and paternal haplotypes. Genotypes are obtained by swapping the s and t indices with equal probability at each locus, to capture the fact that the phase is unknown.

2 A MODEL FOR HAPLOTYPE BLOCK DISCOVERY IN GENOTYPE DATA

The first component of the BlockMapper model allows us to discover sets of haplotype blocks consisting of regions of low recombination. The goal here is not to develop a completely novel model for genotype data, but rather to extend the existing model of (Jojic *et al.*, 2004) to give improved reliability for learning the underlying haplotype block structure of the genotype data. The generative process described by the Jojic *et al.* model for a given individual is illustrated in Fig. 2 and is as follows:

1. Sample a class c (e.g.: a subpopulation) according to $P(c)$
2. Sample a pair of ancestors (s_1, t_1) independently according to initial distributions $\theta_0^c(s_1), \theta_0^c(t_1)$ over the ancestors
3. Given the pair of ancestors $(s_k = i, t_k = n)$, sample a pair of alleles (x_k, y_k) from the ancestral haplotypes \mathbf{R} with probabilities $R_k^i(x_k), R_k^n(y_k)$.
4. To account for two possible orderings of a pair of alleles, select the phase $m_k \in \{0, 1\}$ with equal probability. If $m_k = 1$, swap the values of (x_k, y_k) , otherwise leave them unchanged.
5. Conditioned on the current pair of ancestral states (s_k, t_k) , transition to the next pair of ancestor states (s_{k+1}, t_{k+1}) with transition probabilities $\theta_k^c(s_k, s_{k+1}), \theta_k^c(t_k, t_{k+1})$, increment the locus k and return to 3.

We repeat the above process for all individuals $j = 1, \dots, J$. The graphical representation of the above process is shown in Fig. 3 as a Bayesian network, consisting of a set of nodes encoding random variables in our model and directed edges indicating dependencies between variables.

When we apply the model, we assume the genotype data $x_1 \dots x_N, y_1 \dots y_N$ was generated by this stochastic process and

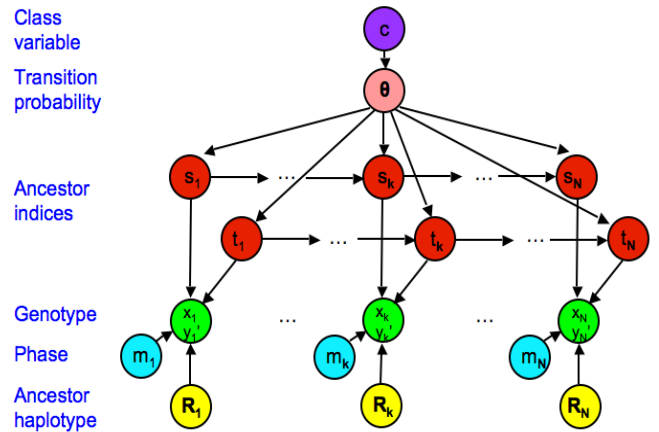


Fig. 3. Bayesian network for the model of genotype data. Nodes correspond to observed and unobserved variables as well as model parameters, with directed edges between nodes representing conditional dependencies encoded by our probability model. For each individual j and at each genomic locus k , we first sample a class, then a pair of ancestors (s_k^j, t_k^j) conditioned on the ancestors (s_{k-1}^j, t_{k-1}^j) at the previous locus, then generate a genotype (x_k^j, y_k^j) from the ancestral haplotypes. The phase m_k^j at locus k determines which genotype measurement is associated with the paternal haplotype and which with the maternal.

aim to infer the values of all the unobserved variables and parameters. Hence, the model allows us both to perform haplotype phasing by inferring the phase variables $\{m_k\}$, and to partition the resulting haplotypes into blocks of SNP markers by inferring the sequences of ancestral indices $s_1 \dots s_N$ and $t_1 \dots t_N$ for each individual. These blocks are extracted from a learned library of A ‘‘ancestral’’ haplotypes $\mathbf{R} = \{R^1 \dots R^A\}$. The model also accounts for the presence of different classes of individuals c , such as those arising from different sub-populations (Gabriel *et al.*, 2002; HapMap Consortium, 2005; Pritchard *et al.*, 2000). Different classes are assigned different initial and transition probabilities $\Theta^c = \{\Theta_1^c \dots \Theta_N^c\}$. We frequently assume that all individuals come from a single class, in which case the c superscript is dropped.

We make a number of key extensions to the model of Jojic *et al.*:

- **‘Soft’ ancestral haplotypes** Previously, Jojic *et al.* used a library of ancestral haplotypes consisting of single alleles, where any difference between these ancestral haplotypes and the observed individual haplotypes was accounted for by a single noise parameter shared across all loci. We modify the model to define ‘soft’ ancestral haplotypes where each locus contains not a single allele, but a probability distribution over alleles corresponding to emission probabilities for each possible allele. This allows for highly localized variability as well as genotyping errors in the data to be accounted for in the learned ancestral haplotype.
- **Parent-child phase constraints** The above model assumes that the genotypes for individuals are generated i.i.d. from one another: in practice however, we are often given genotypes for parent-parent-child trios which provides us with constraints on some of the phase variables m_k for the children. In particular, we can use this parent-child information to infer the phase variables at heterozygous loci, where the phase can often be unambiguously resolved for a child’s genotype given the genotypes of the parents. We can incorporate this parent-child information at these loci into our model by fixing the phase variables at loci where the phase can be unambiguously resolved.
- **Dirichlet prior on emission probability** To allow for the possibility of alleles being observed at particular loci which are not present in the training set, we define a Dirichlet prior over the ancestral allele distribution. This allows a small prior probability for each possible allele, where the probability is given by counts of pseudo-observations (pseudo-counts) β . The quantity β is set to be equal for all alleles and loci.
- **Recombination prior** In a similar fashion, we define a Dirichlet prior over the transition probabilities Θ . However, in this case, we define a different prior probabilities for staying in the same ancestral haplotype and switching between ancestral haplotypes (recombination). The pseudo-counts for switching are set to α , and those for remaining in the same pattern are set to $\alpha\gamma$. Hence, γ acts as a *recombination prior* so that higher values of γ favor staying in the same ancestral state and lower values favor transitions to other ancestral states.

