Grid-LMM: Supplemental Note

Figures

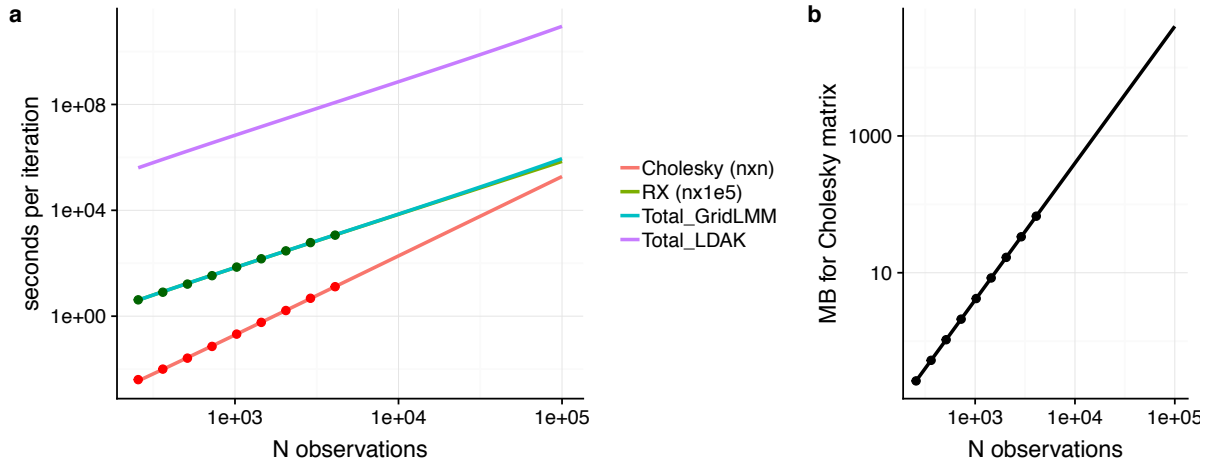

Figure S1. Approximate time and memory requirements of `Grid-LMM` as a function of sample size. (**a**) Computational times for the most costly steps of typical mixed model fitting algorithms: inverting an $n \times n$ covariance matrix (generally using a Cholesky decomposition), and multiplying the inverse matrix by an n-vector (i.e. a vector of marker genotypes). The red curve shows the time required for a Cholesky decomposition using the base `R` function `chol` as a function of n. The green curve shows the time required for a Cholesky matrix-by-marker matrix multiplication with 1×10^5 markers, as a function of n. The blue curve is the sum of the Cholesky decomposition and matrix-vector multiplication operations for a single grid-cell in `Grid-LMM` with 1×10^5 markers. The green and blue curves are barely distinguishable across most of the range because the Cholesky decomposition is generally not limiting. The purple curve would be the expected time for a separate Cholesky decomposition and matrix-vector multiplication for *each marker* in a GWAS with 1×10^5 markers (i.e. the cost of a typical `exact-LMM` method such as `LDAK` for a single iteration). Both the `Grid-LMM` and LDAK times are *per-iteration*. `Grid-LMM` requires this time at each grid cell. `LDAK` requires multiple iterations for the REML optimization separately for each marker. Generally, `Grid-LMM` will evaluate more grid cells than `LDAK` requires iterations per test. However this will not cause a reversal in the relative time requirements unless a very large grid is used. (**b**) Memory requirements for storing an $n \times n$ Cholesky matrix as a function of sample size. In both panels, the curves were extrapolated based on tests with n between 256 and 4096 (actual times shown with points). All timings were estimated using base `R` functions.

