Extraneous markers used for genetic similarity leads to loss of power in GWAS and heritability determination

Christoph Lippert[1*], Gerald Quon[1], Jennifer Listgarten[1*], and David Heckerman[1*]

[1]eScience Group, Microsoft Research, 1100 Glendon Avenue, Suite PH1, Los Angeles, CA, 90024, United States

[*]These authors contributed equally.

ABSTRACT

An important task in the genetic understanding of disease is determining whether there is a significant association between a phenotype and a set of markers. This task is sometimes referred to as a group test in a genome-wide association study. Examples of this task include determining whether a set of rare variants are associated with a phenotype, and whether a set of single nucleotide polymorphisms in a gene are associated with a phenotype. Another important task is determining whether a phenotype is significantly heritable (in the narrow sense) relative to a given set of genetic markers. We show that the two tasks can be formulated as the same statistical test. In addition, we show that the inclusion of extraneous markers in the set under consideration leads to substantial loss in power to detect heritability or association, a phenomenon we call *dilution*.

MAIN TEXT

An important task in the genetic understanding of disease is determining whether there is a significant association between a phenotype and a set of markers. This task is sometimes referred to as a group test in a genome-wide association study (GWAS). Examples of this task include determining whether a set of rare variants are associated with a phenotype, and whether a set of single nucleotide polymorphisms (SNPs) in a gene are associated with a phenotype.

Another important task is determining whether a phenotype is significantly heritable (in the narrow sense) relative to a given set of genetic markers. An example comes from (1) who investigated the role played by stretches of *cis*-DNA sequence in influencing human methylation data for four distinct brain regions, across 150 unrelated individuals. For each methylation locus, they considered sets of SNPs from increasingly larger windows centered on the methylation locus, including a whole chromosome and all SNPs on the chip set, asking whether the locus was heritable with respect to markers in the set. Aggregating over all loci, they found that, as the window size is increased, the number of loci found to be significantly heritable first increased and then decreased, yielding an optimal window size.

Here, we show that the two tasks, group tests in GWAS and heritability determination, can both be formulated as a variance component test in the same statistical model. In addition, we show that the

inclusion of extraneous markers in the set under consideration leads to substantial loss in power to detect heritability or association, a phenomenon we call *dilution*. We relate this effect to the finding of an optimal window size in the analysis of the methylation data and to previous observations in GWAS.

We formulate the two tasks as the same test of significance using linear mixed models (LMMs) (2). Let the vector $\boldsymbol{y}$ of length *N* represent the phenotype for *N* individuals. Using LMMs, we can decompose the variance associated with $\boldsymbol{y}$ as the sum of a linear additive genetic (σ_g^2) and residual (σ_e^2) component,

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

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 covariate effects, $\boldsymbol{I}$ is the $N \times N$ identity matrix, and $\boldsymbol{K}$ is the appropriately scaled realized relationship matrix (RRM) (3) of size $N \times N$. Note that $\boldsymbol{K}$ factors as $\boldsymbol{K} \equiv \boldsymbol{WW}^T$, where $\boldsymbol{W}$ of dimension $N \times s$ contains the s markers in the given set, and that when $s < N$, parameter estimation and computation of the log likelihood becomes extremely efficient (4). Assuming there are no covariates, a SNP-based estimate of narrow-sense heritability for phenotype $\boldsymbol{y}$ is given by (2)

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}.$$

To determine whether the phenotype is heritable with respect to this set of markers, we can compute a *P* value on the hypothesis that h^2 is greater than zero. To determine whether the phenotype is significantly associated with the SNPs, we can compute a *P* value for the hypothesis that σ_g^2 is greater than zero. The latter formulation follows from the equivalence between the linear mixed model and linear regression with random covariates (5). Finally, because h^2 is greater than zero precisely when σ_g^2 is greater than zero, the two tests are identical.

Given this formulation, consider a situation where both causal and extraneous SNPs are used to compute the RRM. As the number of extraneous SNPs is increased, the variance of the estimate of h^2 (and σ_g^2) will increase due to the randomness introduced by the extraneous SNPs. As noted above, we call this phenomenon *dilution*.

To illustrate this increase in variance, we generated mutually independent SNPs from binomial distributions with minor allele frequencies (MAFs) drawn from a uniform distribution on [0.05, 0.4]. We then generated the phenotype variable from an RRM built from a subset of 100 of these SNPs by sampling from a zero-mean Gaussian with covariance set to the RRM. We refer to these generating SNPs as causal SNPs. For each generated data set, rather than compute the variance of the estimate of h^2, we plotted the likelihood of the data as a function of h^2. In other work, we have found such plots to provide a better picture of the variability of h^2, than a single number (in

submission). Holding the number of causal SNPs constant while increasing the number of extraneous SNPs, we see that the variability in the estimate for h^2 increases with the number of extraneous SNPs (Figure 1). Note that the increase in variance is more dramatic for smaller cohort size (N).