2.1 Variational inference and learning

To perform inference and parameter estimation in the generative model, we must compute the posterior distributions over unobserved variables (namely the class variables $\{c^j\}$, the ancestral

indices $\{s_k^j\}, \{t_k^j\}$ and the phase variables $\{m_k^j\}$), as well as estimate the transition probabilities Θ and the ancestral haplotypes \mathbf{R} . Computing the exact posterior over these unobserved variables is intractable: to address this, we follow the learning procedure adopted by Jojic *et al.* (2004) and resort to a tractable variational approximation to the required posterior distributions (Jordan *et al.*, 1999). This variational approximation allows us to infer the chains of ancestral indices $\{s_k\}$ and $\{t_k\}$ separately from one another via the forward-backward algorithm (Rabiner *et al.*, 1989). The transition probabilities Θ and emission probabilities \mathbf{R} are modified to account for the phase variable at each locus, as well as the Dirichlet prior distributions on these parameters, so that

$$\begin{aligned} R_k^i(z) &\propto \sum_j [x_k^j = z]Q(s_k^j = i) + [y_k^j = z]Q(t_k^j = i) + \beta \\ \theta_k(i, n) &\propto \sum_j Q(s_k^j = i, s_{k+1}^j = n) + Q(t_k^j = i, t_{k+1}^j = n) \\ &\quad + \gamma^{[i=n]} \alpha \end{aligned}$$

where Q refers to the approximate posterior distributions computed via the variational method and z refers to a specific observed allele. Thus, $R_k^i(z)$ is updated according to the frequency with which one emits allele z given state i at locus k , and $\theta_k(i, n)$ is updated according to the frequency with which one transitions from state i to state n . Both variational inference and parameter estimation can be accomplished here in a fashion analogous to the standard parameter updates in the Baum-Welch algorithm for learning HMMs (Rabiner *et al.*, 1989), with the basic parameter updates for our model given by Jojic *et al.* (2004). By iteratively updating approximate posterior distributions over phase, ancestral indices and the model parameters via an EM algorithm (Dempster *et al.*, 1977; Neal and Hinton, 1998), we are guaranteed to increase a lower bound on the log-likelihood of the data with each update step and thus improve the fit of our model to the data. Note that our inference procedure can be naturally modified to account for missing genotype data by summing over all possible alleles for a given missing marker.

For a given locus, we could re-arrange the ancestral indices, and the corresponding distributions indexed by these, without affecting the probability distribution defined by the model. Thus, as a post-processing step, we permute ancestral indices at each locus so that the most probable transition is to the same ancestral index. This permutation leads to longer contiguous stretches in which no transitions between ancestors occur. Fig. 4a shows an example of this learning process.

2.2 Discovering haplotype blocks

Our model discovers haplotype blocks implicitly as regions where the probability of transitioning from one ancestral haplotype to another is small. This *recombination probability* can be computed at between any locus k and $k + 1$ as

$$P(s_{k+1} \neq s_k) = \sum_{s_k, s_{k+1} \neq s_k} \pi_k^{s_k} \theta_k(s_k, s_{k+1}) \quad (1)$$

where $\pi_k^{s_k} = \sum_{s_{k-1}} \pi_{k-1}^{s_{k-1}} \theta_{k-1}(s_{k-1}, s_k)$ is the marginal probability under our model of being in state s_k at locus k . Given this recombination probability, we define a haplotype block as a region where the recombination probability is below 5%. Any point

(a) Section of individual unphased genotypes. Colors indicate learned mapping into the ancestral pattern library.

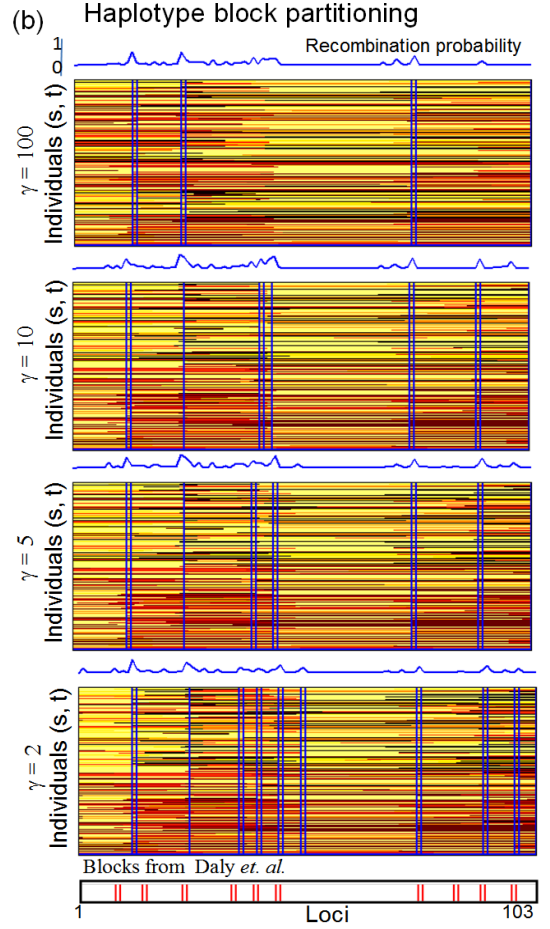
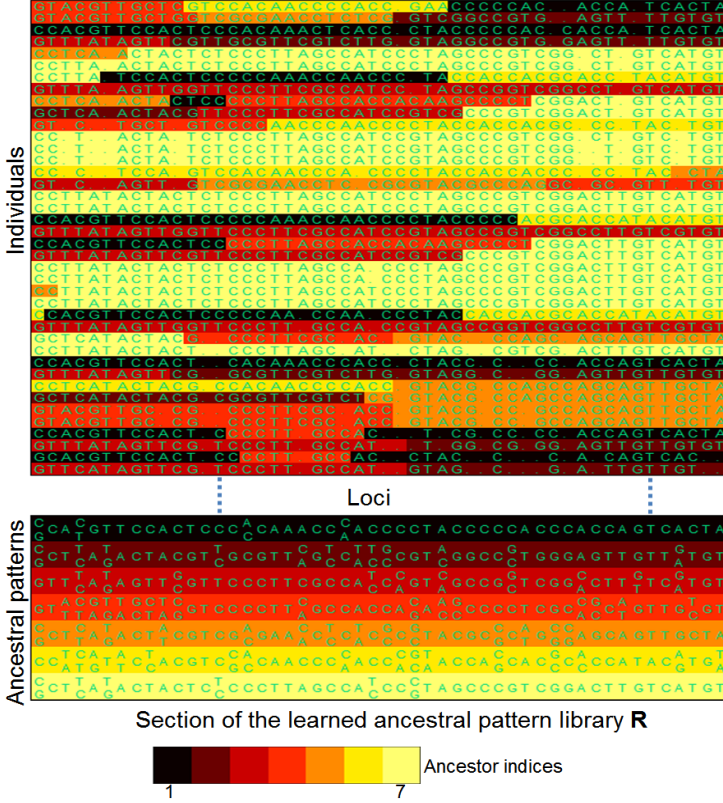