-
-

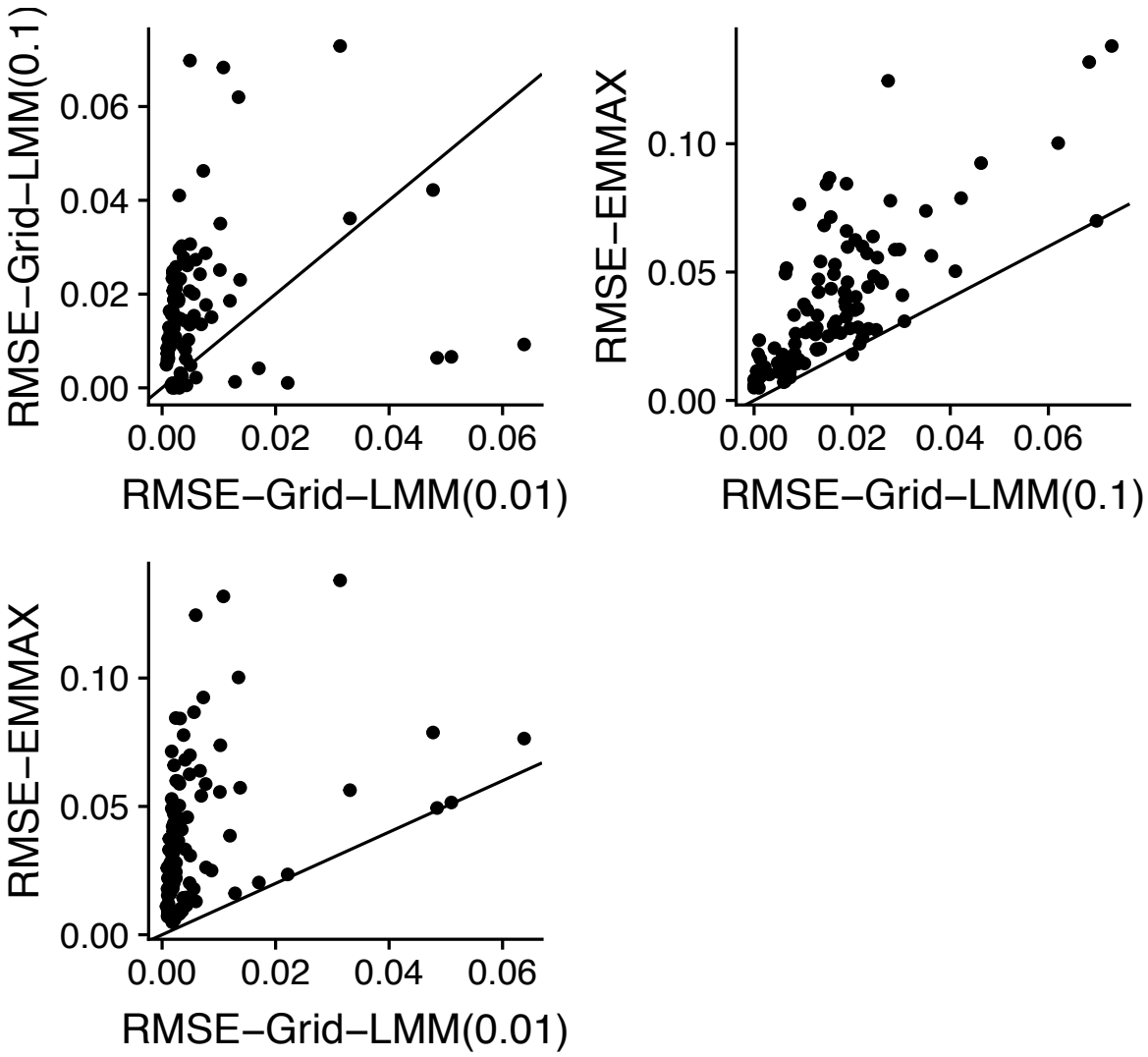

Figure S2. Accuracy of the log-transformed *p*-values across the 107 Arabidopsis phenotypes [1]. GWASs were run for each phenotype using 216,130 markers and up to 199 accessions, with a single random effect controlling for additive genetic relationships among lines. For each phenotype (represented as a single point in the plots), we compared the exact Wald-test $-\log_{10}(p)$ calculated by `GEMMA` to p-values calculated by the approximate methods `EMMAX`, and `Grid-LMM` using either the naive approach with a complete grid of size 0.1 h^2-units, or the fast heuristic algorithm `Grid-LMM-fast` with a fine grid size of 0.01-h^2 units. `Grid-LMM` p-values were always at least as accurate (as measured by root mean-squared-error, RMSE) as those calculated by `EMMAX`. Specifically, p-values calculated with a fine grid-size of 0.01 (using the fast algorithm) were nearly indistinguishable from those of `GEMMA`, except in the rare cases where the REML surface was not unimodal. This was generally restricted to a small subset of rare markers with small-moderate effect sizes, and only occurred for a few traits. In these cases, the complete — but more coarse — grid search of `Grid-LMM` with step sizes of 0.1 was more accurate.

