BIOINFORMATICS

Vol. 26 ISMB 2010, pages i208–i216
doi:10.1093/bioinformatics/btq191

Multi-population GWA mapping via multi-task regularized regression

Kriti Puniyani, Seyoung Kim and Eric P. Xing*

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

Motivation: Population heterogeneity through admixing of different founder populations can produce spurious associations in genome-wide association studies that are linked to the population structure rather than the phenotype. Since samples from the same population generally co-evolve, different populations may or may not share the same genetic underpinnings for the seemingly common phenotype. Our goal is to develop a unified framework for detecting causal genetic markers through a joint association analysis of multiple populations.

Results: Based on a multi-task regression principle, we present a multi-population group lasso algorithm using L_1/L_2-regularized regression for joint association analysis of multiple populations that are stratified either via population survey or computational estimation. Our algorithm combines information from genetic markers across populations, to identify causal markers. It also implicitly accounts for correlations between the genetic markers, thus enabling better control over false positive rates. Joint analysis across populations enables the detection of weak associations common to all populations with greater power than in a separate analysis of each population. At the same time, the regression-based framework allows causal alleles that are unique to a subset of the populations to be correctly identified. We demonstrate the effectiveness of our method on HapMap-simulated and lactase persistence datasets, where we significantly outperform state of the art methods, with greater power for detecting weak associations and reduced spurious associations.

Availability: Software will be available at http://www.sailing.cs.cmu.edu/

Contact: epxing@cs.cmu.edu

1 INTRODUCTION

Association mapping has recently become a popular approach to discover the genetic causes of many complex diseases such as cancer, asthma and diabetes. A typical genome-wide association (GWA) study involves examining genotype data, often single-nucleotide polymorphisms (SNPs), collected over millions of genetic markers in search for an association with the given phenotype such as disease outcome or disease-related quantitative traits, where a very small fraction of the markers are linked to the phenotype. Thus, the main challenge is to maximize the power for identifying causal alleles while suppressing false positives.

We consider the problem of taking advantage of the population structure in the samples to increase the power of an association analysis. It has been observed that population heterogeneity arising from admixing of ancestor populations almost always exists at different levels in any genotype data, and is often correlated with the geographical distribution of the individuals. For example, it has been shown that such heterogeneity is present in the HapMap data (The International HapMap Consortium, 2005) across European, Asian and African populations; and heterogeneity at a finer scale within European ancestry has been found in many genomic regions in the UK samples of Wellcome trust case control consortium (WTCCC) dataset (Wellcome Trust Case Control Consortium, 2007). Although the standard assumption in existing approaches for association mapping is that the effects of causal mutations are likely to be common across multiple populations, the individuals in the same population or geographical region tend to co-evolve, and are likely to possess a population-specific causal allele for the same phenotype. For example, Tishkoff *et al.* (2006) reported that the lactase-persistence phenotype is caused by different mutations in Africans and Europeans. In addition, the same genetic variation has been observed to be correlated with gene-expression levels with different association strengths across different HapMap populations. Our goal is to be able to leverage information across multiple populations, to find causal markers in a multi-population association study.

1.1 Highlights of this article

We propose a novel multi-task-regression-based technique that performs a joint GWA mapping on individuals from multiple populations, rather than separate analysis of each population, to detect associated genome variations. The joint inference is achieved by using a multi-population group lasso (MPGL), with an L_1/L_2 regression penalty (Obozinski *et al.*, 2008; Yuan and Lin, 2006; Zhao *et al.*, 2008) that encourages (but does not enforce) multiple populations to have similar causal markers. We assume that the population label for each individual is either known or has been inferred from the genotype, e.g. by using well-known programs such as Structure (Pritchard *et al.*, 2000) or mStruct (Shringarpure and Xing, 2009).

As illustrated in Figure 1, the MPGL (Fig. 1a) can detect causal SNPs in multiple populations jointly, unlike standard regression techniques applied on individuals in each population separately to infer associations in a population-specific manner (Fig. 1b). Statistically, while the L_1 part of the L_1/L_2 penalty in MPGL plays the role of identifying a small number of SNPs with non-zero association strengths, the L_2 part is applied to the regression coefficients for each SNP across all populations to allow them to have varying association strengths. Thus, association signals that are weak in each population but common to all of the populations are combined across populations, and therefore can be detected with a greater power. At the same time, if a non-causal SNP has weak association in a small subset of populations, the joint inference will not conclude it as being associated with the phenotype, reducing the overall false positive rate.

*To whom correspondence should be addressed.

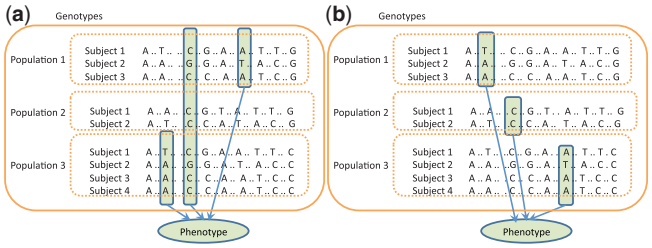

Fig. 1. Illustration of association analysis methods. (**a**) Multi-population group lasso finds causal SNPs via joint inference across multiple populations. (**b**) Standard lasso on each population cannot find SNPs with weak association to phenotype in some populations.

It is worth emphasizing that although MPGL performs a joint selection of SNPs with non-zero association strengths, it still allows each associated SNP to affect different populations with different magnitudes and direction of association. Thus, it recognizes population-specific causal alleles with relatively strong associations in a subset of the populations. Since association strengths are defined for each population separately (but estimated jointly), this technique also effectively corrects for population stratification. Our results demonstrate that the proposed method outperforms various existing methods in terms of reducing the spurious associations due to population structure.

The conventional single-SNP association tests are not able to distinguish between a set of correlated markers, increasing their false positive rates tremendously. Our approach on the other hand, is able to reduce the problem of correlations by analyzing all markers simultaneously in a linear regression framework. Once the marker with the strongest marginal correlation with the phenotype is predicted to be associated, other markers that are also correlated with this marker but do not give additional information about the phenotype are automatically rejected, reducing false positives.

Additionally, using cross-validation, our approach can automatically select the number of markers that are associated with the given phenotype, with high probability. In contrast, the single-SNP association tests involve the problem of selecting a *P*-value cutoff after computing *P*-values to determine significant associations. It is well-known that most phenotypes are complex traits that result from combined effects of many different mutations with small effect sizes, and the true number of causal markers is unknown. Hence, automatically picking the correct number of markers presents a significant advantage to our method.

