We describe FaST-LMM, a linear mixed model for genome-wide association studies that scales linearly in the number of individuals in both runtime and memory use. Our algorithm is an order of magnitude faster than current efficient algorithms (EMMAX/P3D) on Wellcome Trust data with 15,000 individuals. On synthetic data, FaST-LMM can analyze 120,000 individuals in just a few hours, whereas the current algorithms are unable to analyze even 20,000 individuals (http://fastlmm.codeplex.com).

1 Introduction

The problem of confounding by population structure, family structure, and cryptic relatedness in genome-wide association studies (GWAS) is widely appreciated [1, 2, 3, 4, 5, 6, 7]. Statistical methods for correcting these types of confounders have progressed and include approaches such as linear mixed models (LMMs) [8, 2, 9, 3, 4, 5, 6, 7], genomic control, family-based association tests, structured association, and Eigenstrat [7]. While each of these methods has proven useful for correcting certain types of confounding, LMMs have been shown capable of capturing all of these types of confounders simultaneously, without knowledge of which are present, and without the need to tease them apart (e.g. [7]). Unfortunately, LMMs are computationally expensive relative to simpler models. In particular, the runtime and memory footprint required by the model scale as the cube and square of the number of individuals in the data set, respectively. This bottleneck means that LMMs run slowly or not at all on currently or soon-to-be available large data sets having the most power to uncover novel associations.

Roughly speaking, LMMs tackle confounders by using measures of genetic similarity to capture the probabilities that pairs of individuals share causative alleles. Such measures have been based on (e.g.) IBD [11, 10] and the realized relationship matrix (RRM) [9, 10, 12] and have been estimated...
with a relatively small sample (200-2000) of markers [2, 4]. Herein, we take advantage of such sampling to make LMM analysis applicable to much larger data. In particular, we introduce a reformulation of LMMs, called FaST-LMM for factored spectrally transformed linear mixed models. We show that, provided (1) the number of SNPs used to estimate genetic similarity between pairs of individuals is less than the number of individuals in the data set (regardless of how many SNPs are to be tested) and (2) the RRM is used to determine these similarities, then FaST-LMM produces exactly the same results as a standard LMM, but with a runtime and memory footprint that is only linear in the number of individuals. FaST-LMM thus dramatically increases the size of data sets that can be analyzed with LMMs and additionally makes analyses that are currently feasible much faster.

Our FaST-LMM algorithm builds on the insight that the maximum likelihood (ML)—or alternatively, the restricted maximum likelihood (REML)—of a LMM can be rewritten as a function of just a single parameter, $\delta$, the ratio of the genetic variance (a parameter that quantifies the degree to which the random effects influence the phenotype) to the residual variance [13, 3]. Once the ML/REML value has been found for $\delta$, the values for all of the model parameters (i.e., the genetic and residual variances along with the fixed-effect parameters) follow in closed form. Consequently, the identification of the ML/REML parameters becomes an optimization problem over this single variable $\delta$. Additionally, EMMA [3] speeds up the evaluation of the log likelihood for any value of $\delta$, which is ordinarily cubic in the number of individuals, by clever use of spectral decompositions.

Although this approach provides a great improvement in runtime, the approach requires a new spectral decomposition for each SNP tested (a cubic operation). The EMMA and P3D algorithms [5, 4] provide additional computational savings by assuming that variance parameters for each tested SNP are the same. None of these approaches, however, reduce the memory footprint.

Our FaST-LMM approach requires only a single spectral decomposition to test all SNPs, even without assuming variance parameters to be the same across SNPs. Furthermore, when fewer SNPs than individuals are used to estimate genetic similarity, both the runtime and memory footprint of the single required spectral decomposition become linear in the number of individuals rather than cubic and quadratic, respectively. Thus, the computations become markedly less expensive with our approach than those used by EMMA/P3D.

A key insight behind our approach is that the spectral decomposition of the genetic similarity matrix allows the likelihood of the LMM to be refactored in such a way that it is directly analogous to the likelihood of a linear regression model (hence the name factored spectrally transformed linear mixed models). Intuitively, our algorithm algebraically transforms/rotates the target data (the phenotypes) and the input data (the SNPs to be tested and covariates) in such a way that this rotated data effectively contains pseudo-individuals who are uncorrelated, and hence can be analyzed with a linear regression model, which has a runtime and memory footprint linear in the number of individuals.