As mentioned, dilution can be used to understand the observation in Quon *et al.* (2012). Returning to this example, they centered a window symmetrically around each methylation locus, extending the size of this window through 10 kb, 50 kb, 100 kb, 500 kb, 1 Mb, and also tried the entire local chromosome, as well as the entire genome. They then deemed the optimum window to be the one yielding the largest number of heritable methylation loci among those loci which had at least one SNP for every window size. As shown in Figure 2, a window size of 50 kb led to the highest number of heritable methylation loci. While the number of heritable loci was similar for both the 50kb and 100kb windows, it is clear that using too large of a window (*e.g.* the entire genome), or too small of a window (e.g. 10kb or less), dramatically reduced the number of heritable loci.

The effect can be seen as a bias-variance trade-off in the estimation of heritability. In particular, most SNPs influencing a methylation locus are expected to be physically near to the locus (i.e.,are cis-acting), and can therefore be captured by a relatively small window such as the 50 kb. With a smaller window, many influential SNPs are likely to be missed, causing a downward bias in the estimate of $\sigma^2_{\mathrm{i,t,g}}$ and therefore of heritability. With increasing window sizes, more and more extraneous SNPs are included in the RRM, causing an increase in the variance of the estimate of heritability due to dilution. Therefore, in our analysis, as we included more and more SNPs up to and including a window which contained most influential SNPs (i.e., the 50 kb window), the downward bias on heritability decreased (and the estimate of heritability increased). As we went beyond this optimal window size, an increasing proportion of extraneous SNPs were included in the RRM, up until the point where the variance of the estimate of heritability almost completely diminished our power to detect significantly heritable loci. This bias-variance trade-off is evident is plots of likelihood versus h^2 (Figure 3).

Finally, we note that a related phenomenon affects univariate GWAS analyses in the presence of confounding, wherein a *P* value for the association between a single SNP and the phenotype is computed using a linear mixed model with an RRM used to account for the confounding signal. In this case, the addition of extraneous SNPs in the RRM leads to increased variance in the estimate of SNP effect sizes. As a consequence, GWAS *P* values are inflated and power is lost [cite NG].

MATERIALS AND METHODS

The methylation data were prepared as described in Quon *et al* (1). 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.* (6)). 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.* (6), and converted to a 1-of-(M-1) encoding for discrete variables. When scanning *cis* window sizes, they restricted their comparison to the 15,179 methylation loci for which they could find at least one SNP within each of the window sizes considered.

To compute a *P* value for whether a given phenotype ***y*** was heritable—that is, to compute the significance of the genetic variance component in the model—we set $\sigma_g^2 = 0$ to obtain the null model, and then used a likelihood ratio statistic, along with permutations tests to obtain *P* values. The same 420,000 permutations of the individuals were used for each methylation locus. We used the method of Lippert *et al.* (4) to compute likelihoods. Note that have since developed more efficient methods for obtaining *P* values in this setting (in submission).

REFERENCES

1. Quon,G., Lippert,C., Heckerman,D. and Listgarten,J. Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic Acids Research*.

2. 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. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**.

3. Hayes,B.J., Visscher,P.M. and Goddard,M.E. (2009) Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research*, **91**, 47–60.

4. Lippert,C., Listgarten,J., Liu,Y., Kadie,C.M., Davidson,R.I. and Heckerman,D. (2011) FaST linear mixed models for genome-wide association studies. *Nature Methods*, **8**, 833–835.

5. Listgarten,J., Lippert,C. and Heckerman,D. (2012) Fast-LMM-Select tackles confounding from spatial structure and rare variants. *Nature Genetics*, **in press**.

6. 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. (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS genetics*, **6**, e1000952.