We show the effectiveness of our approach on data simulated from HapMap genotypes over a wide variety of conditions, where the set of true causal markers are known. In all our experiments, we assume that the true population structure is unknown, and first use structure to estimate the population structure. Our approach outperforms state-of-the-art methods for association mapping in the presence of population stratification, when the number of causal markers shared across populations vary, the strength of the association in different populations vary and the allele frequency of the causal marker varies across populations. On the WTCCC lactase-persistence datasets, our algorithm is able to directly identify a single marker that has been reported to be associated with lactase persistence, while other approaches either report too many markers or too few.

1.2 Related work

While association mapping is a very well-studied problem in the literature, existing methods either completely ignore population structure, or focus on reducing spurious associations that are linked to the population structure rather than to the phenotype (Devlin *et al.*, 2004; Hoggart *et al.*, 2003; Price *et al.*, 2006; Zhu *et al.*, 2002).

Genomic Control (GC) (Devlin and Roeder, 1999; Devlin *et al.*, 2004) uses supplementary loci (called null markers) to correct for the population effect, which is assumed to be uniform across the genome. Any associations found between these null markers and phenotypes are attributed to the population stratification. GC first estimates an inflation factor using null SNPs, and then correct the *P*-values with this inflation factor. PSAT (Kimmel *et al.*, 2007) uses a novel dynamic programming algorithm, for fast randomized permutations tests to correct for an unknown population structure.

Unlike GC that does not require knowledge of the genealogy of the population or the nature of population heterogeneity, structured association explicitly takes into account the population heterogeneity in the samples. For example, Strat (Pritchard *et al.*, 2000) first learns the population structure using structure, performs an association test within each population and then combines the results across populations. Eigenstrat (Price *et al.*, 2006) on the other hand makes use of principal component analysis to remove ancestry information from the data before performing association tests. Various other approaches have also been proposed to control for population stratification, including likelihood ratio tests (Purcell and Shamb, 2004), logistic-regression-based tests (Epstein *et al.*, 2007) and mixed-model approach (Yu *et al.*, 2005).

Almost all of the association mapping literature is based on performing a statistical test for finding significant correlations between the phenotype and one SNP locus at a time, and correcting for multiple hypothesis testing. While multivariate regression methods such as lasso and ridge regression have been applied for classical association mapping (Hoggart *et al.*, 2008; Malo *et al.*, 2008; Shi *et al.*, 2007; Wu *et al.*, 2009), to our knowledge, our work is the first to consider a multivariate regression framework in an association study involving multiple populations.

2 METHODS

We begin our discussion with a brief overview of lasso for association analysis, and its extension to multi-population association analysis to give a flavor of how multivariate regression techniques can be used for association analysis. We then describe our MPGL algorithm for joint inference over multiple populations, and discuss a procedure for parameter estimation for our algorithm, and a method for selecting the optimal number of association markers automatically.

2.1 Lasso for association mapping

Let $\mathbf{X}$ denote the $n \times p$ matrix of genotype data for a homogeneous population, where n is the number of individuals involved in the study, and p is the number of markers genotyped for each individual. In the case of SNP markers, each element x_{ij} in $\mathbf{X}$ represents the number of minor alleles at the j-th locus of the i-th individual. Let $\mathbf{y}$ be the vector of length n for measurements of the phenotype. We assume a linear model between the genotypes and the phenotype:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta}$ represents a vector of p regression coefficients for association strengths, and ϵ is a vector of length n for zero-mean Gaussian noise with

fixed variance. We normalize $\mathbf{y}$ and each column of $\mathbf{X}$ to have zero mean, so that we do not have to explicitly model the bias term. When n is large and p is small, the regression coefficients $\boldsymbol{\beta}$ can be estimated by minimizing the sum of squared residuals:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\cdot(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \quad (2)$$

When the number of markers p is much larger than the number of individuals n as in a typical association study, the estimate of $\boldsymbol{\beta}$ obtained by solving Equation (2) is unattainable.

In association mapping, we typically expect a small number of loci to be associated with the phenotype, and lasso provides an effective tool to identify those relevant SNPs and set the regression coefficients for irrelevant SNPs to zero (Wu *et al.*, 2009). Lasso obtains an estimate of $\boldsymbol{\beta}$ by minimizing the penalized sum of squared residuals as follows:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\cdot(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\lambda\sum_{i=1}^{p}|\beta_i|. \quad (3)$$

The second term in the above equation indicates an L_1 penalty that encourages a sparsity such that only few SNPs have non-zero regression coefficients. The SNPs with non-zero regression coefficients are then predicted to be associated with the phenotype. The regularization parameter λ controls the amount of such penalty, and thus sparsity. A large value of λ means a greater amount of penalization, leading to more SNPs with zero regression coefficients.

2.2 Association analysis for multiple populations

When lasso is applied to an association mapping with a pooled dataset of all populations, it can effectively detect causal SNPs that have common effects on all of the populations. However, if the SNP influences the phenotype in a subset of the populations, or affects the phenotype with different strengths in different populations, the pooled analysis with lasso is likely to miss the population-specific association signals, since such signals may be outweighed by the information in other populations.

In this section, we assume that the population structure in the samples is known from prior knowledge or analysis, and make use of this information during the association analysis to detect both the population-specific and shared causal mutations. Any of the previously developed methods (Hubisz *et al.*, 2009; Shringarpure and Xing, 2009) for clustering individuals into populations based on allele frequencies can be used to infer the population structure before applying our method. In addition, when the population structure is known, this prior information can be directly used to form groups of individuals corresponding to multiple populations. In this article, we use the allele-frequency admixture model implemented in Structure to infer the admixing proportions for individuals, and apply k-means clustering (Hartigan, 1975) on the estimated admixing proportions to learn the population label for each individual.

Assuming that the population labels for individuals are known, subsequently we discuss how lasso can be applied in this setting, and in the next section, present our new approach that uses an L_1/L_2-regularized regression to maximize power by combining information and estimating the association strengths jointly across multiple populations.

