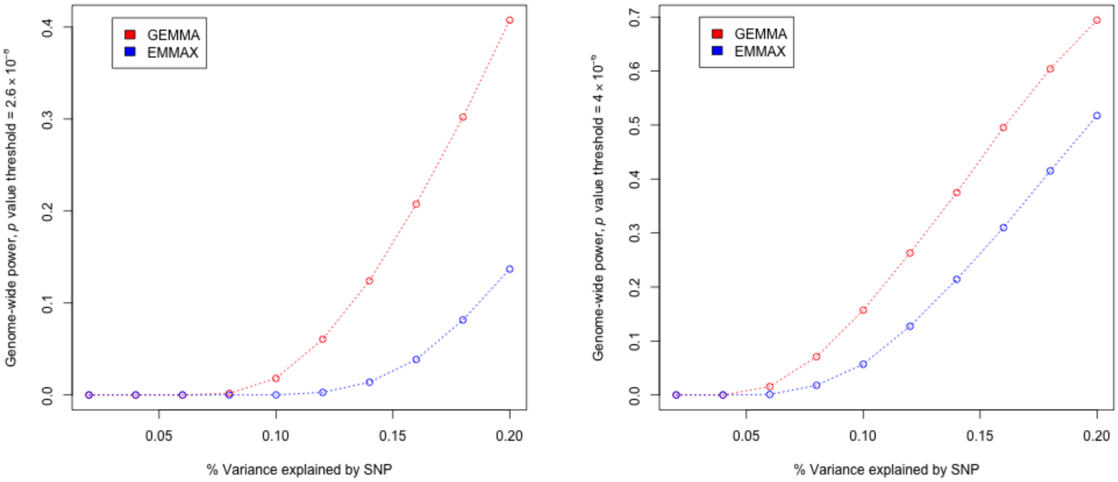
# Supplementary Information

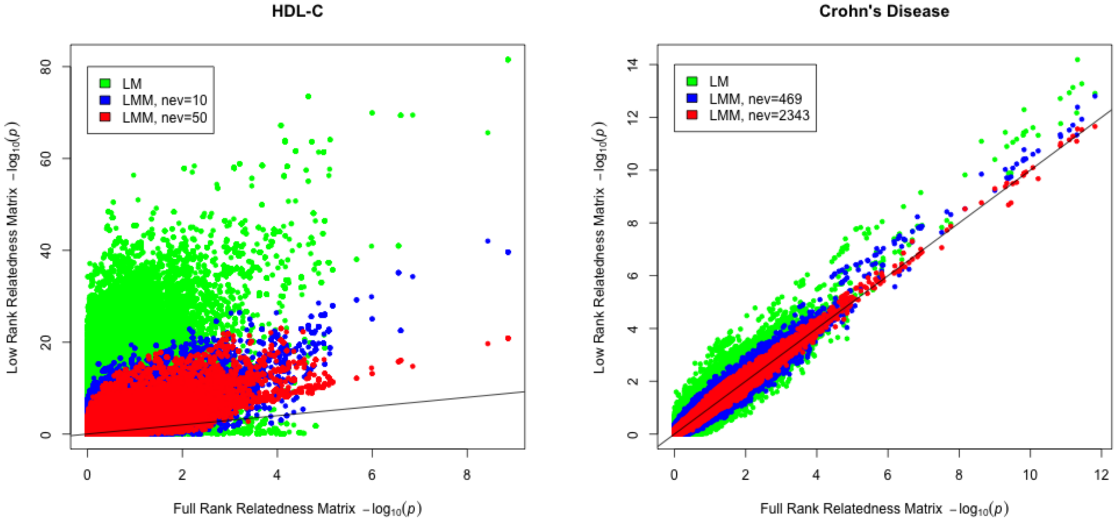
## 1 Supplementary Figures



(a) Statistical power ($p = 2.6 \times 10^{-8}$)

(b) Statistical power ($p = 4.0 \times 10^{-6}$)

Supplementary Figure 1: Statistical power comparison between GEMMA (red) and EMMAX (blue) using simulation with the HMDP data set, at two different genome-wide significance thresholds. The $y$ axis shows how power varies with SNP effect size ($x$ axis).