Fig. 4. a) Individual haplotypes spanning SNP markers 15 to 64 in the Chr5q31 data, along with the mappings to the ancestral haplotypes: each color indicates the ancestral haplotype from which the observed allele is drawn; b) Haplotype block partitioning of the Daly 5q31 data. Each block shows the mapping to the ancestral haplotype library for all 387 genotyped individuals. Colorings within a block for a given individual indicate the most likely ancestral index associated with the block for that individual. The blue graphs above each block show how the probability of recombination varies across the 103 markers. Vertical blue lines indicate recombination hotspots: loci where the probability of recombination exceeds 0.05. A haplotype block is defined as the region between two consecutive recombination hotspots. As the recombination prior parameter γ increases, recombination is made less likely, leading to fewer and longer haplotype blocks. Experiments in the rest of this paper use $\gamma = 2$ which allows for many, short haplotype blocks. For comparison, the haplotype blocks discovered in Daly *et al.*, 2001 are shown in red.

where the probability of recombination exceeds this value is therefore considered to be a boundary between two blocks and is likely to represent a recombination hotspot. Note that the recombination probability is affected by the recombination prior γ discussed earlier. If γ is large, recombination is discouraged, favoring a smaller number of longer haplotype blocks: similarly, if γ is small, the haplotype is divided into a larger number of shorter blocks, as shown in Fig. 4b.

For any individual, we can compactly approximate their haplotype by requiring that only one pair of ancestral haplotypes is used within each haplotype block: so s and t are forced to be constant within a block for that individual. We assign the label $(s, t) = (S_b, T_b)$ to that block as a summary of the genetic variation for that block. As we know that few transitions between ancestral haplotypes occur within a block, this will be a reasonably accurate approximation to the underlying mappings to ancestral haplotypes. For any block b , we can infer the distribution over the ancestral haplotypes used and use these as labels for the given block. Distributions over the block labels S_b and T_b for the maternal and

paternal chromosomes can then be computed from the posterior distributions $Q(s_1 \dots s_N)$ and $Q(t_1 \dots t_N)$ so that $Q(S_b = l) \propto Q(\{s_k = l\}_{k \in b})$ and $Q(T_b = l') \propto Q(\{t_k = l'\}_{k \in b})$, where k ranges over loci within haplotype block b . Thus, our model for genotype data allows us to partition the data into haplotype blocks and determine posterior distributions over the block labels for each individual. These block labels will be used in Section 4 to predict the value of quantitative traits measurements for that individual, such as levels of gene expression or phenotypic measurements.

3 COMPARATIVE RESULTS ON LEARNING HAPLOTYPE BLOCK STRUCTURE

3.1 Learning the block structure of Chromosome 5q31

We applied our model to data from Chromosome 5q31 (Daly *et al.*, 2001), consisting of genotypes of 129 parent-parent-children trios profiled across 103 genomic loci. All 129 children in the data set are patients with Crohn's disease. The model was learned using $A = 7$ ancestral haplotypes and model parameters of $\alpha = 0.001$,

$\beta = 0.01$ for all loci k , ancestral indices s_k and alleles. The probability of recombination at each locus was then computed using Eqn. 1. We repeated this for recombination priors $\gamma = 2, 5, 10, 100$ and threshold the probability of transition at a constant value of 0.05 in each case. The genotypes for a subset of the training individuals and the set of learned ancestral haplotypes are shown in Fig. 4a and haplotype block boundaries are shown in Fig. 4b, along with the haplotype block boundaries reported by Daly *et al.* shown in red. Here, increasing γ effectively smoothes over the probability of recombination, leading to a smaller number of longer haplotype blocks. We found that $\gamma = 2$ gave us the best predictive error on independent test data: this setting of γ also led to a partitioning which gave good agreement with the partitioning reported by Daly *et al.* (Fig. 4b).

3.2 Comparison with original Jojic *et al.* model

We now compare the following three models:

- our model excluding parent-child information,
- our model including parent-child information, so that the child phase is known at many loci,
- the model from Jojic *et al.* (2004) where ‘hard’ ancestral haplotypes are learned and parent-child information is not used.

We wish to compare how well these models represent the true probability density over haplotypes. To achieve this, we tested each model’s ability to predict missing values in the haplotype data. The idea is that, if a model is correctly capturing the correlations between adjacent SNP measurements, then it will be better at predicting SNPs which are not observed. Whenever our data is fully observed, we can artificially remove observations at random locations and then test the ability of each model to fill-in the gaps correctly.

