Mean-Variance QTL Mapping on a Background of Variance Heterogeneity

Robert W. Corty* **and William Valdar***,[1]
*University of North Carolina at Chapel Hill, Department of Genetics

ABSTRACT Most QTL mapping approaches seek to identify "mean QTL", genetic loci that influence the phenotype mean, after assuming that all individuals in the mapping population have equal residual variance. Recent work has broadened the scope of QTL mapping to identify genetic loci that influence phenotype variance, termed "variance QTL", or some combination of mean and variance, which we term "mean-variance QTL". Even these approaches, however, fail to address situations where some other factor, be it an environmental factor or a distant genetic locus, influences phenotype variance. We term this situation "background variance heterogeneity" (BVH) and used simulation to explore its effects on the power and false positive rate of tests for mean QTL, variance QTL, and mean-variance QTL. Specifically, we compared traditional tests, linear regression for mean QTL and Levene's test for variance QTL, with tests more recently developed, namely Cao's tests for all three types of QTL, and tests based on the double generalized linear model (DGLM), which, unlike the other approaches, explicitly models BVH. Simulations showed that, when used in conjunction with a permutation procedure, the DGLM-based tests accurately control false positive rate and are more powerful than the other tests. We also discovered that the rank-based inverse normal transform, often used to corral unruly phenotypes, can be used to mitigate the adverse effects of BVH in some circumstances. We applied the DGLM approach, which we term "mean-variance QTL mapping", to publicly available data on a mouse backcross of CAST/Ei into M16i and, after accommodating BVH driven by father, identified a new mean QTL for bodyweight at three weeks of age.

KEYWORDS
QTL mapping, background variance heterogeneity, vQTL

INTRODUCTION

Quantitative trait locus (QTL) mapping in both model organisms and humans has traditionally focused on finding regions of the genome whose allelic variation influences the phenotypic mean. In the past decade, a number of studies and proposed methods have broadened the scope of QTL mapping to consider effects on the phenotypic variance (Paré *et al.* 2010; Rönnegård and Valdar 2011; Hulse and Cai 2013). These studies and their findings have raised interesting questions and possibilities about underlying biology, evolutionary trajectory, and potential utility in agriculture (Wagner *et al.* 1997; Dworkin 2005; Mulder *et al.* 2015). Nonetheless, consideration of variance effects — whether as the target of inference or as a feature of the data to be accommodated — has thus far remained outside of routine genetic analysis. This may be in part because QTL effects on the variance are sometimes considered of esoteric secondary interest, intrinsically controversial in their interpretation (Sun *et al.* 2013; Shen and Ronnegard 2013), or a priori too hard to detect (Visscher and Posthuma 2010). But it is also likely to be in part because familiar software and procedures are currently lacking, and because the advantages of modeling heterogeneous variance, even when targeting QTL effects on the phenotypic mean, remain underappreciated and largely undemonstrated.

The predominant approach to QTL mapping in model organisms, the focus here, considers each genetic locus in turn, using a standard linear model (SLM) to regress the phenotypes of the mapping population on their genotypes or their inferred genotype probabilities (Lander and Botstein 1989; Haley and Knott 1992). This SLM-based approach is primarily able to detect genomic regions containing a subset of genetic factors of interest — those that drive heterogeneity of phenotype mean. Despite this limited scope,

Manuscript compiled: Friday 16th March, 2018
[1]Corresponding author: William Valdar, 120 Mason Farm Road, Genetic Medicine Building, Suite 5113, Campus Box 7264, Chapel Hill, NC, 27599, USA, william.valdar@unc.edu

however, its use is widespread due to its ease of use, the straightforward interpretation of its detected QTL, its historical importance in the fields of agricultural and livestock genetics, and the fact that many genetic factors truly do influence the expected value of phenotypes. Indeed, SLM-based interval mapping has yielded important insights on commerically- and medically-important traits across many organisms for many years.

The goal of QTL mapping, however, is much broader — to identify genetic factors that influence the phenotype in any way. For example, a genetic factor that influences the sensitivity of the phenotype to micro-environmental variation is of interest, but unless it also affects the mean it is undetectable by the SLM. Similarly, a genetic factor that influences the phenotype through many epistatic interactions but has an average effect near zero is unlikely to be detected by the SLM. These examples motivate the use of statistical tests that can detect genetic effecsts on other aspects of the phenotypic distribution, most notably the phenotype variance (Paré *et al.* 2010).

A number of statistical models and methods have been developed or adapted to identify associations between genotype and phenotypic variance. These include: Levene's test (Struchalin *et al.* 2010), the Fligner-Killeen test (Fraser and Schadt 2010), Bartlett's test (Freund *et al.* 2013), as well as the double generalized linear model (DGLM) and similar (Rönnegård and Valdar 2011; Cao *et al.* 2014) [see also references in (Rönnegård and Valdar 2012)]. Tests have also been developed to detect genotype associations with arbitrary functions of the phenotype, for example higher moments. These include a variant of the Komolgorov-Smirnov test (Aschard *et al.* 2013) and a semiparametric exponential tilt model (Hong *et al.* 2016). The additional flexibility of these latter models makes them promising — a genetic factor that influences, e.g., the kurtosis of a phenotype should be of interest — but at present neither can accommodate covariates and the flexibility that affords them the ability to detect higher order effects brings with it a decreased power to detect mean and variance effects.

The existence of a genetic factor affecting the variance, sometimes termed a "vQTL", or indeed any factor affecting the variance has implications regarding statistical genetic analyses, both those targeting variance effects and those targeted mean-affecting QTL (hereafter, "mQTL"), and these implications have been relatively unexamined.

In particular, if a genetic (or other) factor influences phenotype variance then it follows that examination and testing of any other QTL effect — for example, that of a QTL elsewhere in the genome — must occur against a backdrop of systematically heterogeneous residual variance. The presence of this "background variance heterogeneity" (BVH) when testing for a (foreground) effect simultaneously presents analytic challenges and opportunities, not only for mapping vQTL but also the validity of studies detecting mQTL.

The impact of BVH on mapping mQTL can be thought of as a disruption of the natural observation weights: The SLM assumes the phenotype of every individual is subject to equal noise variance and therefore equal weight; but if it is known that some individuals' phenotypes are inherently less noisy — due to BVH induced by either a vQTL or other factors such as sex, housing, strain or experimenter — then those data should be upweighted, and this would lead to a more powerful test for mQTL detection. Conversely, giving equal weight to subgroups of the data that are inherently noisier than average has the potential to leave outliers with overmuch much influence on the regression, increasing the potential for false positive mQTL detections A case in point is when an mQTL also has variance effects: here the effects on the variance are a type of proximal BVH, and modeling them explicity improves ability to detect effects on the mean Corty *et al* 2018+. Knowledge and appropriate modeling of variance heterogeneity therefore has important implications for making mean-controlling QTL studies sensitive, robust and reproducible.

The impact of BVH on detection of foreground vQTL is more subtle. Parametric methods to identify vQTL typically pit heterogeneous variance alternative models against a homoskedastic, normally distributed null. However, under BVH the null model is not homoskedastic — it is a scale mixture — and this risks the null being rejected too readily. BVH could therefore lead to an inflated vQTL false positive rate.

If BVH is disruptive to QTL mapping generally, it makes sense to incorporate it into the QTL mapping model when its source is known, and to use robustifying techniques to protect against it when its source is unknown. Accommodating BVH of known source is most naturally achieved through modeling covariate effects on the variance, something that is straightforward with the DGLM of Rönnegård and Valdar (2011) but not currently with other proposed methods. Protecting BVH when its source is unknown is less obvious, but since the threat manifests through sensitivity to distributional assumptions, natural contenders include side-stepping such assumptions via non-parametric approaches, *e.g.*, permutation testing, or reshaping the distribution prior to analysis through variable transformation. Both have been considered in the vQTL context, with permutation used in Hulse and Cai (2013) and Yang *et al.* (2012) and transformation in Rönnegård and Valdar (2011), Yang *et al.* (2012), Sun *et al.* (2013), and Shen and Carlborg (2013), but not specifically for controlling vQTL false positives in the presence of BVH.

Here we examine the effect of modeled and unmodeled BVH on power and false positive rate when mapping QTL affecting the mean, the variance or both. In doing so we:

1. Develop a robust, straightforward procedure and software based on the DGLM that can be used for routine mQTL and vQTL analysis;

2. Compare alternative proposed methods for mQTL and vQTL analysis;

3. Show how incorporating BVH can improve power for detecting mQTL and vQTL;

4. Show how sensitivity to model assumptions can be rescued by variable transformation and/or permutation.

5. Illustrate the effect of modeling BVH in existing dataset, an F2 cross from Leamy et al, and discover a new QTL for bodyweight.

In two companion papers, we describe an R package `vqtl` implementing our procedure (Corty and Valdar 2018+ [r/vqtl]), and in Corty *et al* 2018+ apply it to two published QTL mapping experiments detecting a novel mQTL in one and a novel vQTL in the other.

STATISTICAL METHODS

This section reviews four approaches for modeling the effect of a single QTL on the phenotypic mean and/or variance: the standard linear model, Levene's test, Cao's tests, and our preferred procedure based on the DGLM. For each approach we describe a set of alternative procedures for evaluating significance (*i.e.*, calculating

p-values) that provide varying degrees of protection against the impact of BVH and distributional assumptions more generally. The following section, Data and Simulations, then describes a simulation study that assesses the approaches and p-value procedures, and a dataset to which they are applied genomewide.

Definitions

We start by defining three partially overlapping classes of QTL:

mQTL: a locus containing a genetic factor that causes heterogenity of phenotype mean,

vQTL: a locus containing a genetic factor that causes heterogenity of phenotype variance, and

mvQTL: a locus containing a genetic factor that causes heterogenity of either phenotype mean, variance, or both — a generalization that includes the other two classes.

In addition, since we restrict our attention to QTL mapping methods that test genetic association with a phenotype one locus at a time, we distinguish two sources of variance effects:

Foreground Variance Heterogeneity (FVH): effects on the variance that arise from the locus under consideration (the focal locus);

Background Variance Heterogeneity (BVH): effects on the variance that arise from outside of the focal locus, *e.g.*, from another locus or an experimental covariate.

Procedures to evaluate the significance of a single test

In comparing different statistical approaches and their sensitivity to BVH, namely the effect of BVH on power and false positive rate (FPR), it is important to acknowledge that various measures could be taken to make significance testing procedures more robust to model misspecification in general and to BVH specifically. The significance testing methods considered here are frequentist, involving the calculation of a test statistic T on the observed data followed by an estimation of statistical significance based on a conception of T's distribution under the null. However, BVH constitutes a departure of distributional assumptions, and in any rigorous applied statistical analysis when departures are expected it would be typical to consider protective measures such as, for example, transforming the response to make asymptotic assumptions more reasonable, or the use of computationally intensive procedures, such as those based on bootstrapping or permutation, to evaluate significance empirically.

Nominal significance (*i.e.*, the p-value for a single hypothesis test) is evaluated using four distinct procedures. The first two rely on asymptotics:

1. Standard: The test statistic T is computed on the observed data and compared with its asymptotic distribution under the null.

2. Rank-based inverse normal transform (RINT): As for standard, except observed phenotypes $\{y_i\}_{i=1}^n$ are first transformed to strict normality using the function $\mathrm{RINT}(y_i) = \Phi^{-1}[(\mathrm{rank}(y_i) - 3/8)/(n + 1/4)]$, where Φ is the normal c.d.f. and $\mathrm{rank}(y_i)$ is gives the rank (from $1, \ldots, n$) (Beasley *et al.* 2009).

The second two determine significance empirically based on randomization: the test statistic T is recomputed as $T^{(r)}$ under randomizations of the data $r = 1, \ldots, R$, and the resulting set of statistics $\{T^{(r)}\}_{r=1}^R$ is used as the empirical distribution of T under the randomized null. Two alternative randomizations are considered:

3. Residperm: we generate a pseudo-null response $\{y_i^{(r)}\}_{i=1}^n$ based on permuting the residuals of the fitted null model, (Freedman and Lane 1983; Good 2013), a process recently applied in the field of QTL mapping by Cao *et al.* (2014).

4. Locusperm: we leave the response intact, instead permuting the rows of the design matrix (or matrices) that differentiate(s) the null from alternative model.

Procedure to evaluate genomewide significance

In the context of a genome scan, where many hypotheses are tested, we aim to control the genomewide FPR, namely the family-wise error rate (FWER), the probability of making at least one false positive finding across the whole genome. This is done following the general approach of Churchill and Doerge (1994), which is closely related to the locusperm procedure described above, and which we refer to as genomeperm. Briefly, we perform an initial genome scan, recording test statistics $\{T_l\}_{l=1}^L$ for all L loci. Then for each randomization $r = 1, \ldots, R$, and for only the parts of the model that distinguish the null from the alternative model, the genomes are permuted among the individuals; the scan is then repeated to yield simulated null test statistics $\{T_l^{(r)}\}_{l=1}^L$ of which the maximum, $T_{\max}^{(r)}$, is recorded. The collection of $\{T_{\max}^{(r)}\}_{r=1}^R$ from all R such permutations is then used to fit a generalized extreme value distribution (GEV) (Dudbridge and Koeleman 2004), and the quantiles of this are used to estimate FWER-adjusted p-values for each $\{T_l\}_{l=1}^L$.

Standard linear model (SLM) for detecting mQTL

The standard model of quantitative trait mapping uses a linear regression based on the approximation of Haley and Knott (1992) and Martínez and Curnow (1992) to interval mapping of Lander and Botstein (1989). The effect of a given QTL on quantitative phenotype y_i of individual $i = 1, \ldots, n$ is modeled as

$$y_i \sim \mathrm{N}(m_i, \sigma^2) \tag{1}$$

where σ^2 is the residual variance and m_i is a linear predictor for the mean, defined, in what we term the "full model", as

$$\text{Full model:} \quad m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha}\,, \tag{2}$$

where μ is the intercept, $\mathbf{x}_i$ is a vector of covariates with effects $\boldsymbol{\beta}$, and $\mathbf{q}_i$ is a vector encoding the genetic state at the putative mQTL with corresponding mQTL effects $\boldsymbol{\alpha}$. In the case considered here of biallelic loci arising from a cross of two founders, A and B, the genetic state vector $\boldsymbol{q}_i = (a_i, d_i)^{\mathrm{T}}$ is defined as follows: when genotype is known, for genotypes (AA, AB, BB), the additive dosage is $a_i = (0, 1, 2)$ and the dominance predictor is $d_i = (0, 1, 0)$; when genotype is available only as estimated probabilities $p(\mathrm{AA})$, $p(\mathrm{AB})$ and $p(\mathrm{BB})$, following (Haley and Knott 1992; Martínez and Curnow 1992), we use the corresponding expectations, $a_i = 2p(\mathrm{AA}) + p(\mathrm{AB})$ and $d_i = p(\mathrm{AB})$.