(a) Full Rank vs Low Rank (HMDP)

(b) Full Rank vs Low Rank (WTCCC)

Supplementary Figure 2: Comparison of $-\log_{10} p$ values obtained from linear mixed models using low-rank matrices, with those from the usual full-rank matrix. a) shows $p$ values computed from 1,885,197 markers for HDL measurements in the HMDP data set, by either a linear model, linear mixed models using low-rank relatedness matrices constructed from the top 10% (10) or 50% (50) eigenvectors, or linear mixed model using the full-rank relatedness matrix. b) shows $p$ values computed from 442,001 markers for Crohn's disease states in the WTCCC data set, by either a linear model, linear mixed models using low-rank relatedness matrices constructed from the top 10% (469) or 50% (2343) eigenvectors, or linear mixed model using the full-rank relatedness matrix. Black line shows the diagonal line. nev, number of eigenvectors used; LM, linear model; LMM, linear mixed model.

## 2 Supplementary Table

Supplementary Table 1: Comparison of genomic control inflation factors obtained with linear mixed models using relatedness matrices constructed from the top 0% (linear model), 10%, 50% and 100% (full-rank matrix) eigenvectors of the usual relatedness matrix, for both HMDP and WTCCC data sets. EVs, eigenvectors.

| Data Set | Genomic Control Inflation Factor | | | |
|---|---|---|---|---|
| | Linear Model | LMM (10% EVs) | LMM(50% EVs) | LMM(Full) |
| HMDP, HDL-C | 26.392 | 7.575 | 3.879 | 0.955 |
| WTCCC, CD | 1.174 | 1.070 | 1.030 | 1.012 |

# 3 Supplementary Note

## 3.1 Genome-wide Efficient Mixed Model Association, More Details

### 3.1.1 Derivation of the target optimization functions

If $\lambda$ is known, the log-likelihood is maximized at:

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\beta} \end{pmatrix} = ((\mathbf{W}, \mathbf{x})^T \mathbf{H}^{-1} (\mathbf{W}, \mathbf{x}))^{-1} (\mathbf{W}, \mathbf{x})^T \mathbf{H}^{-1} \mathbf{y},$$

$$\hat{\tau} = \frac{n}{(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\alpha}} - \mathbf{x}\hat{\beta})^T \mathbf{H}^{-1} (\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\alpha}} - \mathbf{x}\hat{\beta})} = \frac{n}{\mathbf{y}^T \mathbf{P}_x \mathbf{y}}.$$

The last equation uses the property $\mathbf{P}_x \mathbf{H} \mathbf{P}_x = \mathbf{P}_x$. This can be derived by noticing $\mathbf{P}_x = \mathbf{M}_x (\mathbf{M}_x \mathbf{H} \mathbf{M}_x)^- \mathbf{M}_x$, where $\mathbf{M}_x = \mathbf{I}_n - (\mathbf{W}, \mathbf{x})((\mathbf{W}, \mathbf{x})^T (\mathbf{W}, \mathbf{x}))^{-1} (\mathbf{W}, \mathbf{x})^T$ and $-$ denotes generalized inverse.

Similarly, the log-restricted likelihood is maximized at

$$\hat{\tau} = \frac{n - c - 1}{\mathbf{y}^T \mathbf{P}_x \mathbf{y}}.$$

Therefore, finding MLE and REML estimates is equivalent to optimizing the following functions with respect to $\lambda$:

$$\begin{aligned} l(\lambda) &= \frac{n}{2} \log(\frac{n}{2\pi}) - \frac{n}{2} - \frac{1}{2} \log |\mathbf{H}| - \frac{n}{2} \log(\mathbf{y}^T \mathbf{P}_x \mathbf{y}), \\ l_r(\lambda) &= \frac{n-c-1}{2} \log(\frac{n-c-1}{2\pi}) - \frac{n-c-1}{2} + \frac{1}{2} \log |(\mathbf{W}, \mathbf{x})^T (\mathbf{W}, \mathbf{x})| \\ &\quad - \frac{1}{2} \log |\mathbf{H}| - \frac{1}{2} \log |(\mathbf{W}, \mathbf{x})^T \mathbf{H}^{-1} (\mathbf{W}, \mathbf{x})| - \frac{n-c-1}{2} \log(\mathbf{y}^T \mathbf{P}_x \mathbf{y}). \end{aligned}$$