We applied this comparison method to the three models using the Chr5q31 data described above. Training/test sets were constructed from all individuals who were Crohns’-positive so that the training set contained 115 individuals (100 children and 15 parents) and the test set consisted of 29 children. We trained each model three times with different random initializations and in each case selected the one that achieved best bound on the training data likelihood.

For the test data, we artificially removed a fraction ρ of the SNP measurements and used each learned model to fill-in these missing values. Fill-in was achieved by sampling from the posterior distribution over the missing alleles. Fig. 5 shows the average prediction error rates across 10 training/test splits for each model, as ρ varies from 0 to 0.5. Compared to (Jojic *et al.*, 2004), our model provides a significant improvement in accuracy on this fill-in task, indicating that it more accurately captures the true distribution of haplotypes. Making use of parent-child information leads to a further improvement in performance, so that our model is making a quarter of the number of fill-in errors compared to the Jojic model.

3.3 Comparison with the HaploBlock and FastPHASE haplotype inference algorithms

We also used this fill-in task to compare the predictive accuracy of our model with other haplotype inference algorithms: HaploBlock (Greenspan and Geiger, 2004) and fastPHASE (Scheet and Stephens, 2006). Each algorithm was used to infer missing SNP values, exactly as described in the previous section. To make the comparison fair, we used the version of our model which does not make use of the known phase information for the children’s genotypes. The

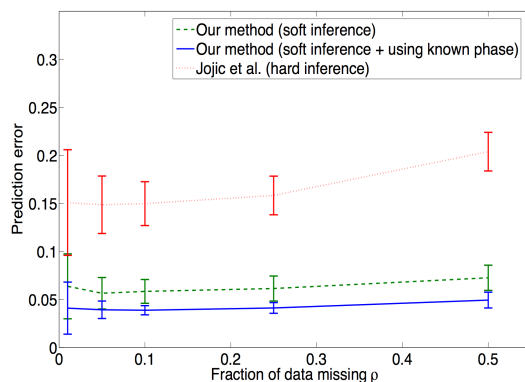


Fig. 5. Mean and standard deviations of prediction error on independent test genotype data as a function of the fraction of data ρ which is missing. The prediction errors shown are obtained from the method of Jojic *et al.* (red), from our method which accounts for variability in the ancestral haplotypes (green) and combined with making use of parent-child information (blue).

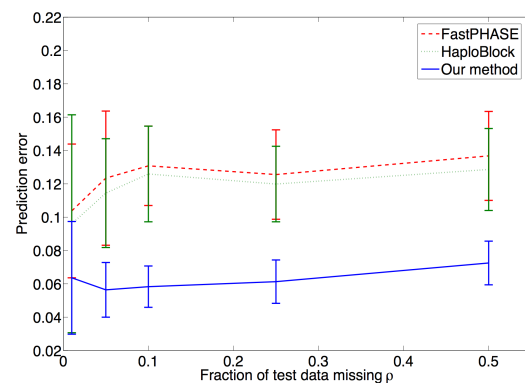


Fig. 6. Mean and standard deviations of test prediction error as a function of the fraction of test genotype locations made missing for our model (blue), fastPHASE (red dotted) and HaploBlock (green dashed).

corresponding predictive accuracy plots are shown in Fig. 6: as can be seen, our model (blue) outperforms both HaploBlock (green dashed) and fastPHASE (red dotted) by a significant margin, where we make about half the number of errors. This result, in tandem with the improved predictive accuracy when using parent-child relationships, indicates that we can more accurately model genotype data by accounting for the underlying haplotype block structure and through the improvements presented above.

3.4 Learning the block structure of the ENm006 region

We applied our model to genotype data consisting of 573 SNPs from the ENm006 region (located on X chromosome) common to the 270 individuals of the HapMap Phase II data (HapMap Consortium, 2005). Briefly, the 270 individuals consist of 30 parent-parent-child trios of Yoruba individuals from Ibadan, Nigeria (YRI), 30 trios of Utah individuals with European ancestry (CEU), 45 Han Chinese individuals from Beijing, China (CHB) and 45 Japanese individuals from Tokyo, Japan (JPT).

Using the same parameters for α, ρ, β and γ as above, we learned our model from this data and partitioned the data into 19 haplotype blocks (Fig. 7). As can be seen, individuals in the three populations (CEU, YRI and CHB+JPT) have noticeably different usage of the ancestral haplotypes. We computed the average block lengths

