

---

**Authors**

Dalton Conley, Rebecca Johnson, Ben Domingue, Christopher Dawes, Jason Boardman, and Mark L Siegal

RESEARCH ARTICLE

# A sibling method for identifying vQTLs

Dalton Conley<sup>1</sup>\*, Rebecca Johnson<sup>1</sup>, Ben Domingue<sup>2</sup>, Christopher Dawes<sup>3</sup>, Jason Boardman<sup>4</sup>, Mark Siegal<sup>5</sup>

**1** Department of Sociology, Princeton University, Princeton, NJ, United States of America, **2** Graduate School of Education, Stanford University, Stanford, CA, United States of America, **3** Wilf Family Department of Politics, New York University, New York City, NY, United States of America, **4** Institute for Behavioral Sciences, University of Colorado, Boulder, Boulder, CO, United States of America, **5** Center for Genomics and Systems Biology, New York University, New York University, New York City, NY, United States of America

\* These authors contributed equally to this work.

\* [dconley@princeton.edu](mailto:dconley@princeton.edu)



**OPEN ACCESS**

**Citation:** Conley D, Johnson R, Domingue B, Dawes C, Boardman J, Siegal M (2018) A sibling method for identifying vQTLs. PLoS ONE 13(4): e0194541. <https://doi.org/10.1371/journal.pone.0194541>

**Editor:** Andre Scherag, University Hospital Jena, GERMANY

**Received:** August 11, 2017

**Accepted:** March 5, 2018

**Published:** April 4, 2018

**Copyright:** © 2018 Conley et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Framingham Heart Study (FHS) genotyping data are available from dbGAP for researchers who meet the criteria for access to the data (url: [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000007.v29.p10](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v29.p10)). Minnesota Twin Family Study (MTFS) are available from dbGAP for researchers who meet the criteria for access to the data (url: [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000620.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000620.v1.p1)).

**Funding:** This work was supported by National Institutes of Health grant R35GM118170 (MS). The

## Abstract

The propensity of a trait to vary within a population may have evolutionary, ecological, or clinical significance. In the present study we deploy sibling models to offer a novel and unbiased way to ascertain loci associated with the extent to which phenotypes vary (variance-controlling quantitative trait loci, or vQTLs). Previous methods for vQTL-mapping either exclude genetically related individuals or treat genetic relatedness among individuals as a complicating factor addressed by adjusting estimates for non-independence in phenotypes. The present method uses genetic relatedness as a tool to obtain unbiased estimates of variance effects rather than as a nuisance. The family-based approach, which utilizes random variation between siblings in minor allele counts at a locus, also allows controls for parental genotype, mean effects, and non-linear (dominance) effects that may spuriously appear to generate variation. Simulations show that the approach performs equally well as two existing methods (squared Z-score and DGLM) in controlling type I error rates when there is no unobserved confounding, and performs significantly better than these methods in the presence of small degrees of confounding. Using height and BMI as empirical applications, we investigate SNPs that alter within-family variation in height and BMI, as well as pathways that appear to be enriched. One significant SNP for BMI variability, in the MAST4 gene, replicated. Pathway analysis revealed one gene set, encoding members of several signaling pathways related to gap junction function, which appears significantly enriched for associations with within-family height variation in both datasets (while not enriched in analysis of mean levels). We recommend approximating laboratory random assignment of genotype using family data and more careful attention to the possible conflation of mean and variance effects.

funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

### From effects of loci on trait means to effects of loci on trait variance

The extent to which a complex trait varies in a population is a product of mutation, genetic drift and natural selection, as well as environmental variation and its interaction with genotype. Moreover, genotypes may differ in their propensities to vary. That is, particular genotypes might be more sensitive than others to changes in the environment or to the effects of new mutations. Sensitivity to the environment could come in the form of phenotypic plasticity, whereby stereotyped phenotypic changes occur under particular circumstances, or in the form of developmental instability, whereby random fluctuations in the internal or external environment lead to different phenotypic outcomes ([1–3]).

