SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
GENOMICS
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
GENOME INFORMATICS
MACHINE LEARNING

Received
8 February 2013

Accepted
24 April 2013

Published
9 May 2013

Correspondence and requests for materials should be addressed to D.H. (heckerma@microsoft.com); C.L. (lippert@microsoft.com) or J.L. (jennl@microsoft.com)

* These authors contributed equally to this work.

The benefits of selecting phenotype-specific variants for applications of mixed models in genomics

Christoph Lippert[1]*, Gerald Quon[2], Eun Yong Kang[1], Carl M. Kadie[3], Jennifer Listgarten[1]* & David Heckerman[1]*

[1]eScience Group, Microsoft Research, 1100 Glendon Avenue, Suite PH1, Los Angeles, CA, 90024, United States, [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Room 32-D516, Cambridge, MA, 02139, United States, [3]eScience Group, Microsoft Research, One Microsoft Way, Redmond, WA, 98052, United States.

Applications of linear mixed models (LMMs) to problems in genomics include phenotype prediction, correction for confounding in genome-wide association studies, estimation of narrow sense heritability, and testing sets of variants (*e.g.*, rare variants) for association. In each of these applications, the LMM uses a genetic similarity matrix, which encodes the pairwise similarity between every two individuals in a cohort. Although ideally these similarities would be estimated using strictly variants relevant to the given phenotype, the identity of such variants is typically unknown. Consequently, relevant variants are excluded and irrelevant variants are included, both having deleterious effects. For each application of the LMM, we review known effects and describe new effects showing how variable selection can be used to mitigate them.

Mixed models are now being used for multiple applications in genomics, including (*i*) phenotype prediction[1–4], (*ii*) genome-wide association studies (GWAS) in the presence of confounding variables such as population structure and family relatedness[5–10], (*iii*) heritability estimation[11,12], and (*iv*) testing sets of variants for association, such as in the analysis of rare variants[13,14]. At their core, mixed models rely on the estimation of a genetic similarity matrix, which encodes the pairwise similarity between every two individuals in a cohort. These similarities are estimated from single nucleotide polymorphisms (SNPs), or other genetic variants. In the mixed model, the degree to which people are genetically similar predicts in part the degree to which their phenotypes are similar. These genetic similarity-based predictions may be based on confounding signal, such as population structure, on causal genetic variants, or on variants tagging causal variants. This variety of predictive variants speaks to the diverse problems that the mixed model is able to tackle in genetics.

Because the variants chosen to estimate genetic similarity influence the quality of the model, they therefore also influence the quality of the solution to all four of the aforementioned applications. In particular, regardless of the task at hand, precisely those variants relevant to the phenotype under investigation should be used in the determination of the genetic similarity matrix[15]. In practice, however, it is not possible to know the relevant variants, and consequently, some relevant variants are excluded while some irrelevant variants are included—both lead to model errors, but with different effects.

In this paper, we review known effects and investigate new effects of these errors across the four applications. We find that for all applications, both the exclusion of relevant variants and the inclusion of irrelevant variants have deleterious effects on the task. In addition, we show how the use of simple and practical variable-selection methods can be used to mitigate these deleterious effects in most applications. In our investigations, we find that the nature of variable selection is different for heritability estimation than for the other three applications. In particular, for heritability estimation, exclusion of relevant variants leads to a biased estimate and, separately, inclusion leads to increased variance in the estimate (with a caveat to be discussed). Consequently, variable selection may not be warranted in some circumstances—for example, when lack of bias is important.

Results

For simplicity, we concentrate on the linear form of the mixed model, the linear mixed model (LMM)[11]. We expect our results to hold for generalized linear mixed models as well. Let the vector $\boldsymbol{y}$ of length N represent the phenotype for N individuals. The LMM decomposes the variance associated with $\boldsymbol{y}$ into the sum of a linear additive genetic (σ_g^2) and residual (σ_e^2) component,

$$p(y) = \mathrm{N}\left(\boldsymbol{y}\middle|\boldsymbol{X\beta}; \sigma_e^2\boldsymbol{I} + \sigma_g^2\boldsymbol{K}\right), \quad (1)$$

where $\boldsymbol{X}$ is the $N \times Q$ matrix of Q individual covariates (e.g., gender, age) and offset term, $\boldsymbol{\beta}$ is the $Q \times 1$ vector of fixed effects, $\boldsymbol{I}$ is the $N \times N$ identity matrix, and $\boldsymbol{K}$ is the genetic similarity matrix of size $N \times N$, determined from a set of SNPs.

One can arrive at Equation 1 from several viewpoints. In one viewpoint, we model the phenotype as a linear regression of fixed SNPs $\boldsymbol{Z}$ ($N \times S$) on the phenotype, with mutually independent effect sizes $\boldsymbol{\alpha}$ distributed $\mathrm{N}\left(\boldsymbol{\alpha}\middle|0; \frac{\sigma_g^2}{S}\boldsymbol{I}\right)$. We assume the values for each SNP are standardized. In this case, the log likelihood can be written as

$$\begin{aligned}
&\log\int \mathrm{N}\left(\boldsymbol{y}\middle|\boldsymbol{X\beta} + \boldsymbol{Z\alpha}; \sigma_e^2\boldsymbol{I}\right)\cdot\mathrm{N}\left(\boldsymbol{\alpha}\middle|\boldsymbol{0}; \frac{\sigma_g^2}{S}\boldsymbol{I}\right)\mathrm{d}\boldsymbol{\alpha} \\
&= \log \mathrm{N}\left(\boldsymbol{y}\middle|\boldsymbol{X\beta}; \sigma_e^2\boldsymbol{I} + \sigma_g^2\frac{1}{S}\boldsymbol{ZZ}\right) \\
&= \log \mathrm{N}\left(\boldsymbol{y}\middle|\boldsymbol{X\beta}; \sigma_e^2\boldsymbol{I} + \sigma_g^2\boldsymbol{K}\right)
\end{aligned},$$

from which it is clear that those SNPs, $\boldsymbol{Z}$, used to estimate $\boldsymbol{K} = \frac{1}{S}\boldsymbol{ZZ}^{\mathrm{T}}$, are one and the same as those SNPs in the regression view of the model (the left-hand expression). Note that, when $\boldsymbol{K} = \frac{1}{S}\boldsymbol{ZZ}^{\mathrm{T}}$, $\boldsymbol{K}$ is called the realized relationship matrix (RRM)[16]. The use of the RRM for $\boldsymbol{K}$ is common in practice[16]; and we do so here.