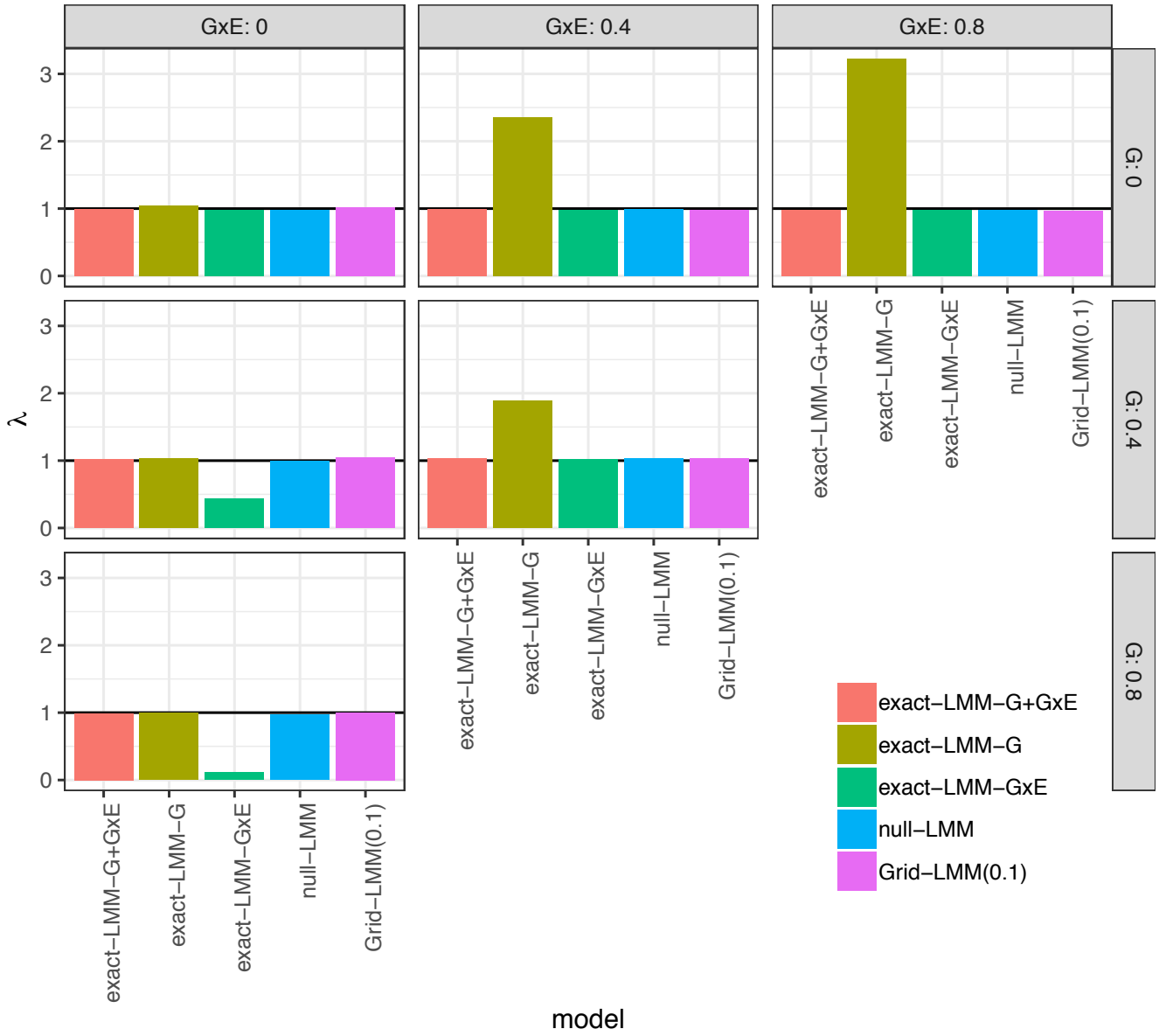

Figure S3. Genomic control inflation factors for simulated data. Simulated datasets were created based on the Atwell genotype data and the G×E analysis. We randomly selected 10,000 markers, generated simulated data with different proportions of additive (G) and gene-environment interaction (G× E) variation for each marker, and calculated Wald F-statistics for an interaction between the marker and the environment. Bars show an estimate of genomic control inflation factors for each of the following five methods. `exact-LMM-G+GxE` is an exact LMM algorithm fit with `LDAK`. This model included both random effects and the marker effect. At a genome-wide scale, it is very slow, with computational complexity $\mathcal{O}(ptn^3)$. `exact-LMM-G` and `exact-LMM-GxE` are exact LMM algorithms, similar to `GEMMA`, which included only one random effect and the marker effect. `null-LMM` is an approximate method similar to `pylmm` that conditions on variance components estimated under a null model with no marker effect. It was run with both random effects. *Grid-LMM* was run with a grid size of 0.1 h^2-units and included both random effects and the marker effect. The λ values were calculated as the ratio between the median value of the the F-statistics returned by each model and the median value of a $F_{1,316-4}$ distribution. The horizontal line shows the expected value $\lambda = 1$ under the true model.