FIGURES

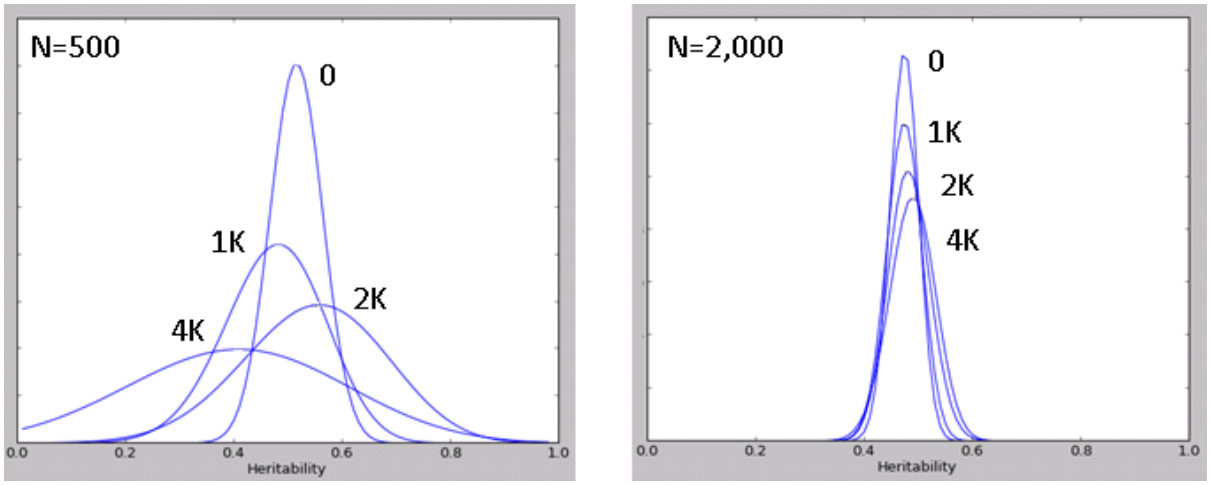

Figure 1. Relative data likelihood as a function of $\boldsymbol{h^2}$ for synthetic data with ten causal SNPs and increasing numbers of extraneous SNPs, indicated adjacent to each curve. Cohort size is N.