2.2.1 Lasso for structured association Given the population labels z_i's, $z_i \in \{1,\ldots,C\}$, for the i-th individual and C populations, we group the genotype and phenotype data according to these labels into $\mathbf{y}^c=\{y_i|z_i=c\}$ and $\mathbf{X}^c=\{\mathbf{x}_i|z_i=c\}$, where $c=1,\ldots,C$, so that within the c-th group the individuals come from the same population. Then, within the c-th group, we can assume that there is no population structure and use lasso in Equation (3) to learn associations, using $\mathbf{y}^c$ and $\mathbf{X}^c$:

$$\min_{\boldsymbol{\beta}^c} \frac{1}{2}(\mathbf{y}^c-\mathbf{X}^c\boldsymbol{\beta}^c)'\cdot(\mathbf{y}^c-\mathbf{X}^c\boldsymbol{\beta}^c)+\lambda||\boldsymbol{\beta}^c||_1, \quad (4)$$

where $\boldsymbol{\beta}^c$ represents within-population association strengths. We repeat the above estimation process for each of the C groups, and examine the sparsity pattern in the $\boldsymbol{\beta}^c$'s to draw conclusions on which SNPs are causal to the phenotype.

2.3 Multi-population group lasso with L_1/L_2 penalty

While applying lasso within an inferred population can detect population-specific causal variants that a pooled analysis of all populations may be unable to identify, this approach analyzes each population separately without taking advantage of the relatedness among the $\boldsymbol{\beta}^c$'s through the shared causal SNPs, and may miss the weak association signals for common SNPs. Building on the multi-task regularized-regression framework recently studied in machine learning and high-dimensional statistics (Meier *et al.*, 2008; Obozinski *et al.*, 2008; Yuan and Lin, 2006; Zhao *et al.*, 2008), we describe our MPGL algorithm using L_1/L_2-regularization that can maximize the power for detecting SNPs that affect more than one population as well as population-specific causal SNPs.

Let us define $\mathbf{B}$ to be a $p\times C$ matrix $[\boldsymbol{\beta}^1,\ldots,\boldsymbol{\beta}^C]$, whose c-th column corresponds to the regression coefficients for the c-th population. Let $\boldsymbol{\beta}_j$ denote the j-th row of $\mathbf{B}$ that corresponds to the regression coefficients for the j-th SNP across the C populations. Then, the L_1/L_2 penalty is defined as follows:

$$||\mathbf{B}||_{L_1/L_2}=\sum_{j=1}^{p}||\boldsymbol{\beta}_j||_2, \quad (5)$$

where $||\mathbf{x}||_2=\sqrt{\sum_{c=1}^{C}x_c^2}$. In this case, the L_1 penalty is applied over the L_2 norms of vectors of regression coefficients $\boldsymbol{\beta}_j$'s, rather than individual elements of regression coefficients as in lasso. Using this penalty, the L_1/L_2-regularized regression for a joint association analysis of multiple populations obtains the estimate of $\mathbf{B}$ by solving the following optimization problem:

$$\min_{\mathbf{B}} \frac{1}{2}\sum_{c=1}^{C}(\mathbf{y}^c-\mathbf{X}^c\boldsymbol{\beta}^c)'\cdot(\mathbf{y}^c-\mathbf{X}^c\boldsymbol{\beta}^c)+\lambda||\mathbf{B}||_{L_1/L_2}, \quad (6)$$

where λ is the regularization parameter that determines the amount of penalization. The L_1/L_2 penalization plays the role of shrinking the regression coefficients $\boldsymbol{\beta}_j$ for the j-th SNP across all populations to zero jointly, if that SNP is not associated with the phenotype, thus reducing the number of false positives. On the other hand, if the SNP is relevant to at least one of the C populations, all of the elements in $\boldsymbol{\beta}_j$ will be selected jointly to have non-zero values, but the L_2 norm still allows the association strengths to be different across the populations for the j-th SNP. Thus, the joint inference made by the L_1/L_2 penalty enables us to infer association between a causal SNP and the phenotype by borrowing strength across populations and setting the corresponding regression coefficients jointly to non-zero values. We notice that a large value of λ will set more rows $\boldsymbol{\beta}_j$'s of $\mathbf{B}$ to zero.

Various block-structured norms in the form of L_1/L_q, $q>0$, to combine information from related inputs or outputs in a regression problem have been previously proposed (Turlach *et al.*, 2005; Zhao *et al.*, 2008). For example, in group lasso (Yuan and Lin, 2006), the grouping structure of the inputs is assumed to be known, and the L_q part of the L_1/L_q norm is defined over regression coefficients for the members in each group, so that they are jointly set to zero or non-zero values. In a multiple-output regression, the L_q part of the L_1/L_q norm is over the regression coefficients for all outputs for each input, and an input is selected to be jointly influencing all of the outputs. Our use of L_1/L_2 norm differs from these previous methods in that we take advantage of the grouping structure among the samples rather than inputs or outputs, where each group corresponds to a population.

Obozinski *et al.* (2008) found that for k regressions, under certain conditions, the sample complexity for L_1/L_2, is up to k times smaller than the lasso sample complexity, with weak assumptions of shared support. Thus, under certain conditions, the L_1/L_2 regression framework will require up to k times fewer samples than lasso to obtain the correct set of associated markers.

2.3.1 Parameter estimation We estimate the regression coefficients $\mathbf{B}$ by solving the optimization problem in Equation (6). The L_1/L_2 penalty is not

smooth at zero, and for this reason, methods based on the first-order gradient cannot be used directly for optimizing it. Hence, we optimize this problem by transforming the problem into a single-output multivariate regression with a group-lasso penalty, and apply a fast optimization method developed for group lasso (Tomioka and Sugiyama, 2009).

In order to transform the problem in Equation (6), we concatenate $\boldsymbol{\beta}^c$'s to form a vector of length $(p \cdot C)$, $\boldsymbol{\beta}_g = [(\boldsymbol{\beta}^1)^T, \ldots, (\boldsymbol{\beta}^C)^T]^T$. Similarly, we concatenate $\mathbf{y}^c$'s to form a vector of length n, $\mathbf{y}_g = [(\mathbf{y}^1)^T, \ldots, (\mathbf{y}^C)^T]^T$, and form $(n) \times (p \cdot C)$ block-diagonal matrix $\mathbf{X}_g$, where $\mathbf{X}^c$'s are placed along the diagonal, and the rest of the elements are set to 0. Then, the problem in Equation (6) can be re-written as:

$$\min_{\mathbf{B}} \frac{1}{2}(\mathbf{y}_g - \mathbf{X}_g\boldsymbol{\beta}_g)^T \cdot (\mathbf{y}_g - \mathbf{X}_g\boldsymbol{\beta}_g) + \lambda ||\mathbf{B}||_{L_1/L_2}. \quad (7)$$

This transformed problem can be viewed as a single-output multivariate regression with $(p \cdot C)$ inputs, where the grouping of the input is defined according to the population structure. This shows that the L_1/L_2-regularized regression for joint analysis of multiple populations is equivalent to group lasso in the transformed space.