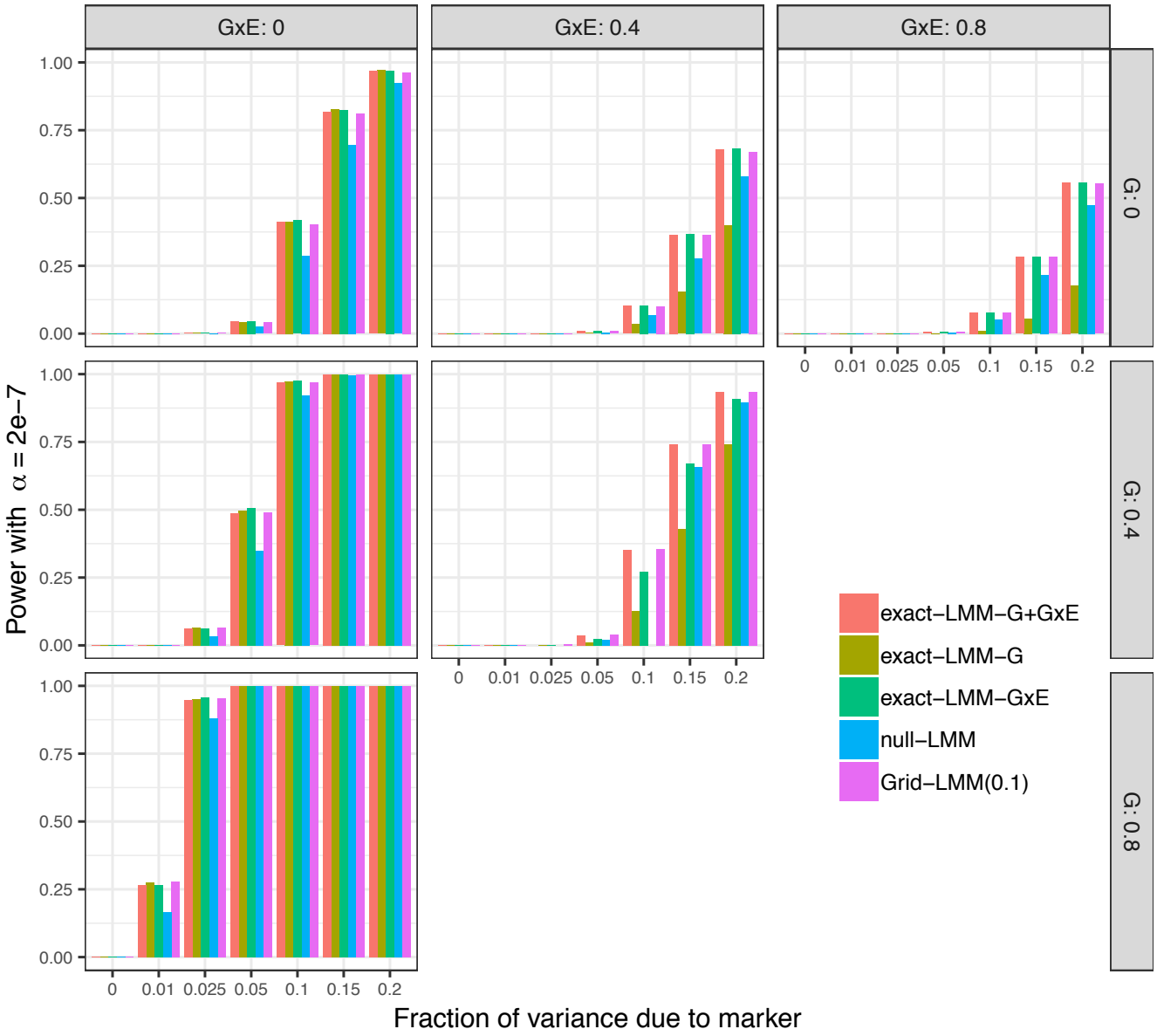

Figure S4. Power analysis for simulated data. Bars show the genome-wide power for randomly selected SNPs in the Atwell genotype data under simulations with different proportions of additive (G) and gene-environment interaction (G×E) variation with different marker effect sizes. Simulations were generated as described in Figure S3, and included only a single marker with zero main effect and G×E effects scaled to a defined percentage of the phenotypic variation. The remaining phenotypic variation was simulated from a multivariate normal distribution constructed by appropriately weighting the additive relationship matrix, the G×E covariance matrix, and the uncorrelated residual variation. Each simulation was run separately for 10,000 randomly selected markers. Wald F-statistics from each method were normalized by dividing by the genomic control inflation factor computed for Figure S3, and then *p-values* were calculated and compared to the Bonferroni corrected threshold $P = 2 \times 10^{-7}$ to determine significance.

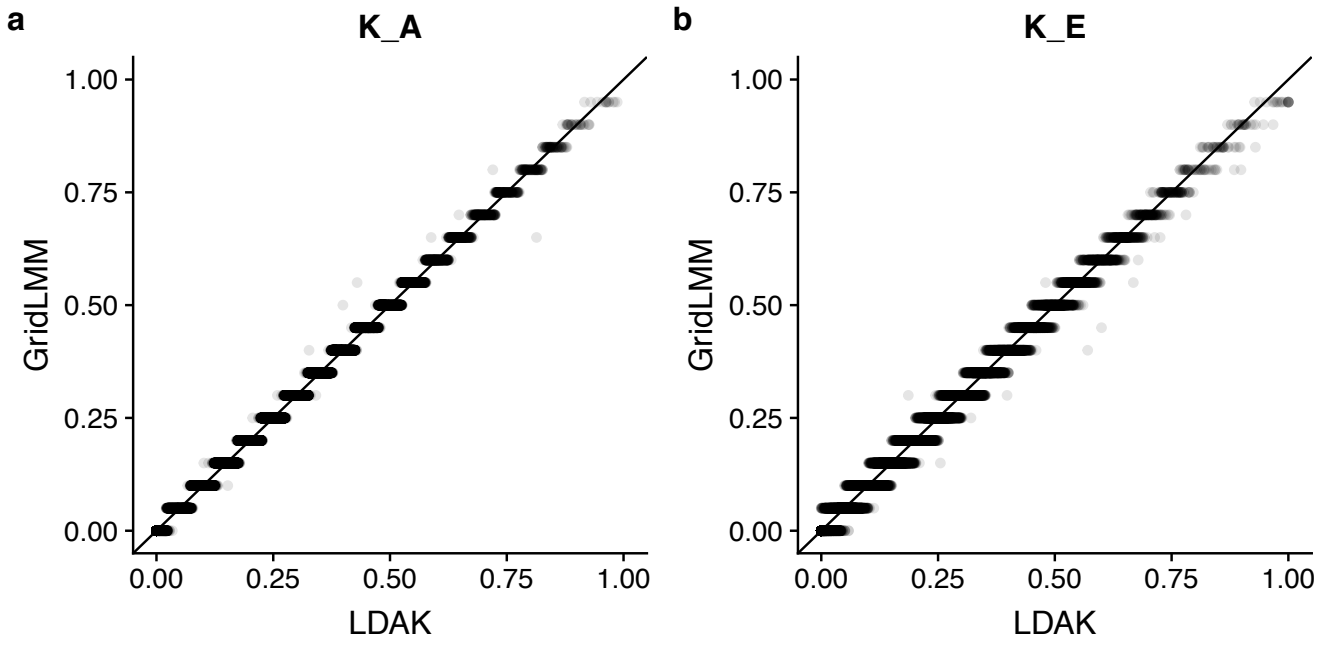

Figure S5. Comparison of REML estimates between Grid-LMM and LDAK for 20,843 *Arabidopsis* genes. (**a**) REML estimates for the additive genetic variance (variance component for $\mathbf{K}_A$). (**b**) REML estimates for the epistatic genetic variance (variance component for $\mathbf{K}_E$).