The test statistic for an mQTL is based on comparing the fit of the full model, acting as an alternative model, with that of a null that omits the locus effect, namely,

$$\text{Null model:} \quad m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\,. \tag{3}$$

Since the regression in each case provides a maximum likelihood fit, the test statistic used here is likelihood ratio (LR) statistic, $T = 2(\ell_1 - \ell_0)$, where ℓ_1 and ℓ_0 are the log-likelihoods under the alternative and the null respectively. For the biallelic model,

the asymptotic test is the likelihood ratio test (LRT) whereby under the null, $T \sim \chi^2_2$. (Note: Alternative evaluation using the F-test is in general more precise but for our purposes provides equivalent results.)

The residperm approach to empirical significance evaluation of T proceeds as follows. We first fit the null model (Equation 3) to obtain predicted values $\widehat{m}_i = \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ and estimated residuals $\widehat{\varepsilon}_i$ such that $y_i = \widehat{m}_i + \widehat{\varepsilon}_i$. Then, for each randomization $r = 1, \ldots, R$, we generate pseudo-null phenotypes $\{y_i^{(r)}\}_{i=1}^n$ as

$$y_i^{(r)} = \widehat{m}_i + \widehat{\varepsilon}_{\pi_r(i)}\,,$$

where if $\boldsymbol{\pi}_r$ is a vector containing a random permutation of the indices $i = 1, \ldots, n$, then $\pi_r(i)$ is its ith element, mapping index i to its rth permuted version. The null and alternative models are then fitted to $\{y_i^{(r)}\}_{i=1}^n$ to yield $\ell_1^{(r)}$ and $\ell_0^{(r)}$, and hence $T^{(r)}$.

In the locusperm approach to empirical significance, the response is unchanged but permutations are applied to the locus genotypes. For each randomization r, the full model m_i is

$$\text{Permuted full model:} \quad m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_{\pi_r(i)}^{\mathrm{T}}\boldsymbol{\alpha} \tag{4}$$

where $\pi_r(i)$ is as defined for residperm above. This full model fit yields $\ell_1^{(r)}$, and then $T^{(r)} = 2(\ell_1^{(r)} - \ell_0)$. Note that $\ell_0^{(r)}$ need not be recomputed after randomization because because only the rows of the design matrices that are unique to the alternative model are permuted and thus $\ell_0^{(r)} = \ell_0$. Genomeperm applies locusperm genomewide: specifically, in each randomization $r = 1, \ldots, R$, the same permutation, $\boldsymbol{\pi}_r$, is applied to all L loci.

Levene's Test (LV) for detecting vQTL

Levene's test is a procedure for differences in variance between groups that can be used to detect vQTL. Suppose individuals are in G mutually exclusive groups $g = 1, \ldots, G$. Let $g[i]$ denote the group to which individual i belongs, denote gth group size as $n_g = \sum_{i=1}^n I_{\{g[i]=g\}}$, and gth group mean as $\bar{y}_g = n_g^{-1}\sum_{i=1}^n y_i I_{\{g[i]=g\}}$. Then denote the ith absolute deviation as $z_i = |y_i - \bar{y}_{g[i]}|$, the group mean of these as $\bar{z}_g = n_g^{-1}\sum_{i=1}^n z_i I_{\{g[i]=g\}}$ and overall mean $\bar{z} = n^{-1}\sum_{i=1}^n z_i$. Levene's W statistic is then

$$W = \frac{\sum_{g=1}^G n_g\left(\bar{z}_g - \bar{z}\right)^2}{(G-1)}\left[\frac{\sum_{i=1}^n (z_i - \bar{z}_{g[i]})^2}{(n-G)}\right]^{-1}, \tag{5}$$

which under the null model of no variance effect follows the F distribution as $W \sim F(N-G, G-1)$ (Levene 1960). Note that replacing means of y with medians gives the related Brown-Forsythe test (Brown and Forsythe 1973), and replacing all instances of z with y in Equation 5 gives the ANOVA F statistic.

Levene's test does not lend itself naturally to the residperm approach because it does not explicitly involve a null model to split the data into hat values and residuals. We therefore use the null model from the SLM (Equation 3) to approximate the residperm procedure with Levene's test. To execute the locusperm procedure, for each randomization r, the group labels are permuted among the individuals, which is equivalent to replacing all instances of $g[i]$ above with $g[\pi_r(i)]$, with $\pi_r(i)$ defined as above. A corresponding genomewide procedure, although not performed here, would ensure that each randomization r applies the same permutation $\boldsymbol{\pi}_r$ across all loci.

Cao's Tests

Cao *et al.* (2014) elaborates the SLM to have a variance parameter that differs by genotype, *i.e.*,

$$y_i \sim \mathrm{N}(m_i, \sigma_i^2), \tag{6}$$

where m_i is the linear predictor, σ_i^2 is the variance of the ith individual. These are defined in what we term the "full model" as

$$\text{Full model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha} \\ \sigma_i^2 &= \phi_{g[i]} \end{cases}, \tag{7}$$

where $g[i]$ indexes the genotype group to which i belongs, and $\{\phi_g\}_{g=1}^G$ are the variances of the $g = 1, \ldots, G$ genotype groups. Thus an individual's variance is entirely dictated by its genotype, and that genotype must be categorically known (or otherwise assigned). Cao *et al.* (2014) fits this model using a two-step, profile likelihood method, which in our applications we observe to be indistinguishable from full maximum likelihood (Figure S7).

Cao *et al.* (2014) describes tests for mQTL, vQTL and mvQTL based on comparing a full model against three different null models; we detail these tests below in our notation, denoting them respectively $\mathrm{Cao_M}$, $\mathrm{Cao_V}$, and $\mathrm{Cao_{MV}}$.

$\mathrm{Cao_M}$ test for detection of mQTL The $\mathrm{Cao_M}$ test involves an LRT between Cao's full model and Cao's no-mQTL model:

$$\text{Cao's no-mQTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \\ \sigma_i^2 &= \phi_{g[i]} \end{cases}, \tag{8}$$

To execute the residperm procedure for $\mathrm{Cao_M}$, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from Cao's no-mQTL model (Equation 8). The locusperm procedure respecifies the full model (Equation 7), leaving the variance model unchanged and specifying the mean predictor as $m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_{\pi_r(i)}^{\mathrm{T}}\boldsymbol{\alpha}$. The genomeperm procedure similarly applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the mean specification across all loci.

$\mathrm{Cao_V}$ for detection of vQTL The $\mathrm{Cao_V}$ test involves an LRT between Cao's full model and Cao's no-vQTL model:

$$\text{Cao's no-vQTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha} \\ \sigma_i^2 &= \sigma^2 \end{cases}, \tag{9}$$

where the unsubscripted σ^2 is a single, overall residual variance. This null model is identical to the alternative model in the SLM (Equation 2).

To execute the residperm procedure for $\mathrm{Cao_V}$, pseudo-null phenotypes are generating using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from Cao's no-mQTL model (Equation 9). The locusperm procedure respecifies the full model (Equation 7), leaving the mean sub-model unchanged and specifying the variance predictor as $\sigma_i^2 = \phi_{g[\pi(i)]}$. The genomeperm procedure applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the variance specification across all loci.

$\mathrm{Cao_{MV}}$ for detection of generalized mvQTL The $\mathrm{Cao_{MV}}$ test involves an LRT between Cao's full model and Cao's no-QTL model:

$$\text{Cao's no-QTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \\ \sigma_i^2 &= \sigma^2 \end{cases}. \tag{10}$$

This null model is identical to the null model in the SLM (Equation 3).

To execute the residperm procedure for Cao_{MV}, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from Cao's no-QTL model (Equation 10). The locusperm procedure specifies the mean predictor as $m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_{\pi(i)}$ and the variance predictor as $\sigma^2_{g[i]} = \phi_{[\pi(i)]}$. The genomeperm procedure applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the mean and variance specifications across all loci.

Double Generalized Linear Model (DGLM)

The DGLM models the phenotype y_i via two linear predictors as

$$y_i \sim \mathrm{N}(m_i, \sigma_i^2), \quad \text{where} \quad \sigma_i^2 = \sigma^2 \times \exp(v_i)$$

where m_i predicts the phenotype mean and v_i predicts the extent to which the baseline residual variance σ^2 is increased in individual i. In what we term the "DGLM full model", these are specified as

$$\text{Full model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha} \\ v_i &= \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\theta} \end{cases}, \tag{11}$$

where μ is the intercept, $\mathbf{z}_i$ is a vector of covariates (which may be identical to $\mathbf{x}_i$), $\boldsymbol{\gamma}$ is a vector of covariate effects on v_i, and $\boldsymbol{\theta}$ is a vector of locus effects on v_i.

As with Cao's full model, the DGLM full model can be compared, in a likelihood ratio test, with various null models to test for mQTL, vQTL (Rönnegård and Valdar 2011), or mvQTL. A full maximum likelihood fitting procedure for the DGLM was provided by Smyth (1989).

DGLM_{M} for detecting mQTL: For detecting mQTL, we use an LRT of the DGLM full model in Equation 11 against the no-mQTL model:

$$\text{No-mQTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \\ v_i &= \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\theta} \end{cases}, \tag{12}$$

where the LR statistic has asymptotic distribution $T \sim \chi^2_2$.

To execute the residperm procedure for DGLM_{M}, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from the Equation 12. The locusperm procedure respecifies the mean predictor as $m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_{\pi(i)}^{\mathrm{T}}\boldsymbol{\alpha}$ and does not modify the variance predictor. The genomeperm procedure similarly applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the mean specification across all loci.

DGLM_{V} for detecting vQTL: For detecting vQTL, we use an LRT of the DGLM full model in Equation 11 against the no-vQTL model:

$$\text{No-vQTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha} \\ v_i &= \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} \end{cases}, \tag{13}$$

where the LR statistic has asymptotic distribution $T \sim \chi^2_2$.

To execute the residperm procedure for DGLM_{V}, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from the Equation 13. The locusperm procedure does not modify the variance predictor and respecifies the mean predictor as $v_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \mathbf{q}_{\pi(i)}^{\mathrm{T}}\boldsymbol{\theta}$. The genomeperm procedure similarly applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the variance specification across all loci.

Category	Test	Description
mQTL	SLM	Conventional test of mean differences; allows neither FVH nor BVH
mQTL	Cao_{M}	Allows FVH, but not BVH
mQTL	DGLM_{M}	Allows FVH and BVH
vQTL	Levene's test	Conventional test of variance differences; detects FVH, does not allow BVH
vQTL	Cao_{V}	Detects FVH, does not allow BVH
vQTL	DGLM_{V}	Detects FVH, allows BVH
mvQTL	Cao_{MV}	Detects FVH, does not allow BVH
mvQTL	DGLM_{MV}	Detects FVH and allows BVH

Table 1 The eight tests that were evaluated in the simulation studies. FVH: foreground variance heterogeneity. BVH: background variance heterogeneity.

DGLM_{MV} for detecting mvQTL: For detecting mvQTL, we use an LRT of the DGLM full model in Equation 11 against the no-QTL model:

$$\text{No-QTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \\ v_i &= \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} \end{cases}, \tag{14}$$

where the LR statistic has asymptotic distribution $T \sim \chi^2_4$.

To execute the residperm procedure for DGLM_{MV}, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from the Equation 14. The locusperm procedure respecifies the mean predictor as $m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_{\pi(i)}^{\mathrm{T}}\boldsymbol{\alpha}$ and the variance predictor as $v_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \mathbf{q}_{\pi(i)}^{\mathrm{T}}\boldsymbol{\theta}$. The genomeperm procedure similarly applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the mean and variance specifications across all loci.

DATA AND SIMULATIONS

Simulation was used to assess the ability of the eight tests described above to distinguish each of the three types of QTL — pure mQTL, pure vQTL, and mixed mvQTL — from a null locus in the presence and absence of background variance heterogeneity (BVH). Tests are distinguished by their ability to accommodate and target foreground variance heterogeneity (FVH) and background variance heterogeneity (Table 1).

In each simulation, $n = 300$ observations were simulated as

$$y_i \sim \mathrm{N}(q_i^{\mathrm{T}}\alpha, \exp(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + q_i^{\mathrm{T}}\theta))$$

where y_i is the phenotype of individual i, q_i is the genotype, and $\mathbf{z}_i$ is the indicator vector for the factor that drives BVH in scenarios where it is present. In each simulation, each row of $\mathbf{q}$ (each q_i) is drawn randomly from [-1, 0, 1] with probability (0.25, 0.5, 0.25) mimicking an F2 intercross. Across all simulations, $\boldsymbol{Z}$ is fixed to be an indicator matrix mapping the first 60 observations to group 1, the next 60 to group 2, ... and the last 60 to group 5.

Values of α, θ, and $\boldsymbol{\gamma}$ differentiate the eight simulation scenarios as described below.

Scenarios

Simulations varied in two dimensions: locus effect (four options) and BVH (two options) for a total of eight simulation scenarios.

Each of the eight scenarios was examined in $S = 10,000$ simulation trials. Locus effect sizes were chosen such that all QTL were detectable with approximately 70% power at a 5% false positive rate for traditional tests in the absence of BVH. Comprehensive details on the simulation setup are described in **Supplementary Materials**.

The four options for a simulated locus effect were as follows:

1. null locus: The locus has no effect on phenotype.

2. pure mQTL: The locus has an additive effect on the phenotype mean that explains 5% of the total variance.

3. pure vQTL: The locus has an additive effect on the log standard deviation that is detectable with approximately 70% power at a 5% false positive rate for traditional tests in the absence of BVH.

4. mixed mvQTL: The locus has both an additive mean effect that explains 3.25% of phenotype variance and an additive variance effect that is approximately equally detectable.

The two options for simulated BVH were as follows:

1. absent: Nothing, except for possibly the locus (if a vQTL or mvQTL), influences the residual variance. ($\gamma = [0, 0, 0, 0, 0]$)

2. present: $\gamma = [-0.4, -0.2, 0, 0.2, 0.4]$, resulting in group-wise standard deviations in the null locus and mQTL scenarios of approximately $[0.67, 0.82, 1, 1.22, 1.49]$. In the vQTL and mvQTL scenarios, these BVH effects combine additively with the locus effects on the log standard deviation scale, yielding 15 distinct standard deviations. These effect sizes generate a spectrum of standard deviations across groups that are consistent with those observed in the real data reanalysis that follows.

Tests and Significance

In each scenario, eleven tests were applied, and four procedures were used to assess the statistical significance of each test, for a total of 32 test-procedures.

The eleven tests comprise four tests for detecting mQTL: LM, Cao_M, and DGLM_M with and without modeling the variance covariate; four for detecting vQTL: Levene's test, Cao_V, DGLM_V with and without modeling the variance covariate; and three for detecting mvQTL: Cao_MV and DGLM_MV with and without modeling the variance covariate, as described in **Methods**.