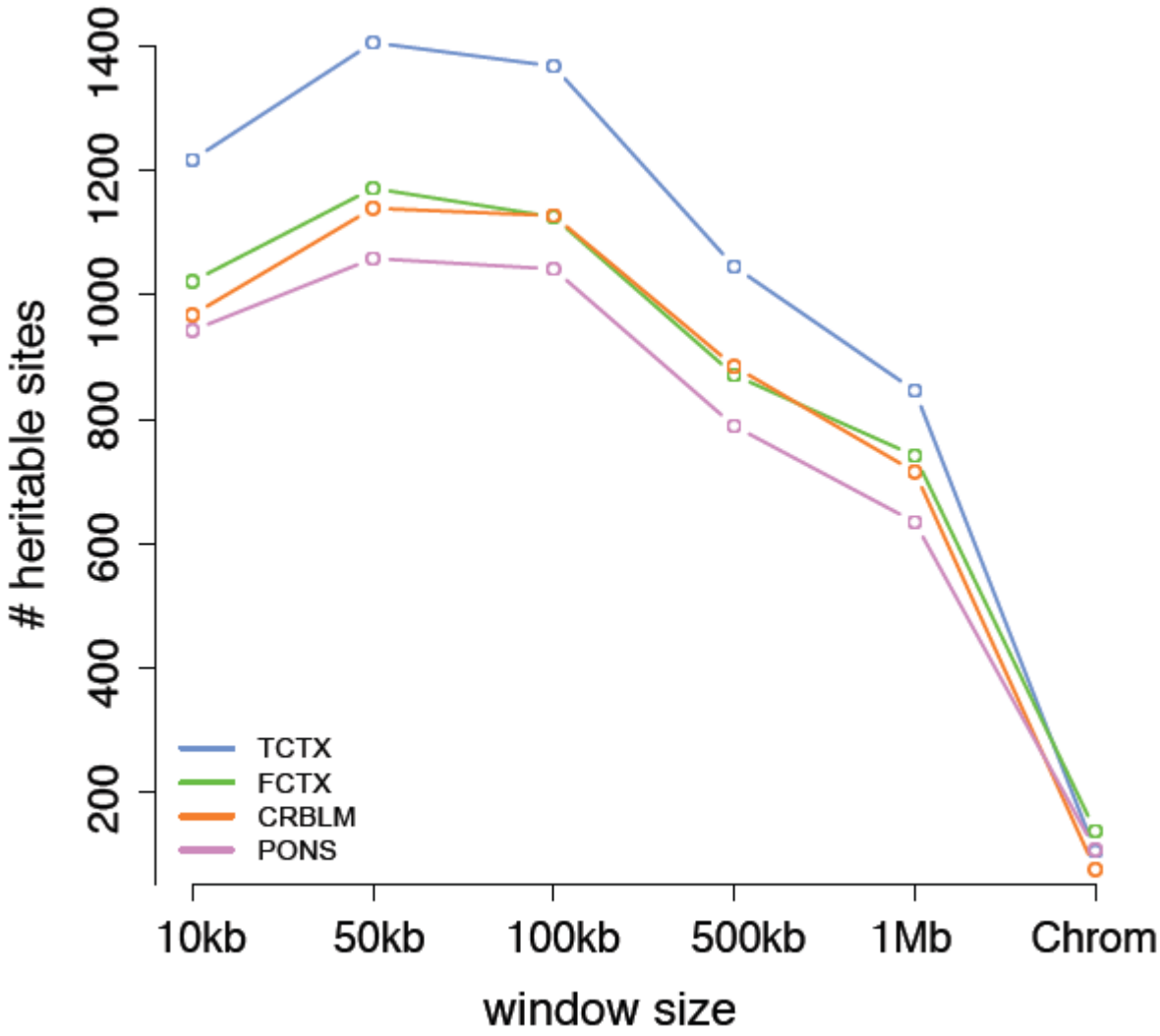

Figure 2. Number of heritable methylation loci in the four brain regions: TCTX, FCTX, CRBLM, and PONS, passing a Bonferroni-corrected *P* value threshold of 0.05, as a function of DNA sequence window size. We used only methylation loci analyzed for all window sizes so as to make them comparable.

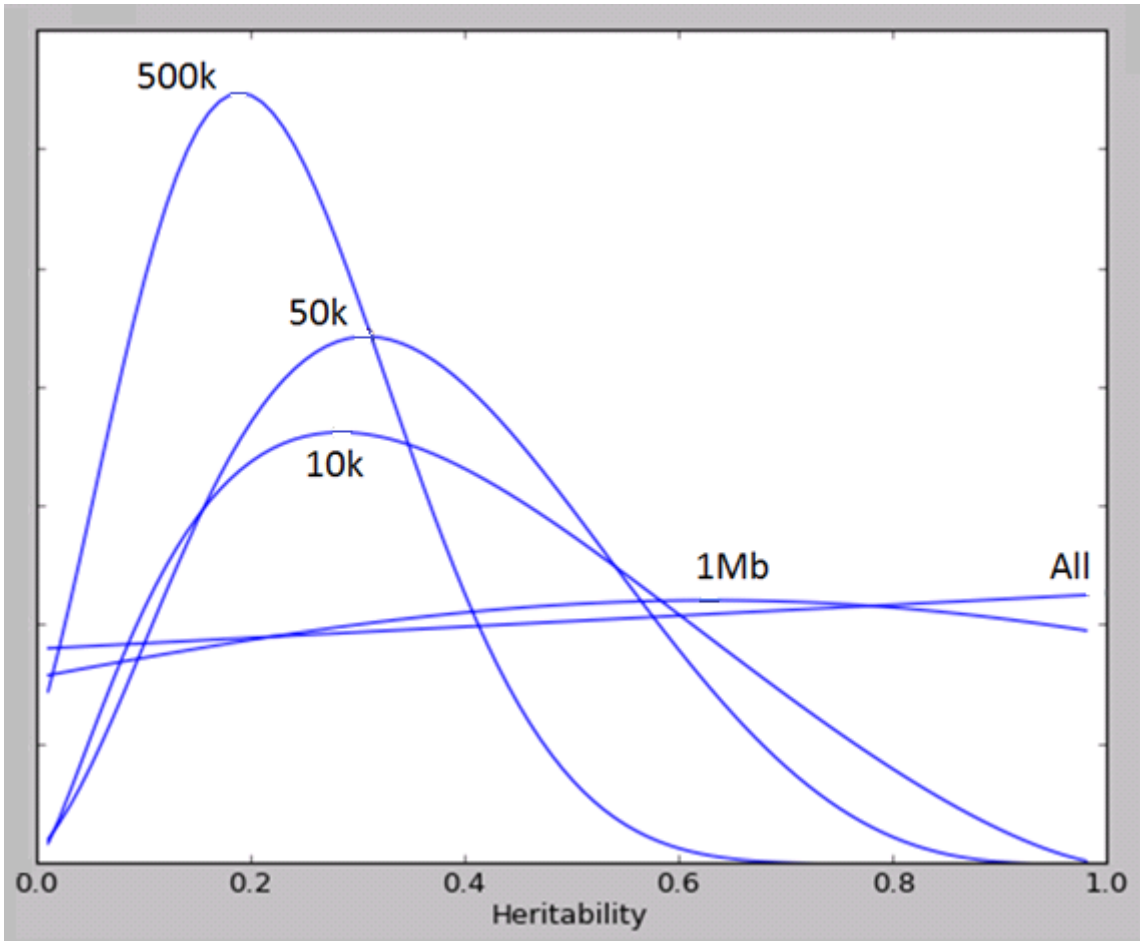

Figure 3. Relative data likelihood as a function of $\boldsymbol{h^2}$ for varying window sizes in the methylation data, as used by Quon *et al.* (1).