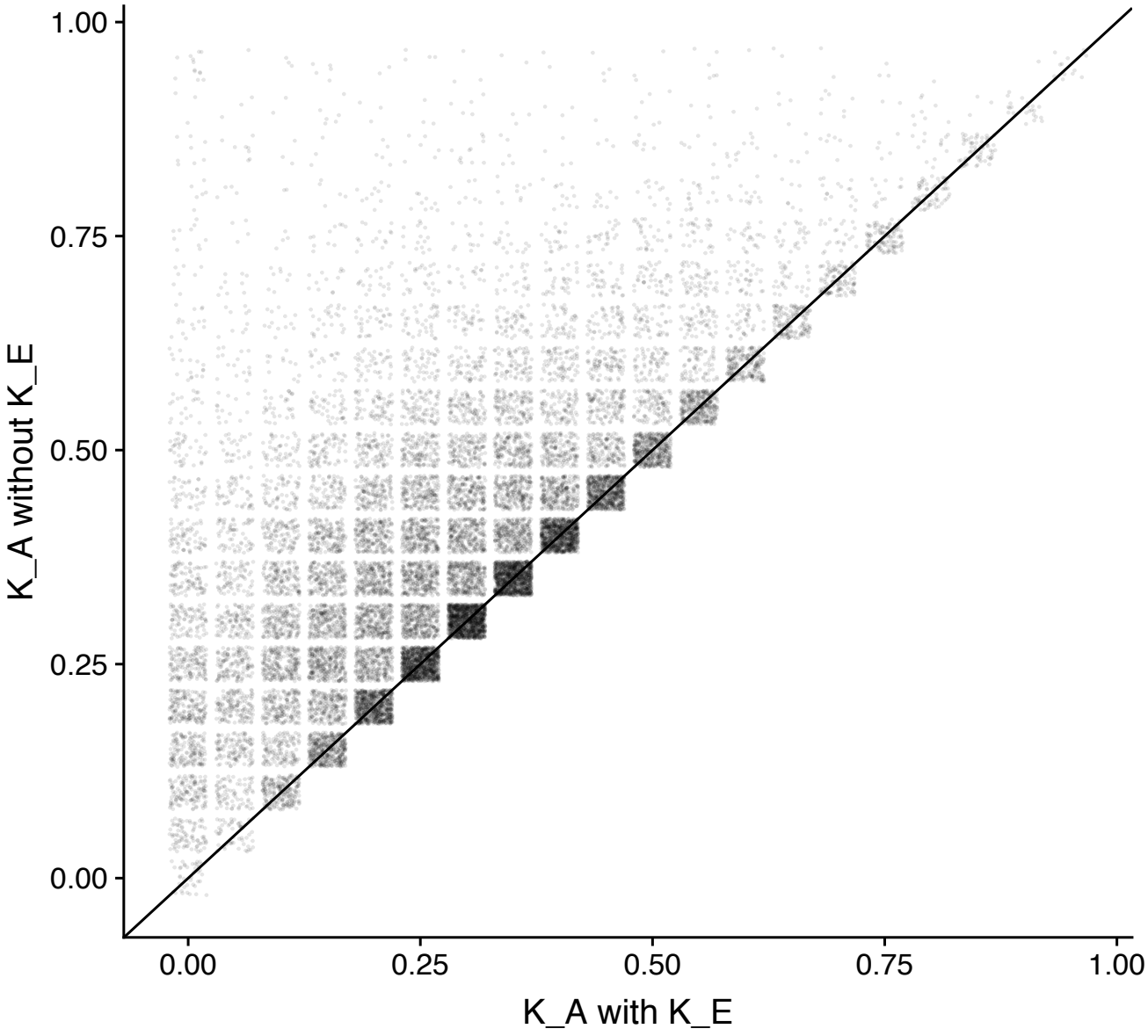

Figure S6. Comparison of REML estimates for the additive genetic variance between models with and without an additional pairwise-epistasis random effect for 20,843 *Arabidopsis* genes. Both models were fit using Grid-LMM with a grid size of 0.05 h^2 units. Point positions are jittered for clarity.

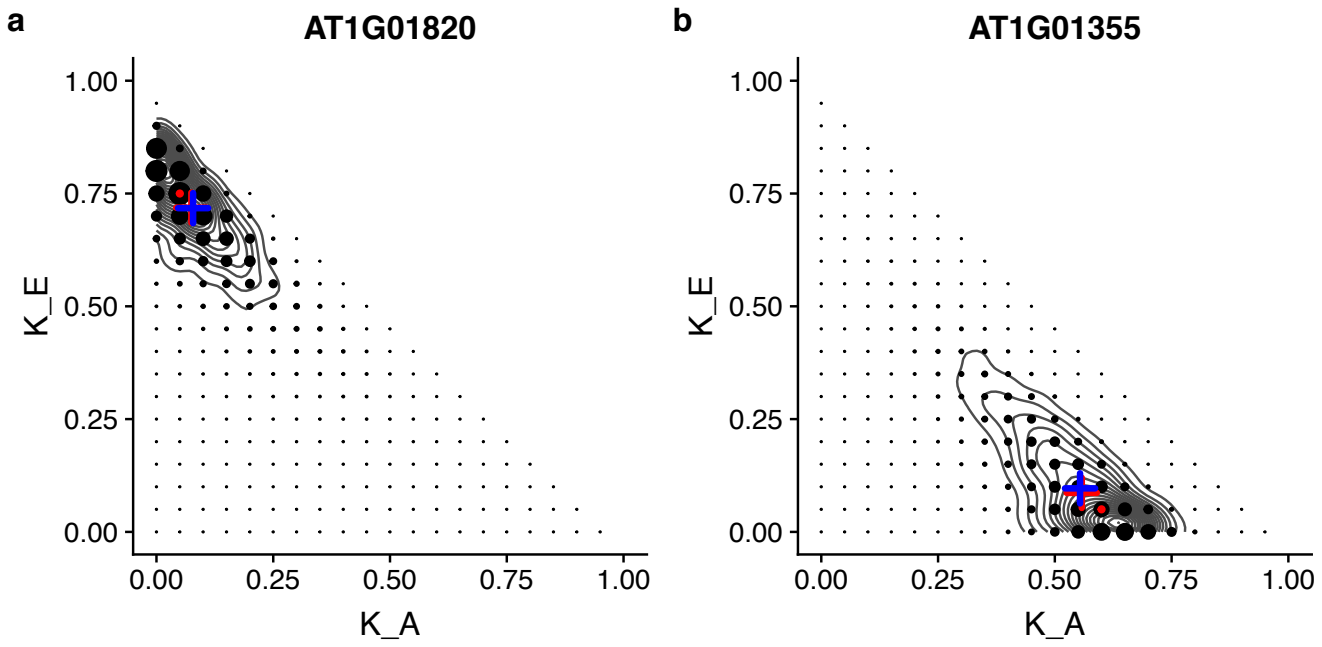

Figure S7. Posterior distributions of variance components for two genes under the half-Student-t(3,0,10) prior. Panels (**b**) and (**c**) of Figure 3 in the main text are repeated, except the half-Student-t(3,0,10) prior on the standard deviation of the variance components of $\mathbf{K}_A$ and $\mathbf{K}_E$ for the random effects was applied to each grid vertex. The prior was approximated by simulating 1×10^4 independent draws for σ_A, σ_E and σ_e, converting these to prior draws for h^2_A and h^2_E, and then measuring the proportion of draws closest to each grid vertex.

