nature genetics

SUPPLEMENTARY INFORMATION

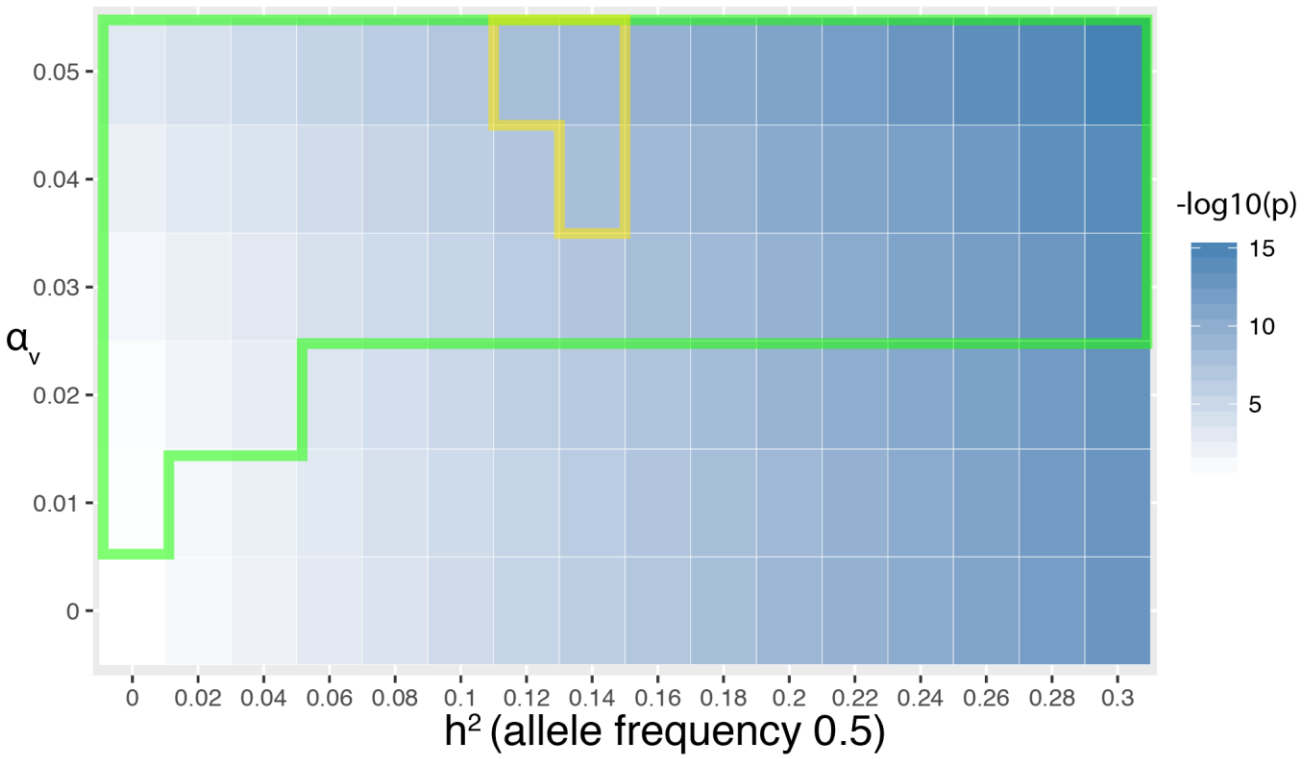
https://doi.org/10.1038/s41588-018-0225-6

In the format provided by the authors and unedited.

Identifying loci affecting trait variability and detecting interactions in genome-wide association studies

Alexander I. Young[1,2]*, Fabian L. Wauthier[1,3] and Peter Donnelly[1,3]*

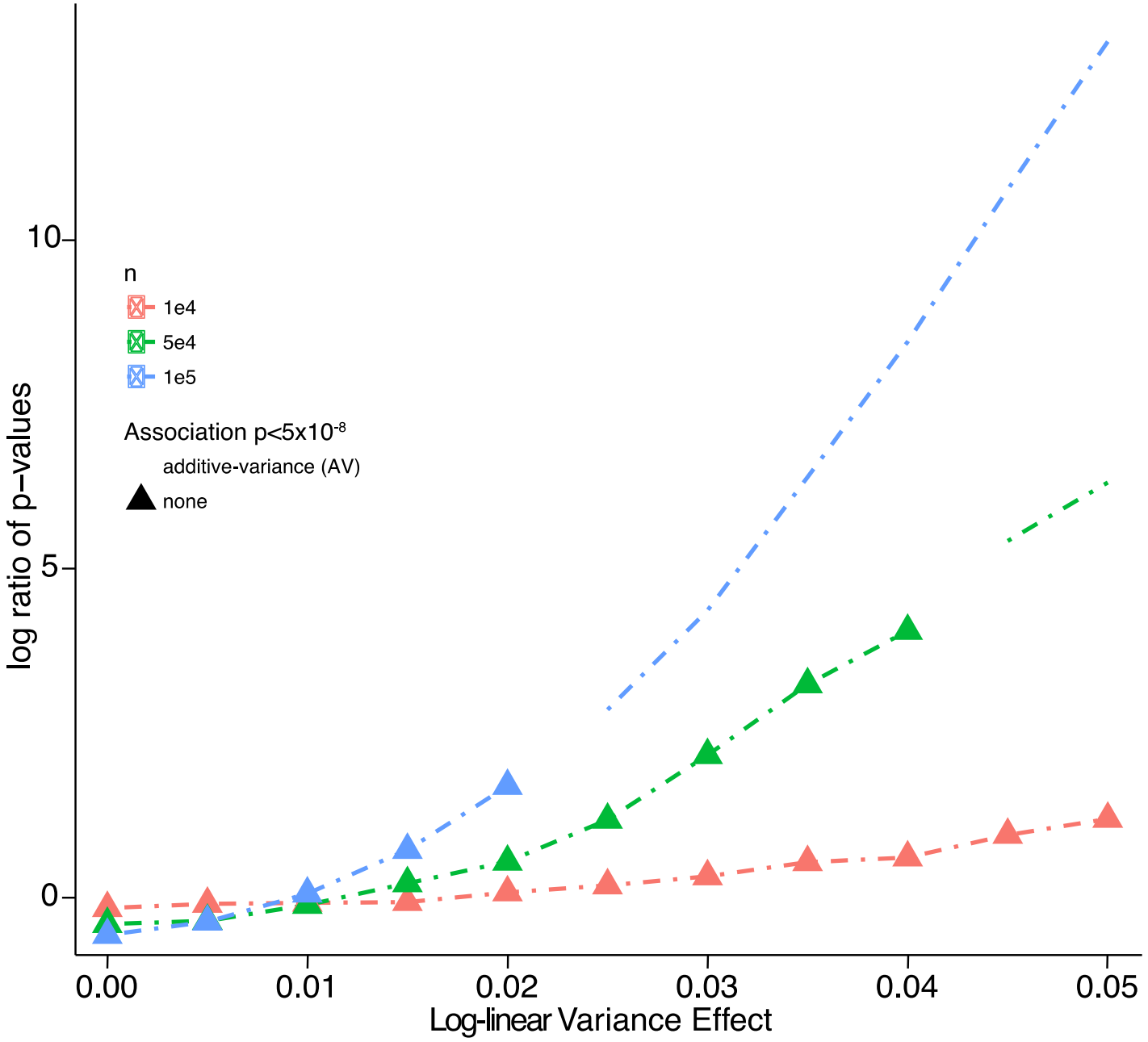
[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [3]Department of Statistics, University of Oxford, Oxford, UK. *e-mail: alexander.young@bdi.ox.ac.uk; donnelly@well.ox.ac.uk



Supplementary Figure 1

Association signal of the additive-variance (AV) test for simulated phenotypes with different parameters.