### 3.1.2 Numeric optimization details

Following EMMA[1], we consider $\lambda$'s ranging from $1 \times 10^{-5}$ (corresponding to almost pure environmental effect) to $1 \times 10^{5}$ (corresponding to almost pure genetic effect). We use Brent's method on the first derivative of the target functions, initialized with the two boundary values, to provide an initial value that is close to a root with relative error of $1 \times 10^{-1}$. We then follow this with Newton-Raphson's method, taking advantage of the second derivative to achieve a relative error of $1 \times 10^{-5}$. This search strategy combines the stability of Brent's method with the efficiency of Newton-Raphson's method. In the implemented software, we also provide the option of dividing the log-scale evaluation interval into equally spaced regions[1]. Brent's method followed by Newton-Raphson's method can be carried out for optimization in each region where the first derivatives change sign. This dividing strategy with ten regions yields identical results, takes less than five minutes longer for both data sets, and is expected to find the maxima in the evaluation interval more easily when the target functions are not well behaved.

### 3.1.3 Matrix calculus properties

The derivation of the first and second derivatives for both target functions uses a few matrix calculus properties:

$$\begin{aligned}
\frac{\partial \log|\mathbf{H}|}{\partial\lambda} &= \mathrm{vec}^T(\mathbf{H}^{-1})\mathrm{vec}(\mathbf{G}) = \mathrm{trace}(\mathbf{H}^{-1}\mathbf{G}),\\
\frac{\partial \mathrm{vec}(\mathbf{P}_x)}{\partial\lambda} &= \frac{\partial \mathrm{vec}(\mathbf{M}_x(\mathbf{M}_x\mathbf{H}\mathbf{M}_x)^{-}\mathbf{M}_x)}{\partial\lambda} = -\mathbf{P}_x\otimes\mathbf{P}_x\mathrm{vec}(\mathbf{G}) = -\mathrm{vec}(\mathbf{P}_x\mathbf{G}\mathbf{P}_x),\\
\frac{\partial \mathrm{vec}^T(\mathbf{H}^{-1})}{\partial\lambda} &= -\mathbf{H}^{-1}\otimes\mathbf{H}^{-1}\mathrm{vec}(\mathbf{G}) = -\mathrm{vec}(\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}),\\
\frac{\partial \mathrm{vec}(\mathbf{P}_x\mathbf{G}\mathbf{P}_x)}{\partial\lambda} &= (\mathbf{I}_n\otimes\mathbf{P}_x\mathbf{G}+\mathbf{P}_x\mathbf{G}\otimes\mathbf{I}_n)\frac{\partial \mathrm{vec}(\mathbf{P}_x)}{\partial\lambda} = -2\mathrm{vec}(\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{G}\mathbf{P}_x),
\end{aligned}$$

where $\otimes$ denotes Kronecker product and vec denotes matrix vectorization (by stacking columns).

### 3.1.4 Simplification of the trace term and the vector-matrix-vector product term

For the trace term, we notice:

$$\begin{aligned}
\mathrm{trace}(\mathbf{H}^{-1}\mathbf{G}) &= \mathrm{trace}(\mathbf{H}^{-1}\frac{(\mathbf{H}-\mathbf{I}_n)}{\lambda}) = \frac{n-\mathrm{trace}(\mathbf{H}^{-1})}{\lambda},\\
\mathrm{trace}(\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}\mathbf{G}) &= \mathrm{trace}(\mathbf{H}^{-1}\frac{(\mathbf{H}-\mathbf{I}_n)}{\lambda}\mathbf{H}^{-1}\frac{(\mathbf{H}-\mathbf{I}_n)}{\lambda}) = \frac{n+\mathrm{trace}(\mathbf{H}^{-1}\mathbf{H}^{-1})-2\mathrm{trace}(\mathbf{H}^{-1})}{\lambda^2},\\
\mathrm{trace}(\mathbf{P}_x\mathbf{G}) &= \mathrm{trace}(\mathbf{P}_x\frac{(\mathbf{H}-\mathbf{I}_n)}{\lambda}) = \frac{n-c-1-\mathrm{trace}(\mathbf{P}_x)}{\lambda},\\
\mathrm{trace}(\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{G}) &= \mathrm{trace}(\mathbf{P}_x\frac{(\mathbf{H}-\mathbf{I}_n)}{\lambda}\mathbf{P}_x\frac{(\mathbf{H}-\mathbf{I}_n)}{\lambda}) = \frac{n-c-1+\mathrm{trace}(\mathbf{P}_x\mathbf{P}_x)-2\mathrm{trace}(\mathbf{P}_x)}{\lambda^2}.
\end{aligned}$$