The eleven tests, however, contain some redundancy, and so in the main text we report results from only eight. Specifically, for a given type of QTL, the DGLM model that omits variance covariates is equivalent to the corresponding Cao's test, and, barring computational errors in fitting, should give equivalent results (as was observed); results from these DGLM models are therefore omitted from the main text, but for completeness are reported in the supplement.

The four procedures for evaluating the statistical significance were: standard, RINT, residperm, and locusperm, as described in the **Methods**.

Evaluation of tests and procedures

Tests and procedures for assessing statistical significance were evaluated based on their receiver operating characteristics (ROC) and their ability to accurately control the FPR to the nominal level. ROC curves display the ability of a test to discriminate between two conditions by plotting FPR against power for all possible cutoffs.

In this case, the ROC curve reflects the ability of a test to discriminate between QTL and null loci. Specifically, for a given method and cutoff c, the FPR was the fraction null simulations in which the nominal p-value p was less than c; the power is the fraction of times this happened in non-null (*i.e.*, QTL) simulations. A test was said to accurately control FPR when, for all c, FPR $= c$; it was said to "dominate" another test when it had higher power across all FPRs.

The ROC curve cannot immediately distinguish between tests that accurately control FPR and those that do not. We added a symbol to each ROC curve at the point where $c = 0.05$. In cases where the point falls on the vertical line at FPR = 0.05, it reflects accurate FPR control. In cases where the point falls to the left or right of the vertical line it reflects a conservative or anticonservative test, respectively. QQ plots, provided in the supplementary material provide a more holistic view on the FPR control of each test and procedure.

Leamy *et al.* Summary of Original Study

Leamy *et al.* (2000) backcrossed mice from strain CAST/Ei, a small, lean strain, into mouse strain M16i, a large, obese strain. Nine F1 males were bred with 54 M16i females to produce a total of 421 offspring (208 female, 213 male), which were genotyped at 92 microsatellite markers across the 19 autosomes and phenotyped for body composition and morphometric traits. We retreived all available data on this cross, which included marker genotypes, covariates, and eight phenotypes (body weight at five ages, liver weight, subcutaneous fat pad thickness, and gonadal fat pad thickness), from the Mouse Phenome Database (Grubb *et al.* 2014), and estimated genotype probabilities at 2cM intervals across the genome using the hidden Markov model in `R/qtl` (Broman *et al.* 2003).

This mapping population has been studied for association with several phenotypes: asymmetry of mandible geometry (Leamy *et al.* 2000), limb bone length (Leamy *et al.* 2002; Wolf *et al.* 2006), organ weight (Leamy *et al.* 2002; Wolf *et al.* 2006; Yi *et al.* 2006), fat pad thickness (Yi *et al.* 2005, 2006, 2007), and body weight (Yi *et al.* 2006). The most relevant prior study to this reanalysis, Yi *et al.* (2006), used standard methods to identify QTL for body weight at three weeks on chromosomes 1 and 18. However, we were not able to reproduce this result, despite following their analysis as described.

Availability of Data and Software

Analyses were conducted in the `R` statistical programming language (R Core Team 2017). The simulation studies used the implementation of the standard linear model from package `stats`, Levene's test from `car`, Cao's tests as published in Cao *et al.* (2014) and the DGLM tests in package `dglm`. Files S1, S2, and S3 contain the R scripts necessary to replicate the simulation studies and their analysis, relying on the `plotROC` package to make ROC plots (Sachs and Others 2017). File S4 contains the data from Leamy *et al.* (2000) that was reanalyzed. File S5 contains the attempted replication of the original analysis (Yi *et al.* 2006) and file S6 contains the new analysis, using package `vqtl`.

The reanalyzed dataset is available on the Mouse Phenome Database (Grubb *et al.* 2014) with persistent identifier MPD:206. The entire project, including data and all analysis scripts, is available as a Zenodo repository at 10.5281/zenodo.1181887.

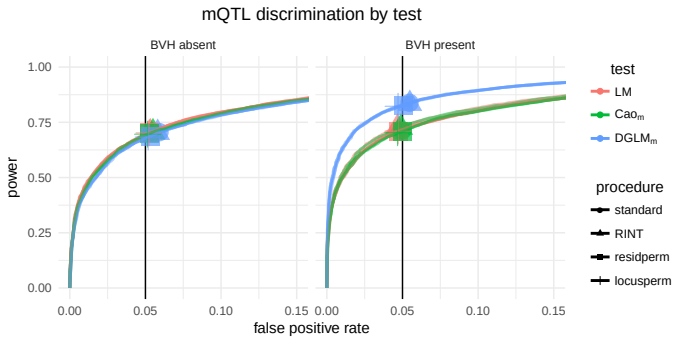


Figure 1 ROC curves for detection of mQTL in presence and absence of BVH. Lines are drawn for three different mQTL tests in Table 1 and four significance procedures, with a point (circle, square, triangle, tick) corresponding to nominal significance at $p = 0.05$ (more details in **Data and Simulations**). DGLM_M dominates Cao_M and SLM in the presence of BVH, accurately controlling FPR with the locusperm and residperm procedures, but not the standard and RINT procedures, which have FPR of 0.058 and 0.060 (Table 2).

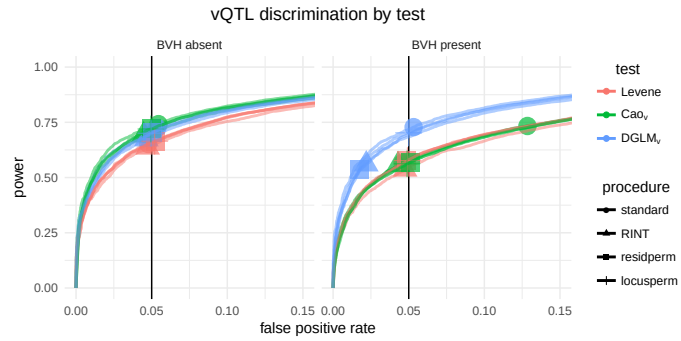


Figure 2 ROC curves for detection of vQTL in presence and absence of BVH. Lines are drawn for three different vQTL tests in Table 1 and four significance procedures, with a point (circle, square, triangle, tick) corresponding to nominal significance at $p = 0.05$ (more details in **Data and Simulations**). DGLM_V dominates Cao_V and Levene's test in the presence of BVH and accurately controls FPR under the standard and locusperm procedures (Table 2). Cao_V suffers a drastic increase in false positives in the presence of BVH under the standard procedure, and DGLM_V would do the same if there were some unmodeled BVH driver, thus DGLM_V under the locusperm procedure is the preferable test.

RESULTS

Simulation study on single locus testing

Simulations were performed to examine the ability of the eight tests listed in Table 1 to detect nonzero effects belonging to their target QTL types (mQTL, vQTL, mvQTL), and to control the number of false positives when no such QTL effects were present. This was done both in the presence and absence of background variance heterogeneity, and for each test, with p-values calculated by each of the four alternative p-value generation procedures (standard, RINT, residperm, locusperm). The full combination of settings is listed in Table 2, which also lists results pertaining to a nominal FPR of 0.05, and described in more detail in **Data and Simulations**.

All three mQTL tests have equivalent performance in the absence of BVH. All three mQTL tests — the standard linear model, the Cao_M test and DGLM_M— accurately controlled FPR under all four significance testing procedures (Figure 1, left panel, Figure S4, left column, and Table 2, column 1, top third). And all twelve test-procedure combinations had indistinguishable power to detect mQTL, in the range [0.692, 0.706] (Table 2, column 2 and Figure S1).

These simulation results do not favor any one test over another, but they do favor the standard and RINT assessment procedures over the residperm and locusperm in the sense that the latter two yeild no additional improvement in FPR control or power for their additional computational cost.

DGLM_M dominates other mQTL tests in the presence of BVH. The SLM and Cao_M accurately controlled FPR under all four procedures to assess statistical significance (Figure 1, right panel, Figure S4, right column, and Table 2, column 5, top third). In contrast, DGLM_M exhibited modest FPR inflation under the standard and RINT procedures, controlling FPR to the nominal level only under the empirical procedures. Nonetheless, despite requiring an empirical procedure to control FPR, in its power to detect an mQTL under BVH, DGLM_M dominated the other two tests, with power in the range of [0.813, 0.816] compared with the power of SLM and Cao_M in the range of [0.689, 0.718] (Figure 1, right panel, Table 2).

Based on the results of these simulations, DGLM_M is the preferable mQTL test in the presence of BVH. But, to accurately control FPR, DGLM_M requires an empirical procedure be used to assess statistical significance; both the residperm and locusperm procedures are capable.

Parametric tests dominate Levene's test for vQTL in the absence of BVH. In null simulations, Cao_V and DGLM_V exhibited slightly anti-conservative behavior using the standard (*i.e.*, asymptotic) significance testing procedure (FPR = 0.053), modestly conservative behavior under the RINT procedure (FPR = 0.043) and slightly conservative behavior under the residperm and locusperm procedures (FPR in the range [0.046, 0.048], Figure 2, left panel, Figure S5, left column, and Table 2, column 2, middle third). Levene's test, in contrast, was overly conservative using the standard and RINT procedures, but accurately controlled FPR under the empirical procedures (Figure S2).

Despite the variation in FPR control among the test-procedure combinations, Cao_V and DGLM_V had more power than Levene's test under all procedures (0.724 vs. 0.667). Thus, the empirical procedures of Cao_V and DGLM_V are the preferred vQTL tests in the absence of BVH, because they have the highest power of the test-procedure combinations that are not anti-conservative. The additional power of Cao_V and DGLM_V relative to Levene's test is consistent with the fact that they make strong parametric assumptions that are exactly true in these simulations and Levene's test does not.

DGLM_V dominates other vQTL tests in the presence of BVH. In the presence of BVH, there were three test-procedure combinations with major departures from accurate FPR control. Cao_V under the standard procedure was drastically anti-conservative, and DGLM_V under both the RINT and residperm procedures was drastically conservative (Figure 2, right panel and Figure S2, and Figure S5, right column). DGLM_V dominated Levene's test and Cao_V, so the standard and locusperm procedure, which accurately control its FPR, seem to be equally preferable and preferable over all other

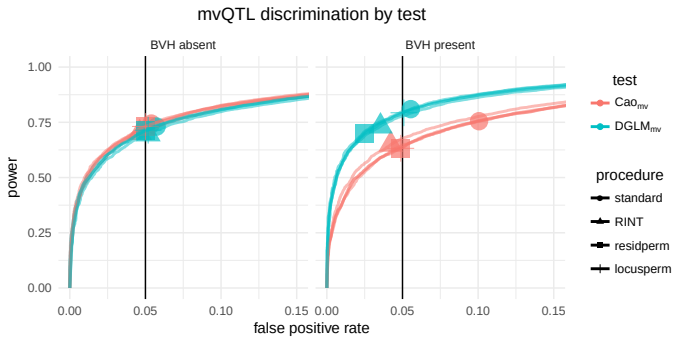


Figure 3 ROC curves for detection of mvQTL in presence and absence of BVH. Lines are drawn for two different mvQTL tests in Table 1 and four significance procedures, with a point (circle, square, triangle, tick) corresponding to nominal significance at $p = 0.05$ (more details in **Data and Simulations**). mvQTL tests combined the responses of mQTL tests and vQTL tests to BVH, yielding a situation in which DGLM_{MV} dominates Cao_{MV}, but only accurately controls FPR under the locusperm procedure (Table 2). The anti-conservative nature of the standard procedure follows from the patterns observed in mQTL tests and the conservative nature of the RINT and residperm procedures follows from the patterns observed in vQTL tests.

test-procedures.

Nonetheless, there is an important caveat that makes locusperm the strongly preferable significance procedure. In this simulation, there are no BVH driving factors unknown to DGLM_{V}. If there were such a factor, DGLM_{V} under the standard procedure would have the same drastic FPR inflation that Cao_{V} showed under the standard procedure in these simulations (Figure S2 (a), third panel). In contrast, the presence of a unknown or unmodeled BVH driving factor does not inflate the FPR of DGLM_{V} under the losucperm procedure. Due to the practical difficulty of excluding the possibility of an unknown BVH driver, the most reliable way to guard against covert FPR inflation without giving up the additional power of DGLM_{V} is to use the locusperm procedure.

mvQTL mirrors vQTL testing; DGLM_{MV} dominates Cao_{MV} in the presence of BVH. As with vQTL tests, there was little to distinguish any test or procedure in the absence of BVH except for the modest conservative nature of the RINT procedure and the concommitant decrease in power (Figure 3, left panel and Table 2, columns 1 and 4, bottom third).

In the presence of BVH, however, DGLM_{MV} dominates Cao_{MV}, with the standard and locusperm procedures accurately controlling FPR (Figure 3, right panel and Figure S6, right column). As with vQTL testing, due to the difficulty in ruling out BVH from an unknown source and the inflated FPR that results from such BVH under the standard procedure, the DGLM_{MV} under the locusperm procedure is the recommended test for mvQTL.

In the presence of BVH, the rank-based inverse normal transformation fails to correct anti-conservative behavior of DGLM_{M} and overcorrects that of DGLM_{V} and DGLM_{MV} A consistent feature of the simulations involving detection of variance effects, whether vQTL or mvQTL, is that FPR control and power is affected, for better or worse, by applying the RINT to the response.

In the presence of BVH, DGLM_{M} under the standard procedure was anti-conservative (FPR = 0.058 at $\alpha = 0.05$). The RINT procedure had no efficacy in returning this test to accurate FPR control

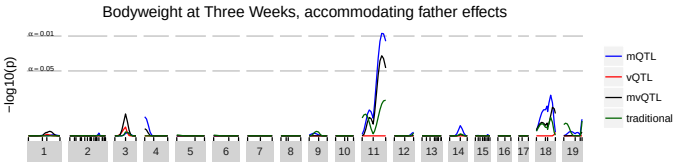


Figure 4 FWER-controlling association statistic at each genomic locus for body weight at three weeks. The linear model (green, "traditional") does not detect any statistically-significant associations. The mQTL test takes into account the heterogeneity of both mean and variance due to which F1 male fathered each mouse in the mapping population and detects one mQTL on chromosome 11.

(FPR = 0.060).

In the case of vQTL detection in the presence of BVH, Cao_{V} under the standard procedure had a drastically inflated FPR (0.123) and the RINT procedure over-corrected it (FPR = 0.044). Similarly, the RINT procedure disrupted DGLM_{V}, which accurately controlled FPR under the standard procedure, causing overly conservative behavior (FPR = 0.021).

As always, in the presence of BVH, the mvQTL tests exhibited a mixture of the patterns obseved in mQTL tests and vQTL tests. Both Cao_{MV} and DGLM_{MV} were anti-conservative under the standard procedure, illustrating their relations to Cao_{V} and DGLM_{M} respectively. And in both cases, the RINT procedure drove an over-correction into the realm of over conservativism (FPR = 0.046 and 0.038 respectively).