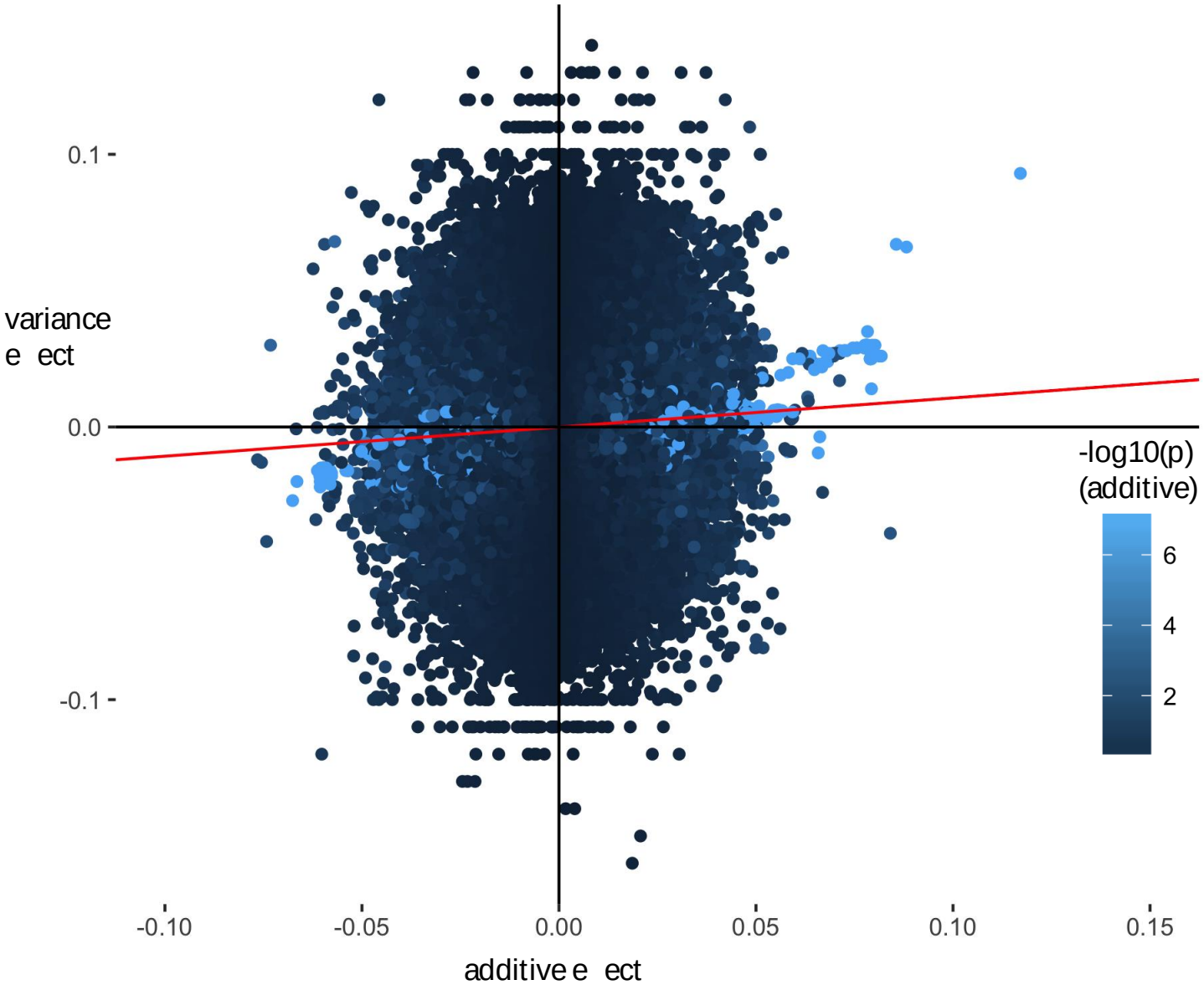
The expected $-\log_{10}$ (*P* value) of the AV test for different additive and log-linear variance effects of the test locus is indicated by shading. Phenotypes were simulated for 100,000 unrelated individuals (Methods). The test locus had frequency 0.05. To make this plot comparable to Fig. 1, we used the same set of additive effects. As in Fig. 1, the strength of the additive effect is parameterized by the amount of variance explained, h^2, if the allele frequency is 0.5. Here the allele frequency is 0.05, so the actual variance explained is 0.19 times the variance explained when the allele frequency is 0.5. The log-linear variance effect is indicated on the *y* axis and corresponds approximately to the proportional change in phenotypic variance per allele. We have highlighted two regions of parameter space: the area inside the green lines is where the association signal is stronger under the AV test than under the additive test, and the area inside the yellow lines is where the AV test is genome-wide significant ($P < 5 \times 10^{-8}$) but the additive test is not.



Supplementary Figure 2

Comparison of association signal for the additive-variance (AV) and additive tests for different sample sizes.

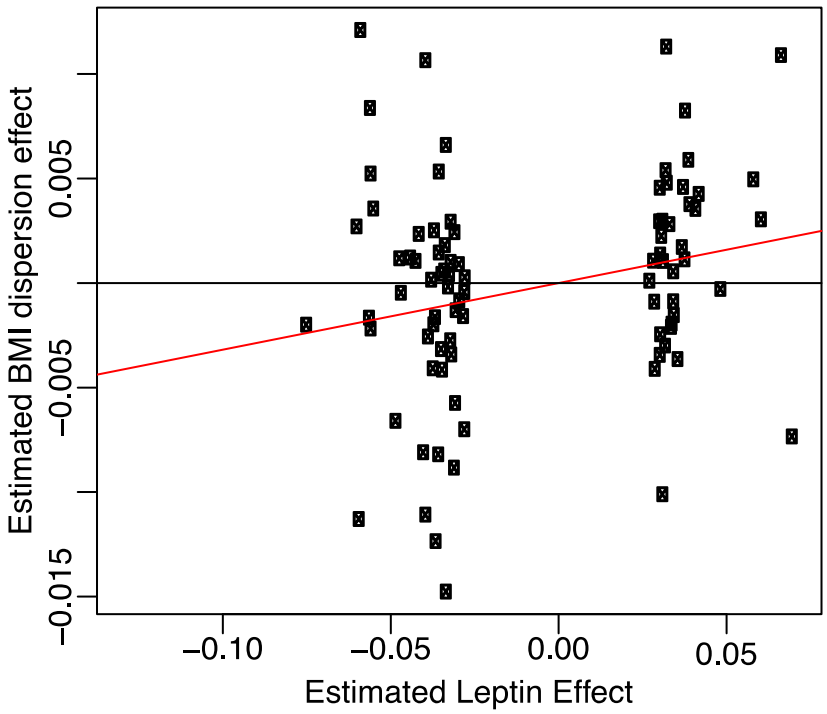
The association signal when testing for both additive and log-linear variance effects (AV test) compared to testing for only an additive effect (additive test) in simulations. The *y* axis gives the expected log ratio (base 10) of the *P* value from the additive test to the AV test for different variance effects of the test SNP (*x* axis), with values above zero indicating a stronger signal from the AV test. The simulations were performed for sample sizes of 10,000 (red), 50,000 (green), and 100,000 (blue), indicated with the different colored curves. The log ratio is plotted as a crossed box if the expected *P* value from the additive-variance test would pass the standard genome-wide significance threshold of 5×10^{-8}, and it is plotted with a triangle if neither of the expected *P* values from the two tests would pass the significance threshold.



Supplementary Figure 3

Relationship between additive and variance effects from GIANT meta-analyses.

Estimated additive (*x* axis) and variance (*y* axis) effects on BMI are plotted for all genome-wide loci, shaded in proportion to the negative $\log_{10}$ (*P* value) for an additive effect, up to a maximum of negative $\log_{10}$ (5×10^{-8}), the conventional boundary for genome-wide significance. The additive effects are taken from Locke et al. (*Nature* **518,** 197–206, 2015), and the variance effects are taken from Yang et al. (*Nature* **490,** 267–272, 2012). Because of the mean–variance relationship of untransformed BMI, any locus with an additive effect is expected to have a variance effect, even after inverse-normal transformation. The red line has slope 0.1071, determined by robust regression of genome-wide variance effects on additive effects, with weights proportional to the inverse square of the standard error of the estimated variance effects.



Supplementary Figure 4

Relationship between estimated leptin effect and estimated dispersion effect on BMI.

Estimated leptin effect (s.d. change in leptin per allele) (*x* axis) and dispersion (*y* axis) effects on BMI are plotted for the top 100 approximately independent SNPs ranked by evidence for a leptin effect (Methods). The leptin effects are taken from Kilpeläinen et al. (*Nat. Commun.* **7,** 2016), and the dispersion effects are taken from our analysis of the UK Biobank. The red line gives the estimated expected dispersion effect for a given leptin effect (Methods).

	general variance effect				
log-linear variance effect	0.01	0.02	0.03	0.04	0.05
0.01	-0.13	0.71	2.06	3.81	6.37
0.02	-0.42	0.43	1.74	3.67	6.17
0.03	-0.56	0.19	1.53	3.53	5.76
0.04	-0.73	0.06	1.53	3.34	5.78
0.05	-0.82	0.00	1.31	3.29	5.74

Supplementary Table 1: **Relative power of two degree of freedom variance test to log-linear variance effect test.** Expected differece between -log10(p-value) from the 2 degree of freedom variance test, which includes both log-linear and general variance effects, and the test for a log-linear variance effect alone. Values were estimated from 1000 independent simulations for each pair of effects (Methods). Values above zero indicate a stronger expected association signal from the 2 degree of freedom test, whereas negative values indicate a stronger expected signal under the test for a log-linear variance effect alone.

SNP	AV α_v	AVD α_v
rs1538749	-0.016	0.017
rs1801282	0.022	0.022
rs900400	0.020	0.021
rs10787472	-0.015	-0.016
rs2303223	-0.015	-0.016
rs1421085	0.027	0.030
rs10423928	0.021	0.023

Supplementary Table 5: **Effect of fitting dominance effects on log-linear variance effect estimates.** We compare the log-linear variance effects (α_v) in Table 1, estimated from the AV model, to log-linear variance effects estimated from the AVD model (Supplementary Note). The AVD model fits additive and dominance effects on the mean of the phenotype, in addition to a log-linear variance effect. We fitted AVD models in the combined related and unrelated samples. We used the same covariates as in the main analysis and the same random effects as used for the related sample (Methods).

Supplementary Note for 'Loci affecting trait variability and detection of interactions in genome-wide association studies'.

July 26, 2018

Alexander I. Young[1,2], Fabian Wauthier[1,3], Peter Donnelly[1,3],

[1] Wellcome Trust Centre For Human Genetics, University of Oxford, Roosvelt Drive, Oxfordshire, U.K., OX3 7BN
[2] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, U.K.
[3] Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, Oxfordshire, U.K., OX1 3LB

Contents

1 Hierarchy of models for mean and variance effects

Assuming the phenotype (Y) distribution is normal conditional on the genotype (G), the most general model relating genotype to phenotype allows for the distribution conditional on each genotype at the locus to be any normal distribution:

$$\mathrm{M_G}: \; Y|G = g \sim \mathcal{N}(\mu_g, \sigma_g^2). \tag{1}$$

To test for association between genotype and phenotype, one could compare the likelihood of model $\mathrm{M_G}$ to the null model,

$$\mathrm{M_0}: \; Y|G = g \sim \mathcal{N}(\mu, \sigma^2). \tag{2}$$

giving a likelihood ratio test on four degrees-of-freedom, which has been previously suggested[1].