Sensitivity to the effects of mutations is related to cryptic genetic variation. In a population in which individuals are relatively insensitive to the effects of mutations, allelic variation may accumulate, while not presenting phenotypic effects. Replacing an allele that confers less sensitivity to mutational effects with one that confers more sensitivity (or introducing an environmental perturbation with similar effect) will cause the accumulated variation to have phenotypic consequences ([4, 5]). This release of cryptic genetic variation might have major implications for the adaptation of populations to environmental change, as well as for the genetics of human complex traits ([6–9]). Indeed, it has been proposed that release of cryptic genetic variation might be responsible for the increased prevalence of human “diseases of modernity” such as diabetes ([7]).

Model organisms have been used to identify genes that control sensitivity to mutations or the environment [5]; [10]). Two approaches have been taken. One approach is to test for differences in phenotypic variance between wild-type and mutant strains. For example, the molecular chaperone Hsp90, when impaired, reveals cryptic genetic variation in the fly *Drosophila melanogaster*, the flowering plant *Arabidopsis thaliana*, the fish *Danio rerio* and the budding yeast *Saccharomyces cerevisiae* ([11–14]).

Evidence on whether Hsp90 controls environmental sensitivity is mixed ([15–17]). However, screens of the *S. cerevisiae* genome identified hundreds of genes that, when mutated, cause increased variation in cell morphology among isogenic cells raised in the same environment. That is, these mutations increased sensitivity to fluctuations in the internal or external environment ([18, 19]). A test of one of these genes, which showed a major increase in environmental sensitivity upon deletion, revealed a high degree of epistasis with new mutations but no net increase in mutational sensitivity upon deletion ([2, 20, 20]). A similar result was found for impairment of Hsp90 ([21]). In general, the relationship between suppression of the effects of environmental variation and suppression of the effects of mutational variation is unclear ([5, 10]).

The second approach to identifying variance-controlling genes is to use linkage or association analysis to map natural genetic variants that confer differential sensitivity. Such searches for variance-controlling quantitative trait loci (vQTLs) have been conducted to identify determinants of microenvironmental sensitivity (i.e., inherent stochasticity or sensitivity to fluctuations within a nominally constant environment) and/or phenotypic plasticity (variation across controlled environments) of various phenotypes, including morphological and life-history traits as well as expression levels of individual genes, in a range of organisms including humans ([22–36]). Natural genetic variation affecting sensitivity to other segregating alleles has been studied as well, and indeed there are natural variants of the Hsp90 gene that appear to reveal cryptic variation ([37–39]).

## Existing methods for detecting vQTL

The first study claiming to map a locus that controls variance of a human trait examined body mass index (BMI) using data from 38 cohorts that participate in the GIANT consortium for genome-wide analysis (GWA) of single-nucleotide polymorphisms (SNPs) that affect human height ([40])(although a prior study had examined variance in order to prioritize the search for gene-by-environment and gene-by-gene interaction effects [41]). Yang et al. computed a Z-score (inverse-normal transformation) for BMI for each of 133,154 individuals then performed GWA on the squared Z-score. This squared Z-score captures the magnitude of each individual's deviation from the mean phenotype and therefore is meant to act as an individual-based measure of variance. In this discovery sample they found SNPs in the FTO gene and the RCOR1 gene that appear to control variation in human BMI. In a replication sample of 36,327 individuals from 13 cohorts, one SNP in the FTO gene was confirmed [40].

## Challenge one: Isolating loci that affect variance in a trait from loci that affect mean of a trait

There are several challenges with association analyses that identify vQTL. The first issue is mean-variance confounding. A given locus can have: 1) effects on mean levels of a trait (mean effects), 2) effects on variance in a trait (variance effects), or 3) effects on both mean levels of a trait and variance in a trait. For isolating variance effects, it is critical to check that any detected effects on "variance" adequately distinguish between the three types of loci. In particular, this conflation of variance and mean effects was a concern in the analysis of Yang et al. ([40]) because normalized variation scores will tend to be higher for populations with higher mean levels and because FTO is one of the most well-established genes that affects the mean level of BMI in human populations ([42–50]). Research examining MMP3 protein levels in cerebrospinal fluid found similar overlap, where SNPs in linkage disequilibrium with a locus well-established in predicting mean levels of the trait (rs679620 of the MMP3 gene) were associated with both higher mean levels and higher variance in the trait ([51]).