The last two equations use the property $\mathrm{trace}(\mathbf{P}_x\mathbf{H}) = \mathrm{trace}(\mathbf{M}_x(\mathbf{M}_x\mathbf{H}\mathbf{M}_x)^{-}\mathbf{M}_x\mathbf{H}) = n-c-1$.

For the vector-matrix-vector product term, we notice:

$$\begin{aligned}
\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y} &= \frac{\mathbf{y}^T\mathbf{P}_x\mathbf{y}-\mathbf{y}^T\mathbf{P}_x\mathbf{P}_x\mathbf{y}}{\lambda},\\
\mathbf{y}^T\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{G}\mathbf{P}_x\mathbf{y} &= \frac{\mathbf{y}^T\mathbf{P}_x\mathbf{y}+\mathbf{y}^T\mathbf{P}_x\mathbf{P}_x\mathbf{P}_x\mathbf{y}-2\mathbf{y}^T\mathbf{P}_x\mathbf{P}_x\mathbf{y}}{\lambda^2}.
\end{aligned}$$

### 3.1.5 Recursions for the trace term and the vector-matrix-vector product term

With blockwise matrix inversion we have:

$$\mathbf{P}_i = \mathbf{P}_{i-1}-\mathbf{P}_{i-1}\mathbf{w}_i(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-1}\mathbf{w}_i^T\mathbf{P}_{i-1}.$$

Nature Genetics: doi:10.1038/ng.2310

This leads to a recursion for the trace terms:

$$
\begin{aligned}
\mathrm{trace}(\mathbf{P}_i) &= \mathrm{trace}(\mathbf{P}_{i-1}) - (\mathbf{w}_i^T \mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T \mathbf{P}_{i-1}\mathbf{w}_i)^{-1}, \\
\mathrm{trace}(\mathbf{P}_i\mathbf{P}_i) &= \mathrm{trace}(\mathbf{P}_{i-1}\mathbf{P}_{i-1}) + (\mathbf{w}_i^T \mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)^2(\mathbf{w}_i^T \mathbf{P}_{i-1}\mathbf{w}_i)^{-2} \\
&\quad - 2(\mathbf{w}_i^T \mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T \mathbf{P}_{i-1}\mathbf{w}_i)^{-1},
\end{aligned}
$$

and a recursion for the vector-matrix-vector product terms, for any vectors **a**, **b** of the right size:

$$
\begin{aligned}
\mathbf{a}^T\mathbf{P}_i\mathbf{b} &= \mathbf{a}^T\mathbf{P}_{i-1}\mathbf{b} - (\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-1}, \\
\mathbf{a}^T\mathbf{P}_i\mathbf{P}_i\mathbf{b} &= \mathbf{a}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{b} \\
&\quad + (\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-2} \\
&\quad - (\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-1} \\
&\quad - (\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-1}, \\
\mathbf{a}^T\mathbf{P}_i\mathbf{P}_i\mathbf{P}_i\mathbf{b} &= \mathbf{a}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{b} \\
&\quad - (\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)^2(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-3} \\
&\quad - (\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-1} \\
&\quad - (\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-1} \\
&\quad - (\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-1} \\
&\quad + (\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-2} \\
&\quad + (\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-2} \\
&\quad + (\mathbf{a}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{b}^T\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{P}_{i-1}\mathbf{w}_i)(\mathbf{w}_i^T\mathbf{P}_{i-1}\mathbf{w}_i)^{-2}.
\end{aligned}
$$

Note that each recursion only requires a few scalar multiplications and does not depend on the number of individuals, as each vector-matrix-vector product in the form of $\mathbf{a}^T\mathbf{P}_i\mathbf{b}$, $\mathbf{a}^T\mathbf{P}_i\mathbf{P}_i\mathbf{b}$ or $\mathbf{a}^T\mathbf{P}_i\mathbf{P}_i\mathbf{P}_i\mathbf{b}$ is a scalar.