In general, the size of the required rotation matrix is quadratic in the number of individuals and computing this matrix (i.e., via a spectral decomposition) is cubic in the number of individuals. When the number of SNPs used to construct the genetic similarity matrix is less than the number of individuals, however, the size of the matrix required to perform the rotations is linear in the number of individuals (and linear in this number of SNPs), and the time required to compute it is linear in the number of individuals (and quadratic in this number of SNPs). Intuitively, these savings can be achieved because the intrinsic dimensionality of the space spanned by the individuals and SNPs used to construct the similarity matrix can never be higher than the smaller of these two values (i.e., the rank of the data matrix used to construct the similarities is at most the smaller of these two values). Thus, we can always choose to perform operations in the smaller space without any loss of information, while the computations remain exact.

To achieve linear runtime and memory footprint, the spectral decomposition of the genetic similarity matrix must be computable without the explicit computation of the matrix itself. The RRM has this property as do other matrices.

1.1 FaST-LMM

In this section, we highlight important points in the development of the maximum likelihood version of FaST-LMM. A complete description, including minor modifications needed for the REML version, is given in [14].
The LMM log likelihood of the phenotype data, \( y \) (dimension \( n \times 1 \)), given fixed effects \( X \) (dimension \( n \times d \)), including the SNP, the covariates, and the column of 1s corresponding to the bias/offset, can be written

\[
LL(\sigma^2, \sigma_g^2, \beta) = \log N(y|X\beta, \sigma^2 K + \sigma_g^2 I),
\]

where \( K \) (dimension \( n \times n \)) is the genetic similarity matrix; \( N(\mathbf{r}|\mathbf{m}, \Sigma) \) denotes a Normal distribution in \( \mathbf{r} \) with mean \( \mathbf{m} \) and covariance matrix \( \Sigma \); \( \mathbf{I} \) denotes the identity matrix; \( \sigma^2_g \) (scalar) is the magnitude of the residual variance; \( \sigma^2 \) (scalar) is the magnitude of the genetic variance; and \( \beta \) (dimension \( d \times 1 \)) are the fixed-effect weights.

To efficiently estimate the parameters \( \beta, \sigma^2, \) and \( \sigma^2_g \) and the log likelihood at those values, we can factor Equation 1 as follows. As in [3], let \( \delta \equiv \sigma^2/\sigma^2_g \) and let \( U S U^T = K \) (where \( U^T \) denotes the transpose of \( U \)) be the spectral decomposition of \( K \), so that Equation 1 becomes

\[
LL(\delta, \sigma^2, \beta) = -\frac{1}{2} \left(n \log(2\pi\sigma^2) + \log(||U(S + \delta I) U^T||) + \frac{1}{\sigma^2} (y - X\beta)^T (U(S + \delta I) U^T)^{-1} (y - X\beta)\right),
\]

where \( |K| \) denotes the determinant of matrix \( K \). The determinant of the genetic similarity matrix, \( |U(S + \delta I) U^T| \) can be written as \( |S + \delta I| \) using the properties that \( |XY| = |X||Y| \), and that \( |U| = |U^T| = 1 \). The inverse of the genetic similarity matrix can be rewritten as \( (U(S + \delta I))^{-1} U^T \) using the property that \( (XY)^{-1} = Y^{-1}X^{-1} \), that \( U^{-1} = U^T \) and \( U^{-T} = U \). Thus, after additionally pushing \( U \) out from the covariance term so that it now acts as a rotation matrix on the inputs \((X)\) and targets \((y)\), we obtain

\[
LL(\delta, \sigma^2, \beta) = -\frac{1}{2} \left(n \log(2\pi\sigma^2) + \log(||S + \delta I||) + \frac{1}{\sigma^2} ((U^T y) - (U^T X) \beta) \right)^2 (S + \delta I)^{-1} ((U^T y) - (U^T X) \beta).
\]