In summary, the RINT procedure is unhelpful in the context of the DGLM_{M}: it inflates the FPR of a test that is appropriately sized under standard procedures. But, in the context of vQTL testing with BVH from an uknown source, it has one useful and important property: pre-processing the phenotype with the RINT, leads to vQTL tests that are conservative rather than anti-conservative, decreasing the probability of false positives at the expense of false negatives.

Genomewide reanalysis of bodyweight in Leamy et al. backcross

To understand the impact of BVH on mean and variance QTL mapping in real data, we applied both traditional QTL mapping, using SLM, and mean-variance QTL mapping, using Cao's tests and the DGLM, to body weight at three weeks in the mouse backcross dataset of Leamy *et al.* (2000).

Analysis with Traditional QTL Mapping Identifies no QTL. We first used a traditional, linear modeling-based QTL analysis, with sex and father as additive covariates and genomewide significance based on 1000 genome permutations (Churchill and Doerge 1994). Although sex was found not to be a statistically significant predictor of body weight ($p = 0.093$ by the likelihood ratio test with 1 degree of freedom), it was included in the mapping model because, based on the known importance of sex in determining body weight, any QTL that could only be identified in the absence of modeling sex effects would be highly questionable. Father was found to be a significant predictor of body weight in the baseline fitting of the SLM ($p = 9.6 \times 10^{-5}$ by the likelihood ratio test with 8 degrees of freedom) and therefore was included in the mapping model.

No associations rose above the threshold that controls family-wise error rate to 5% (Figure 4, green line). One region on the

		BVH absent				BVH present			
test	version	null	mQTL	vQTL	mvQTL	null	mQTL	vQTL	mvQTL
LM	standard	0.050	0.706	0.056	0.510	0.051	0.701	0.050	0.508
	RINT	0.050	0.704	0.052	0.495	0.052	0.718	0.049	0.518
	residperm	0.048	0.700	0.054	0.502	0.049	0.694	0.050	0.504
	locusperm	0.049	0.702	0.054	0.499	0.049	0.695	0.049	0.503
Cao_{M}	standard	0.052	0.705	0.051	0.515	0.050	0.700	0.048	0.516
	RINT	0.051	0.705	0.051	0.503	0.052	0.716	0.047	0.523
	residperm	0.049	0.697	0.049	0.503	0.048	0.695	0.047	0.508
	locusperm	0.048	0.691	0.048	0.500	0.048	0.689	0.044	0.503
DGLM_{M}	standard	0.052	0.705	0.051	0.515	0.058	0.832	0.054	0.649
	RINT	0.051	0.705	0.051	0.503	0.060	0.830	0.054	0.644
	residperm	0.049	0.696	0.049	0.504	0.052	0.816	0.046	0.629
	locusperm	0.048	0.692	0.048	0.500	0.050	0.813	0.046	0.624
Levene's test	standard	0.045	0.049	0.660	0.462	0.046	0.046	0.577	0.393
	RINT	0.045	0.043	0.645	0.422	0.047	0.043	0.546	0.344
	residperm	0.049	0.052	0.667	0.474	0.048	0.050	0.585	0.401
	locusperm	0.049	0.052	0.667	0.474	0.048	0.050	0.583	0.400
Cao_{V}	standard	0.053	0.053	0.750	0.543	0.123	0.127	0.744	0.563
	RINT	0.043	0.042	0.700	0.467	0.044	0.048	0.563	0.364
	residperm	0.047	0.051	0.729	0.519	0.045	0.054	0.572	0.388
	locusperm	0.046	0.049	0.726	0.517	0.047	0.051	0.567	0.382
DGLM_{V}	standard	0.053	0.053	0.750	0.543	0.049	0.056	0.732	0.524
	RINT	0.043	0.042	0.700	0.467	0.021	0.027	0.570	0.340
	residperm	0.048	0.051	0.729	0.520	0.015	0.018	0.542	0.329
	locusperm	0.046	0.049	0.724	0.515	0.046	0.050	0.713	0.498
Cao_{MV}	standard	0.050	0.597	0.643	0.741	0.100	0.642	0.649	0.751
	RINT	0.043	0.585	0.574	0.701	0.046	0.600	0.436	0.646
	residperm	0.046	0.587	0.618	0.728	0.050	0.514	0.507	0.632
	locusperm	0.048	0.589	0.617	0.727	0.050	0.516	0.506	0.630
DGLM_{MV}	standard	0.050	0.597	0.643	0.741	0.057	0.741	0.633	0.807
	RINT	0.043	0.585	0.574	0.701	0.038	0.715	0.445	0.726
	residperm	0.046	0.590	0.618	0.729	0.024	0.621	0.469	0.687
	locusperm	0.046	0.590	0.617	0.728	0.051	0.723	0.601	0.788

Table 2 Positive rates of all tests in all scenarios based on 10,000 simulations, 1,000 permutations each to estimate empirical null distributions (residperm and locusperm), and a nominal false positive rate (FPR) of $\alpha = 0.05$. Entries in column 1 and 5 through all rows, columns 3 and 7 in the top third, and columns 2 and 6 in the middle third represent FPR. The entries in the rest of the table represent power. The largest standard error for an FPR is 0.001. The largest standard error for a power is 0.0025.

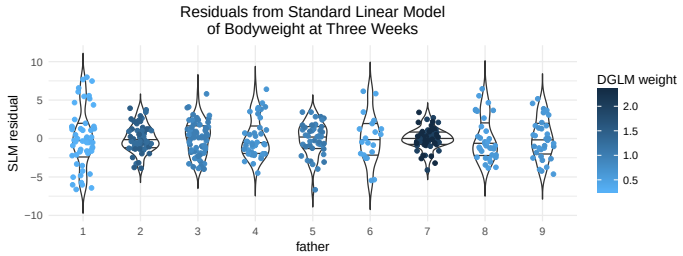


Figure 5 Residuals from the standard linear model for body weight at three weeks, with sex and father as covariates, stratified by father. It is evident that fathers differed in the residual variance of the offspring they produced. For example, the residual variance of offspring from father 1 is greather than that of father 2 and 7. Here, points are colored by their predicted residual variance in the fitted DGLM with sex and father as mean and variance covariates.

distal part of chromosome 11 could be considered "suggestive" with FWER-adjusted $p \approx 0.17$.

To test the sensitivity of the results to the inclusion/exclusion of covariates, the analysis was repeated without sex as a covariate, without father as a covariate, and with no covariates. No QTL were identified in any of these sensitivity analyses.

Analysis with Cao's tests Identifies no QTL The same phenotype was analyzed with Cao's tests, again including sex and father as mean covariates, and using the genome permutation procedures described in **Methods** were used to control FWER. No statistically significant mQTL, vQTL, nor mvQTL were identified (Figure S9).

Analysis with DGLM-based tests Identifies an mQTL The same phenotype was analyzed with the DGLM-based tests. In a baseline fitting of the DGLM, sex was found not to be a statistically significant predictor of mean or residual variance (mean effect $p = 0.18$, variance effect $p = 0.22$, and joint $p = 0.19$ by the LRT with 1, 1, and 2 d.f.). But father was found to be a statistically significant predictor of both mean and variance (mean effect $p = 2.0 \times 10^{-7}$, variance effect $p = 1.8 \times 10^{-11}$, and $p = 4.8 \times 10^{-14}$ by the LRT with 8, 8, and 16 d.f.). Therefore, following the same reasoning as in the mean model described above, both sex and father were included in the mapping model as covariates of both the mean and the variance. As with the other tests, the genome permutation procedures described in **Methods** were used to control FWER.

A genomewide significant mQTL was identified on chromosome 11 (Figure 4, blue line). The peak was at 69.6 cM with FWER-adjusted $p = 0.011$, with the closest marker being D11MIT11 at 75.7 cM with FWER-adjusted $p = 0.016$. Nonparametric bootstrap resampling, using 1,000 resamples (after Visscher *et al.* 1996), established a 90% confidence interval for the QTL from 50 to 75 cM. This region overlaps with the "suggestive" region identified in the traditional analysis.

By the traditional definition of percent variance explained, following from a fitting of the standard linear model, this QTL explains 2.1% of phenotype variance. Though, given the variance heterogeneity inherent in the DGLM that was used to detect this QTL, this quantity is better considered the "average" percent variance explained. The ratio of the QTL variance to the sum of QTL variance, covariate variance, and residual variance ranges from 1% to 6% across the population, based on the heterogeneity of residual variance.

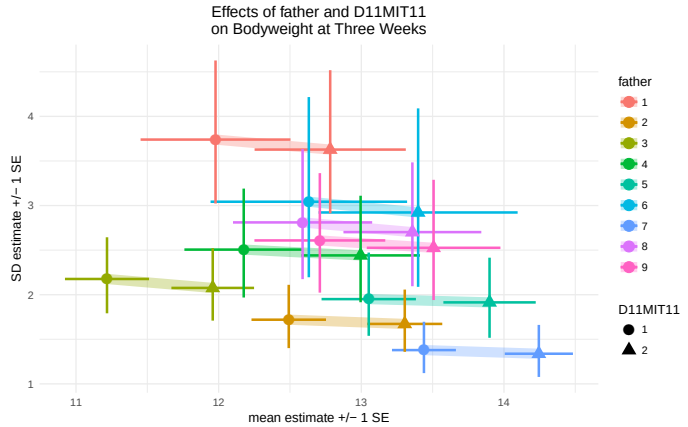


Figure 6 The predictive mean and standard deviation of mice in the mapping population based on father and genotype at the top marker, D11MIT11 on chromosome 11. The genotype effect, illustrated by the colored ribbons is almost entirely horizontal, indicating a difference in means across genotype groups but no difference in variance, consistent with the identification of this QTL as a pure mQTL. The father effects, illustrated by the spread of colored crossbars, have both mean and variance components. For example, father 7 (blue) has the highest predictive mean and lowest predictive standard deviation. His offspring were upweighted in the QTL analysis based on their low standard deviation. Father 1 (red) has an average predictive mean and the highest predictive standard deviation. His offspring were downweighted in the QTL analysis based on their high standard deviation. Note: the effect of sex on phenotype mean and variance was modeled, then marginalized out for readability.

Understanding the Novel QTL The mQTL on chromosome 11 was identified by the DGLM_M test, but not by by the standard linear model or Cao's mQTL test. The additional power of the DGLM_M test over these other tests relates to its accommodation of background variance heterogeneity (BVH).

Specifically, the DGLM reweighted each observation based on its residual variance, according to the sex and F1 father of the mouse. This BVH is visually apparent when the residuals from the standard linear model are plotted, separated out by father (Figure 5).

Some fathers, for example fathers 2 and 7, appear to have offspring with less residual variance than average, whereas others, for example father 1, seem to have offpsring with more residual variance than average. The DGLM captured these patterns of variance heterogeneity, and estimated the effect of each father on the log standard deviation of the observations (Figure 6). Based on these estimated variance effects, observations were upweighted (e.g. fathers 2 and 7) and downweighted (e.g. father 1). This weighting gave the DGLM-based mapping approach more power to reject the null as compared to the SLM.

Other Phenotypes For brevity, we described in detail only the results of the DGLM-based analysis of body weight at three weeks; but, of the eight phenotypes from this cross available on the Mouse Phenome Database, the mean-variance approach to QTL mapping discovered new QTL in four. Five of the eight phenotypes — body weight at twelve days, three weeks, and six weeks, as well sa subcutaneous and gonadal fat pad thickness — exhibited BVH due to father, and for each we performed both traditional QTL mapping using the SLM and mean-variance QTL mapping using the DGLM. For body weight at three weeks and six weeks, we identified one new mQTL and two new vQTL respectively. For subcutaneous fat pad thickness, we "undiscovered" one mean QTL. That is, after reweighting the observations based on the observed variance of each father, two QTL that were detected by the SLM no longer met criteria for statistical significance, as shown in supplementary figures.

DISCUSSION

The simulation studies revealed that in the presence of background variance heterogeneity (BVH), the DGLM-based tests are uniquely powerful in the detection of mQTL, vQTL, and mvQTL.

Our reanalysis of the Leamy et al. dataset demonstrated that the additional power of DGLM_M in the face of BVH can be used to detect an mQTL that was overlooked by all competitor methods.

Detecting and Modeling BVH

To select the right test and procedure to assess significance, it is important to establish whether there is any BVH present. We advocate fitting the DGLM with all potential BVH drivers as variance covariates, then including any that are statistically significant as variance covariates in the mapping model to improve power to detect QTL.

Guidelines for QTL mapping in the presence of BVH

Given that

1. The DGLM-based tests dominate all other tests in the presence of BVH,
2. the locusperm procedure accurately controls the FPR of the DGLM-based tests in the presence of BVH, whether the source is known or not, and
3. the locusperm procedure can be extended into the genomeperm procedure to control FWER,

we advocate for the analysis of experimental crosses that exhibit BVH with the three DGLM-based tests (DGLM_M, DGLM_V, and DGLM_MV) and, where the individuals in the population are exchangeable (as in an F2 or backcross) or where partial exchangeability can be suitably identified [*e.g.*, see (Churchill and Doerge 1994; Zou *et al.* 2006; Churchill and Doerge 2008)], the use of our described genomeperm procedures, which permute the genome in selective parts of the model, to assess genomewide significance.

Because this procedure involves three families of tests rather than one family as would be typical with an SLM-based analysis, an additional correction may be desired to control experiment-wise error rate. DGLM_M and DGLM_V are orthogonal tests (Smyth 1989), but DGLM_MV is neither orthogonal nor identical to either, so the effective number of families is between two and three. One reasonable, heuristic approach to control experiment-wise error rate is simply to lower the acceptable FWER, *e.g.* replacing the standard 0.05 with 0.02.

Data reweighting for mQTL detection

The additional power of mean-variance QTL mapping to detect mQTL in general, and of DGLM_M to detect mQTL in the presence of BVH in particular, can be seen as deriving from how data is reweighted. This reweighting is not based on any prior knowledge on the part of the experimenter, but rather based on patterns of residual variance heterogeneity detected by the DGLM.

The impact of reweighting can be illustrated through consideration of the normal likelihood. For $y_i \sim \text{N}(m_i, \sigma^2/w_i)$, with known weights $w_1, \ldots, w_n$ and known baseline variance σ^2, the log-likelihood can be written as $\ell = \text{const} - \text{WRSS}/2\sigma^2$, where the key quantity to be minimized[1],

$$\text{WRSS} = \sum_{i=1}^{n} w_i (y_i - m_i)^2 \,,$$

is the weighted residual sum of squares, that is, the squared descrepancies between the observed phenotype y_i and its predicted value m_i weighted by w_i. The weights therefore affect how much, relatively speaking, each data point contributes to the likelihood: highly imprecise measurements, such as from individuals whose phenotypes are expected to have high variance, have low weight and diminished contribution, whereas as more precise measurements are correspondingly upweighted. In the DGLM, weights are informed by experimental covariates and the QTL genotype itself, as $w_i = e^{-v_i}$. In the SLM, unless weights are specified externally, there is no such mechanism for phenotype precision to be incorporated and so all weights equal 1. The improvement of the DGLM over the SLM and Cao_M, therefore stems entirely from its greater ability to provide this additional information, and thereby give more credence to phenotype values that are expected to be more precise.