### 3.1.6 Test statistics and $p$ values

To test the null hypothesis $\beta = 0$, we obtain the likelihood ratio test statistic with MLE estimates and the Wald test statistic with the REML estimate as suggested[2,1]:

$$
\begin{aligned}
D_{lrt} &= 2\log\frac{l_1(\hat{\lambda}_1)}{l_0(\hat{\lambda}_0)}, \\
F_{Wald} &= \frac{\hat{\beta}^2}{V(\hat{\beta})}.
\end{aligned}
$$

Nature Genetics: doi:10.1038/ng.2310

where $l_1$ and $l_0$ are the likelihood functions for the null and the alternative models, respectively; $\hat{\lambda}_0$ and $\hat{\lambda}_1$ are the MLE estimates for the null and the alternative models, respectively; $\hat{\beta} = (\mathbf{x}^T \mathbf{P}_c(\hat{\lambda}_r)\mathbf{x})^{-1}(\mathbf{x}^T \mathbf{P}_c(\hat{\lambda}_r)\mathbf{y})$ is the estimate for $\beta$ obtained using the REML estimate $\hat{\lambda}_r$ in the alternative model; and $V(\hat{\beta}) = (n-c-1)^{-1}(\mathbf{x}^T \mathbf{P}_c(\hat{\lambda}_r)\mathbf{x})^{-1}(\mathbf{y}^T \mathbf{P}_x(\hat{\lambda}_r)\mathbf{y})$ is the variance for $\hat{\beta}$. Under the null hypothesis the likelihood ratio test statistic $D_{lrt}$ and the Wald test statistics $F_{Wald}$ come from a $\chi^2(1)$ and a $F(1, n-c-1)$ distribution respectively, and $p$ values can be calculated accordingly.

### 3.1.7 Missing data

We note that the tricks used in GEMMA rely on having complete or imputed genotype data at each SNP. That is, unlike EMMA, which, when testing a particular SNP, can simply ignore individuals with missing genotype data at that SNP, GEMMA requires the user to instead impute all missing genotypes before association testing. Arguably this imputation approach is preferable in any case, since it can improve power to detect associations[3]. In the current implementation of GEMMA, missing genotypes are required to be imputed first. Otherwise, any SNPs with missingness $> 5\%$ will not be analyzed, and other missing genotypes will simply be replaced with the mean genotype of that SNP.

## 3.2 Genotype and Phenotype Data, Details

We analyzed two data sets, one for quantitative traits from mouse and one for binary disease traits from human.

The mouse data set contains measurements of HDL levels for the Hybrid Mouse Diversity Panel (HMDP)[4]. Both phenotypes and genotypes are obtained from `http://mouse.cs.ucla.edu/`. A total of 99 mouse strains (29 classical inbred strains and 70 recombinant inbred strains) and 681 animals with overlapping genotype and phenotype recodes were used. A total of fully imputed 3,918,755 SNPs were used to obtain the identity by state (IBS) matrix as estimates of relatedness[5,4], and a total of 1,885,197 polymorphic SNPs were used for analysis. As in[4], we applied a linear mixed model with an intercept term and tested each SNP in turn (as a fixed effect), without controlling for any other covariates. Following a reviewer's suggestion we also repeated this analysis, but controlling for the top SNP (NES13033708) as a fixed effect in addition to an intercept; comparisons between the methods remained qualitatively similar (data not shown).