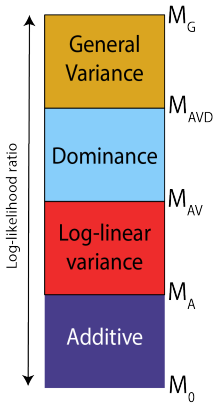


Figure 1: **Nested hierarchy of models**. The hierarchy builds from the null model ($\mathrm{M_0}$ – no mean or variance effects) to the general model ($\mathrm{M_G}$ – arbitrary mean and variance effects). Effects are added successively at each level of the hierarchy (additive, log-linear variance, dominance, and general variance) with the model including all of the effects below it indicated on the right hand side. The overall height of the bar can be seen as the log-likelihood ratio test statistic comparing the general model ($\mathrm{M_G}$) to the null model ($\mathrm{M_0}$), with the heights of the components giving the corresponding log-likelihood ratio test statistics for the specified effects.

While model $\mathrm{M_G}$ can capture any mean and variance effects of a locus, it is possible to fit simpler models that capture mean and variance effects. We introduce a nested hierarchy of models (Figure 1) that allows us to decompose the log-likelihood ratio of model $\mathrm{M_G}$ to the null model into components that give evidence for different mean and variance effects. From this hierarchy, simpler tests can be devised to improve power.

Most genetic association studies fit a model that only allows for an additive effect. The first model above the null model in our hierarchy is the additive model:

$$\mathrm{M_A}: \; Y|G = g \sim \mathcal{N}(\mu + \alpha g, \sigma^2), \tag{3}$$

where μ is the location parameter for the mean, and α is the additive effect of the genetic variant.

We now seek to introduce a variance effect that is analogous to the additive effect on the mean. Because the variance is always positive, one cannot use a linear model, which is unbounded; instead, we use a log-linear model. Let $\sigma_g^2 = \mathrm{Var}(Y|G = g)$. This model has the form

$$\log(\sigma_g^2) = \mu_v + \alpha_v g, \tag{4}$$

where μ_v corresponds to the scale of the variance, and α_v is termed the log-linear variance effect of the locus. The next model in our hierarchy incorporates a log-linear variance effect in addition to an additive effect:

$$\mathrm{M_{AV}}: \; Y|G = g \sim \mathcal{N}(\mu + \alpha g, \exp(\mu_v + \alpha_v g)), \tag{5}$$

which we call the additive-variance model or AV model for short.

We add a dominance effect to this model to allow for non-linearity in the relationship between the conditional means and the number of copies of an allele, giving:

$$\mathrm{M_{AVD}}: \; Y|G = g \sim \mathcal{N}(\mu_g, \exp(\mu_v + \alpha_v g)). \tag{6}$$

Similarly, we add a general variance effect to allow for non-linearity in the relationship between the conditional log-variances and the number of copies of an allele, which takes us to $\mathrm{M_G}$. The log-likelihood ratio between $\mathrm{M_G}$ and $\mathrm{M_0}$ can therefore be decomposed as the sum of the log-likelihood ratio test statistics for each of the mean and variance effects:

$$\begin{aligned} &2[l(\mathrm{M_G}|y,g) - l(\mathrm{M_0}|y,g)] = 2[l(\mathrm{M_G}|y,g) - l(\mathrm{M_{AVD}}|y,g)] + \\ &2[l(\mathrm{M_{AVD}}|y,g) - l(\mathrm{M_{AV}}|y,g)] + 2[l(\mathrm{M_{AV}}|y,g) - l(\mathrm{M_A}|y,g)] + 2[l(\mathrm{M_A}|y,g) - l(\mathrm{M_0}|y,g)]. \end{aligned} \tag{7}$$

The four components individually give evidence for additive ($\mathrm{M_A}$ vs. $\mathrm{M_0}$), log-linear variance ($\mathrm{M_{AV}}$ vs. $\mathrm{M_A}$), dominance ($\mathrm{M_{AVD}}$ vs. $\mathrm{M_{AV}}$), and general variance ($\mathrm{M_G}$ vs. $\mathrm{M_{AVD}}$) effects, which is illustrated in Figure 1.

1.1 Functional form of variance effect for an interacting locus

If the variance effects of loci involved in interactions follow an approximate log-linear form, then a test that includes the log-linear variance effect but not the general variance effect should be more powerful. We now show that this is the case in a simple interaction model between the additive effect of a genetic variant G and an environmental variable E, although the same arguments would apply to interactions with a genetic variant. The model for the phenotype, Y, is

$$Y = G + E + \gamma GE + \epsilon, \tag{8}$$

where G is the number of copies of an allele at a locus, E is an environmental variable, and ϵ is independent noise with variance σ^2_ϵ.

The variance conditional on $G = g$ is

$$\begin{aligned}\mathrm{Var}(Y|G=g) &= \sigma^2_\epsilon + (1+\gamma g)^2 \mathrm{Var}(E) && (9)\\ &= (\sigma^2_\epsilon + \mathrm{Var}(E)) + 2\mathrm{Var}(E)\gamma g + \mathrm{Var}(E)\gamma^2 g^2 && (10)\\ &= (\sigma^2_\epsilon + \mathrm{Var}(E)) + 2\mathrm{Var}(E)\gamma g + \mathrm{O}(\gamma^2 g^2). && (11)\end{aligned}$$

Therefore the conditional variance is a linear function of g up to a correction factor that is proportional to the square of the interaction effect size. Given that the effect sizes of common variants for complex traits in humans are generally small relative to the variance of the trait, the quadratic term is generally going to be too small to detect at current sample sizes. This also applies to the log-conditional variance. If we assume that $(\sigma^2_\epsilon + \mathrm{Var}(E)) = 1$, then

$$\begin{aligned}\log(\mathrm{Var}(Y|G=g)) &= \log(1 + 2\mathrm{Var}(E)\gamma g + \mathrm{O}(\gamma^2 g^2)) && (12)\\ &= 2\mathrm{Var}(E)\gamma g + \mathrm{O}(\gamma^2 g^2). && (13)\end{aligned}$$

This implies that for the effect sizes of common loci on complex traits, a log-linear variance model should be accurate, unless the interaction model involves strongly non-linear functions of the genotype.

2 Relation of test statistics to mutual information

The mutual information between two random variables is a general measure of their dependence which is zero if and only if the two variables are independent, unlike linear correlation. It measures the amount of information that is shared between observations of the variables. The mutual information between a continuous phenotype Y and a genetic variant G is

$$I(Y;G) = H(Y) - H(Y|G); \tag{14}$$

where $H(Y)$ is the differential entropy of the phenotype Y,

$$H(Y) = -\int_{-\infty}^{\infty} f(y)\log(f(y))dy, \tag{15}$$

where $f(y)$ is the density function of the phenotype; and $H(Y|G)$ is the conditional entropy of Y given G,

$$H(Y|G) = \mathbb{E}_G[H(Y|G=g)], \tag{16}$$

where $H(Y|G=g)$ is the entropy of Y given that G takes a particular value, g.

We show that, in the infinitesimal genetic model, the likelihood ratio test statistic comparing $\mathrm{M_G}$ to $\mathrm{M_0}$ at the maximum likelihood parameter estimates is an estimator of the mutual information between Y and G, $I(Y;G)$. The models are as defined in the main text.

2.1 General likelihood ratio test statistic

To derive the maximum likelihood of the data under $\mathrm{M_G}$, we parameterise the model as:

$$Y|G = g \sim \mathcal{N}(\mu_g, \sigma_g^2), \tag{17}$$

where $\mu_g = \mathbb{E}[Y|G = g]$ and $\sigma_g^2 = \mathrm{Var}(Y|G = g)$.

If there are n_g out of n genotypes in category g, and y_{gi} is the i^{th} phenotypic observation in category g, then

$$2l(\mathrm{M_G}|y, g) = -n\ln(2\pi) - \sum_{g=0}^{2} n_g \ln(\sigma_g^2) - \sum_{g=0}^{2}\sum_{i=1}^{n_g} \frac{(y_{gi} - \mu_g)^2}{\sigma_g^2}. \tag{18}$$

This implies that the maximum likelihood estimators are

$$\hat{\mu}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} y_{gi};\ \hat{\sigma}_g^2 = \frac{1}{n_g}\sum_{i=1}^{n_g}(y_{gi} - \hat{\mu}_g)^2. \tag{19}$$

Let $\hat{l}$ be the value of the log-likelihood evaluated at the maximum likelihood estimator, then

$$2\hat{l}(\mathrm{M_G}|y, g) = -n\ln(2\pi) - \sum_{g=0}^{2} n_g \ln(\hat{\sigma}_g^2) - n. \tag{20}$$

For the null model, where $\bar{y}$ is the overall sample phenotype mean,

$$2\hat{l}(\mathrm{M_0}|y, g) = -n\ln(2\pi) - n\ln(\hat{\sigma}^2) - n; \hat{\sigma}^2 = \frac{1}{n}\sum_{g=0}^{2}\sum_{i=1}^{n_g}(y_{gi} - \bar{y})^2. \tag{21}$$

Therefore,

$$2[\hat{l}(\mathrm{M_G}|y, g) - \hat{l}(\mathrm{M_0}|y, g)] = n\ln(\hat{\sigma}^2) - \sum_{g=0}^{2} n_g \ln(\hat{\sigma}_g^2). \tag{22}$$