Yang et al. addressed the effect of mean BMI by showing that there is no global correlation between mean effects of SNPs and their variance effects. However, it must be considered that a lack of correlation between mean and variance effects across the genome could be caused merely by the fact that the vast majority of SNPs have negligible or nonexistent effects on both mean and variation. In the present study, we show through simulation that the squared Z-score method has an inflated type I error rate and detects variance effects for a trait simulated to have mean effects only.

In addition to mean-variance confounding, methods to investigate vQTL must address several other issues, some of which are shared with the estimation of mean effects and others of which, such as mean-variance confounding, are unique to or particularly acute in the case of variance effects ([31]). Methods for vQTL analysis beyond the squared Z-score method can be grouped into: 1) non-parametric methods that test whether the three genotypes at a biallelic locus (minor-allele homozygote, heterozygote, and major-allele homozygote) have equal or unequal variance, and 2) parametric methods that relax the assumption in GWA linear regressions that the residual error is identically distributed across all genotypes.

Non-parametric methods include Levene's test, which uses a test statistic derived from the squared difference between an individual's level of a trait and the genotype-level mean or median, the latter of which is used to make the method more robust to a non-normally distributed trait ([41, 52]) and the Fligner-Killeen (FK) test, which is similar to Levene's test, in that it uses the absolute difference between an individual's level of a trait and the genotype-level median, but then computes the test statistic based on ranks of these differences. The FK test

can be used either as a standalone test for variance effects ([25]), or as the test statistic for the scale component of the Lepage or other joint scale-location test ([53]). More recently proposed non-parametric tests consider not only the variance of the trait distribution but other features as well, such as skew ([54, 55]).

The main drawback of non-parametric tests for detecting vQTL's is that the tests, which compute differences between discrete genotype groups, cannot directly control for important covariates, such as age, sex, and population structure. Some adopt a two-stage regression procedure for including covariate controls: first the trait is regressed on covariates, then the variance test of interest is performed on the residualized dependent variable ([55, 56]). However, two-stage procedures have been shown in simulations to reduce power and induce bias ([51]).

### Challenge two: Controlling for unobserved confounders that can bias the estimate on minor allele count

This problem—how to include covariates in a way that does not induce bias—is acute because of the importance of two types of controls when investigating variance-controlling loci: control for population stratification and controls for nonrandom association between genotype and environment. The former control is needed in all GWA analyses to separate the effect of any particular locus from the effects of all other loci shared by virtue of common ancestry. The latter control might be particularly relevant to vQTL analyses because mean effects of genotypes might impact the environment that is experienced, which might in turn impact variance ([57]).

Parametric approaches that use generalized linear models to jointly estimate the mean and variance of a trait address this problem ([51, 58, 59]); these models can include controls for population stratification as well as controls for observed covariates that influence genetic distribution into variance-affecting environments. The double generalized linear model (DGLM) approach begins with the typical linear model for estimating mean effects where the residual variance is the same across genotypes, then the model is relaxed to allow residual variance to differ by genotype and to incorporate non-genetic covariates that might contribute to residual variance; it iterates between estimating parameters for the mean versus parameters for the variance until convergence ([59]). DGLM thus allows joint estimation of mean effects and variance effects, attempting to address mean-variance confounding, and permits controls for population stratification directly in the model. The main drawback of DGLM is that, similar to other methods that control for confounding by controlling for *observed covariates* (age; sex; the top principal components) correlated with both the individual's genotype and the outcome variable, the method cannot control for unobserved confounding that may bias the estimate of a SNP's effect.