This reweighting can be thought of has having two benefits in the the QTL mapping endeavor. Geneticists are often rightly concerned about high leverage observations, which can cause to false positives. Less often acknowledged is that high leverage observations may also induce false negatives, disrupting an otherwise good statistcal model fit. By bringing the data weights into alignment with their estimated residual variance, the DGLM addresses both of these concerns: by downweighting outliers from systematically noisy subgroups, it reduces the potential for false positives;

[1] Note: $\text{const} = -0.5(n \log 2\pi - \sum_{i=1}^{n} w_i \log \sigma^2)$ can be ignored.

by upweighting outliers from systematically precise subgroups, it reduces the probability of false negatives.

Covariate correction for vQTL detection

Conceptually, the additional power of the $DGLM_V$ to detect vQTL over Cao_V in the presence of BVH, as demonstrated above, derives from its ability to accommodate a covariate, just as any linear regression analysis benefits from accommodating a covariate. The distinction is that, whereas the response for the in a typical regression analysis is the observed data, in the case of BVH and the DGLM, the response is the squared residuals from the mean sub-model.

As with any regression analysis, when the covariate effect is meaningfully large, its inclusion in the model improves the estimation of the effect of interest. The more precise the estimation of the effect of interest allows a greater model improvement from the null to alternative model and ultimately, a more powerful test.

Percent Variance Explained

Variance heterogeneity complicates the notion of percent variance explained (PVE) by a QTL. Assuming the QTL has the same effect on the expected value of the phenotype of all individuals, it will explain a larger percent of total variance for individuals with lower than average residual variance, and vice versa for individuals with higher than average residual variance. In light of this observation, the percent variance explained can either be reported as "average percent variance explained" or can be calculated for some representative sub-groups. For example, if there is variance heterogeneity across sexes, it would be reasonable to report the PVE of a QTL for both males and females, or if a vQTL is known to be present elsewhere in the genome, report the PVE for each vQTL genotype as in Yang *et al.* (2012).

Rank inverse normal transformation: pros and cons for vQTL mapping

In the detection of vQTL, foreground variance heterogeneity (FVH) and BVH come into conflict — the goal is to detect FVH and BVH obscures its detection. Both, however, induce excess kurtosis (fatter tails) in the phenotype distribution. Thus, it is logical that the RINT, which reshapes away excess kurtosis without reference to its source, should have both beneficial and harmful properties.

In the case where there is no known driver of BVH, a scenario represented by the simulations examining Cao_V, the RINT procedure acts like an insurance policy: if there truly is no BVH, the test suffers a modest decrease in power; but if there truly is BVH from an unknown source, it averts the drastic FPR inflation under the standard (*i.e.*, non-empirical) p-value procedure.

In the case where BVH drivers are known, represented by the $DGLM_V$ simulations, the RINT procedure is unnecessary, costing power with its conservatism in the absence of BVH and paradoxically creating even more conservative behavior in the presence of BVH.

The above disadvantages of RINT assume the phenotype data has an underlying normal distribution, either as given or after a simple (*e.g.*, power) transformation. When this is not so, that is, in cases of highly non-normal data, valid inference would be possible by both the RINT and the locusperm procedure, and perhaps the most robust approach would be to use the two in combination. Nonetheless, where normality approximately holds, whether as given or after a simple transformation, we strongly prefer the locusperm procedure without RINT: across all simulation scenarios it exhibited at worst slight conservatism when applied to DGLM-based tests and represents a useful step toward FWER control.

ACKNOWLEDGEMENTS

This work was primarily funded by a National Institutes of General Medical Sciences (NIGMS) grant to WV (R01-GM104125) and a National Institutes of Mental Health grant to RWC (F30-MH108265). RWC received additional support from NIGMS grant T32-GM067553 and National Library of Medicine grant T32-LM012420. The authors thank Larry Leamy for helpful comments.

LITERATURE CITED

Aschard, H., N. Zaitlen, R. M. Tamimi, S. Lindström, and P. Kraft, 2013 A Nonparametric Test to Detect Quantitative Trait Loci Where the Phenotypic Distribution Differs by Genotypes. Genet. Epidemiol. **37**: 323–333.

Beasley, T. M., S. Erickson, R. Public, H. Building, and D. B. Allison, 2009 Rank-Based Inverse Normal Transformations are Increasingly Used, but are they Merited? Behav. Genet. **39**: 580–595.

Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics **19**: 889–890.

Brown, M. B. and A. B. Forsythe, 1973 Robust Tests for the Equality of Variances. J. Am. Stat. Assoc. **69**: 364–367.

Cao, Y., P. Wei, M. Bailey, J. S. K. Kauwe, and T. J. Maxwell, 2014 A versatile omnibus test for detecting mean and variance heterogeneity. Genet. Epidemiol. **38**: 51–59.

Churchill, G. A. and R. W. Doerge, 1994 Empirical Threshold Values for Quantitative Trait Mapping. Genetics **138**: 963–971.

Churchill, G. A. and R. W. Doerge, 2008 Naive application of permutation testing leads to inflated type I error rates. Genetics **178**: 609–610.

Dudbridge, F. and B. P. C. Koeleman, 2004 Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am. J. Hum. Genet. **75**: 424–35.

Dworkin, I., 2005 Canalization, cryptic variation, and developmental buffering: A critical examination and analytical perspective. Variation pp. 131–158.

Fraser, H. B. and E. E. Schadt, 2010 The quantitative genetics of phenotypic robustness. PLoS One **5**.

Freedman, D. and D. Lane, 1983 A nonstochastic interpretation of reported significance levels. J. Bus. Econ. Stat. **1**: 292–298.

Freund, J., A. M. Brandmaier, L. Lewejohann, I. Kirste, M. Kritzler, *et al.*, 2013 Emergence of individuality in genetically identical mice. Science **340**: 756–9.

Good, P., 2013 *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Grubb, S. C., C. J. Bult, and M. A. Bogue, 2014 Mouse Phenome Database. Nucleic Acids Res. **42**.

Haley, C. S. and S. a. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity (Edinb). **69**: 315–24.

Hong, C., Y. Ning, P. Wei, Y. Cao, and Y. Chen, 2016 A semiparametric model for vQTL mapping.

Hulse, A. M. and J. J. Cai, 2013 Genetic variants contribute to gene expression variability in humans. Genetics **193**: 95–108.

Lander, E. S. and S. Botstein, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185.

Leamy, L. J., D. Pomp, E. J. Eisen, and J. M. Cheverud, 2000 Quantitative trait loci for directional but not fluctuating asymmetry of mandible characters in mice. Genet. Res. **76**: 27–40.

Leamy, L. J., D. Pomp, E. J. Eisen, and J. M. Cheverud, 2002 Pleiotropy of quantitative trait loci for organ weights and limb bone lengths in mice. Physiol. Genomics **10**: 21–9.

Levene, H., 1960 Robust tests for equality of variances. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford Univ. Press pp. 278–292.

Martínez, O. and R. N. Curnow, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85**: 480–488.

Mulder, H. a., W. G. Hill, and E. F. Knol, 2015 Heritable environmental variance causes nonlinear relationships between traits: application to birth weight and stillbirth of pigs. Genetics **199**: 1255–69.

Paré, G., N. R. Cook, P. M. Ridker, and D. I. Chasman, 2010 On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the women's genome health study. PLoS Genet. **6**: 1–10.

R Core Team, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rönnegård, L. and W. Valdar, 2011 Detecting major genetic loci controlling phenotypic variability in experimental crosses. Genetics **188**: 435–447.

Rönnegård, L. and W. Valdar, 2012 Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. BMC Genet. **13**: 63.

Sachs, M. C. and Others, 2017 plotROC: A Tool for Plotting ROC Curves. J. Stat. Softw. **79**.

Shen, X. and ö. Carlborg, 2013 Beware of risk for increased false positive rates in genome-wide association studies for phenotypic variability. Front. Genet. **4**.

Shen, X. and L. Ronnegard, 2013 Issues with data transformation in genome-wide association studies for phenotypic variability. F1000Research **2**: 200.

Smyth, G. K., 1989 Generalized linear models with varying dispersion. J. R. Stat. Soc. Ser. B Methodol. **51**: 47–60.

Struchalin, M. V., A. Dehghan, J. C. Witteman, C. van Duijn, and Y. S. Aulchenko, 2010 Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. BMC Genet. **11**: 92.

Sun, X., R. Elston, N. Morris, and X. Zhu, 2013 What is the significance of difference in phenotypic variability across SNP genotypes? Am. J. Hum. Genet. **93**: 390–397.

Visscher, P. M. and D. Posthuma, 2010 Statistical power to detect genetic loci affecting environmental sensitivity. Behav. Genet. **40**: 728–733.

Visscher, P. M., R. Thompson, and C. S. Haley, 1996 Confidence intervals in QTL mapping by bootstrapping.

Wagner, G. P., G. Booth, and H. Bagheri-Chaichian, 1997 A Population Genetic Theory of Canalization. Evolution (N. Y). **51**: 329–347.

Wolf, J. B., D. Pomp, E. J. Eisen, J. M. Cheverud, and L. J. Leamy, 2006 The contribution of epistatic pleiotropy to the genetic architecture of covariation among polygenic traits in mice. Evol. Dev. **8**: 468–476.

Yang, J., R. Loos, M. Goddard, and P. M. Visscher, 2012 FTO genotype is associated with phenotypic variability of body mass index. Nature **490**: 267–272.

Yi, N., D. Shriner, S. Banerjee, T. Mehta, D. Pomp, *et al.*, 2007 An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. Genetics **176**: 1865–1877.

Yi, N., B. S. Yandell, G. A. Churchill, D. B. Allison, E. J. Eisen, *et al.*, 2005 Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics **170**: 1333–1344.

Yi, N., D. K. Zinniel, K. Kim, E. J. Eisen, A. Bartolucci, *et al.*, 2006 Bayesian analyses of multiple epistatic QTL models for body weight and body composition in mice. Genet. Res. **87**: 45–60.

Zou, F., Z. Xu, and T. Vision, 2006 Assessing the significance of quantitative trait loci in replicable mapping populations. Genetics **174**: 1063–1068.

APPENDIX

Calculation of an additive effect to explain a given proportion of total variance in an F2 intercross

The variance attributable to a genetic factor with alleles (AA, AB, BB) at frequency (0.25, 0.5, 0.25), additive effect a and no dominance effect is:

$$\mathrm{V}_a = \frac{1}{4}\left(-a\right)^2 + \frac{1}{2}\left(0\right) + \frac{1}{4}\left(a\right)^2 = \frac{1}{2}a^2\,.$$

For a genetic factor that explains a fraction p of total phenotype variance:

$$\mathrm{V}_a = p\mathrm{V}_y = p(\mathrm{V}_a + \sigma^2) = \frac{p\sigma^2}{(1-p)}\,.$$

Combining and solving for a gives $a = \sqrt{2p\sigma^2/(1-p)}$.

SUPPLEMENTARY MATERIALS

Simulation Details:

In simulation with BVH present, the group-wise effects on the log standard deviation were $\gamma = [-0.4, -0.2, 0, 0.2, 0.4]$. Though $\overline{\gamma} = 0$, the exponential transform connecting these effects to the standard deviation results in a simulated phenotype with slightly more total variance than one without BVH. Therefore, the additive effect of the locus on phenotype mean was adjusted when BVH was introduced, in order to maintain a constant percent variance explained by the mean effect. The following values were used in the simulation.

	no BVH	yes BVH
null	$\alpha = 0, \theta = 0$	$\alpha = 0, \theta = 0$
mQTL	$\alpha = 0.22, \theta = 0$	$\alpha = 0.25, \theta = 0$
vQTL	$\alpha = 0, \theta = 0.17$	$\alpha = 0, \theta = 0.17$
mvQTL	$\alpha = 0.18, \theta = 0.14$	$\alpha = 0.2, \theta = 0.136$

null locus and mQTL in the absence of BVH: All observations have standard deviation 1.

vQTL in the absence of BVH: The genotype-wise standard deviations implied by the additive effect of 0.17 on the log standard deviation are approimately: [0.84, 1.00, 1.19].

mvQTL in the absence of BVH: The genotype-wise standard deviations implied by the additive effect of 0.14 on the log standard deviation are approimately: [0.87, 1.00, 1.15].

null locus and mQTL in the presence of BVH: The covariate-wise standard deviations implied by the effects of [-0.4, -0.2, 0, 0.2, 0.4] on the log standard deviation are approximately: [0.67, 0.82, 1.00, 1.22, 1.49].

vQTL in the presence of BVH: Locus and covariate effects on the residual variance combine additively on the log standard deviation scale, yielding 15 distinct standard deviations:

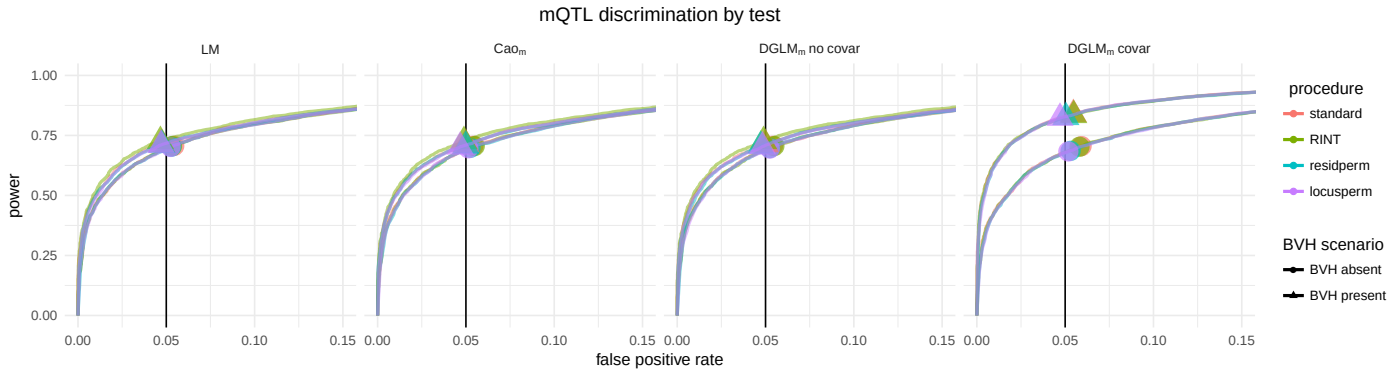
	genotype		
covar	-1	0	1
1	0.57	0.67	0.79
2	0.69	0.82	0.97
3	0.84	1.00	1.19
4	1.03	1.22	1.45
5	1.26	1.49	1.77

mvQTL in the presence of BVH: Locus and covariate effects on the residual variance combine additively on the log standard deviation scale, yielding 15 distinct standard deviations:

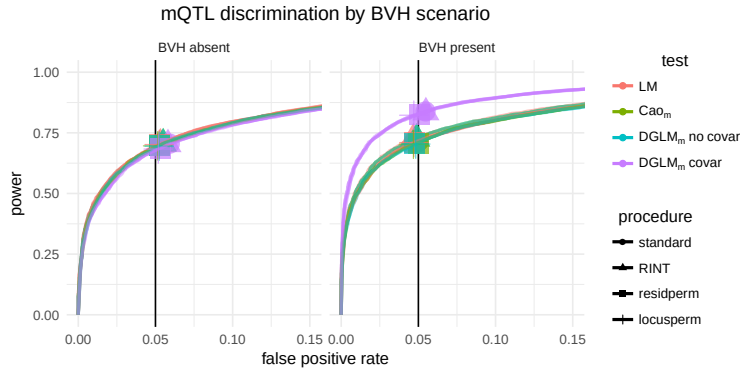
	genotype		
covar	-1	0	1
1	0.59	0.67	0.77
2	0.71	0.82	0.94
3	0.87	1.00	1.15
4	1.07	1.22	1.40
5	1.30	1.49	1.71

ROC Curves

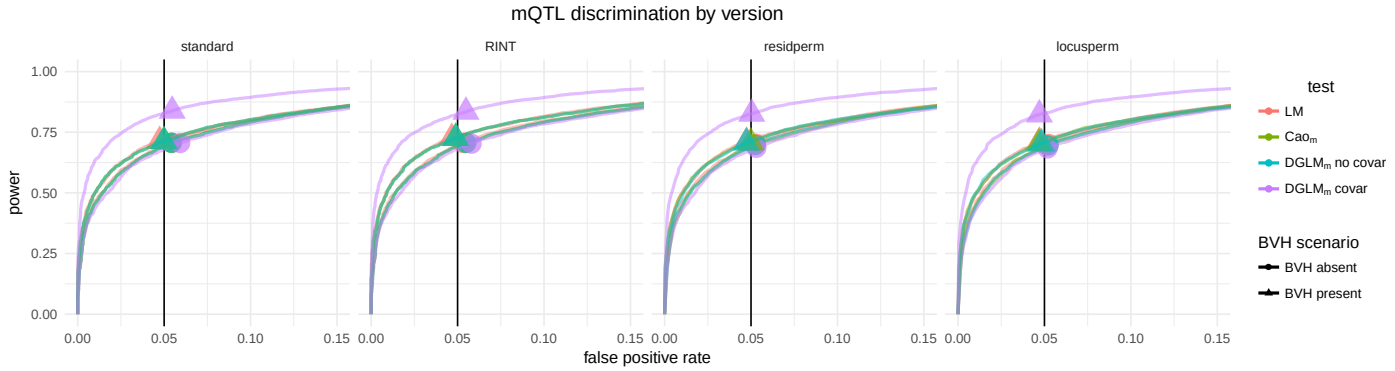
Figure S1 ROC Curves for mQTL tests in the detection of mQTL. The same 28 ROC curves are plotted three times, organized by (a) test, (b) BVH scenario, and (c) version to allow for comparisons across all dimensions.



(a) All test-evaluations accurately control FPR. DGLM$_M$ with BVH of known source is the most powerful test.

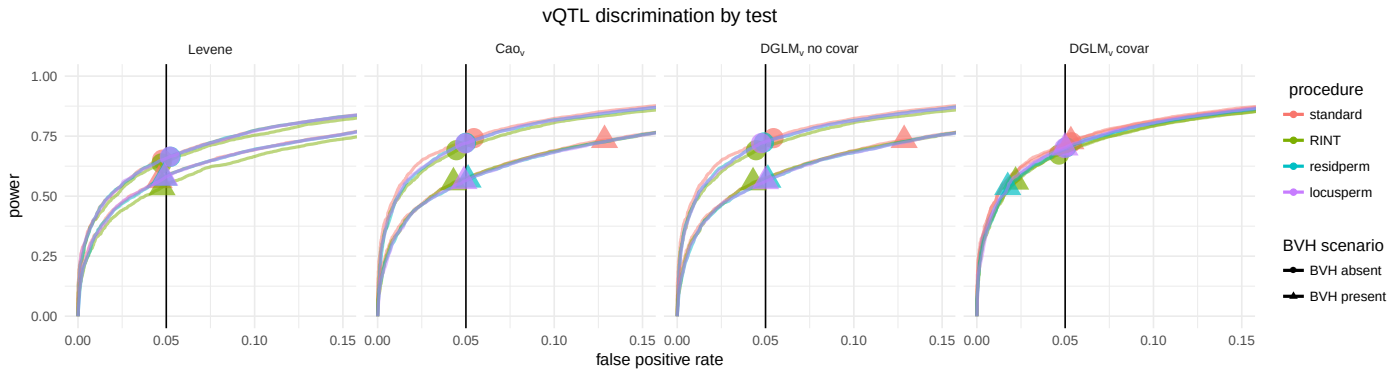


(b) Within BVH scenarios, all mQTL tests perform equivalently.

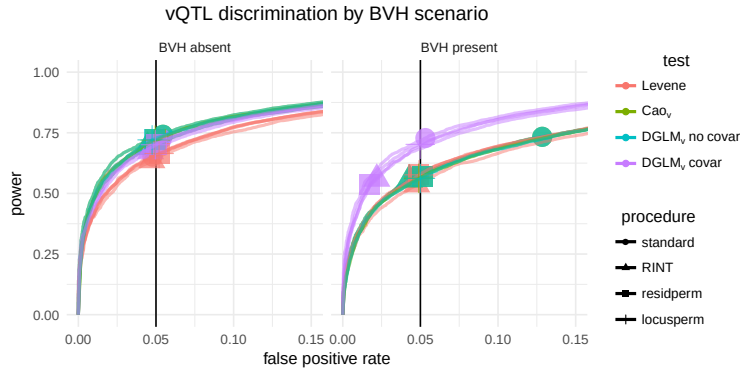


(c) DGLM$_M$ outperforms all other tests across all evaluation methods.

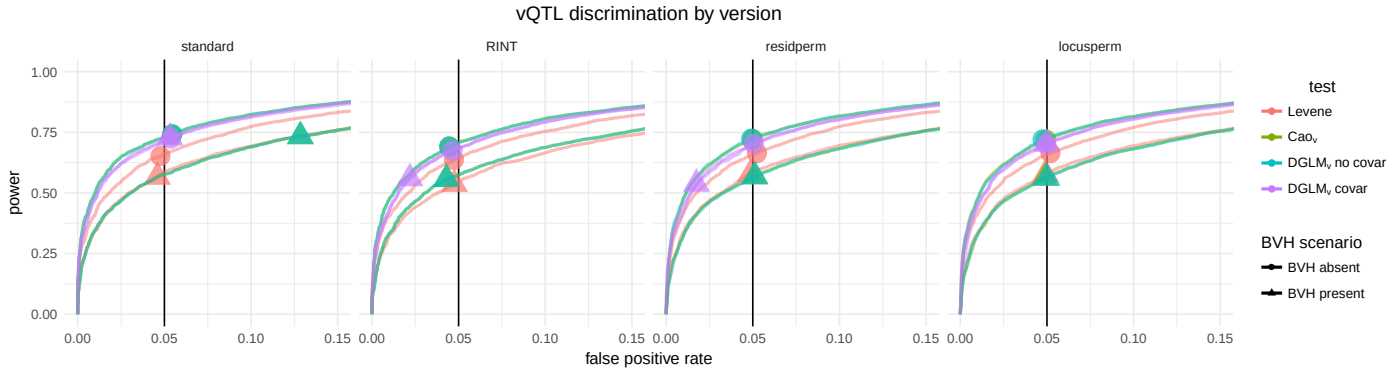
Figure S2 ROC Curves for vQTL tests in the detection of vQTL. The same 28 ROC curves are plotted three times, organized by (a) test, (b) BVH scenario, and (c) version to allow for comparisons across all dimensions.



(a) Levene's test accurately conrols FPR in all scenarios.Cao$_V$ and DGLM$_V$ have inflated FPR in the presence of BVH of uknnown source. DGLM$_V$'s RINT and residperm versions are anti-conservative in the presence of BVH of known source.

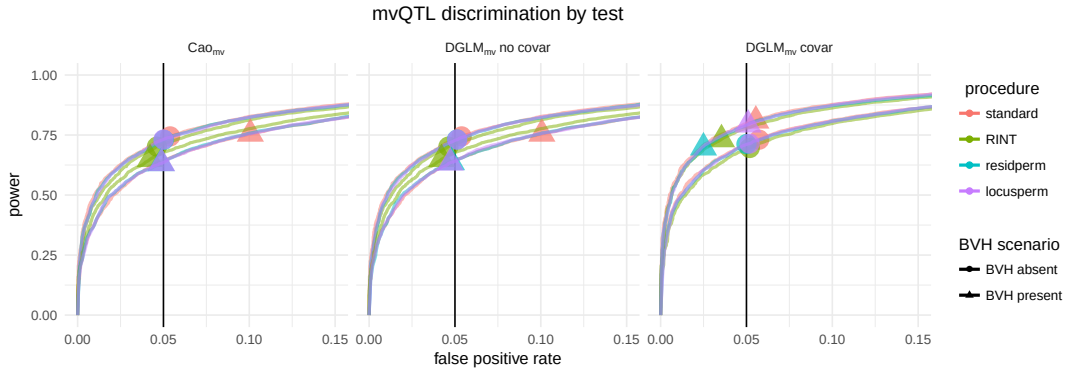


(b) In the absence of BVH, Levene's test is less powerful than Cao$_V$ and DGLM$_V$. In the face of BVH of uknown source, all tests suffer decreased power, except DGLM$_V$'s standard version, which fails to accurately control FPR. In the scenario with BVH of known source, DGLM$_V$ recovers most of the power lost with introduction of BVH, but its RINT and residperm versions are anti-conservative.

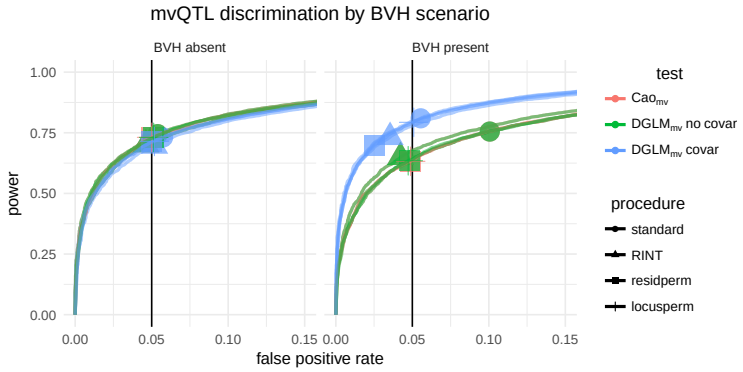


(c) The only version of DGLM$_V$ that accurately controls FPR across all BVH scenarios is locusperm. Its standard version is anticonservative in the presence of BVH of unknown source and its RINT and residperm versions are conservative in the presence of BVH of known source.

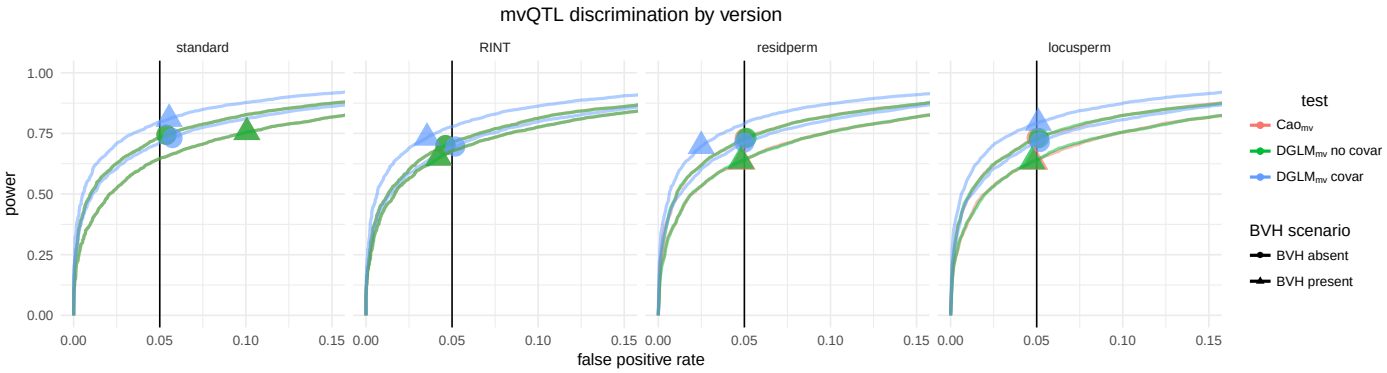
Figure S3 ROC Curves for mvQTL tests in the detection of mvQTL. The same 28 ROC curves are plotted three times, organized by (a) test, (b) BVH scenario, and (c) version to allow for comparisons across all dimensions.



(a) Cao_{MV} and DGLM_{MV} both suffer a decrease in discrimiation (down and right shift of ROC curve) in the presence of BVH of unknown (or unmodeled) source. Only DGLM_{MV} can accommodate the source when it is known and therefore can achieve superior discrimination in that case. The standard and locusperm versions of DGLM_{MV} accurately control FPR.