The human data set contains population controls and cases with Crohn's diseases from the WTCCC study[6]. Quality controlled genotypes were obtained from WTCCC and missing genotypes were imputed with BIMBAM[3]. 4686 individuals (1748 cases and 2938 controls) and a total of 442,001 SNPs were used for analysis. The Balding-Nichols matrix was used as estimates of relatedness[5] and binary variables were treated as quantitative traits as suggested[3,5]. As in[5] we applied a linear mixed model with an intercept term and tested each SNP in turn (as a fixed effect), without controlling for any other covariates.

## 3.3 Supplementary Results

### 3.3.1 Power simulation

We performed a power comparison between GEMMA and EMMAX using simulations similar to ref[7], in the HMDP data set. (We did not perform comparisons on the WTCCC data set because the empirical $p$ value comparisons show that EMMAX produces almost identical results to GEMMA in this case.)

For the power simulation we simulated phenotypes by adding effects to the original phenotype observations as in "Scheme 1" from ref[7]. Specifically, we first identified polymorphic SNPs unassociated with the original phenotype (exact Wald test $p$ value $> 0.05$). We ordered the 990,841 SNPs satisfying this criteria by their genomic location, and selected from them 10,000 evenly spaced SNPs to act as causal SNPs. For each causal SNP, we specified its effect size so that it explained a particular percentage of the phenotypic variance (proportion of variance explained, or PVE), and the effect was added back to the original phenotype to form the new simulated phenotype. For each pre-specified PVE (ranged from 2% to 20%), we simulated 10000 sets of phenotype, one for each causal SNP, and calculated $p$ values for each SNP-phenotype pair. We calculated statistical power as the proportion of (Wald test) $p$ values exceeding the genome-wide significance level, either at the conventional 0.05 level after Bonferroni correction ($p = 2.6 \times 10^{-8}$), or at a $p$ value known to achieve the same family-wise error rate in the HMDP panel ($p = 4.0 \times 10^{-6}$)[4].

Supplementary Figure 1 shows the genome-wide statistical power of GEMMA versus EMMAX in the HMDP data set, at two different genome-wide significance $p$ values, for SNPs with different effect sizes. The results suggest that GEMMA can be several times more powerful than EMMAX for SNPs with various effect sizes in this case.

### 3.3.2 Effects of modifying relatedness matrix

We explored the effects of using a low-rank relatedness matrix by computing $p$ values from linear mixed models with different relatedness matrices. In particular we considered replacing the full relatedness matrix $G$ with a rank $r$ approximation of the form $\hat{G}_r = \sum_{k=1}^{r} \delta_k u_k u_k^T$, where $\delta_k$ are the eigenvalues of $G$ (ordered to be decreasing) and $u_k$ are the corresponding eigenvectors. Standard theory ensures that $\hat{G}_r$ is the best rank $r$ approximation to $G$ in Frobenius norm.

We considered the relatedness matrices $\hat{G}(r)$ that corresponded to using either the top 0%, 10%, 50% or 100% eigenvectors of $G$. Note that 0% corresponds to the linear model, and 100% corresponds to the LMM with the usual relatedness matrix $G$. Supplementary Figure 2 shows comparison of $p$ values, and Supplementary Table 1 shows comparison of genomic control inflation factors, obtained with different relatedness matrices in both HMDP and WTCCC data sets. The results show that, compared with using $G$, using a lower-rank relatedness matrix can cause much larger changes in $p$ values than approximation methods such as EMMAX (Figure 1). Moreover, the genomic control inflation factors also suggest that, for these data sets, using a lower-rank relatedness matrix compromises the ability of the linear mixed model to control for sample structure.

# References

1. Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

2. Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208 (2006).

3. Guan, Y. and Stephens, M. Practical issues in imputation-based association mapping. *PLoS Genetics* **4** (2008).

4. Bennett, B. J., Farber, C. R., Orozco, L., Kang, H. M., Ghazalpour, A., Siemers, N., Neubauer, M., Neuhaus, I., Yordanova, R., Guan, B., Truong, A., Yang, W.-P., He, A., Kayne, P., Gargalovic, P., Kirchgessner, T., Pan, C., Castellani, L. W., Kostem, E., Furlotte, N., Drake, T. A., Eskin, E., and Lusis, A. J. A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Research* **20**, 281–290 (2010).

5. Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354 (2010).

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

7. Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., and Buckler, E. S. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355–360 (2010).