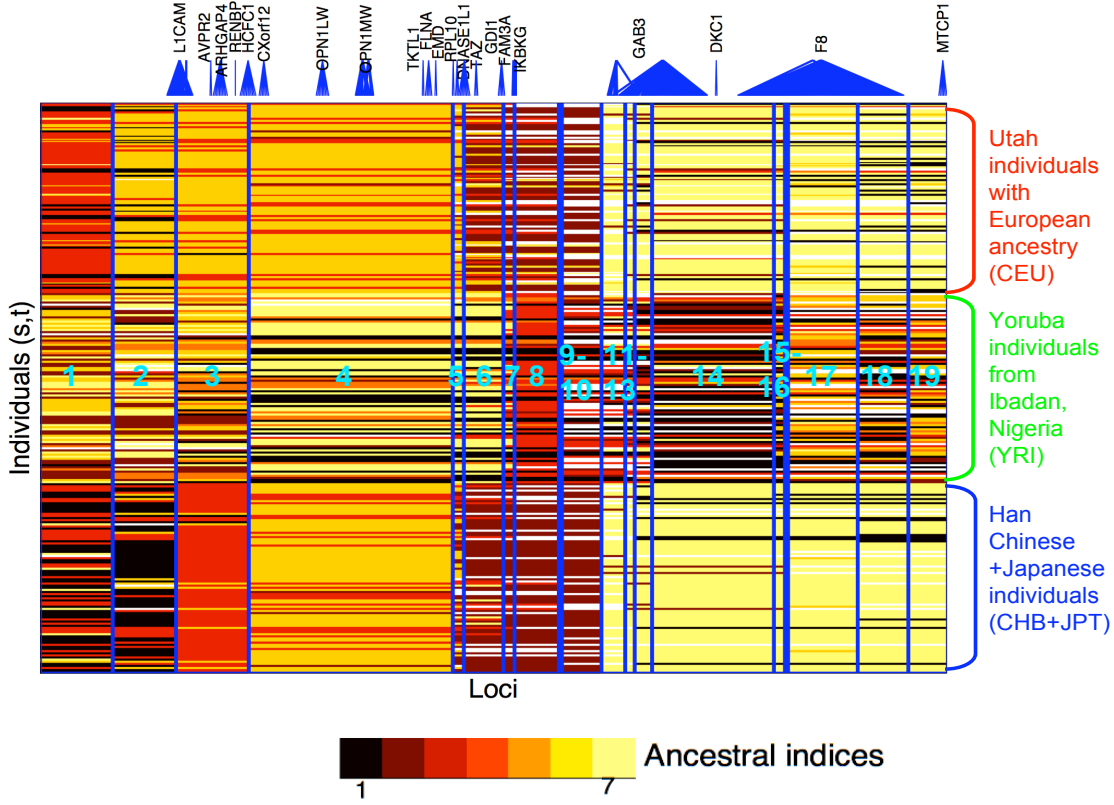


Fig. 7. Haplotype block partitioning of the ENm006 data. Each block shows the mapping to the ancestral haplotype library for all 270 HapMap genotyped individuals. Vertical blue lines indicate recombination hotspots discovered by our model. A haplotype block is defined implicitly as the region between two recombination hotspots. Colorings within a block for a given individual indicate the most likely ancestral index labeling associated with the block for that individual. Population labels are shown on the right of the diagram and known protein-coding genes in the ENm006 genomic region are shown at the top, each linked to SNPs in the genomic regions spanned by that gene.

for each sub-population separately using the block partitioning derived from each individual's recombination probability. We find the average block lengths are 41.6, 21.3 and 39.0 markers for CEU, YRI and CHB+JPT sub-populations respectively. Consistent with previous results (Gabriel *et al.*, 2002; HapMap Consortium, 2005), we find that the average block length for the YRI sub-population to be the lowest, followed by the CHB+JPT and CEU sub-populations. Having presented our method and results for learning a set of haplotype blocks, we now present our model for relating blocks to trait measurements.

4 A JOINT BAYESIAN MODEL OF HAPLOTYPE AND REGULATORY/PHENOTYPIC VARIATION

We now turn to the problem of linking the states of these haplotype blocks to patterns of quantitative trait variation. To achieve this, we introduce the model which is shown as a Bayesian network in Fig. 8. This model assumes that a subset of haplotype blocks affect each quantitative measurement of regulatory and phenotypic variation through a linear relation. These quantitative measurements may include phenotypes, high throughput gene expression data or any other continuous measurement that may be putatively linked to genetic variation.

Our model considers the blocks jointly and is able to suppress spurious relations due to correlations between block states across different blocks. For example, if variation in a single haplotype block b completely explains variation in the expression of a particular gene g , then a model that would consider blocks separately (for example, using multiple linear regressions) would also find relationships between the gene g and all blocks whose states are correlated with the state of b . Such spurious relations are ignored by our approach and the remaining relations which are most informative are discovered by the model.

Let us assume the haplotypes for J individuals have been partitioned into B non-overlapping haplotype blocks using the method described above. For each individual j , we define the label for a particular block b as an ordered pair $l_b^j = (S_b, T_b)^j$, representing the pair of ancestral haplotype indices inferred for this block, with the ancestral indices ranging from $1 \cdots A$. We represent S_b and T_b as indicator vectors of length A so that, if a block is associated with the a^{th} ancestor, the a^{th} element of the indicator vector is one and the remaining elements are zero.

Let \mathbf{Z} be a $J \times G$ matrix, where z_g^j is the measurement of the g^{th} quantitative trait of the j^{th} individual. We can represent the measurements of a particular trait g for all individuals by \mathbf{z}_g . We would

now like to infer associations between haplotype block labels and quantitative trait measurements whilst rejecting noisy or spurious associations. To optimize predictive accuracy, we will encourage sparsity in our model to favor only a small number of blocks influencing any given trait. We model this explicitly using a binary *relevance variable* w_{bg} associated with each combination of block and trait, such that $w_{bg} = 1$ indicates an association between the block b and measurement g , and $w_{bg} = 0$ indicates no association. We place a Bernoulli prior distribution over each w_{bg} in which $P(w_{bg} = 1) = p$ so that smaller values of p favor more sparse solutions.

We model a particular trait g for an individual j as a weighted sum of contributions from the subset of B blocks which are considered relevant according to \mathbf{W} , so that $z_g^j = \sum_b w_{bg} (S_b^j + T_b^j)^T \boldsymbol{\mu}_{bg} + \text{noise}$, where $\boldsymbol{\mu}_{bg}$ is a A -dimensional vector of weights for a particular block-trait pair. As the influence of a particular haplotype block on a trait is independent of the individual, the relevance variables \mathbf{W} and the block contributions $\boldsymbol{\mu}_{bg}$ are shared across all individuals. Thus, the A weights for any given block-trait pair allow us to model the joint patterns of variation in block label and trait measurement across individuals.