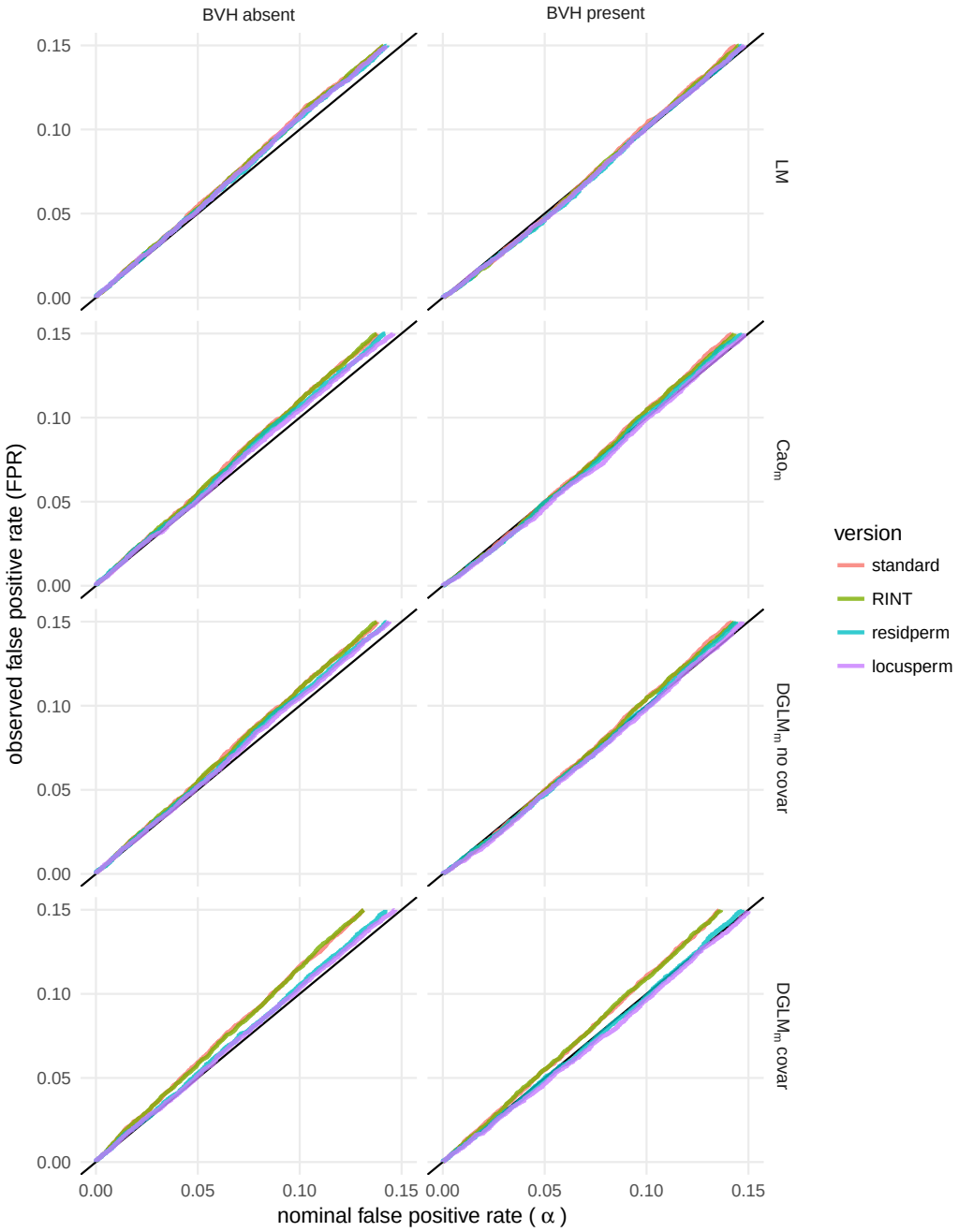
(b) In the absence of BVH, both mvQTL tests accurately control FPR and have similar power. In the presence of BVH of uknown source, the standard version of both mvQTL tests is anti-conservative and the other three versions maintain FPR control but suffer a decrease in power compared to the no-BVH scenario. Only DGLM_{MV} can incorporate information on the BVH-driving covariate. It achieves increased power and accurately controls FPR in its standard and locusperm versions and is conservative in its RINT and locusperm versions.



(c) Only the locusperm version accurately controls FPR in all scenarios. The standard version of both tests are anti-conservative in the presence of BVH of known source and the RINT and residperm versions are conservative in DGLM_{MV} the presence of BVH of known source.

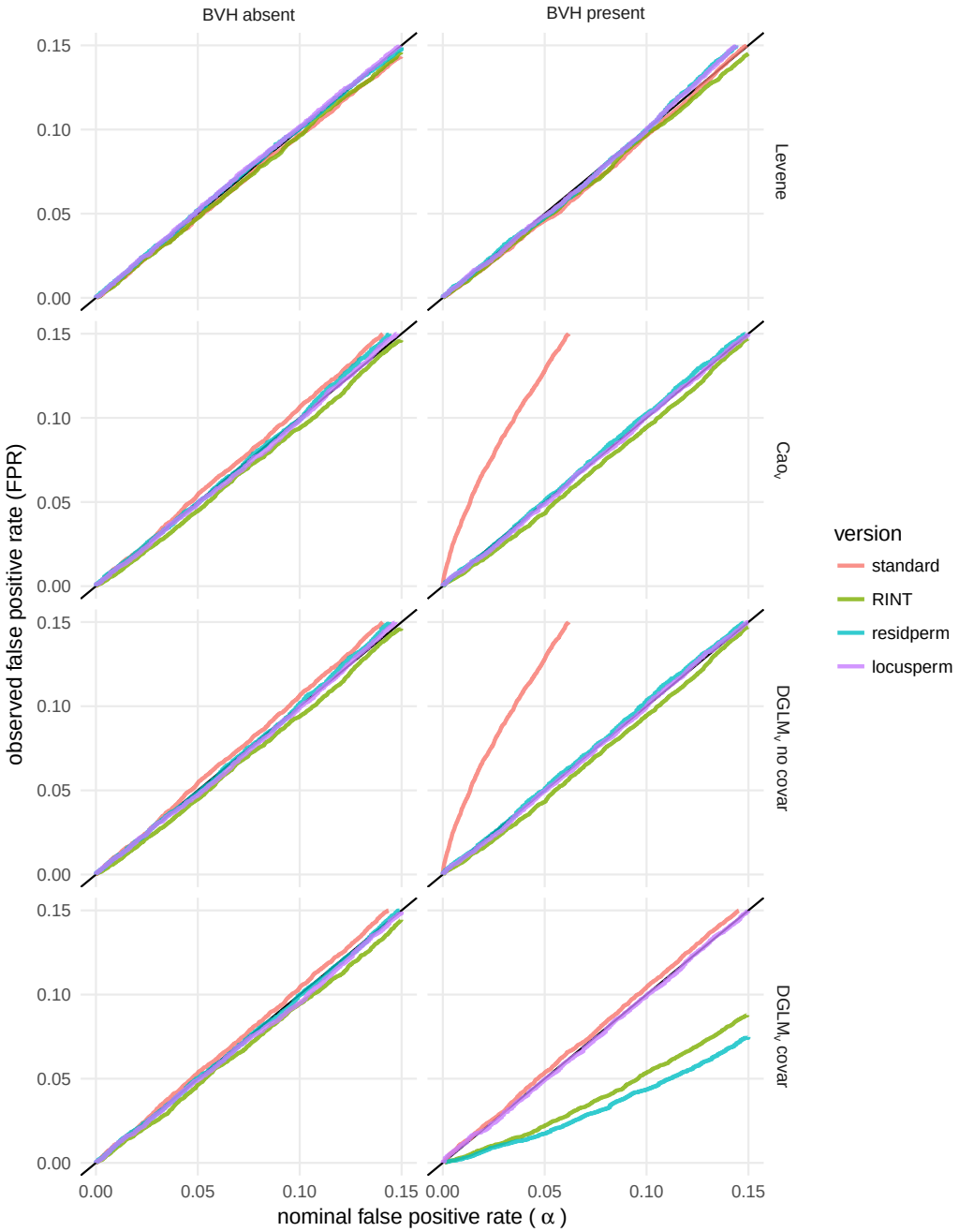
QQ Plots

Figure S4 The empirical false positive rate of each test-version for each nominal false positive rate, α, in $[0, 0.1]$.



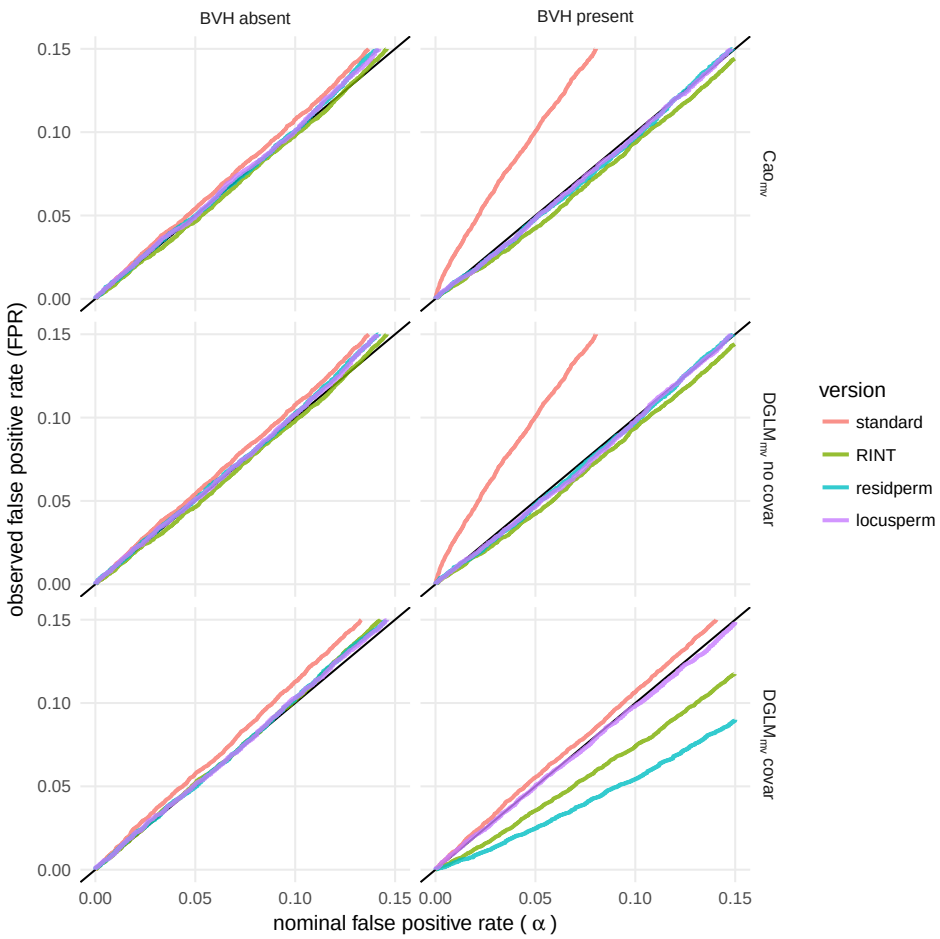
A test that accurately controls FPR will have empirical FPR $= \alpha$ for all value of α. All mQTL tests accurately control FPR.

Figure S5 The empirical false positive rate of each test-version for each nominal false positive rate, α, in $[0, 0.1]$.



A test that accurately controls FPR will have empirical FPR $= \alpha$ for all value of α. Amongst vQTL tests, Cao$_V$ has conservative behavior in the presence of BVH when the standard procedure is used, and DGLM$_V$ has anti-conservative behavior when the RINT and residperm procedures are used.

Figure S6 The empirical false positive rate of each test-version for each nominal false positive rate, α, in $[0, 0.1]$.



A test that accurately controls FPR will have empirical FPR $= \alpha$ for all value of α. mvQTL tests show the same pattern of deviation from accurate FPR control as vQTL tests (Figure S5), but to a lesser extent.

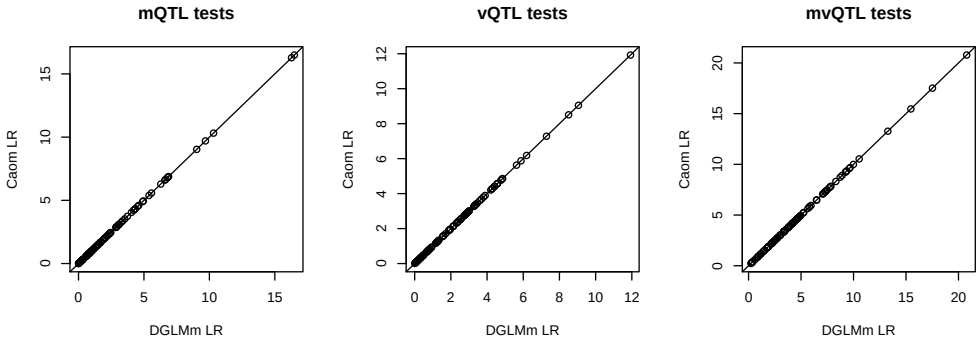
False Positive Rate Table

Table S1 Positive rates of all eleven tests in all scenarios based on 10,000 simulations, 1,000 permutations each to estimate empirical null distributions (residperm and locusperm), and a cutoff of $p = 0.05$. Note that in all cases the DGLM test without the covariate had identical or very nearly identical FPR to the Cao test that tests for the same kind of QTL.

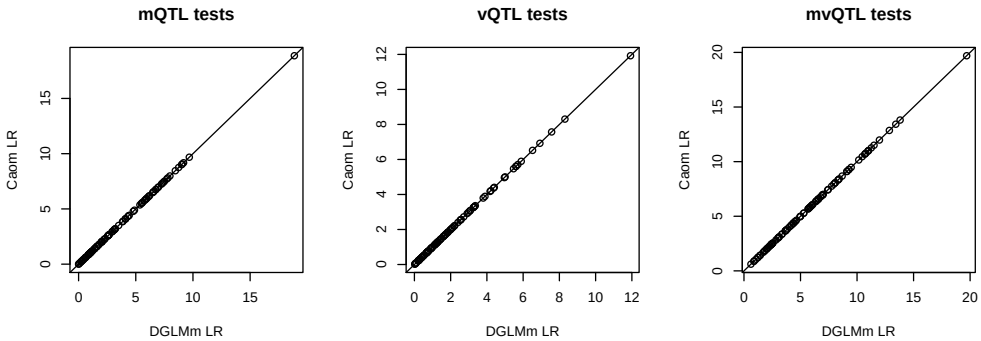
		BVH absent				BVH present			
test	version	null	mQTL	vQTL	mvQTL	null	mQTL	vQTL	mvQTL
LM	standard	0.054	0.706	0.055	0.504	0.047	0.713	0.051	0.510
	RINT	0.053	0.706	0.053	0.490	0.047	0.727	0.047	0.522
	residperm	0.052	0.700	0.053	0.495	0.046	0.708	0.050	0.506
	locusperm	0.051	0.698	0.054	0.494	0.046	0.708	0.050	0.506
Cao_{M}	standard	0.055	0.707	0.049	0.511	0.050	0.713	0.048	0.522
	RINT	0.055	0.706	0.047	0.501	0.049	0.726	0.047	0.534
	residperm	0.051	0.696	0.046	0.504	0.049	0.704	0.046	0.512
	locusperm	0.050	0.693	0.046	0.499	0.046	0.699	0.045	0.510
DGLM_{M} no covar	standard	0.055	0.707	0.049	0.511	0.050	0.713	0.048	0.522
	RINT	0.055	0.706	0.047	0.501	0.049	0.726	0.047	0.534
	residperm	0.052	0.695	0.046	0.502	0.047	0.703	0.048	0.513
	locusperm	0.051	0.693	0.044	0.500	0.047	0.700	0.044	0.509
DGLM_{M} with covar	standard	0.059	0.705	0.052	0.512	0.055	0.838	0.057	0.658
	RINT	0.058	0.703	0.050	0.503	0.055	0.835	0.056	0.651
	residperm	0.052	0.683	0.044	0.490	0.049	0.823	0.051	0.634
	locusperm	0.051	0.680	0.045	0.485	0.046	0.821	0.049	0.630
Levene's test	standard	0.048	0.045	0.653	0.466	0.046	0.048	0.566	0.387
	RINT	0.048	0.040	0.637	0.422	0.047	0.043	0.536	0.339
	residperm	0.052	0.049	0.661	0.477	0.048	0.051	0.573	0.394
	locusperm	0.051	0.050	0.661	0.475	0.048	0.050	0.573	0.393
Cao_{V}	standard	0.054	0.051	0.742	0.543	0.128	0.124	0.733	0.571
	RINT	0.045	0.040	0.691	0.468	0.043	0.047	0.557	0.365
	residperm	0.049	0.048	0.720	0.520	0.050	0.050	0.565	0.388
	locusperm	0.048	0.048	0.718	0.517	0.048	0.048	0.559	0.382
DGLM_{V} no covar	standard	0.054	0.051	0.742	0.543	0.128	0.124	0.733	0.571
	RINT	0.045	0.040	0.691	0.468	0.043	0.047	0.557	0.365
	residperm	0.048	0.049	0.721	0.522	0.050	0.050	0.565	0.388
	locusperm	0.047	0.048	0.718	0.519	0.049	0.049	0.559	0.381
DGLM_{V} with covar	standard	0.053	0.053	0.724	0.525	0.054	0.054	0.729	0.531
	RINT	0.046	0.041	0.673	0.453	0.022	0.022	0.560	0.341
	residperm	0.050	0.047	0.699	0.501	0.017	0.015	0.533	0.325
	locusperm	0.049	0.048	0.698	0.501	0.049	0.050	0.700	0.498
Cao_{MV}	standard	0.054	0.594	0.637	0.745	0.100	0.651	0.644	0.755
	RINT	0.046	0.585	0.570	0.703	0.042	0.608	0.435	0.649
	residperm	0.049	0.587	0.609	0.726	0.048	0.523	0.505	0.628
	locusperm	0.049	0.588	0.611	0.728	0.047	0.523	0.504	0.630
DGLM_{MV} no covar	standard	0.054	0.594	0.637	0.745	0.100	0.651	0.644	0.755
	RINT	0.046	0.585	0.570	0.703	0.042	0.608	0.435	0.649
	residperm	0.050	0.588	0.610	0.728	0.047	0.524	0.503	0.632
	locusperm	0.050	0.587	0.612	0.728	0.046	0.522	0.505	0.631
DGLM_{MV} with covar	standard	0.058	0.596	0.620	0.732	0.055	0.745	0.621	0.809
	RINT	0.052	0.587	0.554	0.696	0.036	0.720	0.432	0.732
	residperm	0.049	0.576	0.582	0.710	0.025	0.625	0.457	0.694
	locusperm	0.051	0.577	0.582	0.712	0.049	0.731	0.591	0.790