The “Fa” in FaST-LMM gets its name from this factorization. As the covariance matrix of the Normal distribution is now a diagonal matrix \( S + \delta I \), the log likelihood can be rewritten as the sum over \( n \) terms, yielding

\[
LL(\delta, \sigma^2, \beta) = -\frac{1}{2} \left(n \log(2\pi\sigma^2) + \sum_{i=1}^n \log(|S|_i + \delta) + \frac{1}{\sigma^2} \sum_{i=1}^n \left((|U^T|_i - |U^T X|_i) \beta \right)^2 (|S|_i + \delta)\right),
\]

where \( |X|_i \) denotes the \( i \)th row of \( X \). Note that this expression is equal to the product of \( n \) univariate Normal distributions on the rotated data, yielding the equation

\[
LL(\delta, \sigma^2, \beta) = \log \prod_{i=1}^n N(|U^T y|, |U^T X|_i \beta, \sigma^2, |S|_i + \delta).
\]

To determine the values of \( \delta, \sigma^2, \) and \( \beta \) that maximize the log likelihood, we first differentiate Equation 2 with respect to \( \beta \), set it to zero, and analytically solve for \( \beta = \hat{\beta}(\delta) \) (i.e., as a function of \( \delta \)). We then substitute this expression for \( \beta \) into Equation 2, differentiate the resulting expression with respect to \( \sigma^2 \), set it to zero, and solve analytically for \( \hat{\sigma}^2(\delta) = \hat{\sigma}^2(\delta) \). Next, we plug in \( \hat{\sigma}^2 \) and \( \hat{\beta} \) into Equation 2 so that it is a function only of \( \delta \), \( LL(\delta, \hat{\sigma}^2(\delta), \hat{\beta}(\delta)) = LL(\delta) \). Finally, we optimize this function of \( \delta \) using a one-dimensional numerical optimizer involving Brent’s method (see [14] for details on the optimization strategy used).

Note that, given \( \delta \) and having pre-computed the spectral decomposition of \( K \), each evaluation of the likelihood has a runtime that is linear in \( n \). Consequently, when testing \( s \) SNPs in a GWAS, the time complexities are \( O(n^2) \) for finding all eigenvalues (S) and eigenvectors (U) of \( K \), \( O(n^2 s) \) for rotating the phenotype vector \( y \), and all of the SNP and covariate data (i.e., computing \( U^T y \) and \( U^T X \)), and \( O(Cn) \) for performing \( C \) evaluations of the log likelihood during the one-dimensional optimization over \( \delta \). Therefore, the total time complexity of FaST-LMM, given \( K \), is \( O(n^3 + n^2 s + Cn) \).

By keeping \( \delta \) fixed to its value from the null model (analogously to EMMAx/P3D), this complexity reduces to \( O(n^3 + n^2 s + Cn) \). The size of both \( K \) and \( U \) is \( O(n^2) \), which dominates the space complexity, as each SNP can be processed independently so that there is no need to load all SNP data into memory at once. In most applications, the number of fixed effects per test, \( d \), is a single digit integer and is omitted in these expressions because its contribution is negligible.

Next we consider the case where \( K \) is of low rank—that is, \( k \), the rank of \( K \) is less than \( n \). This case will occur when the RRM is used and the number of (linearly independent) SNPs used to estimate it,
\( s_k = k \) is smaller than \( n \). For a more general exposition, wherein \( K \) is of low rank for other reasons (e.g., by forcing some eigenvalues to zero, as mentioned in the text), see [14].

Let \( USU^T = K \) be the complete spectral decomposition of \( K \). Thus, \( S \) is an \( n \times n \) diagonal matrix containing the non-zero eigenvalues on the top-left of the diagonal, followed by \( n - k \) zeros on the bottom-right, and \( U \) is an \( n \times n \) matrix of eigenvectors. Now, write the \( n \)-by-\( n \) orthonormal matrix \( U \) as \( U \equiv [U_1, U_2] \), where \( U_1 \in \mathbb{R}^{n \times k} \) contains the eigenvectors corresponding to non-zero eigenvalues, and \( U_2 \in \mathbb{R}^{n \times n-k} \) contains the eigenvectors corresponding to zero eigenvalues.