The solution for the problem in Equation (7) can be obtained by re-writing it in an equivalent form, as a constrained optimization problem, converting it into the dual form, and then solving this dual problem (Tomioka and Sugiyama, 2009). Since the problem in Equation (7) is convex, we are guaranteed that the solution obtained by optimizing this dual problem is equivalent to the solution obtained by optimizing the original primal problem in Equation (7). The dual can now be expressed in the augmented lagrangian form (Bertsekas, 1982), to develop a much faster optimization technique than conventional gradient-based techniques. Further, the dual augmented lagrangian can update both primal and dual variables, and exploit the known sparsity in the primal solution. This provides this method with two major advantages. By tracking the solution in both primal and dual space, the stopping criteria used is that the primal and dual objective values are close to each other, which is a direct measure of how close the current solution is to the true optimal solution. Since the stopping criteria directly measures how close the current solution is to the optimal solution, the algorithm is more stable and precise than other optimization techniques. Secondly, since the optimization occurs in the dual space, it is efficient when the number of SNPs is much more than the number of individuals, as is the case in a typical association analysis. For these reasons, the implementation of our method is extremely fast and takes < 1 second on an Intel Core-2 CPU with 1 GB memory to estimate the association strengths of 2000 SNPs in two populations with 257 individuals. When $n \ll p$ as in a typical association analysis, we found that this implementation of L_1/L_2-regularized regression is as efficient as lasso.

2.3.2 Selecting the number of association markers One of the main advantages of regularized-regression approaches is that we can tune the regularization parameter λ automatically to select the correct number of association markers, with high probability. In contrast, we note that all methods based on single-SNP association tests require the user to input an arbitrary P-value cutoff that is used to determine which of the markers are significantly associated.

We hold out a small number of individuals in the entire dataset as a validation set and select the value of the regularization parameter λ that gives the optimal level of sparsity, by minimizing prediction errors on the validation set. Our procedure involves two steps. In step one, we fit a suite of candidate models with different λ values on the training data, as described in Section 2.3.1. Since we only change the value of λ and the training data remains the same, warm starts can be used for fast optimization. As λ increases, the number of markers with non-zero association strengths reduces. In step two, we evaluate each model using the phenotype prediction error on the validation set, and select the λ that gives the lowest prediction error on the validation set. We repeat this process 10 times by randomly splitting the data into training and validation sets, and select the value for λ that gives the lowest error on average over the 10 runs.

To evaluate the phenotype prediction error, we need to first reduce the bias introduced by penalized regression. Hence, once we learn the sparsity pattern of non-zero regression coefficients, for each model, we re-estimate the regression coefficients for those non-zero elements using a standard least square method without penalty terms (Hastie *et al.*, 2003). Thus, for a particular λ, if the set of predicted associated markers is S_λ, then $\hat{\beta}(\lambda)$ is computed as the least square estimator for the regression with markers restricted to S_λ. In other words, we have $\hat{\beta}(\lambda) = (X_\lambda^T X_\lambda)^{-1} X_\lambda^T Y$, where $X_\lambda = (X_{\bullet j} : j \in S_\lambda)$ is the matrix for all individuals, but only those markers predicted as associated with the phenotype by our model. The $\hat{\beta}(\lambda)$ is then extended to length p by setting $\hat{\beta}_j(\lambda) = 0$ for $j \notin S_\lambda$. The phenotype prediction error is then simply computed using this estimate of $\hat{\beta}(\lambda)$ as $(Y_v - X_v\hat{\beta}(\lambda))^2$, where (X_v, Y_v) is the validation set. For lasso regression, Wasserman and Roeder (2009) prove that the true association markers will be included in the predicted associated markers found by this validation procedure with high probability. We expect that a similar proof will also apply for the L_1/L_2 regression.

Finally, the optimal λ selected by the procedure outlined above is used to train the model on the entire dataset, and we report results on these estimates.

3 RESULTS

In this section, we compare the performance of the MPGL with those of previously developed methods such as single-SNP association tests without controlling for population stratification, GC (Devlin and Roeder, 1999), Eigenstrat (Price *et al.*, 2006), lasso for a pooled dataset of individuals from all populations (Wu *et al.*, 2009), and lasso for structured association (lasso SA) as discussed in Section 2.2.1. We perform an extensive simulation study under various scenarios, using datasets simulated from HapMap genotypes, and demonstrate our method on a real dataset, the lactase-persistence phenotype with the WTCCC genotypes.

3.1 Simulation study

To provide a realistic setting for simulations, we use genotypes of 257 unrelated individuals in two HapMap populations, 87 individuals from Maasai in Kinyawa, Kenya (MKK) and 170 individuals from the Asian population comprising of Han Chinese and Japanese (JPT+CHB), and simulate the phenotypes based on these genotypes. After discarding SNPs with variance <0.5% that corresponds to a minor allele frequency of <1%, we select every 10th SNP to reduce the effects of linkage disequilibrium, and use a block of 2000 SNPs as inputs. To simulate phenotypes from these genotypes, we randomly select 20 SNPs that are causal in at least one of the two populations with non-zero regression coefficients, and set their association strengths to values sampled from a uniform distribution of $[0, 5]$ with the directions of the associations assigned randomly to either positive or negative. For each individual, given the regression coefficients corresponding to the population that the individual belongs to, we generate the phenotype using the linear relationship in Equation (1) with noise distributed as $N(0, 1)$. We generate 50 such datasets based on different regions in the autosomal chromosomes, and report the results averaged over these datasets. We run structure on each set of genotypes with the number of populations set to two to learn the admixing proportions of each individual. An example of the population structure learned by structure from a single set of 2000 SNPs is shown in Figure 2.