In particular, two types of unobserved confounders may be correlated both with an individual's genotype and mean or variance effects in a trait. First is population stratification. Population stratification is typically controlled for in these methods by inclusion of principal components of the sample population's genotypes among the vectors of covariates predicting the mean and/or residual variance of a trait. This control is especially important for traits, such as BMI, that are expected to show considerable environment-dependence. Especially when pooling such traits across cohorts, there is a risk that systematic differences in environment correlate regionally with systematic differences in genetic variation. That is, it is plausible, due to population stratification, that any genetic signal is merely acting as a proxy for culture and environment—a potential confounder that has been well-illustrated by the “chopsticks gene” example ([60]). Controlling for population stratification using principal components (PCs) addresses some confounding, but there is residual between-family confounding

even with these controls. This residual confounding can occur when environmentally influential factors are not randomly distributed across families but also do not correlate with the eigenvectors in the genetic matrix.

The other critical covariate that might bias vQTL estimates is genotype-environment correlation (rGE). Genotype-environment correlations may be caused by niche construction, whereby individual organisms shape the environment (in a genotype-dependent way) ([61–65])—dynamics we might expect to occur for phenotypes that have significant behavioral and environmental etiologies, such as BMI. As a result, genotypes may be associated with variance in BMI and other traits not through direct genetic effects but through interaction with alternative environments associated with variance such as more versus less sedentary lifestyles. An analogous situation that illustrates this potentially confounding genotype-by-environment interaction effect is that of caffeine consumption. A variant in a gene that encodes a caffeine-metabolizing enzyme can lead to greater variation with no effect on the mean through a mechanism of niche construction—i.e. individuals with the minor allele avoid coffee altogether, or if they are unable to do so, they end up drinking more than those with the major allele, thus leading to greater variance thanks to the coffee “environment” ([57]).

Approaches such as DGLM can control for observed covariates that are correlated with genotype and influence construction of variance-affecting environments. However, these methods cannot control for unobserved differences that produce these correlations. The present paper uses a family-based model to control for these unobserved differences. Two existing family-based models allow for investigations of variance-based loci in samples among related as opposed to unrelated individuals, but do not leverage the family-based structure of the data to control for unobserved confounders that vary between families ([58]; [66]). First is a family-based version of the likelihood ratio test that adds a random effect meant to capture familial correlation in a trait. Although the family-level random effect helps control for unobserved variation between families that may influence variance in a trait through pathways other than genotype, the model relies on the strong assumption of independence between these unobserved features of family and the observed covariates. We show via simulation (see [Results](#)) that when there is non-zero correlation between unobserved features of a family and observed covariates, random effects approaches generate biased estimates of a SNPs’ effects, confirming results shown in non-genetics contexts ([67]). Similarly, DGLM in a sample containing monozygotic and dizygotic twins, which has the advantage of isolating non-genetic from genetic sources of variance, does not control for unobserved features that vary between families and that affect construction of variance-affecting environments ([66]).

### Proposed solution: The sibling standard deviation method

We offer an alternative methodology—comparisons of variation within sibling sets while controlling for parental genotype—that does not assume independence between observed covariates and unobserved between-family differences. As a result, the method better approximates random assignment of genotype in a laboratory. Utilizing a regression-based framework, the approach retains the advantages of the DGLM and Bayesian regression approaches: the ability to include covariates and control for mean effects when estimating variance effects through estimation of parameters capturing both. The model uses sibling pairs as the unit of analysis and regresses the standard deviation of the sibling pair’s trait on the pair’s count of minor alleles with sibling pair-level controls that include controls for the mean level of the trait in the sibling pair, parental genotype, pair sex (MM or FM or FF), mean pair age, and the within-pair age difference (for the full model specification, as well as alternative specifications tested, see

**Methods**). The control for the mean level of the trait in a sibling pair avoids inflated Type I error rates for SNPs that affect the mean of a trait but not the trait's variance.