Thus, we have
\[
K = USU^T = [U_1, U_2]\begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} [U_1, U_2]^T = U_1S_1U_1^T + U_2S_2U_2^T.
\]

As \( S_2 = [0] \), \( K \) can be recovered by \( K = U_1S_1U_1^T \), the \( k \)-spectral decomposition of \( K \), so-called because it contains only \( k \) eigenvectors and arises from taking the spectral decomposition of a matrix of rank \( k \). The expression \( (K + \delta I) \) appearing in the LMM likelihood, however, is always of full rank (because \( \delta > 0 \)).

\[
K + \delta I = U(S + \delta I)U^T = U\begin{bmatrix} S_1 + \delta I & 0 \\ 0 & \delta I \end{bmatrix}U^T.
\]

Therefore, it is not possible to ignore \( U_2 \) as it enters the expression for the log likelihood. Furthermore, directly computing the complete spectral decomposition does not exploit the low rank of \( K \). Consequently, we use an algebraic trick involving the identity \( U_2U_2^T = I - U_1U_1^T \) to rewrite the likelihood in terms not involving \( U_2 \) (see [14] for details). As a result, we incur only the time and space complexity of computing \( U_1 \) rather than \( U \).

Given the \( k \) spectral decomposition of \( K \), the likelihood of the model can be evaluated with time complexity \( O(nsk) \) for the required rotations and \( O(C(n + k)s) \) for the \( C \) evaluations of the log likelihood during the one-dimensional optimizations over \( \delta \). By keeping \( \delta \) fixed to its value from the null model, as in EMMAX/P3D, \( O(C(n + k)s) \) is reduced to \( O(C(n^2)) \). In general, \( k \)-spectral decompositions can be computed by first constructing the genetic similarity matrix from \( k \) SNPs at a time complexity of \( O(n^2 s_k) \) and space complexity of \( O(n^2) \), and then finding its first \( k \) eigenvalues and eigenvectors at a time complexity of \( O(n^2 k) \). When the RRM is used, however, we can perform the \( k \)-spectral decomposition more efficiently by circumventing the construction of \( K \), because the singular vectors of the data matrix are the same as the eigenvectors of the RRM constructed from that data (e.g., [15]). In particular, we can obtain the \( k \)-spectral decomposition of \( K \) from the singular value decomposition of the \( n \times s_k \) SNP matrix directly, which is an \( O(nsk) = O(n^2 s_k) \) operation. Therefore, the total time complexity of low-rank FaST-LMM using \( \delta \) from the null model is \( O(nsk + nsk + C(n + k)) \). Assuming SNPs to be tested are loaded into memory in small blocks, the total space complexity is \( O(n s_k) \).

Finally, we note that, for both the full and low-rank versions of FaST-LMM, the rotations (and, if performed, the search for \( \delta \) for each test) are easily parallelized. Consequently, the runtime of the LMM analysis is dominated by the spectral decomposition (or singular value decomposition for the low-rank version). Although parallel algorithms for singular-value decomposition exist, improvements to such algorithms should lead to even greater speedups.

2 Experiments

We compared runtimes and memory footprint for non-parallelized implementations of the FaST-LMM and EMMAX/P3D algorithms. For the latter, we used the EMMAX implementation, which was no less efficient than P3D (in TASSEL) in terms of runtime and memory use. In the comparison, we used GAW14 data to construct synthetic datasets having roughly 1, 5, 10, 20, 50 and 100 times as many individuals, and always the same number of SNPs (approximately 8K) as the original data (see Methods). The largest such data set contained 123,800 individuals. All SNPs were used to estimate genetic similarity. Figure 1a shows the memory footprints for FaST-LMM and EMMAX when all SNPs were analyzed. EMMAX would not run on the 20x, 50x, or 100x datasets, because the memory required to store the large matrices exceeded the 32 gigabytes available. In contrast, FaST-LMM, which did not require these matrices (because it bypassed their computation, using them only
Figure 1: Memory footprint (a) and runtimes (b) of FaST-LMM running on a single processor as a function of the number of individuals in synthetic data sets based on GAW14. In each run, 7,579 SNPs were used both to estimate genetic similarity (RRM for FaST-LMM and IBS for EMMAX) and to test for association. FaST-LMM full refers to an analysis where the variance parameters were re-estimated for each test, whereas FaST-LMM refers to estimating these parameters only once for the null model, as in EMMAX/P3D. FaST-LMM and FaST-LMM full had the same memory footprint. EMMAX would not run on the datasets that contained 20 or more times the number of individuals in the GAW14 data, because the memory required to store the large matrices exceeded the 32 gigabytes available.