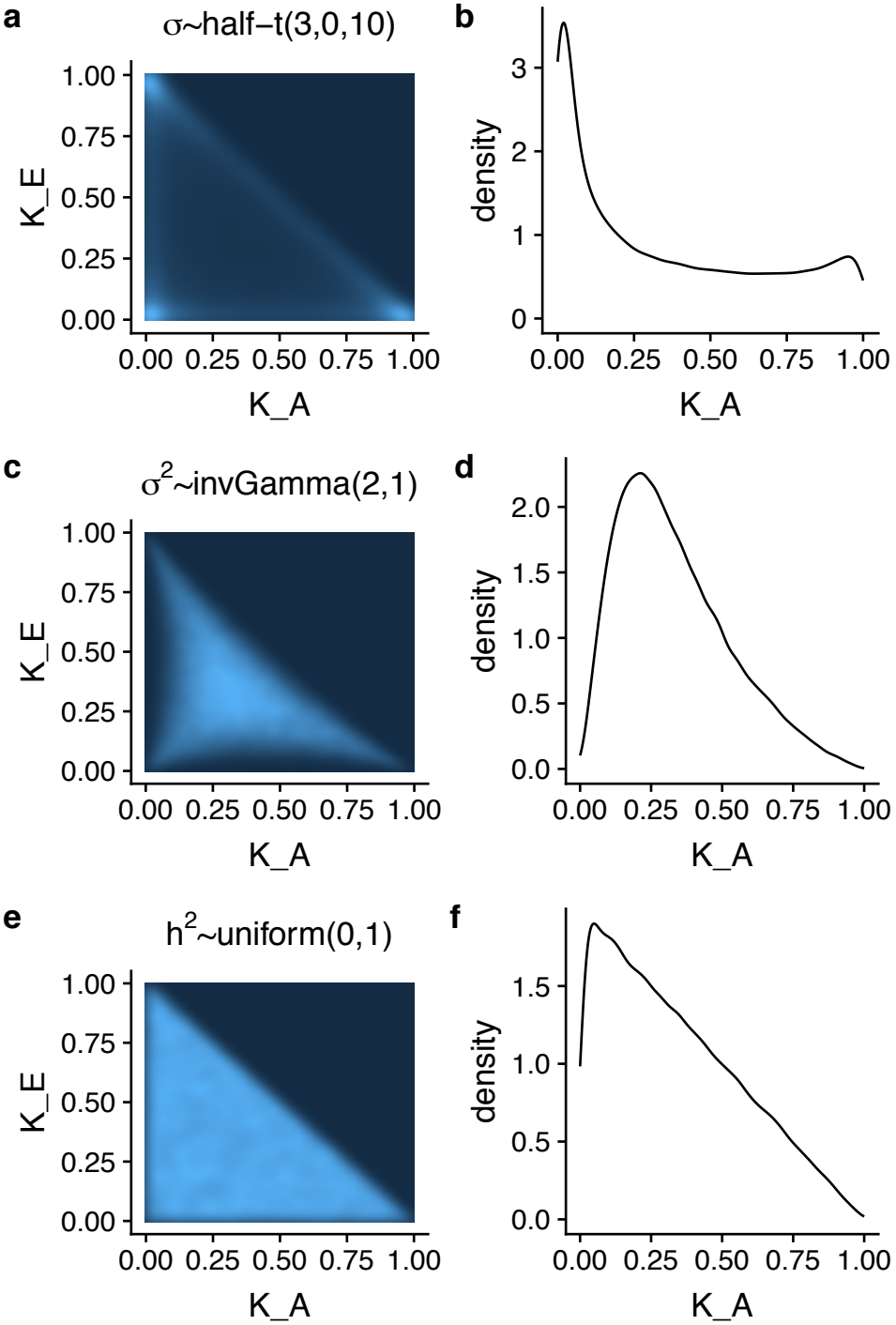

Figure S8. Inverse-Gamma and half-Student-t priors are informative for variance component proportions. We compare the implied prior distributions on variance component proportions for three classes of priors in a two-random effect model (e.g. $\mathbf{K}_A$ and $\mathbf{K}_E$ as random effects plus uncorrelated random error). (**a**)-(**b**) Each standard deviation parameter was assigned a half-Student-t prior with 3 degrees of freedom and scale parameter of 10. (**c**)-(**d**) Each variance parameter was assigned an inverse-Gamma prior with shape parameter 2 and scale parameter 1. (**e**)-(**g**) A uniform prior was applied to the 2-dimensional simplex of $[h^2_A, h^2_E, h^2_e]$. This is the default prior in GridLMM and equivalent to all analyses reported in the main text. (**a**)-(**c**)-(**e**) 2D-density plots for the two variance component proportions. Lighter blue denotes higher prior density. (**b**)-(**d**)-(**f**) Marginal densities for the $\mathbf{K}_A$ variance component proportion under each prior. The half-Student-t prior implies high probability that only one variance component is important. The inverse-Gamma prior implies high probability that all variance component proportions are non-zero.