2.2 Maximum likelihood estimator of the mutual information

We now derive the mutual information between a genotype and a continuous phenotype in the infinitesimal genetic model with Gaussian error[2]. In the infinitesimal model, the unconditional distribution of Y is normal:

$$Y \sim \mathcal{N}(\mu, \sigma^2). \tag{23}$$

The differential entropy of Y is therefore $H(Y) = 0.5\ln(2\pi e\sigma^2)$. The mutual information between Y and G, $I(Y; G)$, can be expressed as

$$I(Y; G) = H(Y) - \mathbb{E}_G[H(Y|G = g)]. \tag{24}$$

Under the infinitesimal genetic model, the conditional distribution $Y|G = g$ is also normal with variance σ_g^2. Therefore

$$\mathbb{E}_G[H(Y|G = g)] = 0.5\mathbb{E}_G[\ln(2\pi e\sigma_g^2)]. \tag{25}$$

The mutual information is therefore

$$I(Y;G) = H(Y) - H(Y|G) = 0.5\ln(2\pi e\sigma^2) - 0.5\mathbb{E}_G[\ln(2\pi e\sigma_g^2)]; \tag{26}$$

$$= 0.5\ln(\sigma^2) - 0.5\sum_{g=0}^{2}\mathbb{P}(G = g)\ln(\sigma_g^2). \tag{27}$$

If we estimate the mutual information with the maximum likelihood estimators of the parameters, this gives

$$2n\hat{I}(Y;G) = n\ln(\hat{\sigma}^2) - \sum_{g=0}^{2} n_g \ln(\hat{\sigma}_g^2) = 2[\hat{l}(\mathrm{M_G}|y,g) - \hat{l}(\mathrm{M_0}|y,g)]. \tag{28}$$

We have shown that the maximum likelihood estimator of the mutual information between a genotype and a phenotype in the infinitesimal genetic model is proportional to a general likelihood ratio test for dependence, which is on four degrees of freedom. Therefore, when there is no association, the asymptotic distribution of the maximum likelihood estimator of the mutual information between genotype and phenotype is $(2n)^{-1}\chi_4^2$. This can be seen as a case of the known relationship between mutual information and log-likelihood ratio test statistics in parametric models[3].

The mutual information between a genotype and phenotype is zero if and only if they are independent. We have therefore shown that, under the infinitesimal genetic model with Gaussian residual error, the likelihood ratio test comparing $\mathrm{M_G}$ to $\mathrm{M_0}$ will, for all fixed significance levels greater than zero and less than one, have power to detect an association that tends to 100% with sample size if and only if Y and G are dependent.

3 The Heteroskedastic Linear Model

All of the models in the above hierarchy can be incorporated into a class of models called heteroskedastic linear models, which allow for an arbitrary vector of covariates to influence the residual variance of the response. Similar models and algorithms have a long history in the fields of heteroskedastic regression models[4] and econometrics.

Consider a phenotype Y with multivariate normal distribution:

$$Y \sim \mathcal{N}(X\alpha, D), \tag{29}$$

for some diagonal matrix D. A natural and simple way to model heteroskedasticity is to use a log-linear model. We can thereby model the diagonal elements of D as

$$D_{ii} = \exp(V_i\beta), \ \rightarrow \ \log(D_{ii}) = V_i\beta, \ i = 1,\ldots,n; \tag{30}$$

where V_i is a vector of v covariates measured for observation i, and β is a $[v \times 1]$ vector of coefficients which models the linear change in the log-residual-variance with that covariate vector. We can arrange the vectors of covariates V_i, $i = 1, \ldots, n$, into a design matrix for the residual variance, V, of dimension $[n \times v]$. We can then express D as

$$D = \exp(\text{diag}(V\beta)), \tag{31}$$

where $\text{diag}(V\beta)$ is the diagonal matrix with diagonal entry i equal to $V_i\beta$, and $\exp(\text{diag}(V\beta))$ is the matrix exponential of the diagonal matrix $\text{diag}(V\beta)$. A column of $1's$ models the scale of the residual variance. Alternatively, without a column of $1's$ in V, one could express D as

$$D = \sigma^2 \exp(\text{diag}(V\beta)), \tag{32}$$

which makes clear the effect of changing an element of V is to scale the residual variance up or down by some factor that depends on β.

The heteroskedastic linear model is therefore

$$Y \sim \mathcal{N}(X\alpha, \exp(\text{diag}(V\beta))). \tag{33}$$

3.1 Inference algorithm

We give the inference steps first, referencing the relevant equations where necessary, with detailed derivations in the relevant subsections. The approach we take is to optimise over the profile likelihood, $L_{\text{prof}}(\beta) = L(\hat{\alpha}_\beta, \beta)$, where $\hat{\alpha}_\beta$ is the value of α that maximises the likelihood for a particular β, the solution to (39).

1: $\alpha_{\text{OLS}} = (X^T X)^{-1} X^T y$. {Initialise α}
2: set $\hat{\beta}_*$ to the solution to (42) with $\alpha = \alpha_{\text{OLS}}$. {Initialise β}
3: Find $\hat{\beta} = \underset{\beta}{\text{argmax}}\, L_{\text{prof}}(\beta)$ using the L-BFGS-B algorithm, with $\hat{\beta}_*$ as the initial value.
4: Find $\hat{\alpha}$ as the solution to (39) for $\beta = \hat{\beta}$.
5: Compute the inverse Fisher Information Matrix (50) at $(\alpha, \beta) = (\hat{\alpha}, \hat{\beta})$ to obtain standard error estimates.

The gradients and Fisher Information Matrix used are derived below.

3.2 Likelihood

For convenience, instead of the full likelihood, we work with

$$L(\alpha, \beta | y, X, V) = 2 \log \mathcal{N}(y | X, V, \alpha, \beta) + n \log(2\pi), \tag{34}$$

where $\mathcal{N}(y|X, V, \alpha, \beta)$ is the multivariate normal density of the heteroskedastic linear model at y given X, V, α, β. Therefore, if y_i is the i^{th} observation of the phenotype and

X_i is the i^{th} row of X,

$$L = L(\alpha, \beta | y, X, V) = -\sum_{i=1}^{n} V_i\beta - \sum_{i=1}^{n}(y_i - X_i\alpha)^2 \exp(-V_i\beta). \tag{35}$$

3.3 Gradient

For notation for the first derivative of a scalar function L of a $[k \times 1]$ vector x, we express the first (partial) derivative in terms of x^T:

$$\frac{\partial L}{\partial x^T} = \left[\frac{\partial L}{\partial x_1}, \ldots, \frac{\partial L}{\partial x_k}\right] \text{ is } [1 \times k]. \tag{36}$$

This has the advantage of writing the linear approximation from the Taylor series of the scalar function as

$$L(x) \approx L(x_0) + \frac{\partial L}{\partial x^T}(x - x_0). \tag{37}$$

3.3.1 With respect to mean effects

$$\frac{\partial L}{\partial \alpha^T} = 2(y - X\alpha)^T D^{-1} X \tag{38}$$

This implies the MLE for α, $\hat{\alpha}$ must satisfy the linear system:

$$X^T D^{-1} X \hat{\alpha} = X^T D^{-1} y, \tag{39}$$

for a given β.

3.3.2 With respect to variance effects

$$\frac{\partial L}{\partial \beta^T} = \sum_{i=1}^{n}((y_i - X_i\alpha)^2 \exp(-V_i\beta) - 1)V_i \tag{40}$$

If we assume that $|V_i\beta|$ is small $\forall i$, then

$$\frac{\partial L}{\partial \beta^T} \approx \sum_{i=1}^{n}[(y_i - X_i\alpha)^2(1 - V_i\beta) - 1]V_i \tag{41}$$

If we set this approximation to zero, we can solve a linear system for an approximate MLE for β, $\hat{\beta}_*$:

$$V^T \text{diag}(e^2) V \hat{\beta}_* = \sum_{i=1}^{n}[(y_i - X_i\alpha)^2 - 1]V_i, \tag{42}$$

where e^2 is the element-wise square of the residuals: $e_i^2 = (y_i - X_i\alpha)^2$.

3.4 Second derivative and asymptotic covariance

The second derivative of L with respect to α is

$$\frac{\partial^2 L}{\partial \alpha \partial \alpha^T} = -2X^T D^{-1} X. \tag{43}$$

With respect to β, it is

$$\frac{\partial^2 L}{\partial \beta \partial \beta^T} = -\sum_{i=1}^{n} (y_i - X_i\alpha)^2 \exp(-V_i\beta) V_i^T V_i \tag{44}$$

$$= -V^T \delta V, \tag{45}$$

where δ is a diagonal matrix with diagonal element $\delta_{ii} = (y_i - X_i\alpha)^2 \exp(-V_i\beta)$. Delta is positive semi-definite $\forall\beta$ as $\delta_{ii} \geq 0\ \forall\beta$, implying that $-L$ is a convex function of β.

To complete the Hessian of the log-likelihood, we need

$$\frac{\partial^2 L}{\partial \alpha \partial \beta^T} = -2\sum_{i=1}^{n} (y_i - X_i\alpha) \exp(-V_i\beta) X_i^T V_i \tag{46}$$

$$= -2X^T \mathrm{diag}((y - X\alpha) \circ \exp(-V\beta)) V, \tag{47}$$

where $(y - X\alpha) \circ \exp(-V\beta)$ is the element-wise product of $(y - X\alpha)$ and $\exp(-V\beta)$.