Cao's Profile-Likelihood Approximation is Extremely Accurate

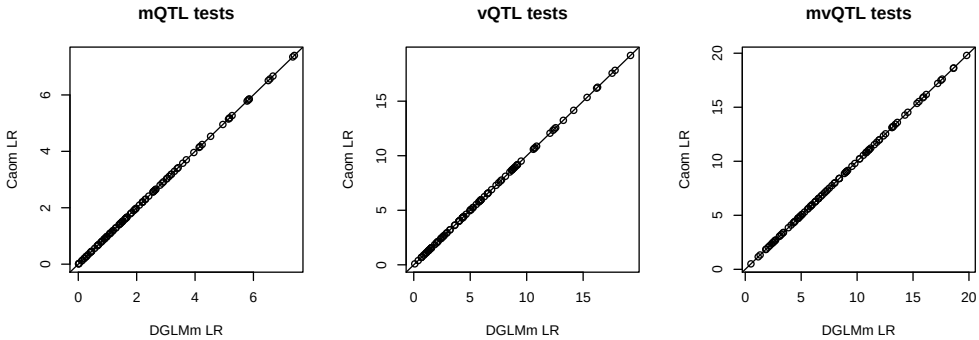
(a) null simulations



(b) mQTL simulations



(c) vQTL simulations



(d) mvQTL simulations

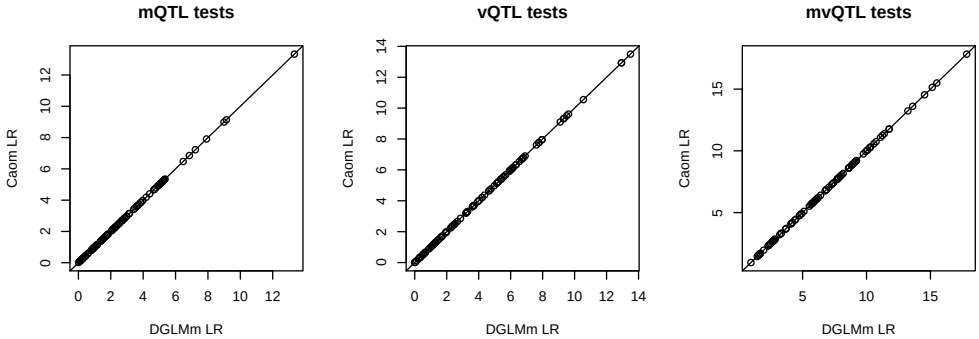


Figure S7 On simulated null loci, mQTL, vQTL, and mvQTL, Cao's profile likelihood method had identical likelihood ratio to DGLM when DGLM does not use any variance covariates.

Cao's Tests for All Phenotypes with BVH

These scans were conducted with the DGLM, without accounting for effects of sex and father on variance, shown by simulation to be identical to Cao's tests (Figure S7 and Table S1).

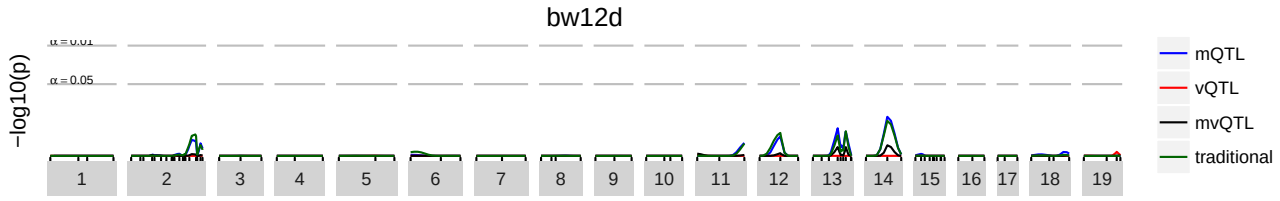


Figure S8

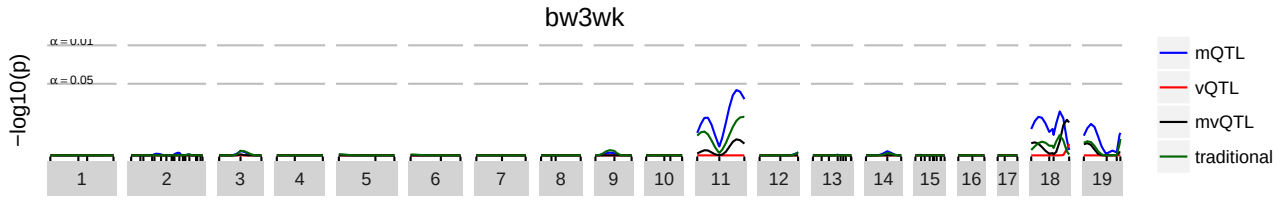


Figure S9

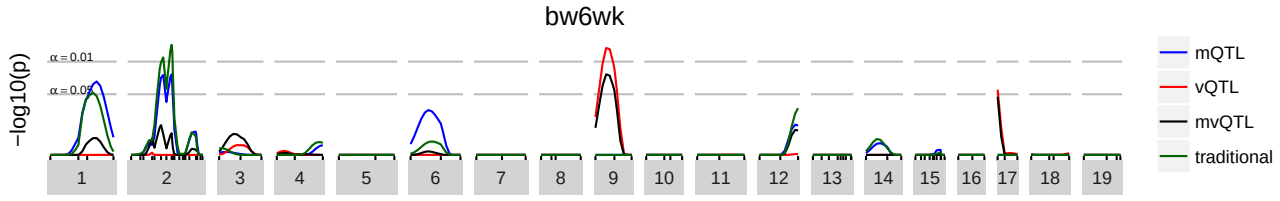


Figure S10

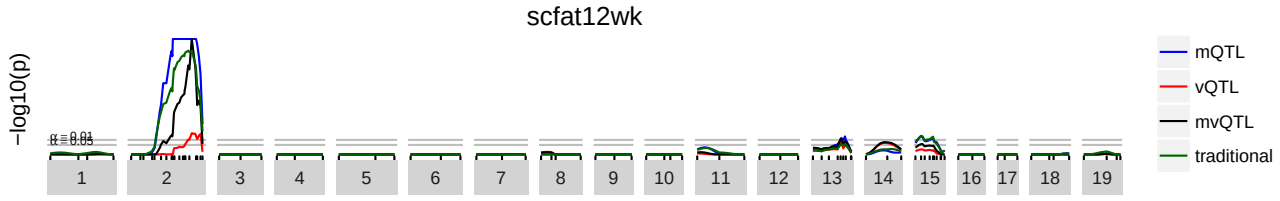


Figure S11

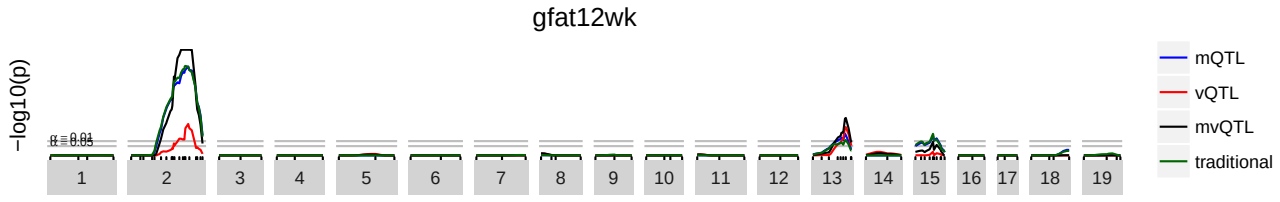


Figure S12

DGLM Tests for All Phenotypes with BVH

These scans were conducted with the DGLM, accounting for effects of sex and father on variance.

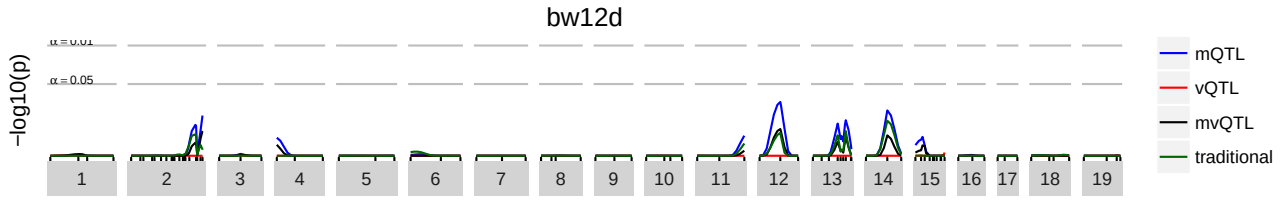


Figure S13

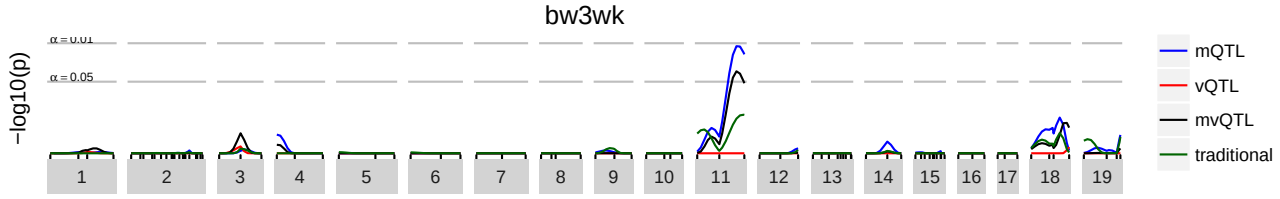


Figure S14

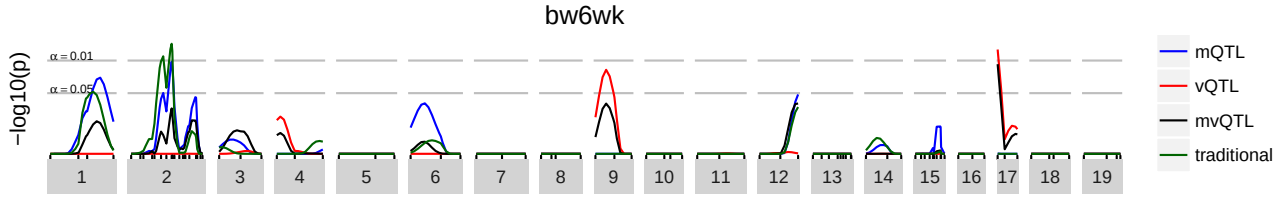


Figure S15

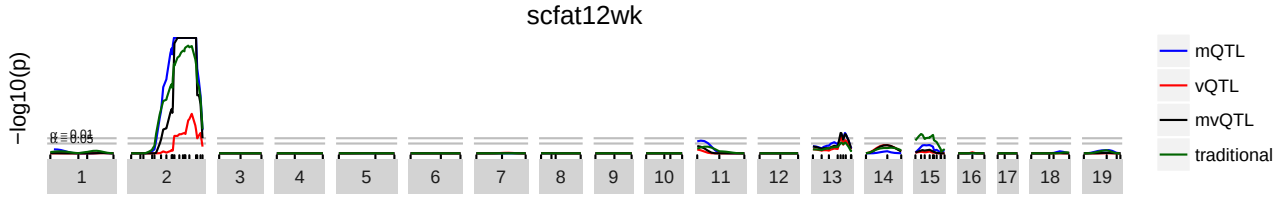


Figure S16

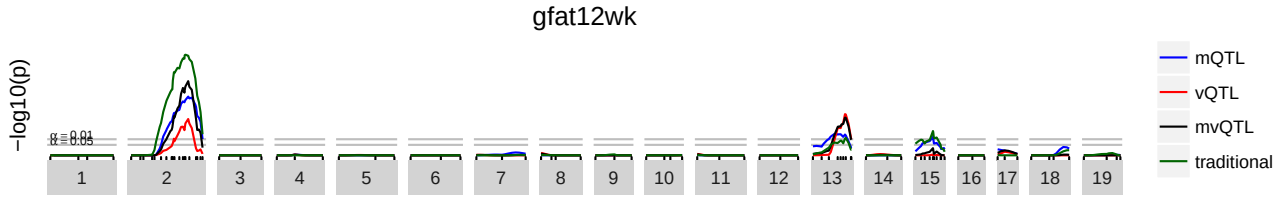


Figure S17