In Figure 3, we compare the performances of different association methods using a single simulated dataset. We first plot the true association strengths in red for Populations 1 and 2 in Figure 3a

and 3b, with a blue marker representing causal SNPs shared across multiple populations. As can be seen, some of the SNPs have very weak association with the phenotype, making detection of this association very hard. Figures 3c–j show the association strengths detected by various methods. In each panel, we mark the true association with a green circle, and a predicted association with a red '+'. For the multivariate regression methods (Figures 3c–g), we used cross-validation to select the association SNPs, and hence, we plot the association strengths for the predicted causal SNPs. Thus, an overlap between a red and green marker is a true association, a green marker without a corresponding red marker is a false negative, and a red marker without the green marker is a false positive. In Figures 3c–f, we observe that MPGL has 69% precision and 45% recall, and clearly outperforms lasso for structured association with 49.6% precision and 37.5% recall. Lasso (Fig. 3g) detects not only many of the causal SNPs correctly, but also has a very high false positive rate, since it does not use any population structure information. For the methods based on single-SNP hypothesis testing (Fig. 3h–j), we plot a line corresponding to a P-value of 0.05. Thus, the green markers above the P-value cutoff line are true positives, the green markers below the P-value cutoff line are false negatives, and the red markers above the cutoff line representing false positives. Eigenstrat (Fig. 3h) in spite of having structural information, and identifying many true positives, also has a very high false positive rate, and overall has similar performance as lasso. Both GC and single-SNP analysis have a very large number of false positives as well, and it is very hard to derive any meaningful conclusions from these results. This only serves to emphasize that detecting and correcting for the presence of population structure is very essential for correct association analysis.

We evaluate various association methods in terms of how successfully they identify the true causal SNPs with few false positives, and summarize the results by plotting the partial receiver operating characteristic (partial ROC) curves and reporting the partial areas under the curves (PAUC). A partial ROC curve plots the true positive rate for recovering true causal SNPs on the y-axis and the false positive rate on the x-axis, over a range of small values of false positive rates. When the number of SNPs is large and the

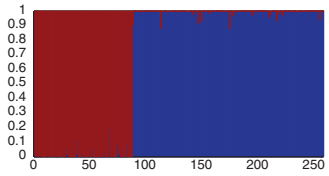

Fig. 2. Admixing proportions inferred by structure assuming two populations on a single simulated dataset.

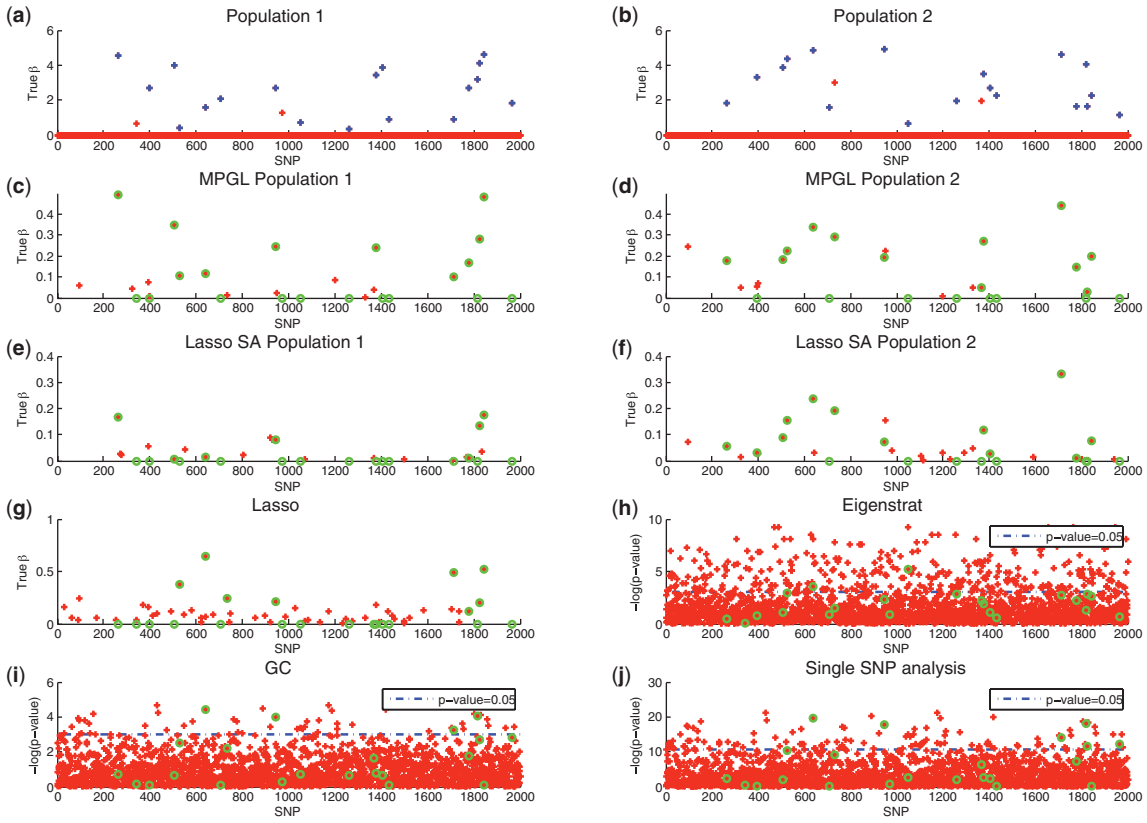

Fig. 3. Results from association analysis of a single simulated dataset from HapMap. (**a**–**b**) The red markers show the true association strengths for populations 1 and 2, respectively (the shared causal SNPs are highlighted in blue). Association strengths are shown for (**c**–**d**) MPGL for populations 1 and 2, (**e**–**f**) lasso for structured association for populations 1 and 2, (**g**) lasso, (**h**) Eigenstrat, (**i**) GC and (**j**) single-SNP association tests. We plot the absolute values of the regression coefficients in (**a**–**g**), and $-\log(P\text{-value})$'s in (**h**–**j**). Red markers show the predicted value, and green circles show the true causal SNPs.