Figure 2: Accuracy of association p-values resulting from SNP sampling on WTCCC data for the CD phenotype. Each point in the plot shows the negative log p-values of association for a particular SNP from a LMM using a 4K SNP sample (y-axis) and all SNPs (x-axis) to measure genetic similarity. The complete set used all 338K SNPs from all but chromosome 1, whereas the 4K sample used equally spaced SNPs from these chromosomes. RRM was used for genetic similarity. All 28K SNPs in chromosome 1 were tested. Dashed lines show the genome-wide significance threshold \((5 \times 10^{-7})\). To better emphasize the importance of small p-values, correlation was computed for log p-values. A corresponding plot for an 8K sample looks essentially the same (\(\rho = 0.976\)).

implicitly), completed the analyses using 28 gigabytes of memory on the largest dataset. Figure 1b, which shows runtimes for this analysis, highlights the linear dependence of the computations on the number of individuals when the numbers of individuals exceeds the 8K SNPs used to construct the
RRM. Also note that computations remain practical within our approach even when the variance parameters are re-estimated for each test.

It is known that the LMM with no fixed effects using an RRM constructed from a set of SNPs is equivalent to a linear regression of the SNPs on the phenotype, with linear weights (i.e., SNP effects) integrated over independent Normal distributions having the same variance [9, 10]. In this view, SNP sampling can be seen as the omission of regressors, and hence an approximation. Nonetheless, linkage disequilibrium should diminish the effects of sampling. To examine this issue, we compared association p-values with and without sampling on the WTCCC data for the CD phenotype (see Methods). Specifically, we tested all SNPs on chromosome 1 while using SNP sets of various sizes from all but this chromosome—the complete set (338K), 8K, and 4K—to compute the RRM. The p-values resulting from the complete and sampled sets were quite similar on inspection (see Figure 2).

More important, the two algorithms made near identical calls of significance, using the genome-wide significance threshold of $5 \times 10^{-7}$. Namely, 24 SNPs were called significant when the complete set was used, whereas the 8K and 4K analyses labeled only one additional SNP significant and missed none. By comparison, the Armitage trend test (ATT) labeled seven additional SNPs significant and missed none. Furthermore, the $\lambda$ statistics were similar for the complete, 8K, and 4K analyses—1.137, 1.168, and 1.199, respectively—in contrast to $\lambda = 1.333$ for the ATT. Finally, using these samples to construct genetic similarity, FaST-LMM ran an order of magnitude faster than EMMAX: 23 and 53 minutes for the 8K and 4K FaST-LMM analyses, and 260 and 290 minutes for the respective EMMAX analyses.

The calibration of p-values was assessed using the $\lambda$ statistic, pervasive in GWAS. The value $\lambda$ is defined as the ratio of the median observed to median theoretical p-value, after conversion from p-value space to log-likelihood space by way of an inverse chi-square mapping [1, 16]. Values of $\lambda$ substantially greater than (less than) 1.0 are indicative of inflation (deflation).

The Genetic Analysis Workshop (GAW) 14 data [17] consisted of autosomal SNP data from an Affymetrix SNP panel and a phenotype of whether an individual smoked a pack of cigarettes a day or more for six months or more. In addition to the curation provided by GAW, we excluded a SNP when either (1) its minor allele frequency was less than 0.05, (2) its values were missing in more than 0.05 of the population, or its allele frequencies were not in Hardy-Weinberg equilibrium ($p < 0.0001$). In addition, we excluded an individual when more than 0.1 of his or her SNP values were missing. After filtering, there were 7,579 SNPs across 1,261 individuals. The data consisted of multiple races and numerous close family members—1,034 individuals in the dataset had parents or children also in the dataset.

The GAW14 data was used as the basis for creating large synthetic datasets to evaluate runtimes and memory use. Datasets GAW14.x, with $x = 1, 5, 10, 20, 50,$ and 100 were generated. Roughly, the synthetic GAW14.x dataset was constructed by “copying” the original dataset $x$ times. For each white, black, and Hispanic individual in the original data (1,238 individuals), $x$ individuals were created in the copy. The family relationships among these individuals were similarly copied from the pedigree on the real data. For each individual with no parents, data for each SNP was sampled using the race-based marginal frequency of that SNP in the original dataset. The SNPs for the remaining individuals were determined from the parental SNPs assuming a recombination rate of 38 per genome. A phenotype for each individual was then sampled from a generalized linear mixed model (GLMM) with a logistic link function whose parameters were adjusted to mimic that of the real data. In particular, the offset and genetic-variance parameters of the GLMM were adjusted so that (1) the phenotype frequency in the real and synthetic data were almost the same, and (2) the genetic variance parameter of a LMM fit to the real and synthetic data were comparable. It was assumed that there were no fixed effects. GAW14 and GAW14.1 had almost identical runtimes and memory footprints.