Tables

Table S1. Markers associated with heterogenous stock mice body weight only when accounting for non-additive genetic covariance. Markers with $-\log_{10}(p) > 3$ on chromosomes 4 and 11 are shown. Position information for each marker is derived from [2]. Other studies that identified the same marker associations are listed. WTCHG: Wellcome Trust Center Heterogeneous Stock Mice. LG,SM Advanced Intercross: F_9 and F_{10} generation mice from of an intercross between LG/J and SM/J strains selected for differences in body size.

marker	Chr	bp	cM	Effect Size	-log10(p)	Genes	References	Phenotype	Population
rs4224463	4	44561913	20.8151	-0.2307899	5.827733	Trmt10b, Exosc3	[3]	Starting weight	WTCHG
rs6313392	4	44568868	20.81513	-0.2307899	5.827733	Exosc3	[3]	Starting weight	WTCHG
rs3665393	4	44721094	20.81515	-0.2313425	5.941238	Shb	[3]	Starting weight	WTCHG
rs13477678	4	44734525	20.81518	-0.1715087	3.978578	Trim14, Coro2a, Shb	—	—	—
rs3668228	4	44764150	20.8152	-0.2307148	5.794518	Shb	[3]	Starting weight	WTCHG
rs13477682	4	45785592	21.78908	-0.1339141	3.199823	Gm16731	—	—	—
rs13481023	11	51651037	29.3193	0.1682438	4.543509	Gm26551, Cdkl3	[4]	Tibia length	LG,SM Advanced Intercross
rs8243055	11	51893791	29.31933	0.1682438	4.543509	Gm26551, Tcf7	—	—	—
rs6173994	11	52168577	29.51397	0.1599403	4.19647	—	—	—	—
rs3668680	11	55712245	31.26964	-0.1584908	3.386805	Gm12239	—	—	—

References

1. Susanna Atwell, Yu S Huang, Bjarni J Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M Tarone, Tina T Hu, Rong Jiang, N Wayan Muliyati, Xu Zhang, Muhammad Ali Amer, Ivan Baxter, Benjamin Brachi, Joanne Chory, Caroline Dean, Marilyne Debieu, Juliette de Meaux, Joseph R Ecker, Nathalie Faure, Joel M Kniskern, Jonathan D G Jones, Todd Michael, Adnane Nemri, Fabrice Roux, David E Salt, Chunlao Tang, Marco Todesco, M Brian Traw, Detlef Weigel, Paul Marjoram, Justin O Borevitz, Joy Bergelson, and Magnus Nordborg. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, June 2010.
2. Sagiv Shifman, Jordana Tzenova Bell, Richard R Copley, Martin S Taylor, Robert W Williams, Richard Mott, and Jonathan Flint. A High-Resolution Single Nucleotide Polymorphism Genetic Map of the Mouse Genome. *PLoS Biology*, 4(12):e395, November 2006.
3. Jin Liu, Kai Wang, Shuangge Ma, and Jian Huang. Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method. *Statistics and its interface*, 6(1):99–115, January 2013.
4. Elizabeth A Norgard, Joseph P Jarvis, Charles C Roseman, Taylor J Maxwell, Jane P Kenney-Hunt, Kaitlin E Samocha, L Susan Pletscher, Bing Wang, Gloria L Fawcett, Christopher J Leatherwood, Jason B Wolf, and James M Cheverud. Replication of long-bone length QTL in the F9-F10 LG,SM advanced intercross. *Mammalian genome : official journal of the International Mammalian Genome Society*, 20(4):224–235, March 2009.

Supplemental Methods

Bayesian posterior and Bayes Factor derivation

We describe our approach for approximating Bayesian posteriors of variance component proportions, as well as calculating Bayes Factors for linear mixed model (LMM) comparisons, using Grid-LMM. Consider the following Bayesian hierarchical regression model specification:

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\ \mathbf{e} &\sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{V}) \\ \boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Psi}) \\ \sigma^2 &\sim \mathrm{IG}(a,b),\end{aligned} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is an $p \times 1$ coefficient vector, and $\mathbf{e}$ is an $n \times 1$ vector of independent and identically distributed errors. The joint prior distribution $p(\boldsymbol{\mu}, \sigma^2)$ is said to be the normal-inverse-gamma $\mathrm{NIG}(\boldsymbol{\mu}, \boldsymbol{\Psi}, a, b)$ distribution. The residual covariance matrix $\mathbf{V} = \sum_{l=1}^{L} h_l^2 \mathbf{K}_l + h_e^2 \mathbf{I}$ is the result of integrating over all random effects, and $h_e^2 + \sum_{l=1}^{L} h_l^2 = 1$ with the h_l^2 parameters giving the proportion of the total random effect variation contributed by each random effect. Our main objective is to conduct inference on these variance component proportions $\boldsymbol{\theta} = [h_1^2, \ldots, h_L^2, h_e^2]$. Secondarily, we may also want to compute the posterior summaries for $\boldsymbol{\beta}$, or for the variance components themselves (i.e. $\sigma_l^2 = \sigma^2 h_l^2$). Therefore, we must calculate the following:

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}) \propto p(\boldsymbol{\theta}) p(\mathbf{y} \,|\, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \int_{\boldsymbol{\beta}, \boldsymbol{\sigma^2}} p(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) p(\boldsymbol{\beta}, \sigma^2) \, d\boldsymbol{\beta} \, d\boldsymbol{\sigma^2} \tag{2}$$

According to Equation 1, the term inside the integral is equal to:

$$\begin{aligned} p(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) p(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} |\mathbf{V}|^{-1/2} \times \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\ \times \frac{b^a}{\Gamma(a)} (2\pi)^{p/2} (\sigma^2)^{p/2+a+1} |\boldsymbol{\Psi}|^{1/2} \times \exp\left\{ -\frac{1}{\sigma^2} \left[b + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu})^\intercal \boldsymbol{\Psi}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}) \right] \right\}. \end{aligned} \tag{3}$$

where $\Gamma(\bullet)$ is the gamma function. Using the following linear algebra identity for arbitrary vectors $(\boldsymbol{\gamma}, \boldsymbol{\alpha})$ and square matrix $\mathbf{A}$,:

$$\boldsymbol{\gamma}^\intercal \mathbf{A} \boldsymbol{\gamma} - 2\boldsymbol{\alpha}^\intercal \boldsymbol{\gamma} = (\boldsymbol{\gamma} - \boldsymbol{A}^{-1}\boldsymbol{\alpha})^\intercal \mathbf{A} (\boldsymbol{\gamma} - \mathbf{A}^{-1}\boldsymbol{\alpha}) - \boldsymbol{\alpha}^\intercal \mathbf{A}^{-1} \boldsymbol{\alpha},$$