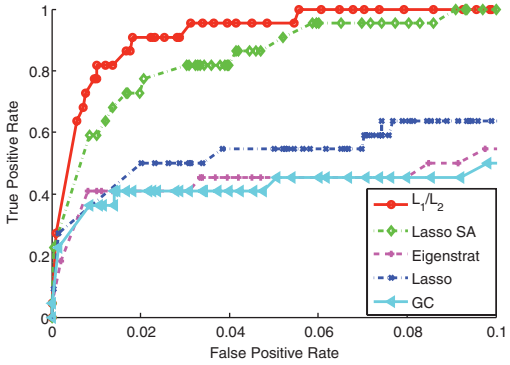

Fig. 4. Partial ROC curves of various association methods based on a single simulated dataset. Out of the 20 causal SNPs, 16 SNPs have non-zero association strengths in both of the two populations, while the remaining four SNPs are causal in only one of the two populations.

number of SNPs with true associations is small, we are primarily concerned with the area of low false positive rates in the ROC curves, and thus, we focus our attention to this range of false positive rates $[0, 0.05]$. The PAUC is defined as the area under the ROC curves in this range of false positive rate, normalized by the length of the partial range being considered, so that the maximum PAUC possible is one. A higher value for PAUC represents better performance.

Figure 4 shows the partial ROC curves comparing the various methods in the range of false positive rates $[0, 0.1]$ obtained from a single simulated dataset. We select 16 SNPs as causal in both of the two populations and four SNPs as causal for one of the two populations. As can be seen in Figure 4, both the MPGL and lasso for structured associations significantly outperform all of the existing association methods, demonstrating the effectiveness of these approaches. In addition, MPGL significantly improves the performance of lasso applied to each population separately, because it can borrow strength across multiple populations to detect true associations. Interestingly, we observe from Figure 4 that lasso for a pooled dataset of all populations performs nearly as well as the single-SNP analyses controlling for the population stratification such as Eigenstrat and GC, even though it does not make use of any information on population structure. This reaffirms that in general the sparse multivariate regression provides a powerful tool for association analysis, compared to traditional single-SNP hypothesis tests. In general, we found that the range of false positive rates $[0, 0.05]$ summarizes the overall trend in performances of different methods across all ranges of false positive rates, so in the remainder of this section, we show the PAUC values based on this choice of interval for the false positive rate.

3.1.1 Effects of varying number of shared causal SNPs We vary the number of shared causal SNPs across populations to see how the amount of shared sparsity pattern affects the performance of various methods. In Figure 5, we vary the percentage of shared causal SNPs in the two populations, while keeping the total number of causal SNPs fixed at 20, and show the results averaged over 50 datasets. Figure 5 shows that when there is a large overlap between the sets of causal SNPs of the two populations, MPGL significantly outperforms all other methods, since it is able to borrow strength across different populations to determine the shared SNPs as causal or non-causal. In addition, we notice that as the number of shared

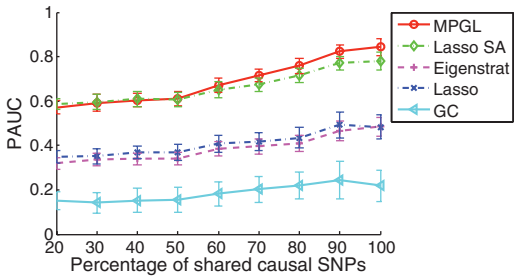

Fig. 5. PAUC in simulated data, when the numbers of causal SNPs shared across sub-populations varies. Assuming 20 causal SNPs, we vary the percentage of these 20 SNPs that are causal in both of the sub-populations.

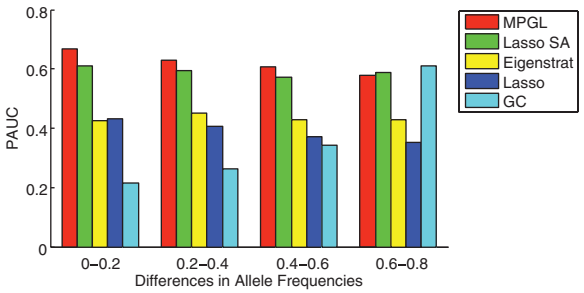

Fig. 6. PAUC in simulated data, when the amount of differences in allele frequencies of causal SNPs across sub-populations varies.

causal SNPs decreases, the performance of MPGL approaches that of lasso for structured association. Thus, even when there is little overlap in association SNPs between the different populations, MPGL does not compromise its performance, compared to the method that does not combine information across populations. The paired t-test showed that the improvement in PAUC scores in MPGL over lasso for structured association is significant when the percentage of overlap in causal SNPs between associated SNPs across populations is >50% with 99% confidence interval.

3.1.2 Effect of varying allele frequency of causal SNPs across populations Next, we explore how the amount of differentiation across populations that is present in the causal SNPs affects the performance of various association methods. We use the absolute value of the difference of the minor allele frequency between the two populations as a measure of the amount of differentiation of a SNP across populations, and consider four different levels of differentiations given as intervals of [0,0.2], [0.2,0.4], [0.4,0.6] and [0.6,0.8]. In each simulated dataset, we randomly select five SNPs with the same level of differentiation as causal SNPs, and show the results averaged over 50 datasets in Figure 6. We observe that for almost all levels of differentiation, MPGL significantly outperforms all other methods. Although the performance of GC improves as the amount of differentiation increases, the variance of the performance of GC is more than twice the variance of the other methods, which is not desirable.

3.1.3 Effect of varying association strength of causal SNPs In order to see how the signal-to-noise ratio affects the performance of the various methods, we vary the strength of associations for

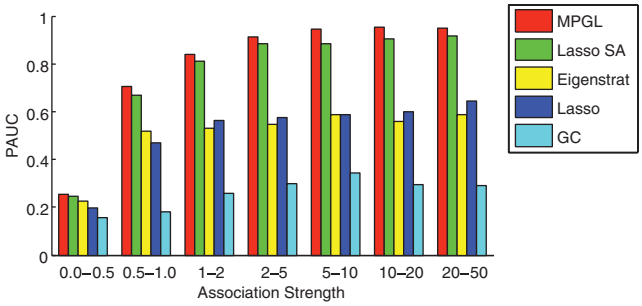

Fig. 7. PAUC in simulated data, when the association strengths for all causal SNPs are randomly sampled from one of the seven different intervals.