The Wellcome Trust Case Control Consortium (WTCCC) 1 data consisted of the SNP and phenotype data for seven common diseases: bipolar disorder (BP), coronary artery disease (CAD), hypertension (HT), Chon’s disease (CD), rheumatoid arthritis (RA), type-I diabetes (T1D), and type-II diabetes (T2D) [18]. Each phenotype group contained about 1,900 individuals. In addition, a set of approximately 1,500 controls from the UK Blood Service Control Group (NBS) was included. A second control group from the 1958 British Birth Cohort (58C) was not included, as permissions for the data precluded use by a commercial organization. Our analysis for a given disease phenotype used data from the NBS group and the remaining six phenotypes as controls. In our initial analysis,
we excluded individuals and SNPs as described both in [5] and in the primary analysis [18]. The difference between values of \( \lambda \) from an uncorrected analysis (logistic regression) and those from [5] averaged 0.01 across the phenotypes with a standard deviation of 0.01, indicating that the absence of the 58C data in our analysis had little effect on inflation/deflation. In these initial analyses, we found a substantial over-representation of \( p \)-values equal to one, and traced this to the existence of more than two thousand non-varying SNPs or single-nucleotide constants (SNCs). In addition, we found (not surprisingly) that SNPs with very low minor-allele frequencies led to skewed \( p \)-value distributions. Consequently, we employed a more conservative SNP filter, also described by the WTCCC in [18], wherein a SNP was excluded if either its minor-allele frequency was less than 0.01, it was missing in greater than one percent of individuals, or it was in the extended MHC region. After filtering, 366,940 SNPs remained. The 8K and 4K SNP sets used to estimate genetic similarity from all but chromosome 1 were created by including every forty-second and every eighty-fourth SNP, respectively, along each chromosome. In the sampling experiments, we included non-white individuals and close family members to increase the potential for confounding and thereby better exercise the LMM. In total, there were 14,925 individuals across the seven phenotypes and control. Only the phenotype CD was used, because it was the only one that had appreciable apparent inflation according to ATT \( p \)-values.

All analyses assumed a single additive effect of a SNP on the phenotype, using a 0/1/2 encoding for each SNP. All FaST-LMM runs used the RRM, whereas EMMAX used the IBS kinship matrix. Missing SNP data was mean imputed. A likelihood ratio test was used to compute \( p \)-values for FaST-LMM. Runtimes were measured on a dual AMD six core Opteron machine with a 2.6GHz clock and 32GB of RAM. Only one core was used. FaST-LMM used the AMD Core Math Library.

3 Future work

A future direction would be to apply FaST-LMM to multivariate analyses. Once the rotations have been applied to the SNPs, covariates, and phenotype, multivariate additive analyses, including ones with using penalized/regularized estimation methods, can be achieved in time linear in the number of individuals with no additional spectral decompositions or rotations. Time complexity can be further reduced by using only the top eigenvectors of the spectral decomposition to rotate the data (those with the largest eigenvalues). On the WTCCC data, use of less than 200 eigenvectors yielded univariate \( p \)-values comparable to those obtained from many thousands of eigenvectors. Combining the ideas behind FaST-LMM with extensions of LMMs may also prove worthwhile. For example, CMLM [4] is a compressed LMM wherein similarities between groups of individuals rather than individuals are used. The approach can lead to improved power. As described in [14], spectral transformations can also be used to improve the computational efficiency of compressed LMMs. Finally, the identification of associations between genetic markers and gene expression—eQTL analyses—can be thought of as multiple applications of GWAS [19], making our FaST-LMM approach applicable to such analyses.


Acknowledgments

We thank Erin Renshaw for help with implementation of Brent’s method and the \( \chi^2 \) distribution, Jonathan Carlson for help with tools used to manage the data and deploy runs on our computer cluster, and Nico Pfeifer for use of his Armitage trend test implementation. The GAW14 data were provided by the Collaborative Study on the Genetics of Alcoholism (U10 AA008401). This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.
References