In summary, the LMM using realized relationship genetic similarities constructed from a set of SNPs is mathematically equivalent to linear regression of those SNPs on the phenotype with effect sizes integrated over independent normal distributions having the same variance $\frac{\sigma_g^2}{S}$.

This viewpoint helps to understand why the LMM is able to address several of its applications. In particular, when using an LMM to predict a phenotype from a given set of SNPs, we construct the RRM using this set of SNPs, effectively using them as covariates in a linear-regression predictive model. For GWAS, we should use SNPs associated with the phenotype as covariates, which is effectively accomplished by constructing the RRM from these SNPs[9]. When testing sets of SNPs for association with an LMM, we test whether those SNPs acting jointly as covariates (or equivalently as SNPs in the RRM) improve the predictive power of the model. When we examine heritability estimation, we will consider a different viewpoint that also leads to Equation 1.

We now consider each of these applications of the LMM, examining how the exclusion of relevant SNPs and the inclusion of irrelevant SNPs affect each application, and how variable selection may mitigate the effects of exclusion and inclusion.

Prediction. In the prediction setting, we are interested in predicting an unobserved phenotype for some individuals for which we have SNP data, using a "training" dataset that includes both SNPs and phenotypes for a different set of individuals. As we mentioned, in the linear-regression view of LMMs, the SNPs used to estimate genetic similarity act as fixed effects for prediction. Furthermore, integration over the sizes of these effects acts to regularize the predictive model[17,18]. Such a regularized linear regression model is also known as Gaussian process regression[17] and Bayesian linear regression[17].

The performance of LMMs for phenotype prediction has been studied heavily in the breeding community[1,3], where it is well known that the exclusion of relevant SNPs and the inclusion of irrelevant SNPs leads to model misspecification, which in turn leads to a decrease in predictive accuracy. While these effects are well known (see Supplementary Information), we conducted experiments using synthetic data to investigate their magnitude and potential practical impact for this application. We generated several sets of synthetic SNPs and phenotypes for 3,000 individuals. For the first dataset, we generated 100 undifferentiated (mutually independent) SNPs with minor allele frequencies (MAFs) drawn from a uniform distribution on [0.05, 0.4]. We then generated the phenotype variable directly from an LMM using these SNPs. In particular, we constructed an RRM from these 100 undifferentiated SNPs, and then sampled phenotypes across individuals from a zero-mean, multivariate Gaussian with covariance set to this RRM, and covariance parameters set to $\sigma_g^2 = \sigma_e^2 = 0.1$. We refer to these 100 SNPs as causal (and relevant). Finally, we added to the dataset 80,000 undifferentiated SNPs, irrelevant to the generated phenotype.

Recent evidence has shown that highly polygenic traits are more common than originally believed[19]. Therefore, for the second dataset, we used many more causal SNPs such that the generated phenotype would be highly polygenic. In particular, we generated data as just described, except we now used 10,000 undifferentiated causal SNPs instead of 100. We refer to the first and second datasets as low and high polygenicity datasets, respectively.

After generating the synthetic data, we varied which SNPs were excluded or included in the RRM to gauge how these changes would affect prediction accuracy. In particular, we excluded increasing numbers of relevant SNPs and included increasing numbers of irrelevant SNPs at random, using ten-fold cross-validation to measure prediction accuracy by way of the out-of-sample log likelihoods. We also used a squared-error criterion, yielding similar results on all experiments. As is known[18], we found that as the number of excluded relevant SNPs or the number of irrelevant SNPs increased, out-of-sample prediction accuracy decreased substantially (Figure 1).

To study the effects of variable selection on prediction, we used a simple approach. As discussed in the next section, this approach will also be useful for GWAS. Our approach searched over various sets of SNPs to identify those that maximized cross-validated prediction accuracy. To keep the search practical, we ordered SNPs for each fold by their univariate linear-regression P values on the training data for that fold. We then used increasing numbers of SNPs by this ordering, measuring prediction accuracy on the out-of-sample test set. Next, we averaged the prediction accuracy over each fold. Finally, we identified the number of SNPs k that optimized this average. This method for variable selection (using either the log-likelihood or squared-error criterion) chose 40 for the low and all 80,100 SNPs for the high polygenicity cases, respectively (Figure 2, left column).

We next explored exclusion and inclusion of SNPs on two additional datasets with population structure, roughly emulating what might be found in typical GWAS datasets. For these generated datasets, we used an identical set of SNPs, except that we additionally included a set of either 100 or 10,000 SNPs (for the low and high polygenicity cases, respectively) from two populations, using the Balding-Nichols model[20] with a 60 : 40 population ratio, an $F_{ST} = 0.1$, and parent population MAFs drawn from a uniform distribution on [0.05, 0.4]. Finally, to inject population structure into the phenotype, we modified the earlier phenotype generation process by adding or subtracting the quantity π to the LMM-generated phenotype variable, depending on the population membership of the individual (as in the Balding-Nichols model). We used $\pi = 0.32$, corresponding to a variance of 0.1 explained by the confounding population