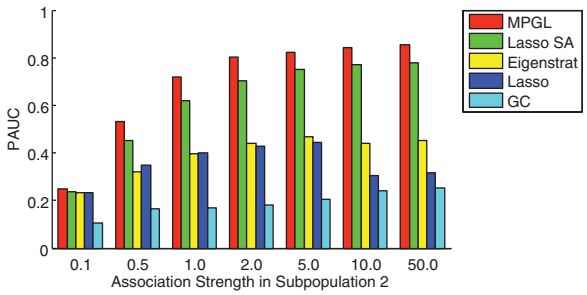

Fig. 8. PAUC in simulated data, when the ratio of association strengths of each causal SNP between the two populations varies. The association strength of the causal SNP in population 1 is sampled from a uniform distribution of [0,1], and the association strength of the same SNP in population 2 is multiplicative of seven different values.

causal SNPs and show the results in Figure 7. In each simulated dataset, the absolute values of the non-zero regression coefficients are generated from one of the seven uniform distributions, [0,0.5], [0.5,1.0], [1.0,2.0], [2.0,5.0], [5.0,10.0], [10.0,20.0] and [20.0,50.0]. For each of the intervals we show results averaged over 50 datasets. For all of the ranges of association strengths, MPGL significantly outperforms all of the other methods. As the signal-to-noise ratio increases, the performance of MPGL improves at a faster rate than any of the other methods. Especially, when the association strengths is >5.0, its PAUC reaches 0.9, which is close to the perfect recovery of true relevant SNPs. This shows that we can significantly benefit from the use of the L_1/L_2 penalty that combines information on sparsity pattern across multiple populations.

3.1.4 Effect of varying association strengths across populations Finally, we consider the scenario in which the same causal SNP influences the phenotype with different strengths in different populations. For example, the association strength of a SNP might differ in different populations due to genetic drift, growth, or contraction of the two populations. In order to replicate this scenario in our simulation, we generate the magnitude of association strength for each causal SNP in one of the two populations from a uniform distribution over [0,1], and then set the association strength of the same SNP in the other population to a value that is larger by a multiplicative factor of 0.1, 0.5, 1.0, 2.0, 5.0, 10.0 or 50.0. The PAUC values averaged over 50 datasets are shown in Figure 8. The results show that even when the sparsity pattern of causal SNPs is shared, but the values of association strengths are different across populations, MPGL has the flexibility of allowing the association strengths to differ for multiple populations and performed significantly better than other methods.

3.2 WTCCC dataset with lactase-persistence phenotype

We perform an association analysis of lactase-persistence phenotype with genotypes in the WTCCC dataset, and compare the results from our method with various other approaches.

While lactase activity typically disappears in childhood after weaning, some individuals have the ability to digest lactose during the adulthood. This trait, known as lactase-persistence, has been shown to be completely determined by a particular mutation near the *LCT* gene that encodes the lactase-phlorizin hydrolase (Enattah *et al.*, 2002). In addition, it has been observed that the lactase activity is widely different across populations (Bersaglieri *et al.*, 2004). In particular, the geographic distribution of lactase persistence is highly correlated with the distribution of dairy farming, and this phenotype is more commonly observed in northern Europe. Since the lactase activity is correlated with the population structure, it is necessary to control for the population stratification to correctly identify the mutation that determines lactase persistence.

We use the genotypes of 1504 individuals in the control group of WTCCC dataset, and perform an association analysis, assuming that the lactase-persistence phenotype is completely determined by SNP rs4988243 in chromosome 2. Although the known causal variant with 100% association with the lactase persistence has not been typed in this dataset, SNP rs4988243 lies in a high linkage disequilibrium region ($r^2 > 0.9$) with this known genetic variant in HapMap dataset. The previous analysis of population structure in WTCCC dataset has shown that the 135.16–136.82 Mb region on chromosome 2 that includes the *LCT* gene and SNP rs4988243 at 136.32 M exhibits geographical variation, and we include the 2500 SNPs in this region in our analysis. Although the UK populations in WTCCC datasets consist of immigrants from various parts of Europe in history, the previous analysis of this data found that in many of the genomic regions, there was not a significant differentiation, and that the associations for case-control populations were not significantly affected by population stratifications. Since our focus in this article is an association analysis under population stratification, we perform an analysis with lactase-persistence as phenotype rather than case–control labels for diseases.

We use Structure to learn groupings of individuals according to populations, before applying structured association methods with lasso or MPGL. We determine the number of ancestor populations K based on approximate posterior probabilities, as was suggested in Pritchard *et al.* (2000), and obtain $K = 4$ as the optimal number of ancestor populations. Then, we run k-means algorithm to cluster the individuals into four populations, based on the admixture proportions for individuals estimated by structure. Figure 9 shows the admixture proportions of individuals as columns using four different colors for each of the four ancestor populations, after clustering individuals into four groups.

For MPGL as well as the other regression-based methods, the value of the regularization parameter λ is selected as described in Section 2.3.2, with 50 individuals in the validation set, and the remaining 1454 individuals in the training set. Since the phenotype

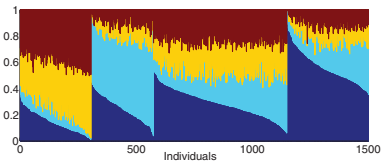

Fig. 9. Population structure in the genomic region around gene *LCT* in WTCCC dataset. Each column represents the admixture proportion of an individual estimated by structure. The colors represent different ancestor populations.

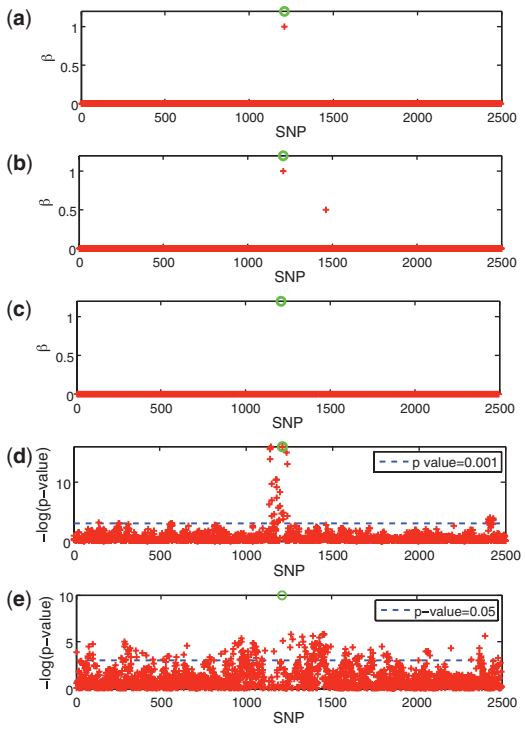

Fig. 10. Results from association analysis of lactase-persistence dataset. Association strengths are shown for (**a**) MPGL, (**b**) lasso for structured association, (**c**) lasso for a pooled analysis of all populations, (**d**) Eigenstrat and (**e**) GC. We plot the absolute values of the regression coefficients in (**a**), (**b**) and (**c**), and $-\log(P\text{-value})$ in (**d** and **e**). The locus with the causal SNP rs4988243 is marked with a green circle at the top of the plot.

is binary, we use MPGL with logistic regression instead of linear regression.