Therefore the Hessian of the log-likelihood with respect to (α, β) is

$$H = \begin{bmatrix} -X^T D^{-1} X & -X^T \mathrm{diag}((y - X\alpha) \circ \exp(-V\beta)) V \\ -V^T \mathrm{diag}((y - X\alpha) \circ \exp(-V\beta)) X^T & -V^T \delta V/2, \end{bmatrix} \tag{48}$$

where we have divided by 2 to make it correspond to the true log-likelihood, not L.

The negative expectation of the Hessian, the Fisher Information Matrix, is therefore

$$-\mathbb{E}[H] = \begin{bmatrix} X^T D^{-1} X & 0 \\ 0 & V^T V/2 \end{bmatrix}, \tag{49}$$

because $\mathbb{E}[\delta_{ii}] = \exp(-V_i\beta)\mathbb{E}[(y_i - X_i\alpha)^2] = 1\ \forall\ i$, and $\mathbb{E}[(y - X\alpha)] = 0$. Therefore the inverse of the information matrix is

$$I((\alpha, \beta))^{-1} = \begin{bmatrix} (X^T D^{-1} X)^{-1} & 0 \\ 0 & 2(V^T V)^{-1} \end{bmatrix}. \tag{50}$$

This matrix will be invertible as long as X and V are of full column rank, which is also enough to ensure that the negative log-likelihood is asymptotically strictly convex, so that the maximum likelihood solutions are unique. Therefore, the asymptotic covariance of the maximum likelihood estimator of (α, β) is given by the inverse Fisher Information Matrix if X and V are of full column rank.

4 The heteroskedastic linear mixed model

We first consider a linear mixed model which allows for heteroskedasticity in both the random effects and the residual error:

$$Y = X\alpha + Z\gamma + \epsilon; \tag{51}$$

where X is the $[n \times c]$ design matrix for the fixed effects, α; Z is the $[n \times l]$ design matrix for the random effects, γ; and ϵ is the residual error vector. We define the covariance matrices:

$$H = \mathrm{Cov}(\gamma); \text{ and } D = \mathrm{Cov}(\epsilon). \tag{52}$$

We are interested in modelling the heteroskedasticity in both the random effects and the residual error. We model the heteroskedasticity in the residual error as in the previous section:

$$D = \exp(\mathrm{diag}(V\beta)). \tag{53}$$

The l random effects will in general have different variances, and the difference in variance between different random effects may depend on known covariates. If the random effects represent allelic substitution effects on a phenotype, we might expect non-synonymous coding variants to contribute more to the phenotypic variance than synonymous coding variants. In random effects models, heteroskedasticity is usually modelled by considering a partition of the l random effects into k discrete categories, with each random effect in each category having equal variance. This results in

$$ZHZ^T = \sum_{j=1}^{k} \sigma_j^2 Z_j Z_j^T. \tag{54}$$

While this can model heteroskedasticity coming from discrete, non-overlapping categories, it cannot model heteroskedasticity that follows continuous variables or multiple, overlapping variables.

The log-linear variance model offers greater flexibility in modelling the heteroskedasticity in the random effects. We consider uncorrelated random effects, so that H is diagonal:

$$H = \exp(\mathrm{diag}(W\lambda)), \tag{55}$$

where W is a $[l \times w]$ design matrix for the log-variance of the random effects, with coefficient vector λ.

One disadvantage of this model is that it becomes impossible to test the hypothesis that a particular category of random effects contributes nothing to the variance, as a zero contribution to the variance corresponds to a coefficient in λ of negative infinity. If the random effects represent different allele substitution effects, however, this may not matter, as all variants can be expected to 'contribute' a small amount to the phenotypic

variance due to population stratification and confounding with shared environment. The interpretation of the coefficients in λ for particular covariates then becomes a variance contribution above or below the background level, which may be a more meaningful question than whether there is any contribution above zero.

Assuming that the random effects and the residual error are Gaussian, this gives

$$Y|X, Z, \alpha, \beta, \lambda \sim \mathcal{N}(X\alpha, ZHZ^T + D); \tag{56}$$

$$D = \exp(\text{diag}(V\beta));\ H = \exp(\text{diag}(W\lambda)). \tag{57}$$

In the empirical analyses and software implementation, we consider a simplification of this with $H = h^2\text{I}$.

5 Efficient inference for the low-rank heteroskedastic linear mixed model

5.1 Algorithm overview

We implement the algorithm in Python using NumPy for linear algebra operations. Facilities for defining heteroskedastic linear mixed models and finding maximum likelihood estimates of parameters are provided in the Python package HLMM, which is freely available on an MIT license. We impute missing observations in the random effects matrix, Z, with the relevant column mean. We analyse only those individuals with complete observations of all the other model variables. We note that our approach has similarities to computational approaches previously used in general linear mixed models[5].

Let $\theta = (\beta, h^2)$ be the vector of variance parameters of the simplified model with $H = h^2\text{I}$. To fit the simplified model, we optimise over the profile likelihood, $L_{\text{prof}}(\theta) = L(\hat{\alpha}_\theta, \theta)$, where $\hat{\alpha}_\theta$ is the value of α that maximises the likelihood for a particular θ, the solution to (72).

1: Input an initial guess for h^2, h_0^2. {Initialise h^2}
2: Find $\hat{\beta}_{HLM}$, the maximum likelihood estimate of β in the model without the random effects, by application of algorithm in Section 3.1. {Initialise β}
3: Initialise θ as $\theta_0 = (\hat{\beta}_{HLM}, h_0^2)$.
4: Use the L-BFGS-B algorithm to find the $\hat{\theta}$ that maximises $L_{\text{prof}}(\theta)$, using θ_0 as the initial value. The likelihood and its gradient are computed using the expressions derived below.
5: Find $\hat{\alpha}$, the α that maximises $L(\alpha, \hat{\theta})$.
6: Estimate standard errors from the negative inverse of a numerical approximation to the Hessian of the log-likelihood at $(\hat{\alpha}, \hat{\beta}, \hat{h}^2)$.

For each chromosome, we first fit a null model using a user input initial guess for h^2, and we use the resulting maximum likelihood estimate for h^2 as the initial guess for h^2 for all locus specific models.

5.1.1 Overall complexity

The overall time complexity for the computation of the likelihood and gradients is

$$\mathrm{O}(nl^2 + l^3 + ncl + cl^2 + nc^2 + c^3 + nv + lw). \tag{58}$$

The overall space complexity is

$$\mathrm{O}(nl + l^2 + nc + c^2 + nv + lw). \tag{59}$$

Both the time and space complexity are linear in n when the other parameters are fixed.

5.2 Computation of the likelihood in $\mathrm{O}(n)$ Operations

As D is not proportional to the identity matrix, a rotation defined by the eigenvectors of Z does not diagonalise the system. However, the likelihood and its derivative can still be computed in $\mathrm{O}(n)$ by taking advantage of the structure of the covariance of Y, which is a diagonal matrix plus a low rank matrix.

Let

$$\mathrm{Cov}(Y) = \Sigma = ZHZ^T + D, \tag{60}$$

then the log-likelihood is

$$l = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(y - X\alpha)^T\Sigma^{-1}(y - X\alpha) \tag{61}$$

Instead of maximising l, we equivalently maximise $L = 2l + n\log(2\pi)$:

$$L = -\ln|\Sigma| - (y - X\alpha)^T\Sigma^{-1}(y - X\alpha). \tag{62}$$

To naively compute the likelihood, one needs the inverse of Σ, computation of which requires $\mathrm{O}(n^3)$ operations. By application of the Woodbury Matrix Identity, the inverse of Σ can be reduced to the inverse of D, which is diagonal, plus a low rank correction:

$$\Sigma^{-1} = D^{-1} - D^{-1}Z(H^{-1} + Z^TD^{-1}Z)^{-1}Z^TD^{-1}. \tag{63}$$

Let $\Lambda = H^{-1} + Z^TD^{-1}Z$, then we also have, by the Matrix Determinant Lemma,

$$\log|\Sigma| = \log|\Lambda| + \log|H| + \log|D| \tag{64}$$

$$\log|\Sigma| = \log|\Lambda| + \sum_{j=1}^{l} W_j\lambda + \sum_{i=1}^{n} V_i\beta. \tag{65}$$

Therefore,

$$L = -\sum_{i=1}^{n} V_i\beta - \sum_{i=1}^{n}(y_i - X_i\alpha)^2\exp(-V_i\beta) - \sum_{j=1}^{l} W_j\lambda \tag{66}$$

$$-\log|\Lambda| + [Z^TD^{-1}(y - X\alpha)]^T\Lambda^{-1}[Z^TD^{-1}(y - X\alpha)] \tag{67}$$

This can be computed in $\mathrm{O}(nl^2 + l^3)$ operations by precomputing the $[l \times 1]$ vector

$$r = Z^T D^{-1}(y - X\alpha). \tag{68}$$

The likelihood can thereby be expressed as

$$\begin{aligned} L = &- \sum_{i=1}^{n} V_i\beta - \sum_{i=1}^{n}(y_i - X_i\alpha)^2 \exp(-V_i\beta) - \\ &\sum_{j=1}^{l} W_j\lambda - \log|\Lambda| + r^T\Lambda^{-1}r, \end{aligned} \tag{69}$$

where the first line is the likelihood for the diagonal system without the random effect, and the second line is the contribution to the likelihood of the random effects. The computation is dominated by calculation of Λ in $\mathrm{O}(nl^2)$ operations and its inverse and determinant in $O(l^3)$ operations.