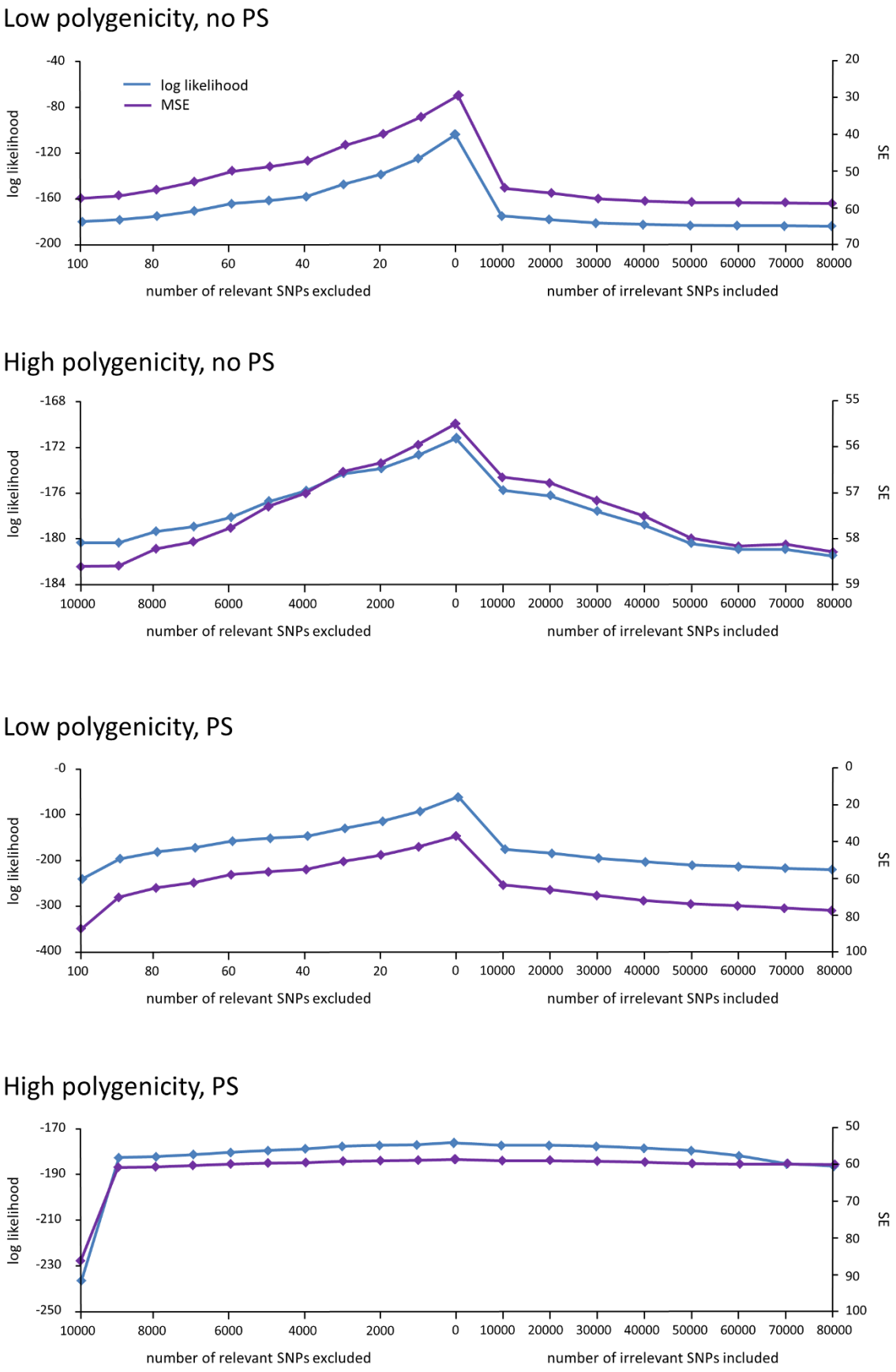

Figure 1 | The effects of excluding relevant SNPs and including irrelevant SNPs on phenotype prediction. Out-of-sample log likelihood $\log p(y|X)$ (blue) and squared error (purple) averaged over the folds of cross validation are plotted as a function of the number of relevant SNPs randomly excluded (left) and number of irrelevant SNPs randomly included (right) in the RRM.

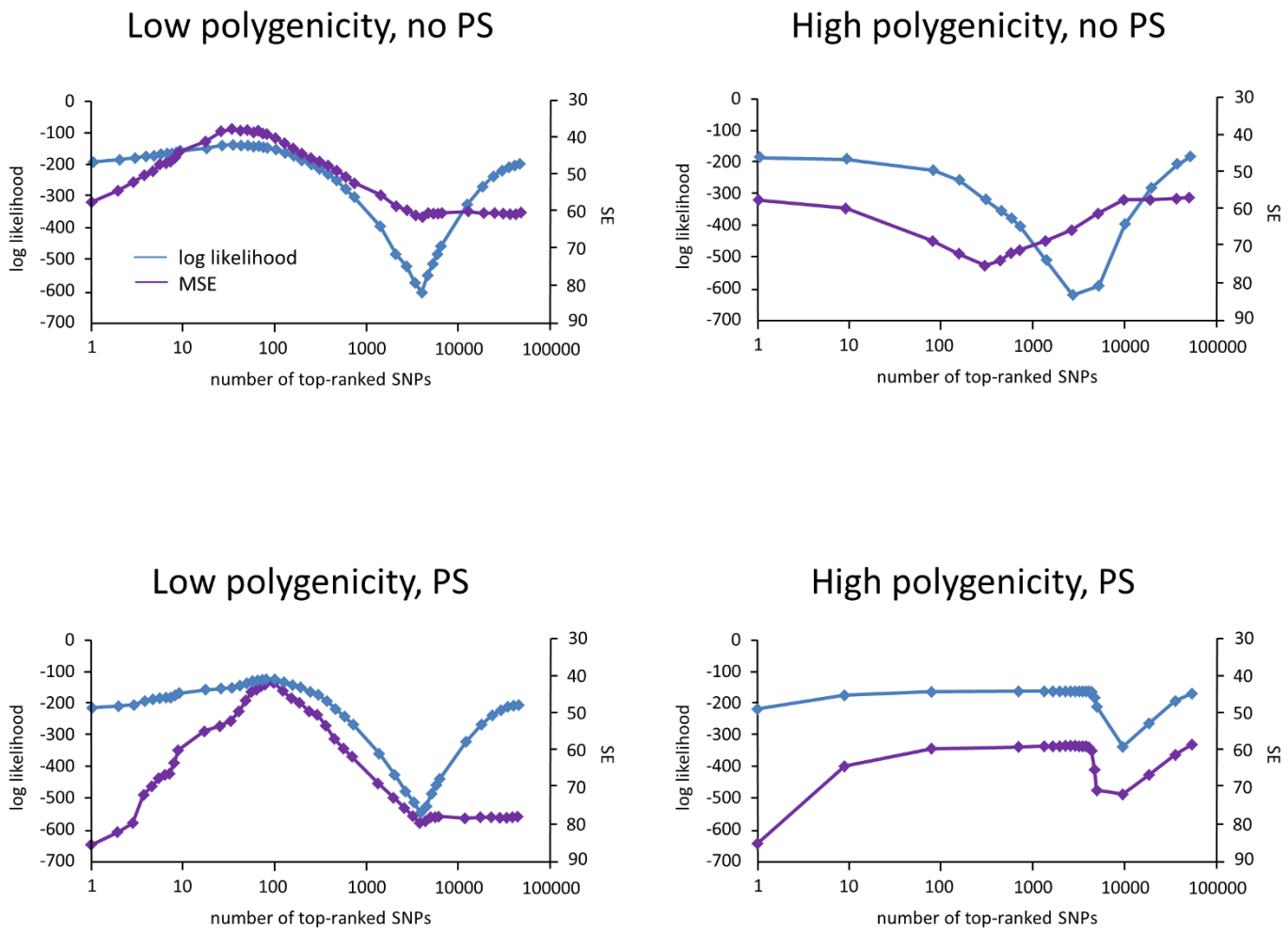

Figure 2 | Variable selection for phenotype prediction. For each fold in 10-fold cross-validation, SNPs are sorted by their univariate *P* values on the training data. Then, the top *k* SNPs are used to train the LMM. Finally, the out-of-sample log likelihood $\log p(\boldsymbol{y}|\boldsymbol{X})$ and squared error are computed using the LMM and averaged over the folds. The plots show the averaged log likelihood (blue) and squared error (purple) as a function of *k*.

variable. Note that, given this generation process, the differentiated SNPs are not causal, but are relevant to the phenotype.