We define a matrix $\mathbf{L}_b \in \mathbf{R}^{J \times A}$, where $(j, a)^{th}$ element is given by $[S_b^j = a] + [T_b^j = a]$. Assuming a zero-mean isotropic Gaussian noise with precision ρ_g , we have

$$P(\mathbf{Z}|\mathbf{W}, \mathbf{L}, \boldsymbol{\mu}, \rho) = \prod_g N(\mathbf{z}_g; \sum_b w_{bg} \mathbf{L}_b \boldsymbol{\mu}_{bg}, \frac{1}{\rho_g} \mathbf{I}) \quad (2)$$

We place a Normal-Gamma prior distribution on the weights $\boldsymbol{\mu}$ and the inverse variances ρ_g such that:

$$P(\boldsymbol{\mu}, \rho) = \prod_{b,g} N(\boldsymbol{\mu}_{bg}; \boldsymbol{\mu}_0, \frac{1}{\rho_g \tau_0} \mathbf{I}) \text{Gamma}(\rho_g; \alpha_0, \beta_0), \quad (3)$$

where the parameters of the prior distributions are shared across all blocks and traits. Given the quantitative measurements and the haplotype block structure across J individuals, we then aim to learn a model that maximizes $P(\mathbf{Z})$:

$$P(\mathbf{Z}) = \sum_{\mathbf{W}, \mathbf{L}} \int_{\boldsymbol{\mu}, \rho} P(\mathbf{Z}|\mathbf{W}, \mathbf{L}, \boldsymbol{\mu}, \rho) P(\mathbf{W}) P(\boldsymbol{\mu}, \rho) P(\mathbf{L}) d\boldsymbol{\mu} d\rho$$

4.1 Variational Bayes learning of associations

The marginalization over the random variables required to compute $P(\mathbf{Z})$ in the above equation cannot be computed analytically, and hence we cannot maximize this quantity exactly. As in Section 2, we resort to a variational approximation (Jordan *et al.*, 1999) where instead of maximizing $P(\mathbf{Z})$, we instead maximize a lower bound on $\log P(\mathbf{Z})$,

$$\log P(\mathbf{Z}) \geq \sum_{\mathbf{W}, \mathbf{L}} \int_{\boldsymbol{\mu}, \rho} Q(\mathbf{W}, \mathbf{L}, \boldsymbol{\mu}, \rho) \log \frac{P(\mathbf{Z}, \mathbf{W}, \mathbf{L}, \boldsymbol{\mu}, \rho)}{Q(\mathbf{W}, \mathbf{L}, \boldsymbol{\mu}, \rho)}.$$

Here, Q approximates the required posterior distributions over the latent variables: in particular, we use the approximation

$$Q(\mathbf{W}, \mathbf{L}, \boldsymbol{\mu}, \rho) = Q(\mathbf{W}) Q(\mathbf{L}) Q(\boldsymbol{\mu}|\rho) Q(\rho) = \prod_{b,g} q_{bg}^{w_{bg}} (1 - q_{bg})^{1-w_{bg}} Q(\mathbf{L}) N(\boldsymbol{\mu}_{bg} | \boldsymbol{\mu}_{bg}^*, \sigma_{bg}^*) \text{Gamma}(\rho_g | \alpha_g^*, \beta_g^*)$$

where the distribution over block labels $Q(\mathbf{L}) = \prod_b Q(S_b) Q(T_b)$ is obtained from Section 2.2 and is held fixed.

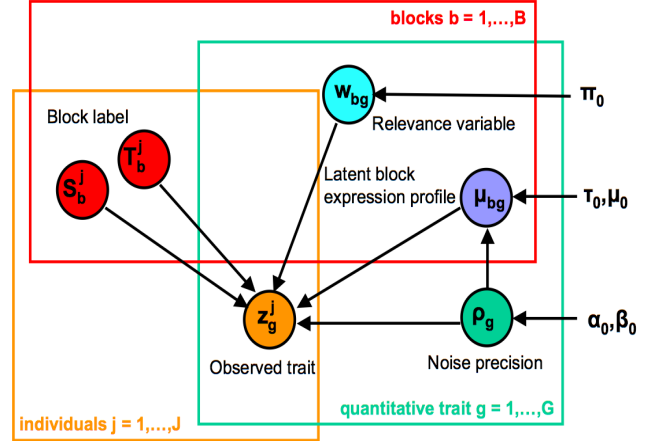


Fig. 8. Bayesian network for the model linking haplotype block structure to regulatory and phenotypic traits. The boxes, or “plates”, indicate that the structures contained within them are replicated a number of times indicated at the top of each plate. This replication allows for multiple haplotype blocks to contribute to particular a gene expression/phenotypic measurement for a given individual. Each gene/phenotype in our network is assigned a set of indicator variables which select haplotype blocks that are likely to be relevant given the model parameters and the data.

To carry out the optimization, we use variational Bayes (Attias, 1999) that performs approximate Bayesian inference by iteratively updating $Q(\mathbf{W})$, $Q(\boldsymbol{\mu}|\rho)$ and $Q(\rho)$ so that each update increases the bound on $\log P(\mathbf{Z})$. To prevent the premature switching off of haplotype blocks for certain genes, we use an annealing strategy in which we progressively decrease the sparsity prior probability p over the relevance variables for a fixed number of variational Bayes iterations, until it reaches the target sparsity value.

5 RESULTS ON LEARNING ASSOCIATIONS BETWEEN HAPLOTYPE BLOCKS AND TRAITS

5.1 Linking haplotype block structure in Chr5q31 to Crohn’s disease phenotype

Here, we applied our BlockMapper model to predict Crohn’s disease phenotype in the Chr5q31 data (Daly *et al.*, 2001). The data consists of the genotypes and the phenotypes (presence (+1) /absence (-1) of Crohn’s disease) for 387 individuals. We first inferred the haplotype blocks from the genotype data for all 387 individuals (see Sec.3.1). Then, we used the Bayesian learning technique described above to learn the association between the haplotype block labels and the phenotypes of a randomly chosen subset of 287 individuals. We evaluated the accuracy of our model by measuring the error when predicting presence/absence of the disease on the remaining 100 individuals. We repeated this experiment for 10 independent train-test splits. Since our model predicts a real-valued trait corresponding to the input haplotype block labels, we used the sign of the predicted mean trait value to indicate presence (positive) or absence (negative) of the disease. To investigate the sensitivity of our model to the sparsity prior p , we repeated the above experiment for a range of values and found that the error rate was consistently similar for $0.1 \leq p \leq 0.6$. For example, $p = 0.3$ yielded an error rate of 23.1% with a standard deviation of 3.45% computed over the 10

splits. We found that the genotype information relevant to predicting Crohn’s was highly localized, with haplotype blocks 2 and 10 (Fig. 4b) being the most informative towards predicting Crohn’s disease ($p < 4.76 \times 10^{-5}$, Wilcoxon-Mann-Whitney test, alpha value of 1×10^{-3} Bonferroni-corrected). These results indicate that patterns of genetic variation in Chr5q31 are informative towards predicting the Crohn’s disease phenotype and that our haplotype blocks manage to capture this variation.