5.3 Efficient computation of the maximum likelihood estimator of the fixed effects

The derivative of the log likelihood with respect to α is

$$\frac{\partial L}{\partial \alpha^T} = 2(y - X\alpha)^T D^{-1}X - 2(y - X\alpha)^T D^{-1}Z\Lambda^{-1}Z^T D^{-1}X \tag{70}$$

$$= 2(y - X\alpha)^T [D^{-1}X - D^{-1}Z\Lambda^{-1}Z^T D^{-1}X] \tag{71}$$

To find the MLE, we equate the derivative to zero and solve for α. We get that the MLE for α, $\hat{\alpha}$ must satisfy the linear system:

$$[X^T D^{-1}X - X^T D^{-1}Z\Lambda^{-1}Z^T D^{-1}X]\hat{\alpha} = X^T D^{-1}y - X^T D^{-1}Z\Lambda^{-1}Z^T D^{-1}y \tag{72}$$

$X^T D^{-1}X$ can be computed in $\mathrm{O}(nc^2)$ operations; $X^T D^{-1}Z$ is a $[c \times l]$ matrix which can be computed in $\mathrm{O}(ncl)$ operations; so assuming that Λ^{-1} has already been computed, and that $[X^T D^{-1}X - X^T D^{-1}Z\Lambda^{-1}Z^T D^{-1}X]$ is full rank, $\hat{\alpha}$ can be computed in $\mathrm{O}(nc^2 + ncl + cl^2 + c^3)$ operations.

5.4 Derivative with respect to variance parameters

In Appendix A, we derive the derivatives of the likelihood with respect to the variance parameters. We give the results here.

5.4.1 Derivative with respect to λ

$$\frac{\partial L}{\partial \lambda^T} = \sum_{j=1}^{l} [(\Lambda_{jj}^{-1} + \Gamma_{jj}) \exp(-W_j \lambda) - 1] W_j;\ \Gamma = \Lambda^{-1} r r^T \Lambda^{-1}. \tag{73}$$

Computing this gradient requires computation of Γ in $\mathrm{O}(l^3)$ operations, then an $O(lw)$ operation. It is therefore linear in the number of heteroskedasticity parameters for the random effect, w.

5.4.2 Derivative with respect to β

$$\frac{\partial L}{\partial \beta^T} = \sum_{i=1}^{n} (k_i \exp(-V_i \beta) - 1) V_i, \tag{74}$$

where k is a function of Λ, X, Z, D, and the residuals. To compute k requires $\mathrm{O}(nl^2)$ operations, then to complete the gradient computation requires an $\mathrm{O}(nv)$ operation, so the gradient computation is linear in the number of log-linear variance parameters.

6 Detecting dispersion effects

6.1 Dispersion effects

Consider a phenotype Y and a bi-allelic genetic variant G_l, then the effect of the genotype on the mean of Y is captured by the conditional means:

$$\mathbb{E}[Y|G_l = g] = \mu_{lg},\ g = 0, 1, 2. \tag{75}$$

Here the subscript l indicates the particular genetic variant examined. Furthermore, the effect of the genotype on the variance of Y is captured by the conditional variance:

$$\mathrm{Var}[Y|G_l = g] = \sigma_{lg}^2,\ g = 0, 1, 2. \tag{76}$$

If the effect of the genotype on the phenotypic variance can be explained by the mean-variance relation of the phenotype distribution, then, for some function h,

$$\mathrm{Var}[Y|G_l = g] = h(\mu_{lg}),\ g = 0, 1, 2. \tag{77}$$

This relationship between genotype and variance can be removed (approximately) by a variance stabilising transform, η,

$$\mathrm{Var}[\eta(Y)|G_l = g] \approx 1;\ g = 0, 1, 2;\ \text{where } \eta(x) = \int \frac{dx}{\sqrt{h(x)}}. \tag{78}$$

We aim to detect changes in phenotypic variance that cannot be explained by a change in mean with genotype and are approximately invariant under transformations of the phenotype, which we term 'dispersion effects'. If the conditional variances are, for some function Δ_l,

$$\mathrm{Var}[Y|G_l = g] = h(\mu_{lg})\Delta_l(g),\ g = 0, 1, 2. \tag{79}$$

where $\Delta_l(g)$ is not a function of μ_{lg}, then the conditional variances of a transformed phenotype $\tau(Y)$, where τ is a differentiable function, are

$$\mathrm{Var}[\tau(Y)|G_l = g] \approx h(\mu_{lg})[\tau'(\mu_{lg})]^2\Delta_l(g),\ g = 0, 1, 2. \tag{80}$$

The relationship between the genotype and conditional variances of $\tau(Y)$ is not removed because $\Delta_l(g)$ is not a function of the conditional mean, μ_{lg}. The function $\Delta_l(g)$ can therefore be taken to represent the 'dispersion effect' of the genotype on the phenotype. We consider a log-linear functional form for the dispersion effect of a genotype: $\Delta_l(g) = e^{d_l(g-2f_l)}$, where d_l is the log-linear dispersion effect and f_l is the allele frequency.

Our variance model is therefore,

$$\mathrm{Var}[Y|G_l = g] = h(\mu_{lg})e^{d_l(g-2f_l)},\ g = 0, 1, 2. \tag{81}$$

Transformation of the phenotype by a function τ, to a first-order approximation, affects only the mean-variance relation, and not the log-linear dispersion effect:

$$\mathrm{Var}[\tau(Y)|G_l = g] \approx h_\tau(\mu_{lg})e^{d_l(g-2f_l)},\ g = 0, 1, 2, \tag{82}$$

where $h_\tau(\mu_{lg}) = h(\mu_{lg})[\tau'(\mu_{lg})]^2$.

6.2 Mean and variance effects after transformation

Assuming that there is a linear relation between phenotypic mean and genotype for the untransformed phenotype, Y, then $\mu_{lg} = \mu + a_l(g - 2f_l)$, where a_l is the additive effect of the genotype on the phenotypic mean. The conditional means of the transformed phenotype are:

$$\mathbb{E}[\tau(Y|G_l = g)] \approx \tau(\mu_{lg}) + \tau''(\mu_{lg})\sigma^2_{lg}/2,\ g = 0, 1, 2. \tag{83}$$

Assuming that the additive effect on the untransformed phenotype is small, so that $\mathrm{O}(a_l^2)$ terms can be ignored,

$$\tau(\mu_{lg}) \approx \tau(\mu) + \tau'(\mu)a_l(g - 2f_l). \tag{84}$$

Assuming that the dispersion effect is also small, so that $\mathrm{O}(d_l^2)$ terms can be ignored, we have

$$\tau''(\mu_{lg})\sigma^2_{lg} \approx \tau''(\mu)h(\mu) + [(\tau'''(\mu)h(\mu) + \tau''(\mu)h'(\mu))a_l + \tau''(\mu)h(\mu)d_l](g - 2f_l). \tag{85}$$

Therefore,

$$\mathbb{E}[\tau(Y|G_l = g)] \approx \mu_\tau + a_{\tau l}(g - 2f_l) + \tau''(\mu)h(\mu)d_l(g - 2f_l)/2,\ g = 0, 1, 2, \tag{86}$$

where $\mu_\tau = \tau(\mu) + \tau''(\mu)h(\mu)/2$ is the approximate mean of the transformed phenotype, and

$$a_{\tau l} = [\tau'(\mu) + (\tau'''(\mu)h(\mu) + \tau''(\mu)h'(\mu))/2]a_l. \tag{87}$$

The additive effect on the transformed phenotype, α_l, is therefore

$$\alpha_l \approx a_{\tau l} + \tau''(\mu)h(\mu)d_l/2 \tag{88}$$

This shows that a component of the additive effect on the transformed phenotype is due to the dispersion effect on the untransformed phenotype, with the magnitude of this component increasing with the magnitude of the second derivative of the transformation function at the untransformed phenotype mean.

Assuming that h is differentiable and ignoring $\mathrm{O}(a_l^2)$ terms:

$$\log(\mathrm{Var}[\tau(Y)|G_l = g]) \approx \log(h_\tau(\mu)) + \frac{h'_\tau(\mu)}{h_\tau(\mu)}a_l(g - 2f_l) + d_l(g - 2f_l),\ g = 0, 1, 2. \tag{89}$$

Therefore the log-linear variance effect on the transformed phenotype is

$$\alpha_{vl} \approx r_{\mathrm{av}}a_{\tau l} + d_l, \tag{90}$$

where

$$r_{\mathrm{av}} = \frac{h'_\tau(\mu)}{h_\tau(\mu)[\tau'(\mu) + (\tau'''(\mu)h(\mu) + \tau''(\mu)h'(\mu))/2]}. \tag{91}$$

The log-linear dispersion effect is therefore

$$d_l \approx \left(\frac{2}{2 - r_{\mathrm{av}}\tau''(\mu)h(\mu)}\right)(\alpha_{vl} - r_{\mathrm{av}}\alpha_l). \tag{92}$$

In practice, unless there is a strong relationship between mean and variance effects after transformation (large r_{av}) and the transformation function is highly non-linear around the phenotypic mean (large $\tau''(\mu)$),

$$d_l \approx \alpha_{vl} - r_{\mathrm{av}}\alpha_l. \tag{93}$$

This suggests dispersion effects can be estimated from estimates of α_{vl}, α_l, and r_{av}.