For these two datasets with population structure, we again found that, as the number of (randomly) excluded relevant SNPs or the number of included irrelevant SNPs increased, out-of-sample prediction accuracy decreased substantially (Figure 1). In addition, the simple variable-selection method described previously (using either the log-likelihood or squared-error criterion) chose 125 SNPs and 900 SNPs for the low and high polygenicity case, respectively (Figure 2, right column).

Interestingly, for all four datasets (low and high polygenicity and with or without population structure), predictive accuracy as a function of *k*, the number of SNPs used to train the LMM, showed a large dip followed by an increase as *k* was increased (Figure 2). The large dip is due to overfitting of the ratio σ_e^2/σ_g^2, which is fit by an in-sample maximization of the restricted likelihood (*i.e.,* REML). When this ratio is fit with out-of-sample maximization, the magnitude of the dip is greatly decreased. For example, in the high polygenicity dataset with population structure using REML, the prediction accuracy as measured by negative squared error at its maximum (900 SNPs) and local minimum (16,000 SNPs) is −58.9 and −70.5, respectively (Figure 2). In contrast, when this ratio is fit out-of-sample, the negative squared error at 16,000 SNPs is −59.2.

Genome-wide association studies. The next application of LMMs to genomics that we consider is GWAS. LMMs have been used for both univariate[6–10,21] and set tests[13,14], but here we concentrate on univariate testing for simplicity. In this application, the strength of association between a test SNP and the phenotype is determined with, for example, an *F* test, wherein the null model is the LMM described above, and the alternative model additionally has the test SNP as a fixed effect.

Looking at GWAS from the linear-regression viewpoint, we should use covariates that include causal SNPs, SNPs that tag unavailable causal SNPs, and SNPs that are associated with the phenotype via hidden or confounding variables[9]. The inclusion of causal or tagging SNPs as covariates can improve power by reducing the model misspecification that would otherwise result from their exclusion. (Note that, when there is ascertainment bias, their inclusion can reduce power[22]. In our experiments, we generate the phenotype without ascertainment bias. For real data, care should be taken.) The inclusion of SNPs that are associated with confounding variables helps to correct for the confounding by effectively conditioning on them. In particular, typically more than one SNP is associated with a confounder. Therefore, when testing such a SNP, there will be another SNP correlated to it that is being used as a covariate, reducing the strength of association of the tested SNP and therefore avoiding a potential false-positive association. Finally, as we have discussed, using SNPs as covariates is equivalent to using them in the RRM with an LMM. Herein, we take this approach.

The extent to which these conditioning effects are beneficial is closely related to the extent to which the RRM contributes to phenotype prediction accuracy. That is, the more predictive these