5.2 Linking genetic variation in ENm006 to GDI1 expression

We then evaluated our model’s predictive accuracy on continuous gene expression measurements from the HapMap project. These measurements consist of the expression profiles for 47,294 genes profiled in EBV-transformed lymphoblastoid cell lines (Stranger *et al.*, 2007). In particular, we focused on a set of 28 genes located in the ENm006 region, allowing us to search for cis- associations with the haplotype blocks established in Section 3.

We randomly split the set of haplotype block labels and gene expression data across to the 270 individuals into three sets - 170 for training, 50 for validation and 50 for testing. We learned the model using the training set, while optimizing the sparsity prior parameter p on the validation set using prediction error as the criterion. The prediction error used was the sum of squared residuals between the model’s estimate of the gene expression and the true level of expression. We found $p = 0.2$ minimized the error on the validation set and subsequently used this value for evaluating prediction error on the test set. We repeated the above analysis for 5 different train-validation-test splits with 3 random initializations in each split. Fig. 9a shows the frequency (over 5 data splits and 3 random initializations) with which each gene is associated to each haplotype block according to the average magnitude of the relevance variables w_{bg} .

To assess the impact of accounting for non-genetic factors in explaining gene expression variation, we made use of all the available non-genetic traits (gender, population and parental information). We use the parent-child relationships in the HapMap data to recast age as a categorical variable (child/parent/unknown). We incorporated these additional factors into our probability model from Eqn. 2 by encoding the states for each non-genetic factor in a fashion analogous to the encoding for terms L_b for the haplotype blocks, effectively treating the three additional variables as three ‘virtual’ blocks. Using the training/testing methodology described above, we learned a model which also accounts for both the influence of non-genetic factors on gene expression as well as the influence of haplotype blocks. Fig.9b shows the frequency (over 5 data splits and 3 random initializations) with which each gene is associated with haplotype blocks and the non-genetic attributes: in contrast to Fig.9a, we can see that the model has eliminated many spurious associations between genes and haplotype blocks which are in fact due to population-specific differences in gene expression. Indeed, we find that gene expression is well predicted by the population variable, consistent with recent reports that support this hypothesis (Spielman *et al.*, 2007).

In particular, Fig. 9b also highlights significant cis-associations between blocks 2 and 5 in the ENm006 region and the GDI1 gene, which has been implicated in X-linked mental retardation (Shisheva *et al.*, 1994). Figs. 9c,d show the relationship between GDI1 expression levels and the haplotype block labels for blocks 2 and 5 across all individuals: as can be seen, the particular label assigned

to a haplotype block is informative of variation in GDI1 expression, as given by variations in the median GDI1 expression with respect to haplotype block label within the YRI sub-population in block 2 ($p < 2.22 \times 10^{-4}$, Wilcoxon-Mann-Whitney test, alpha value of 0.01 Bonferroni-corrected for all pairwise tests of 6 assigned ancestral indices over 3 sub-populations) and the CHB+JPT sub-population in block 5 ($p < 3.33 \times 10^{-4}$). We only tested associations between genes and which were discovered by our relation model and thus we did not use a Bonferroni correction for the number of genes or haplotype blocks.

To see the extent to which gene expression is explained by non-genetic information, we also learned a baseline model that explains the gene expression using *only* non-genetic factors consisting of age, gender and population. In the case of GDI1, we found that incorporating the haplotype block information in addition to the non-genetic attributes used by the baseline model reduced the population-specific test prediction errors (i.e.: squared prediction error computed over subsets of individuals from a particular subpopulation) by 4.9% for CHB+JPT, 1.1% for YRI and -2.1% for CEU, suggesting a contribution from both subpopulation and genetic variation within populations to the variation of GDI1 across individuals. We computed test error for other genes in the ENm006 region, such as ARHGAP4, RENBP, HCFC1 and TKTL1 and did not observe corresponding reductions in prediction error for these genes. Although the number of individuals here is relatively small, the fact that we are able to infer significant associations between haplotype blocks and gene expression suggests that considering larger data sets would allow the prediction of more reliable associations.

6 DISCUSSION AND CONCLUSIONS

In this paper, we presented BlockMapper, a Bayesian model for learning haplotype block structure from genotype data and associating the discovered blocks with quantitative traits. We have presented the two components of the model, where the first component infers a set of reliable haplotype blocks from genotype data and the second learns associations between these blocks and quantitative traits, such as gene expression. We have shown that our model for genotype data significantly outperforms standard haplotype inference algorithms on the task of predicting missing alleles in test data. We also demonstrated that BlockMapper has good predictive accuracy when predicting Crohn’s disease phenotype in the Chr5q31 region. Using our model, we discovered novel cis-associations between haplotype blocks 2 and 5 in the ENm006 region of ChrX and GDI1, a gene implicated in X-linked mental retardation.