we can rearrange Equation 3 to be:

$$p(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) p(\boldsymbol{\beta}, \sigma^2) = C^* \times (\sigma^2)^{-(a+p/2+n/2+1)} \times \exp\left\{ -\frac{1}{\sigma^2} \left[b^* + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}^*)^\intercal \boldsymbol{\Psi}^{*-1} (\boldsymbol{\beta} - \boldsymbol{\mu}^*) \right] \right\}, \tag{4}$$

where we define the terms:

$$\begin{aligned} C^* &= \frac{b^a}{\Gamma(a)} (2\pi)^{-(n+p)/2} |\mathbf{V}|^{-1/2} |\boldsymbol{\Psi}|^{-1/2} \\ \boldsymbol{\Psi}^* &= (\boldsymbol{\Psi}^{-1} + \mathbf{X}^\intercal \mathbf{V}^{-1} \mathbf{X})^{-1} \\ \boldsymbol{\mu}^* &= \boldsymbol{\Psi}^* (\boldsymbol{\Psi}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} + \mathbf{X}^\intercal \mathbf{V}^{-1} \mathbf{y}) \\ b^* &= b + \frac{1}{2} \left[\boldsymbol{\mu}^\intercal \boldsymbol{\Psi}^{-1} \boldsymbol{\mu} + \mathbf{y}^\intercal \mathbf{V}^{-1} \mathbf{y} - \boldsymbol{\mu}^{*\intercal} \boldsymbol{\Psi}^{*-1} \boldsymbol{\mu}^* \right]. \end{aligned}$$

Note that if we assume $\boldsymbol{\mu} = \mathbf{0}$, then $b^* = b + \mathrm{RSS}_{\mathbf{V},\boldsymbol{\Psi}}/2$, where $\mathrm{RSS}_{\mathbf{V},\boldsymbol{\Psi}} = \mathbf{y}^\intercal (\mathbf{V}^{-1} - \mathbf{V}^{-1} \widetilde{\mathbf{P}} \mathbf{V}^{-1}) \mathbf{y}$ is the (generalized) residual sum of squares given $\mathbf{V}$ and $\boldsymbol{\Psi}$, with $\widetilde{\mathbf{P}} = \mathbf{X}\boldsymbol{\Psi}^* \mathbf{X}^\intercal \mathbf{V}^{-1} \mathbf{X} \boldsymbol{\Psi}^* \mathbf{X}^\intercal$. Equation 4 is the kernel of the $\mathrm{NIG}(\boldsymbol{\mu}^*, \boldsymbol{\Psi}^*, a^*, b^*)$ distribution with $a^* = a + n/2$. Therefore, we can evaluate the integral as in Equation 2 as:

$$\int_{\boldsymbol{\beta}, \boldsymbol{\sigma^2}} p(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) p(\boldsymbol{\beta}, \sigma^2) \, d\boldsymbol{\beta} \, d\sigma^2 = \frac{b^a}{(2\pi)^{(n+p)/2} |\mathbf{V}|^{1/2} |\boldsymbol{\Psi}|^{1/2} \Gamma(a)} \times \frac{(2\pi)^{p/2} |\boldsymbol{\Psi}^*|^{1/2} \Gamma(a^*)}{(b^*)^{a^*}}. \tag{5}$$

In order to evaluate the posterior $p(\boldsymbol{\theta} \,|\, \mathbf{y})$, we must calculate the quantity in Equation 5 for each value in $\boldsymbol{\theta}$ with $p(\boldsymbol{\theta}) > 0$. Using Grid-LMM, we assume this only occurs at a finite (and generally small) set of values on the L-dimensional simplex. At

each value, we must calculate the determinant of the $n \times n$ dimensional matrix $\mathbf{V}$. However, because of the grid, we efficiently traverse the posterior, reducing the number of evaluations of $\mathbf{V}^{-1}$. All terms on the right hand side of Equation 5 that do not depend on $\boldsymbol{\theta}$ (or $\mathbf{V}$) cancel out in the posterior and Equation 2 can be written in a simplified form as:

$$p(\boldsymbol{\theta}\,|\,\mathbf{y}) = \frac{|\mathbf{V}_{\boldsymbol{\theta}}|^{-1/2}|\boldsymbol{\Psi}^*_{\boldsymbol{\theta}}|^{1/2}(b^*_\theta)^{-a^*}\,p(\boldsymbol{\theta})}{\sum|\boldsymbol{\Psi}_{\boldsymbol{\theta}}|^{-1/2}|\mathbf{V}^*_{\boldsymbol{\theta}}|^{1/2}(b^*_\theta)^{-a^*}\,p(\boldsymbol{\theta})} \tag{6}$$

where the subscripts are added to denote each term that is a function of $\boldsymbol{\theta}$. Given this analytical form of $p(\boldsymbol{\theta}\,|\,\mathbf{y})$, we can now calculate Bayes factors (BF) comparing mixed effect models with different terms in $\mathbf{X}$ [e.g. 1]. In this case, the term $|\boldsymbol{\Psi}|^{1/2}$ will differ among models (i.e. with or without the marker of interest), and therefore will not cancel in the BF ratio. However, as previously shown, if $\boldsymbol{\Psi}$ is diagonal, we can still take the limit of the BF as the hyperparameters $a, b \to 0$ and $\sigma^2_{\boldsymbol{\beta}} \to \infty$ for the covariates (but not for the testing marker which is instead a scale parameter that is specified by the user). This choice of prior distribution class is informative only for the testing SNP, and is otherwise scale-free [1].

References

1. Bertrand Servin and Matthew Stephens. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLOS Genetics*, 3(7):e114, July 2007.