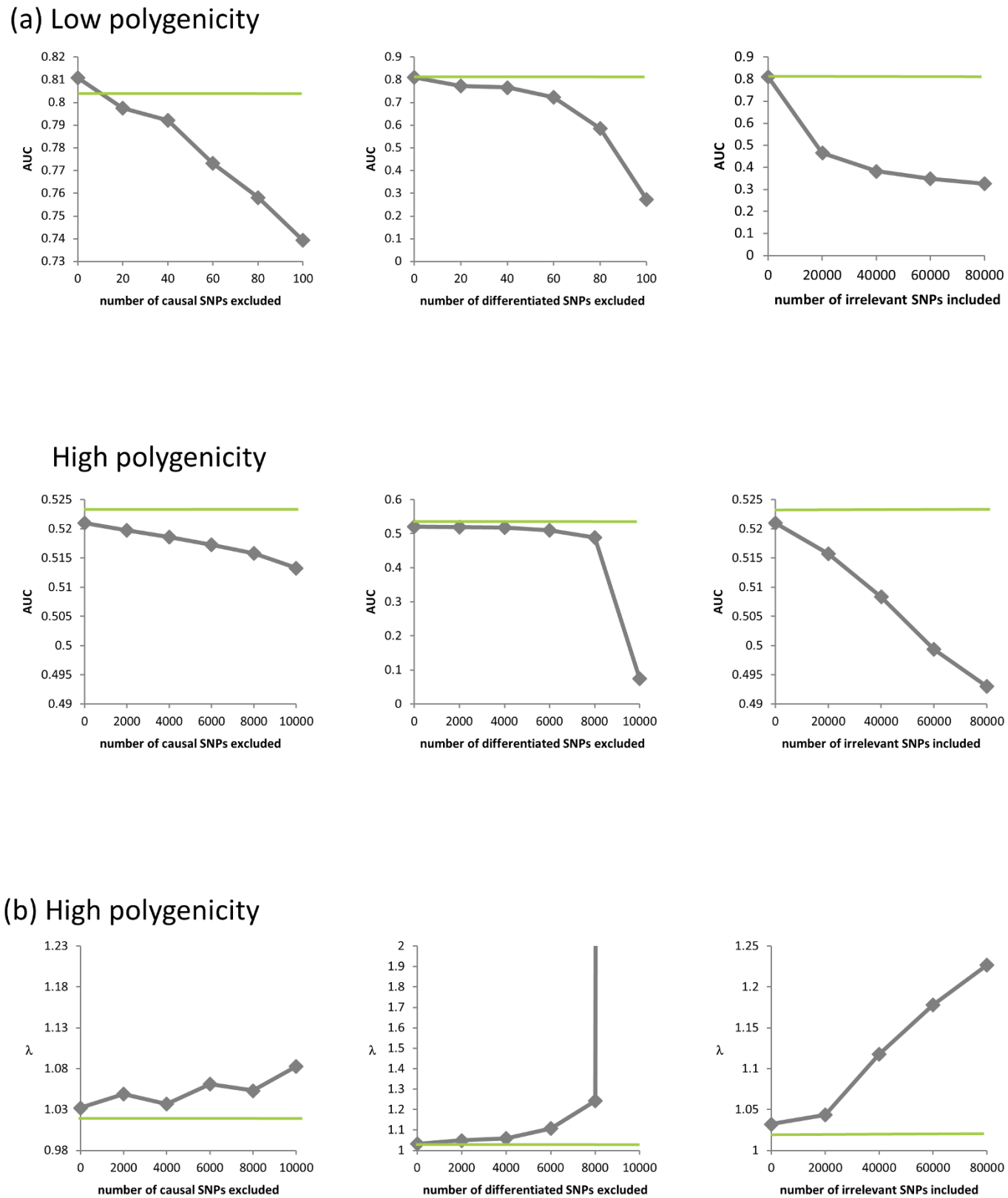

Figure 3 | The effects of excluding relevant SNPs and including irrelevant SNPs on power and inflation. (a) AUC as a function of the number of the causal SNPs excluded (with no irrelevant SNPs included), the number of differentiated SNPs excluded (with no irrelevant SNPs included), and the number of irrelevant SNPs included for the low and high polygenicity cases (including all relevant SNPs). (b) The genomic control factor λ as a function of the number of causal SNPs excluded (with no irrelevant SNPs included), the number of differentiated SNPs excluded (with no irrelevant SNPs included), and the number of irrelevant SNPs included for the high polygenicity case (including all relevant SNPs). The performance of the simple variable-selection method is indicated with green lines. The only plot with a non-monotonic pattern is the one showing λ as a function of the number of causal SNPs excluded (lower left). Nonetheless, the effect is significant in that, with 6,000 or more causal SNPs excluded, the GWAS *P* value distributions differ significantly from uniform according to a two-sided KS test (*P* values 0.047, 0.021, and 0.002 for 6,000, 8,000, and 10,000 SNPs excluded, respectively).

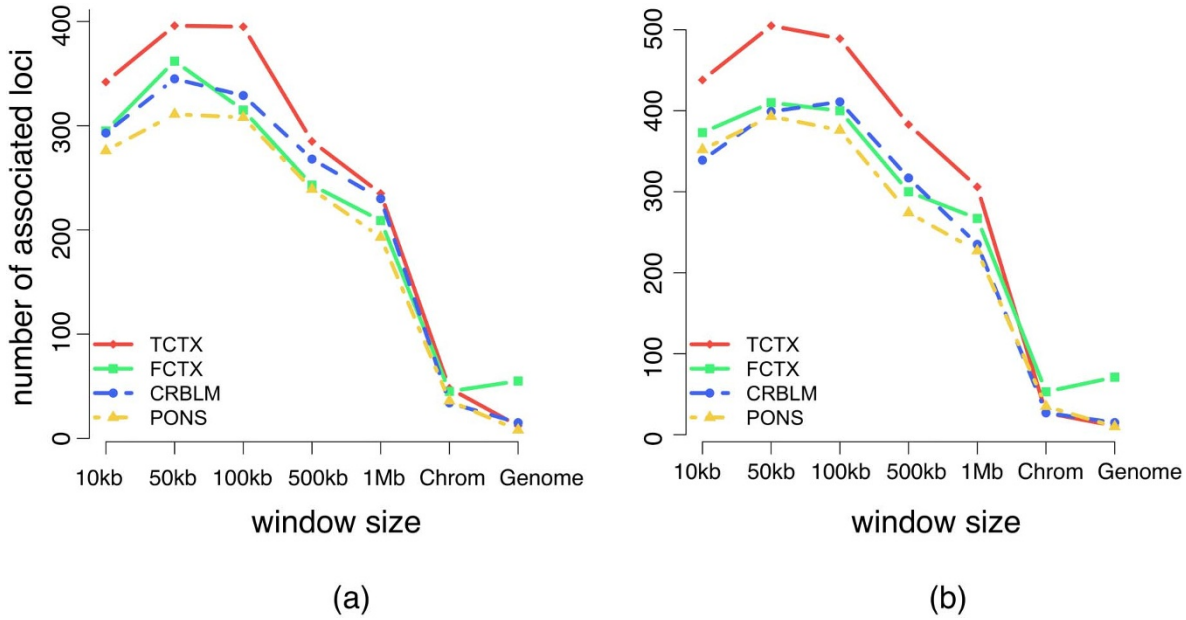

Figure 4 | Number of associated methylation loci in the four brain regions (TCTX, FCTX, CRBLM, and PONS) that pass a Bonferroni-corrected *P* value threshold of 0.05 as a function of DNA sequence window size. Only methylation loci that had at least one SNP in every window were included in the analysis so as to make the windows comparable. The plots are divided into those for even (a) and odd (b) chromosomes.

covariates are of the phenotype, the more so conditioning on them yields increased power and less inflation of the test statistic. Consequently, exclusion of relevant SNPs (causal or differentiated) and inclusion of irrelevant SNPs in the RRM, both of which create model misspecification, should diminish the beneficial conditioning effects, thereby leading to a decrease in power and an increase in inflation.

To examine these phenomena empirically, we applied the LMM to the low and high polygenicity datasets with population structure used earlier to study prediction. We computed a *P* value for the degree of association between each SNP and the phenotype. Then, to assess the effects of exclusion and inclusion of SNPs in the RRM on power, we measured the area under the curve (AUC) of the receiver operator characteristic (ROC) curve. When we measured AUC using every SNP in the dataset (including the 80,000 irrelevant SNPs), the AUC did not vary substantially with exclusion of relevant SNPs or inclusion of the irrelevant SNPs. These 80,000 irrelevant SNPs were randomly distributed in the ranking and weakened the effect of our experimental conditions on AUC. Therefore, we measured AUC using only the causal and differentiated SNPs (omitting the irrelevant SNPs), yielding AUC values that varied significantly. As expected, the exclusion of causal SNPs, the exclusion of differentiated SNPs, and the inclusion of irrelevant SNPs, all lead to decreased power (Figure 3a).

To assess the effects of test-statistic inflation from exclusion and inclusion, we measured the genomic control factor, λ, for the differentiated SNPs, which are non-causal and should have $\lambda = 1$ in expectation. Furthermore, we limited our experimental conditions to the high polygenicity dataset, as the low polygenicity dataset contained relatively few differentiated SNPs, leading to high variance estimates of λ. As expected, exclusion of causal SNPs and differentiated SNPs, as well as inclusion of irrelevant SNPs led to inflation (Figure 3b).