In this paper, we learned the two parts of the model in sequence: in future we would like to extend this so that the block structure and relationships are learned in tandem, allowing haplotype blocks to be discovered which are more relevant to the traits of interest. In addition, we primarily focused on finding cis-associations in this work: we intend to scale-up our analysis to a larger set of genes to find both cis- and trans- associations. Also, we currently treat each gene expression as an independent observation but we can also take into account co-expression/co-regulation between genes by learning these relationships from data (Segal *et al.*, 2003). Finally, it will be important to study alternative non-linear relationships between the haplotype blocks and traits. This work also calls for access to large data sets of genotypes and traits measured across a larger number of individuals than used here. Our model and results, in tandem

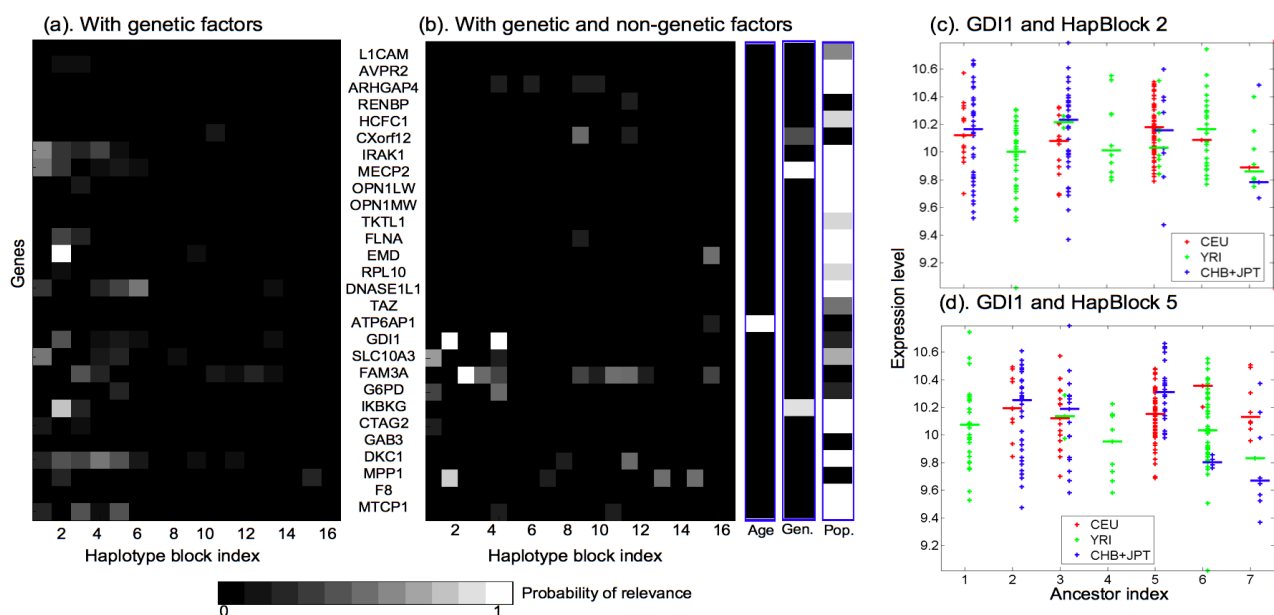


Fig. 9. a) Relevance variables linking haplotype blocks and gene expression: each entry represents how frequently a given haplotype block is associated with a given gene under our model; b) Relevance variables learned when non-genetic factors are included in the model – accounting for non-genetic factors has removed many spurious associations which can be better explained by non-genetic factors, such as population; c,d) GDI1 expression versus haplotype block labels for haplotype blocks 2 and 5 in the ENm006 region, with measurements broken down according to the 3 sub-populations (CEU, YRI, CHB+JPT). Lines indicate the median of gene expression for a given block label. Each gene expression measurement is displayed twice, showing the expression value against the ancestral index for each of the maternal and paternal haplotype blocks. The plots indicate that the ancestral index in haplotype blocks 2 and 5 is informative of variability in GDI1 expression, particularly in the YRI and CHB+JPT sub-populations.

with the impending improvements in high throughput genotyping and profiling technologies, underscores the potential of discovering thousands of previously uncharacterized associations between genetic variants and quantitative traits.

7 ACKNOWLEDGEMENTS

We would like to thank Paul Scheet for discussions about fastPHASE, and Richard Durbin and Manolis Dermitzakis of the Wellcome Trust Sanger Institute for valuable feedback on the model. JCH was supported by an internship at Microsoft Research Cambridge. We also thank Nebojsa Jovic for helpful discussions.

REFERENCES

Ardlie, K.G., Kruglyak, L. and Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**, 299-309 (2002).

Attias, H. Inferring parameters and structure of latent variable models by variational Bayes. *Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 21-30 (1999).

Botstein, D. and Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33**, Suppl. 228-237 (2003).

Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365-1369 (2005).

Daly, M.J. *et al.* High-resolution haplotype structure in the human genome. *Nat. Genetics* **29**, 229-232 (2001).

Dempster, A., Laird, N. and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1), 138 (1977).

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225-2229 (2002).

Greenspan, G. and Geiger, D. High density linkage disequilibrium mapping using models of haplotype block variation. *Bioinformatics* **20**, Suppl 1:i137-144 (2004).

The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

Jovic, N., Jovic, V., and Heckerman, D. Joint discovery of haplotype blocks and complex trait associations from SNP sequences without family data. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, (2004).

Jordan MI, Ghahramani Z, Jaakkola TS and Saul LK. An introduction to variational methods for graphical models. *Learning in Graphical Models*, Cambridge: MIT Press (1999).

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G. *et al.* Genetic analysis of genome-wide variation in gene expression. *Nature* **430**, 743-747 (2004).

Neal, R.M. and Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. in *Learning in Graphical Model M.I. Jordan (editor)* (1998).

Pritchard, J.K., Stephens, M. and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959 (2000).

Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257-286 (1989).

Scheet, P. and Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**, 629-644 (2006).

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and N. Friedman. A module map showing conditional activity of expression modules in cancer. *Nat. Genetics* **34**, 166-76, (2003).

Shisheva, A. *et al.* Cloning, characterization, and expression of a novel GDP dissociation inhibitor isoform from skeletal muscle. *Mol Cell Biol* **14**, 3459-68 (1994).

Spielman, R.S., Bastone, L.A., Burdick, J.T., Morley, M., Ewens, W.J. *et al.* Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genetics* (2007).

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).

Wang, W.Y., Barratt, B.J., Clayton, D.G. and Todd, J.A. Genome-wide association studies: Theoretical and practical concerns. *Nat Rev Genet* **6**, 109-118 (2005).