6.3 Estimating dispersion effects

To ensure test statistics are properly calibrated, and that the sampling distributions of additive and log-linear variance effects are uncorrelated, phenotypes are inverse-normally transformed. For large samples from continuous phenotypes, inverse-normal transformation corresponds closely to transformation by a differentiable function τ (Appendix B). For all genetic variants $l = 1, \ldots, L$ without large phenotypic effects,

$$\tau(Y)|G_l = g \sim \mathcal{N}\left(\mu_\tau + \alpha_l(g - 2f_l), h_\tau(\mu_{lg})e^{d_l(g-2f_l)}\right). \tag{94}$$

This is true because, although the transformation function τ is specified to ensure normality of Y, rather than $Y|G_l = g$, if the phenotypic distribution function is only altered slightly by conditioning on G_l, then $\tau(Y)|G_l = g$ should also be approximately normally distributed. (Note that this would not be true for genetic loci with very large phenotypic effects, where the residuals would no longer be normal[6]).

Let $\tau''(\mu)h(\mu)/2 = k$. By fitting the heteroskedastic linear model to estimate additive (α_l) and log-linear variance effects (α_{vl}), asymptotically

$$\begin{bmatrix}\hat{\alpha}_l \\ \hat{\alpha}_{vl}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}a_{\tau l} + kd_l \\ r_{\text{av}}a_{\tau l} + d_l\end{bmatrix}, \begin{bmatrix}\text{Var}(\hat{\alpha}_l) & 0 \\ 0 & \text{Var}(\hat{\alpha}_{vl})\end{bmatrix}\right), \tag{95}$$

where $\text{Var}(\hat{\alpha}_l)$ and $\text{Var}(\hat{\alpha}_{vl})$ are the asymptotical variances of the maximum likelihood estimators (50).

We now give a procedure for estimation of r_{av} from genome-wide estimates of α_l and α_{vl}. Assuming that α_l is known without error, then r_{av} is (approximately) the regression coefficient of $\hat{\alpha}_{vl}$ on α_l across the loci. Assuming that k is small, so terms involving k^2 can be ignored,

$$\frac{\text{Cov}_l(\hat{\alpha}_{vl}, \alpha_l)}{\text{Var}_l(\alpha_l)} \approx r_{\text{av}} + k\frac{\text{Var}_l(d_l)}{\text{Var}_l(\alpha_l)} + (1 + kr_{\text{av}})r_{\text{ad}}\sqrt{\frac{\text{Var}_l(d_l)}{\text{Var}_l(\alpha_l)}}, \tag{96}$$

where the l subscript on $\text{Cov}_l(\hat{\alpha}_{vl}, \alpha_l)$ indicates the covariance is across the loci, not the sampling covariance within a locus, and $r_{\text{ad}} = \text{Corr}_l(a_{\tau l}, d_l)$. For most traits, one would expect that $\text{Var}_l(d_l) << \text{Var}_l(\alpha_l)$. Therefore,

$$\frac{\text{Cov}_l(\hat{\alpha}_{vl}, \alpha_l)}{\text{Var}_l(\alpha_l)} \approx r_{\text{av}}. \tag{97}$$

However, we only have noisy estimates of α_l, so we regress $\hat{\alpha}_{vl}$ on $\hat{\alpha}_l$. This regression coefficient is

$$\frac{\text{Cov}_l(\hat{\alpha}_{vl}, \hat{\alpha}_l)}{\text{Var}_l(\hat{\alpha}_l)} \approx r_{\text{av}}\left(1 + \frac{\mathbb{E}_l[\text{Var}(\hat{\alpha}_l)]}{\text{Var}_l(\hat{\alpha}_l) - \mathbb{E}_l[\text{Var}(\hat{\alpha}_l)]}\right)^{-1}. \tag{98}$$

This regression gives a biased estimate of r_{av} due to noise in the estimation of α_l, an example of regression dilution. The bias decreases with the signal to noise ratio in the

distribution of α_l over the loci: to what degree variation in α_l represents variation in real additive effects versus sampling error. Therefore,

$$r_{\text{av}} \approx \frac{\text{Cov}_l(\hat{\alpha}_{vl}, \hat{\alpha}_l)}{\text{Var}_l(\hat{\alpha}_l)} \left(1 + \frac{\mathbb{E}_l[\text{Var}(\hat{\alpha}_l)]}{\text{Var}_l(\hat{\alpha}_l) - \mathbb{E}_l[\text{Var}(\hat{\alpha}_l)]}\right), \tag{99}$$

which can be estimated from the genome-wide distributions of $\hat{\alpha}_{vl}$ and $\hat{\alpha}_l$.

When very many (not strongly linked) loci with varying additive effects have been analysed, sampling error in r_{av} will be very small. Ignoring sampling variation in the estimation of r_{av}, we have

$$\hat{d}_l = \hat{\alpha}_{vl} - r_{\text{av}}\hat{\alpha}_l;\ \text{Var}(\hat{d}_l) = \text{Var}(\hat{\alpha}_{vl}) + r_{\text{av}}^2\text{Var}(\hat{\alpha}_v). \tag{100}$$

This gives a test statistic for a non-zero dispersion effect:

$$\frac{\hat{d}_l^2}{\text{Var}(\hat{\alpha}_{vl}) + r_{\text{av}}^2\text{Var}(\hat{\alpha}_v)} \sim \chi_1^2. \tag{101}$$

We note that sample estimates of r_{av} may be sensitive to a few strong effect loci with both additive effects and dispersion effects. We recommend estimating r_{av} using robust regression techniques to reduce this effect (Online Methods).

References

[1] Cao, Y., Wei, P., Bailey, M., Kauwe, J. S. K., and Maxwell, T. J. A versatile omnibus test for detecting mean and variance heterogeneity. *Genetic Epidemiology*, **38**(1):51–59 2014. ISSN 07410395. doi:10.1002/gepi.21778.

[2] Barton, N. H., Etheridge, A. M., and Véber, A. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, **118**:50–73 2017. ISSN 10960325. doi:10.1016/j.tpb.2017.06.001.

[3] Brillinger, D. R. Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics*, **18**(2000):163–182 2004.

[4] Harvey, A. A. C. Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, **44**(3):461–465 1976.

[5] Wolfinger, R., Tobiast, R., Sall, J., Tobias, R., and Sall, J. Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models. *SIAM Journal on Scientific Computing*, **15**(6):1294–1310 1994. doi:10.1137/0915079.

[6] Beasley, T. M., Erickson, S., and Allison, D. B. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics*, **39**(5):580–595 2009. ISSN 00018244. doi:10.1007/s10519-009-9281-0.

[7] Magnus, J. R. and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley Series in Probability and Statistics: Texts and References Section. Wiley 1999. ISBN 9780471986331.

A Computation of the derivative with respect to the variance parameters

To compute the derivative with respect to the variance parameters, we use the method of differentials[7] to compute the infinitesimal change in the log-likelihood, dL, with respect to infinitesimal changes in λ or β.

We illustrate this with an example. For a scalar function, L, of a column vector, x, one computes the infinitesimal change in L, dL, with an infinitesimal change in a the vector, dx:

$$dL = \frac{\partial L}{\partial x^T} dx, \tag{102}$$

which corresponds to the linear term in the Taylor expansion of L:

$$L = L(0) + \frac{\partial L}{\partial x^T} dx + \dots \tag{103}$$

Note that the infinitesimal change, dx, is of the same dimension as x, whereas $\frac{\partial L}{\partial x^T}$ has dimension equal to x^T.

In deriving the differentials with respect to the variance parameters, we utilise the differential formulae[7]:

$$d\Lambda^{-1} = -\Lambda^{-1} d\Lambda \Lambda^{-1},\ d\log|\Lambda| = \mathrm{tr}(\Lambda^{-1} d\Lambda). \tag{104}$$

A.1 Derivative with respect to λ

We compute the infinitesimal change in L with respect to an infinitesimal change in λ, $d\lambda$. The differential of the log-likelihood with respect to λ relies upon the differential of Λ with respect to λ:

$$dL = -\sum_{j=1}^{l} W_j d\lambda - \mathrm{tr}(\Lambda^{-1} d\Lambda) - r^T \Lambda^{-1} (d\Lambda) \Lambda^{-1} r. \tag{105}$$

The differential of Λ with respect to λ is

$$d\Lambda = dH^{-1} = -H^{-1} \mathrm{diag}(W d\lambda) \tag{106}$$

Therefore,

$$-\mathrm{tr}(\Lambda^{-1}d\Lambda) = \mathrm{tr}(\Lambda^{-1}H^{-1}\mathrm{diag}(Wd\lambda)) \tag{107}$$
$$= \sum_{j=1}^{l} \Lambda_{jj}^{-1} \exp(-W_j\lambda)W_j d\lambda.$$

For the other component of the differential, we have

$$-r^T\Lambda^{-1}(d\Lambda)\Lambda^{-1}r = r^T\Lambda^{-1}H^{-1}\mathrm{diag}(Wd\lambda)\Lambda^{-1}r \tag{108}$$
$$= \mathrm{tr}(\Lambda^{-1}rr^T\Lambda^{-1}H^{-1}\mathrm{diag}(Wd\lambda)). \tag{109}$$

Let $\Gamma = \Lambda^{-1}rr^T\Lambda^{-1}$, then

$$-r^T\Lambda^{-1}(d\Lambda)\Lambda^{-1}r = \sum_{j=1}^{l} \Gamma_{jj} \exp(-W_j\lambda)W_j d\lambda. \tag{110}$$

Therefore,

$$dL = -\sum_{j=1}^{l} W_j d\lambda + \sum_{j=1}^{l} \Lambda_{jj}^{-1} \exp(-W_j\lambda)W_j d\lambda + \sum_{j=1}^{l} \Gamma_{jj} \exp(-W_j\lambda)W_j d\lambda. \tag{111}$$

Therefore,

$$\frac{\partial L}{\partial \lambda^T} = \sum_{j=1}^{l} [(\Lambda_{jj}^{-1} + \Gamma_{jj}) \exp(-W_j\lambda) - 1]W_j. \tag{112}$$

A.2 Derivative with respect to β

To aid differentiation, we rewrite Λ to make its reliance on β explicit:

$$\Lambda = H^{-1} + Z^T D^{-1} Z = H^{-1} + \sum_{i=1}^{n} Z_i^T Z_i \exp(-V_i\beta), \tag{113}$$

where Z_i is the i^{th} $[1 \times l]$ row vector of Z.