Given the connection between prediction and the beneficial condition effects previously discussed, we also expect the LMM to perform well for GWAS when using variable selection. To select SNPs for the RRM, we again identified the number of SNPs k yielding the best out-of-sample predictions as described previously, and then selected the k SNPs that had the smallest linear-regression *P* values on the entire dataset. This selection procedure yielded better power and control for inflation than the use of all SNPs (Figure 3). We also note that, on data from spatially structured populations with rare variants[23], this selection procedure yielded similar improvements, in line with earlier results[10]. Perhaps most interesting, for the high polygenicity dataset, variable selection outperformed the use of precisely all relevant SNPs (Figure 3). That is, excluding some relevant SNPs proved to be beneficial. This result is not surprising as some relevant SNPs will have such a small effect size that they act like irrelevant SNPs, interfering with the proper modelling of the hidden confounder.

Finally, we note that GWAS performance (power and control for inflation) was more sensitive to variable selection than was out-of-sample prediction accuracy. For example, using the high polygenicity dataset with population structure, prediction accuracy with 900 and all 100,000 SNPs was about the same (Figure 2), whereas GWAS performance was quite different (Figure 3). This difference in sensitivity could possibly lead to a situation where, due to a near tie in prediction accuracy for different sets of SNPs, our method would select a set of SNPs with suboptimal GWAS performance. Although we have yet to encounter such a situation in our experiments (including many not reported here), this sensitivity could be mitigated by running GWAS multiple times, each time using the number of ordered SNPs corresponding to one of the near ties in prediction accuracy, and then using the analysis that yields the smallest λ.

Heritability estimation. In this section, we consider the estimation of narrow-sense heritability of a phenotype in a population: the fraction of variance explained by additive genetic effects, excluding effects of dominance and epistasis. We begin with the assumption that the phenotype for individual j, denoted y_j, is the sum of additive SNP effects:

$$y_j = \mu + \sum_i \alpha_i z_{ij} + \epsilon_j, \qquad (2)$$

where μ is an offset, z_{ij} represents the value of SNP i for individual j, α_i is the fixed-effect size relating the value of SNP i to the phenotype, and ϵ_j is the effect of the environment, assumed to be random with distribution $N\left(0,\sigma_e^2\right)$. Let $\boldsymbol{z}_i$ denote the vector of z_{ij} for all individuals

j. In contrast to our previous viewpoint, we now consider the $\boldsymbol{z}_i$ to be random, such that the collection of $\boldsymbol{z}_i$ and ϵ_j are mutually independent, and where the distribution of each $\boldsymbol{z}_i$ is $N(0,\boldsymbol{K}_c)$. Here, $\boldsymbol{K}_c$ is the matrix containing the probability of identity by descent between pairs of individuals at each causal variant.

From Equation 2, using the fact that the diagonal entries of $\boldsymbol{K}_c$ are one, it follows that

$$\mathrm{var}(y_j) = \sum\nolimits_i \alpha_i^2 + \sigma_e^2.$$

Letting $\sigma_g^2 \stackrel{\mathrm{def}}{=} \sum_i \alpha_i^2$, we then obtain the standard formula for narrow-sense heritability,

$$h^2 = \frac{\sum_i \alpha_i^2}{\mathrm{var}(y_j)} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}.$$

Now consider the relationship between Equation 2 and an LMM. Integrating out the z_{ij} from Equation 2, we obtain

$$p(y) = \mathrm{N}\left(y \middle| 0; \sigma_e^2 \boldsymbol{I} + \sigma_g^2 \boldsymbol{K}_c\right).$$

Also, the scatter matrix $\widehat{\boldsymbol{K}_c} = \frac{1}{S}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$, where $\boldsymbol{Z}$ are the observed values of the SNPs, is a consistent estimate for $\boldsymbol{K_c}$. Thus, in this new viewpoint, we return to Equation 1 (with no fixed effects). Consequently, we can apply Equation 1 to estimate σ_g^2 and σ_e^2, and hence h^2. We denote such an estimate $\widehat{h^2}$. Finally, note that the two key measures of performance of heritability estimation are the bias and variance of $\widehat{h^2}$.

This approach for estimating narrow-sense heritability is different from the more traditional pedigree-based approach[24]. An advantage of using the LMM over the pedigree-based approach is that we can use SNP data for distantly related individuals, thus mitigating confounding effects from the environment[11,12]. In fact, when estimating heritability with an LMM, Yang *et al.*[11] advocate the explicit removal of closely related individuals based on the similarity of their SNPs. Also note that using only distantly related individuals mitigates confounding by epistatic and dominance effects[12,25].

Because variances are non-negative, h^2 is bounded by 0 and 1. Estimation procedures that enforce these bounds (such as constrained REML) yield biased estimates with lower variance, when compared with unconstrained estimation[26]. These effects can easily be seen for the case where the true heritability is zero. In this case, for finite data, estimates of heritability will be near zero but always non-negative, and hence the expected estimate of heritability will be greater than zero. Similarly, the variance of constrained estimates will be lower than for unconstrained estimates. To quantify this effect, we generated datasets as described above with no population structure, 100 causal SNPs, and total variance of 0.2 for 500 individuals. First, we generated 1,000 datasets with a heritability of 0.5, far from the bounds. When using constrained REML, the average difference between h^2 and $\widehat{h^2}$ was 0.01, well within the standard deviation of $\widehat{h^2}$ equal to 0.05, consistent with no bias. Next, we generated 1,000 datasets with a heritability of 0 (on the boundary), yielding an average difference between h^2 and $\widehat{h^2}$ of 0.015, larger than the standard deviation of $\widehat{h^2}$ equal to 0.013, indicating bias and lower variance.

As in the other applications of the LMM, we do not know which SNPs are relevant in practice. Consequently, it is likely that relevant SNPs will be excluded and irrelevant SNPs will be included when performing the estimation. To examine bias and variance due to excluding causal (relevant) SNPs, we used the 1,000 datasets with $h^2 = 0.5$ described in the previous paragraph, but used 100, 80, 60, 40, 20, and 1 randomly selected SNPs for the RRM. The resulting heritability estimates had averages and standard deviations (in parentheses) of 0.49 (0.05), 0.39 (0.05), 0.29 (0.05), 0.19 (0.05), 0.09 (0.04), and 0.01 (0.02), respectively. Here, there was downward bias from exclusion, and a fairly constant level of variance, except for a decrease in variance near the boundary $h^2 = 0$. Downward bias with fairly constant variance has been demonstrated previously[11].