The association strengths for lactase-persistence estimated by different methods are shown in Figure 10. In each panel, we mark the true association SNP rs4988243 as a green circle. As can be seen in Figure 10a, MPGL correctly identifies SNP rs4988243 as the sole SNP with a non-zero association with lactase persistence. In Figure 10b, lasso with structured association also detects the true causal SNP, although there is one false positive that is also found to have a non-zero association with the phenotype. We find that lasso for structured association predicted this SNP to be associated with the phenotype in one of the populations, but not in the other three populations. We observe that the lack of signal from this SNP in the other populations successfully allows MPGL to conclude that there is no signal, and reject this SNP automatically.

In comparison to these two structured association methods, lasso that assumes no population structure completely misses the association signal, as can be seen in Figure 10c. Although in Figure 10d Eigenstrat is able to detect the true causal SNP at $P < 0.001$, there are additional 65 SNPs that are also found associated with lactase persistence. Out of these 65 false positives in Eigenstrat, 52 SNPs are in the highly differentiated subregion near SNP rs4988243, and the other 13 SNPs are found across the region of 2500 SNPs that we analyze. Only 8 of the 52 SNPs within the differentiated region are in a high LD with the true causal SNP ($r^2 \geq 0.8$). Thus, Eigenstrat finds significantly greater number of false-positives due to population stratification than MPGL. We notice that unlike Eigenstrat, the sparse regression methods in Figures 10a and b are able to exclude SNPs that are in LD with the true causal SNP and detect the true causal SNP as the associated SNP. Finally, we find in Figure 10e that GC has a large number of false positives due to the large confounding effect produced by the population structure. Overall, our results in Figure 10 shows that MPGL is a powerful method that detects association signals with no false positives in the presence of population stratification, and clearly outperforms the existing methods.

4 DISCUSSION

In this article, we proposed a multi-population group lasso using L_1/L_2 regression for joint association analysis of multiple populations. Our method assumes that population labels are known or can be learned from a separate analysis, and performs an association analysis within each population while borrowing information across populations. The L_1/L_2 penalty in our method allows us to detect population-specific causal alleles as well as causal alleles that are common across all populations with greater power and fewer false positives. Our experiments on HapMap-simulated and lactase-persistence datasets showed that our method is significantly more powerful than other previous approaches, and at the same time, can control for population stratification to reduce spurious associations. Possible future directions include incorporating geographical and spatial distribution over populations instead of assuming that all of the populations are jointly related as in L_1/L_2 regularization.

Funding: National Science Foundation (DBI-0546594), (DBI-0640543); National Institutes of Health (1R01GM087694); Alfred P. Sloan Fellowship (to E.P.X).

Conflict of Interest: none declared.

REFERENCES

Bersaglieri,T. *et al*. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.*, **74**, 1111–1120.

Bertsekas,D.P. (1982) *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, Boston.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Devlin,B. *et al*. (2004) Genomic control to the extreme. *Nat. Genet.*, **36**, 1129–1130.

Enattah,N.S. *et al*. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.*, **30**, 233–37.

Epstein,M.P. *et al*. (2007) A simple and improved correction for population stratification in case-control studies. *Am. J. Hum. Genet.*, **80**, 921–930.

Hartigan,J.A. (1975) *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA.

Hastie,T. *et al*. (2003) *The Elements of Statistical Learning*. Springer, New York.

Hoggart,C. *et al*. (2003) Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.*, **72**, 1492–1504.

Hoggart,C.J. *et al*. (2008) Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.

Hubisz,M.J. *et al*. (2009) Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Res*.

Kimmel,G. *et al*. (2007) A randomization test for controlling population stratification in whole-genome association studies. *Am. J. Hum. Genet.*, **81**, 895–905.

Malo,N. *et al*. (2008) Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am. J. Hum. Genet.*, **82**, 375–85.

Meier,L. *et al*. (2008) The group lasso for logistic regression. *J. Roy. Stat. Soc. B*, **70**, 53–71.

Obozinski,G. *et al*. (2008) High-dimensional union support recovery in multivariate regression. In *Advances in Neural Information Processing Systems 21*. Vancouver, B.C., Canada.

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

Pritchard,J. *et al*. (2000) Association mapping in structured populations. *Am. J. Hum. Genet.*, **67**, 170–181.

Purcell,S. and Shamb,P. (2004) Properties of structured association approaches to detecting population stratification. *Hum. Heredity*, **58**, 93–107.

Shi,W. *et al*. (2007) Detecting disease-causing genes by LASSO-Patternsearch algorithm. *BMC Proceedings*, **1**(Suppl. 1), S60.

Shringarpure,S. and Xing,E.P. (2009) mstruct: inference of population structure in light of both genetic admixing and allele mutations. *Genetics*, **182**, 575–593.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1399–1320.

Tishkoff,S.A. *et al*. (2006) Convergent adaptation of human lactase persistence in africa and europe. *Nat. Genet.*, **39**, 31–40.

Tomioka,R. and Sugiyama,M. (2009) Dual augmented lagrangian method for efficient sparse reconstruction. *IEEE Signal Proccesing Lett.*, **16**, 1067–1070.

Turlach,B. *et al*. (2005) Simultaneous variable selection. *Technometrics*, **47**, 349–363.

Wasserman,L. and Roeder,K. (2009) High-dimensional variable selection. *Ann. Stat.*, **37**, 2178–2201.

Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Wu,T.T. *et al*. (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Yu,J. *et al*. (2005) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. B*, **68**, 49–67.

Zhao,P. *et al*. (2008) Grouped and hierarchical model selection through composite absolute penalties. *Technical Report 703*, Department of Statistics, University of California, Berkeley.

Zhu,X. *et al*. (2002) Association mapping, using a mixture model for complex traits. *Genetic Epidemiol.*, **23**, 181–196.