We also rewrite r to make its dependence on β explicit.

$$r = \sum_{i=1}^{n} Z_i^T (y_i - X_i\alpha) \exp(-V_i\beta). \tag{114}$$

The differential of the likelihood with respect to a change in β is

$$dL = -\sum_{i=1}^{n} V_i d\beta + \sum_{i=1}^{n} (y_i - X_i\alpha)^2 \exp(-V_i\beta)V_i d\beta - \tag{115}$$
$$\mathrm{tr}(\Lambda^{-1}d\Lambda) + d(r^T\Lambda^{-1}r).$$

The differential of Λ with respect to β is

$$d\Lambda = -\sum_{i=1}^{n} Z_i^T Z_i \exp(-V_i\beta) V_i d\beta. \tag{116}$$

It can therefore be shown that,

$$-\mathrm{tr}(\Lambda^{-1} d\Lambda) = \sum_{i=1}^{n} Z_i \Lambda^{-1} Z_i^T \exp(-V_i\beta) V_i d\beta. \tag{117}$$

We use the fact that

$$d(r^T \Lambda^{-1} r) = 2r^T \Lambda^{-1} dr - r^T \Lambda^{-1} d\Lambda \Lambda^{-1} r \tag{118}$$

to derive the differential of the inner product $r^T \Lambda^{-1} r$.

The differential of r with respect to β is

$$dr = -\sum_{i=1}^{n} Z_i^T (y_i - X_i\alpha) \exp(-V_i\beta) V_i d\beta. \tag{119}$$

The differential of Λ^{-1} is

$$\begin{aligned} d\Lambda^{-1} &= -\Lambda^{-1} d\Lambda \Lambda^{-1} \\ &= \sum_{i=1}^{n} \Lambda^{-1} Z_i^T Z_i \Lambda^{-1} \exp(-V_i\beta) V_i d\beta. \end{aligned} \tag{120}$$

It can then be shown that

$$d(r^T \Lambda^{-1} r) = \sum_{i=1}^{n} r^T \Lambda^{-1} Z_i^T \left(Z_i \Lambda^{-1} r - 2(y_i - X_i\alpha) \right) \exp(-V_i\beta) V_i d\beta. \tag{121}$$

This can be calculated efficiently by realising that $Z_i \Lambda^{-1} r = r^T \Lambda^{-1} Z_i^T$, and that this is the i^{th} element of the vector

$$a = Z\Lambda^{-1}[Z^T D^{-1}(y - X\alpha)]. \tag{122}$$

Therefore, the differential is

$$d(r^T \Lambda^{-1} r) = \sum_{i=1}^{n} a_i \left(a_i - 2(y_i - X_i\alpha) \right) \exp(-V_i\beta) V_i d\beta \tag{123}$$

Therefore,

$$dL = -\sum_{i=1}^{n} V_i d\beta + \sum_{i=1}^{n} (y_i - X_i\alpha)^2 \exp(-V_i\beta) V_i d\beta + \tag{124}$$
$$\sum_{i=1}^{n} Z_i \Lambda^{-1} Z_i^T \exp(-V_i\beta) V_i d\beta + a_i \left(a_i - 2(y_i - X_i\alpha)\right) \exp(-V_i\beta) V_i d\beta$$

Therefore,

$$\frac{\partial L}{\partial \beta^T} = \sum_{i=1}^{n} \{(y_i - X_i\alpha)^2 + Z_i\Lambda^{-1}Z_i^T + a_i\left(a_i - 2(y_i - X_i\alpha)\right)\} \exp(-V_i\beta) V_i - \sum_{i=1}^{n} V_i. \tag{125}$$

Let

$$k_i = (y_i - X_i\alpha)^2 + Z_i\Lambda^{-1}Z_i^T + a_i\left(a_i - 2(y_i - X_i\alpha)\right), \tag{126}$$

then

$$\frac{\partial L}{\partial \beta^T} = \sum_{i=1}^{n} (k_i \exp(-V_i\beta) - 1) V_i. \tag{127}$$

B Inverse normal transformation

If a phenotype Y admits a continuous cumulative distribution function, F_Y, then

$$\tau(Y) = \Phi^{-1}(F_Y(Y)) \sim \mathcal{N}(0,1), \tag{128}$$

where Φ is the cumulative distribution function for the standard normal distribution. If F_Y is absolutely continuous, then τ is differentiable, with derivative

$$\tau'(y) = \frac{f_Y(y)}{\phi(\tau(y))}, \tag{129}$$

where f_Y is the density function of the phenotype, and ϕ is the standard normal density function.

Without knowing the true distribution function of Y, one can replace F_Y with the empirical cumulative distribution function,

$$\hat{F}_Y(y) = \frac{1}{1+n} \sum_{i=1}^{n} \mathbf{1}_{y_i \leq y}, \tag{130}$$

where $\mathbf{1}_{y_i \leq y}$ indicates whether observation i of Y is less than or equal to y or not. The empirical inverse-normal transformation function becomes

$$\hat{\tau}(Y) = \Phi^{-1}(\hat{F}_Y(Y)). \tag{131}$$

By the Glivenko-Cantelli theorem, $\hat{F}_Y$ converges uniformly to F_Y, so, by the Continuous Mapping Theorem, $\hat{\tau}(Y)$ converges almost surely to $\tau(Y)$. The rate of convergence of $\hat{F}_Y$ to F_Y is the standard $\sqrt{n}$ rate of convergence (Donsker's Theorem), so for the large sample sizes seen in contemporary human genetics, $\hat{\tau}(Y)$ should closely approximate $\tau(Y)$. Therefore results about the moments of $\tau(Y)$ should accurately approximate results about the moments of $\hat{\tau}(Y)$.

C Population structure control

The mean of a phenotype may differ between populations that are genetically different, which can generate spurious additive associations. To reduce this effect, genetic principal components are often included in the model as covariates (affecting the mean).

Analogously, the variance of a phenotype may differ between populations that are genetically different. For example, in the UK Biobank, the variance of BMI in people of self-declared British ethnicity is higher than those of self-declared Chinese ethnicity. This could lead to inflation of log-linear variance test statistics if not properly controlled for. We argue, by analogy to population structure affecting the mean, that using genetic principal components as variance covariates in a log-linear variance model can reduce the inflation of log-linear variance test statistics.

To simulate geographic structure in the mean and variance of a phenotype distribution, we used variables from the UK Biobank that give the north (Data-Field 129) and east (Data-Field 130) co-ordinates of the individuals' place of birth in the UK as mean and variance covariates. We simulated phenotypes for the British subsample of the UK Biobank interim data release. We used the following model to simulate phenotypes:

$$Y \sim \mathcal{N}(\text{north} - \text{east}, \exp(0.2[\text{north} - \text{east}])). \tag{132}$$

This created a trait where both the mean and variance of the trait differed greatly between different regions of the UK, despite there being no genetic component to the trait. We fitted models with linear mean and log-linear variance effects for each SNP, with and without the top 20 principal components and the genotyping array as mean and variance covariates. The mean log-likelihood ratio test statistic under the null should be 1, which is what could be achieved with perfect control of population structure in this simulation with no real genetic effects.

Without fitting any mean and variance covariates, the mean log-likelihood ratio test statistics across loci on chromosome 22 were: 5.78 for the additive test, and 4.03 for the log-linear variance test. This indicates very strong mean and variance population

structure. We saw no evidence for a correlation between allele frequency and log-linear variance test statistic (sample correlation 0.003).

For this analysis, we used the top 20 principal components and the genotyping array as mean and variance covariates. This reduced the mean test statistics to 1.19 (additive test), and 1.13 (log-linear variance test). Here, fitting principal components as variance covariates is clearly effective in reducing the inflation of log-linear variance test statistics.

For computational efficiency in additive genome-wide association studies, the maximum likelihood estimates of the mean covariates from the null model can be used to 'project out' their effects before fitting models for specific SNPs. This enables the fitting of SNP-specific models with only a couple of mean parameters. If $\hat{\alpha}_0$ is the maximum likelihood estimate of the mean effects in the null model, then one transforms the phenotype, Y, to

$$Y - X\hat{\alpha}_0. \tag{133}$$

In a similar fashion, the phenotype can be rescaled so as to remove the influence of the variance covariates in the null model, reducing the number of variance parameters to fit for each SNP. If $\hat{\beta}_0$ is the maximum likelihood estimate of the variance effects in the null model, the transform performed is

$$Y \to \exp\left(\mathrm{diag}(-0.5V\hat{\beta}_0)\right)(Y - X\hat{\alpha}_0). \tag{134}$$

In our simulation of a trait with mean and variance structure, we tested if performing this transform and fitting only SNP specific mean and variance effects was effective at controlling for population structure. We used the same mean and variance covariates as above (top 20 principal components and the genotyping array). The mean additive test statistic was 1.12, and the mean log-linear variance test statistic was 1.24. Performing this transform is therefore approximately as effective at controlling for the effects of structure on the test statistics as fitting the full model at each SNP, while being computationally more efficient. There may be a loss in power, however, for causal SNPs that are correlated with mean and variance covariates in the null model.