To examine bias and variance due to including irrelevant SNPs, we report the experiments of Zaitlen and Kraft[12], who generated data as we did, except they generated 10 causal SNPs using a MAF range of [0.05,0.5] and a cohort size of 1,500. Adding 100, 1,000, and 5,000 irrelevant SNPs to the RRM, consistently yielded an average $\widehat{h^2}$ value of 0.50, but the standard deviations were 0.018, 0.025, and 0.067, respectively. The variance increased from inclusion, but bias remained at zero.

In summary, exclusion of relevant SNPs leads to a biased estimate of heritability with little effect on variance, whereas the inclusion of irrelevant SNPs leads to an increase in the variance of the estimate with little effect on bias. These effects are in contrast to the applications of phenotype prediction and GWAS. Roughly, for heritability estimation, exclusion affects bias and inclusion affects variance separately, whereas, in the other applications, exclusion and inclusion both degrade common measures of performance. Consequently, although variable selection is typically useful for phenotype prediction and GWAS, it may not be useful for heritability estimation. For example, when lack of bias is extremely important, variable selection should be avoided. Of course, if one is looking to balance the bias and variance of a heritability estimate, variable selection can still be useful. In this case, variable selection is not simple, because, for a given set of SNPs, one typically obtains a single estimate of $\widehat{h^2}$ and its variance[27]. Nonetheless, variable selection could lead to scenarios where the trade-off between bias and variance can reasonably be made. For example, when two sets of variables both yield $\widehat{h^2}$ that are relatively close together, but one has much lower estimated variance, then the set of SNPs leading to lower variance may be preferred. In contrast, when two sets of variables both yield $\widehat{h^2}$ variances are that relatively close together, but one has much lower $\widehat{h^2}$ (due to the downward bias from exclusion), then the set of SNPs leading to higher $\widehat{h^2}$ may be preferred.

Set tests. The last use of the mixed model that we examine is determining whether there is a significant association between a phenotype and a set of variants. This task is sometimes referred to as a set test in GWAS[13,14]. Examples of this task include determining whether a set of rare variants are associated with a phenotype, and whether a set of SNPs in a gene or pathway are associated with a phenotype.

As in our discussion of heritability estimation, let us assume that there is no population structure in the data. In this case, to test for association, we build an RRM with the given set of SNPs, and test whether σ_g^2 (or equivalently h^2) is greater than zero. As we have seen in the previous section, as the number of relevant SNPs excluded is increased, $\widehat{h^2}$ will become increasingly downwardly biased, thereby decreasing the power to detect $\sigma_g^2 > 0$. In addition, as the number of irrelevant SNPs is increased, the variance of $\widehat{h^2}$ will increase, thereby again decreasing the power to detect $\sigma_g^2 > 0$. This last effect can be understood by noting that the increased variance in $\widehat{h^2}$ arises from a flatter likelihood surface with a lower maximum, which in turn translates to smaller differences between the null and alternative model maximum values of the restricted likelihood.

Often, the set to be tested for association is identified through prior information, leaving no opportunity for variable selection. In some situations, however, there is an opportunity for variable selection. For example, consider the work of Quon *et al.*[28], where they investigated the role played by stretches of *cis*-DNA sequence in influencing human methylation levels for four distinct brain regions, across 150 unrelated individuals. Although they were interested in the influence of *cis*-DNA on methylation loci, they had little prior knowledge on the width of the *cis* window that contained the bulk of causal SNPs. Consequently, they performed variable selection by way of a window-size selection procedure—considering windows centered on each methylation locus of increasing size. As they increased the

window size, the number of loci associated with SNPs first increased and then decreased, yielding a maximum at the window size of 50 kb. This pattern is consistent with our understanding. Namely, the initial increase can be attributed to increasing power due to decreasing downward bias in $\widehat{h^2}$, and the subsequent decrease can be attributed to decreasing power due to increasing variance in $\widehat{h^2}$.

When performing variable selection for set tests, one should be careful to guarantee that the selection criterion and the test statistic (σ_g^2 in our case) are independent under the null hypothesis. For the tests with methylation data, this condition can be achieved by using half of the methylation loci (*e.g.,* on one set of chromosomes) to identify the window size that yields the most associated loci, and then applying that window size to the second half of methylation loci. When we applied this procedure to the methylation data processed by Quon *et al.*, we found the window size that maximized the number of identified loci to be the same for the two sets of loci when aggregated across all tissues: 50 kb. When performing the analysis on each brain region separately, among the eight sets tested (four regions for each of the two partitions), there was only one that did not choose 50 kb (Figure 4).

Discussion

We have examined the deleterious effects of excluding relevant variants and including irrelevant variants specific to a given phenotype in the estimation of the genetic similarity matrix, and shown how simple variable-selection methods can mitigate these effects. For the problem of narrow-sense heritability estimation, we have shown that exclusion of relevant variants leads to a biased estimate of heritability, whereas inclusion of irrelevant variants separately leads to an increase in the variance of the estimate, making variable selection undesirable in some circumstances.

As mentioned, the effects of excluding relevant variants and including irrelevant variants for phenotype prediction has been studied intensely by the breeding community. This community has studied variable-selection methods for balancing these effects, including a method known as "SNP pre-selection"[3], which is somewhat related to the method we examined here. The breeding community has also developed methods related to, but beyond variable selection, involving Bayesian model averaging[1,3]. We have concentrated on variable selection here, as model averaging is less likely to scale to the extremely large cohort sizes anticipated in human genomics. The effects of including irrelevant variants for handling confounding variables in GWAS has also been studied (see the discussion of "dilution" in Listgarten *et al.*[9]), although the effects are better characterized here.

Finally, in order to obtain a basic understanding of exclusion, inclusion, and variable selection across these four applications of the LMM, we have ignored issues including non-linear effects, epistasis, non-Gaussian distributions associated with case-control studies, linkage disequilibrium among variants, other forms of genetic relatedness such as family and cryptic relatedness, and ascertainment bias. Although work has addressed some of these aspects for some of the four applications of the LMM, e.g., Ref. 12,22,29–31, further study across all applications are a source of promising investigation.

Methods

All analyses assumed an additive effect of a SNP on the phenotype, using a 0/1/2 encoding for each SNP (indicating the number of minor alleles for an individual). For the real data, missing SNP data were mean imputed.

Publically available FaST-LMM software[8,9] (http://mscompbio.codeplex.com/) was used for all computations. All inference was performed using REML. To compute a *P* value for whether a set of SNPs was associated with a given phenotype, we set $\sigma_g^2 = 0$ to obtain the likelihood of the null model, and then used a likelihood ratio statistic along with permutation tests to obtain *P* values[28]. The same 420,000 permutations of the individuals were used for each methylation locus.

For GWAS, *P* values were computed using an F test. In all experimental conditions, the SNP tested was excluded from the RRM to avoid proximal contamination[8,9].

The calibration of *P* values was assessed using the genomic control factor, λ[32]. The value λ is defined as the ratio of the median observed to median theoretical test statistic. When there is no signal in the data, a calibrated result corresponds to $\lambda = 1.0$, and values of λ substantially greater than 1.0 are indicative of inflation.

Methylation data were prepared as described in Quon *et al.*[28]. Briefly, individual SNP data and chromosomal coordinates were downloaded from dbGAP Study Accession phs000249.v1.p1. Normalized methylation levels across four brain regions (cerebellum (CRBLM), frontal cortex (FCTX), caudal pons (PONS), and temporal cortex (TCTX)) from 150 individuals were obtained from GEO accession GSE15745. This data profiled methylation levels of 27,578 CpG loci assayed using an Illumina HumanMethylation27 BeadChip. Methylation locus chromosome coordinates were obtained from GEO (GPL8490). All SNPs missing in more than 1% of the individuals, or those whose minor allele frequency was less than 0.01 were discarded. All individuals missing more than 5% of their SNP data were removed. Several methylation loci and individual samples were removed due to data quality concerns (see Supplementary Information of Gibbs *et al.*[33]). Individual covariate data, including age, gender post mortem interval, region source, and methylation assay batch, was obtained from Supplementary Table S1 from Gibbs *et al.*[33], and converted to a 1-of-(M-1) encoding for discrete variables.

1. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
2. Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B. & De los Campos, G. Beyond missing heritability: prediction of complex traits. *PLoS Genetics* **7**, e1002051 (2011).
3. Moser, G., Tier, B., Crump, R. E., Khatkar, M. S. & Raadsma, H. W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics, Selection, Evolution: GSE* **41**, 56 (2009).
4. Goddard, M. E., Wray, N. R., Verbyla, K. & Visscher, P. M. Estimating Effects and Making Predictions from Genome-Wide Marker Data. *Statistical Science* **24**, 517–529 (2009).
5. Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M. & Holland, J. B. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208 (2006).
6. Kang, H. M., Zaitlen, N. a., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. & Eskin, E. Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
7. Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. a., Kong, S.-Y., Freimer, N. B., Sabatti, C. & Eskin, E. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354 (2010).
8. Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. & Heckerman, D. FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011).
9. Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E. & Heckerman, D. Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525–526 (2012).
10. Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature Genetics* **45**, 470–471 (2013).
11. Yang, J., Benyamin, B., Mcevoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G. & Montgomery, G. W. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, (2010).
12. Zaitlen, N. & Kraft, P. Heritability in the genome-wide association era. *Human Genetics* **131**, 1655–1664 (2012).
13. Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* **89**, 82–93 (2011).
14. Listgarten, J., Lippert, C., Kang, E. Y., Xiang, J., Kadie, C. & Heckerman, D. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics*, doi:10.1093/bioinformatics/btt177 (2013).
15. Vilhjálmsson, B. J. & Nordborg, M. The nature of confounding in genome-wide association studies. Nature Reviews. *Genetics* **14**, 1–2 (2012).
16. Hayes, B. J., Visscher, P. M. & Goddard, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research* **91**, 47–60 (2009).
17. Review, P. A., Random, G. & Tech-, C. F. C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X. c 2006 Massachusetts Institute of Technology. www. GaussianProcess.org/gpml (2006).
18. Bernardo, J. & Smith, A. *Bayesian Analysis* (Chichester: John Wiley) (1994).
19. Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., Ingelsson, E., O'Connell, J. R. & Mangino, M. *et al.* Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics: EJHG* **19**, 807–812 (2011).
20. Balding, D. J. & Nichols, R. a. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12 (1995).
21. Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M. & Holland, J. B. *et al.* A unified

mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208 (2006).
22. Zaitlen, N., Lindström, S., Pasaniuc, B., Cornelis, M., Genovese, G., Pollack, S., Barton, A., Bickeböller, H., Bowden, D. W. & Eyre, S. *et al.* Informed Conditioning on Clinical Covariates Increases Power in Case-Control Association Studies. *PLoS Genetics* **8**, e1003032 (2012).
23. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics* **44**, 243–246 (2012).
24. Lange, K. Mathematical and statistical methods for genetic analysis (New York: Springer) (2002).
25. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* (2012).
26. Searle, S. R., Casella, G. & McCulloch, C. *Variance Components,* Volume **631**, (Wiley-Interscience) (2006).
27. Cramér, H. *Mathematical methods of statistics* (Princeton: Princeton University Press) (1946).
28. Quon, G., Lippert, C., Heckerman, D. & Listgarten, J. Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic Acids Research*.
29. Li, Q. & Ã, K. Y. Improved Correction for Population Stratification in Genome-wide Association Studies by Identifying Hidden Population Structures. *October* 1–12 (2007).
30. Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A better coefficient of determination for genetic profile analysis. *Genetic Epidemiology* **36**, 214–224 (2012).
31. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* **91**, 1011–1021 (2012).
32. Devlin, A. B., Roeder, K. & Devlin, B. Genomic Control for Association. **55**, 997–1004 (2008).
33. Gibbs, J. R., Van der Brug, M. P., Hernandez, D. G., Traynor, B. J. & Nalls, M. a, Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I. P., Troncoso, J. et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genetics* **6**, e1000952 (2010).

Acknowledgments

We thank Alkes Price, Peter Visscher, and Noah Zaitlen for useful comments regarding variable selection in genome-wide association studies. Funding support for the Brain eQTL Study (dbGaP phs000249.v1.p1) was provided through the Division of Aging Biology and the Division of Geriatrics and Clinical Gerontology, NIA. The Brain eQTL Study includes a genome-wide association study funded as part of the Intramural Research Program, NIA.

Author contributions

C.L. and J.L. designed research, contributed analytic tools, and wrote the paper. G.Q. conducted experiments. E.K. and C.K. contributed analytic tools. D.H. designed research, conducted experiments, contributed analytic tools, analyzed data, and wrote the paper. All authors reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Lippert, C. *et al.* The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Sci. Rep.* **3**, 1815; DOI:10.1038/srep01815 (2013).