The proposed methodology, although restricted in applicability to datasets that have a family-based design with at least two offspring, makes a trade-off. The method has reduced statistical power because the sample size is halved when we treat siblings rather than individuals as the unit of analysis. However, the advantage is an estimate of a minor allele's contribution to variance that, in the presence of unobserved confounders, correctly fails to find variance effects when a locus only has mean effects.

This is a particularly important trade-off to make when we wish to rule out gene-environment correlations across populations as well as population stratification as alternative explanations to variance-locus associations. As noted above, this is especially critical for a phenotype such as BMI and a locus such as FTO given that environment and behavior (such as sedentari-ness) alter the FTO-BMI relationship and may vary significantly across cohorts/societies ([68]). In cases where the goal is not to study control of variance per se, but instead is to probe the existence of gene-environment or gene-gene interactions in a way that avoids the high-dimensional parameter space problems of traditional approaches, the use of vQTL approaches that have higher statistical power but also a higher rate of false positives for SNPs that affect the mean of a trait but not the variance might be warranted.

In addition to power, another important feature of an approach to detect vQTL is the flexibility to capture non-linear effects of alleles, which the DGLM, the parametric bootstrap-based likelihood ratio test, and Bayesian regressions allow for by allowing genotypes to be specified using three indicator variables to capture non-linearities. The present family-based approach is potentially susceptible to confounding of variance effects by non-linear effects of alleles, because the association mapping is done on the sibship unit so the genotype is represented as the total number of major or minor alleles of each sib pair. Therefore, if dominance were at play with respect to mean levels, then this might generate spurious effects on variance ([69]). That is, if among heterozygotes there was no effect of an allele on mean levels but among homozygotes there was, then this itself would generate apparent, but spurious, effects on variance even when controlling for linear mean effects. However, by comparing subgroups among sibship pairs that have two minor alleles (out of four possible in total), we are able to check for this possibility. Specifically, by comparing those 2-minor allele pairs where both siblings are heterozygotes (i.e. each individual has one minor allele) with those where both are homozygotes (where one sibling has zero minor alleles and the other has two), we can rule out this possible statistical artifact of non-linear effects on mean levels. Although the possibility of non-linearities that do not reflect true variance effects can never be totally eliminated ([69]), this approach guards against a primary form of non-linearity—dominance.

## Preview of the results

Below we report two sets of analyses. In simulations, we show how unobserved confounding can bias estimates of a SNP's effect. Two approaches to estimating variance effects—the squared Z-score method and DGLM—display inflated type I error rates in the presence of this confounding. In contrast, the sibling standard deviation approach that we propose detects variance effects when these effects are present but, by controlling for the mean of the trait across siblings, correctly fails to find variance effects when only mean effects are present.

In an empirical application, we then use the sibling SD approach to perform genome-wide analyses of variance effects on two phenotypes: height and BMI. We replicate one genome-



wide statistically suggestive hit for variation in BMI from the Framingham Heart Study (FHS) data, our discovery sample, in our replication sample, the Minnesota Twin Family Study (MTFS). We also test whether our potential variance-related alleles are merely reflecting dominance effects among heterozygotes; we find no evidence for this. Finally, we perform gene-based and pathway enrichment analysis. We find one pathway, related to gap junction function, that is significantly enriched in both our discovery and replication samples for associations with variance in height. We then discuss the implications of our findings for prior and future research.

## Results

### Simulations

The simulations are divided into two parts. First, we show that while we can address unobserved confounding when estimating the *mean* of a trait using a fixed effects estimator that identifies the effect of an allele off of between-sibling variation, combining this approach with two current approaches to variance detection—the squared Z-score method and DGLM—fails to correct for this bias. This introduces the challenge: how can we estimate the effect of an allele on trait variance in the presence of unobserved confounding? Part two evaluates the sibling SD method as a solution.

**Part one: Two approaches to variance estimation in the absence versus presence of an unobserved confounder.** When estimating mean effects, we can remove bias in the estimate of the effect of a minor allele caused by unobserved confounding by using a fixed effects estimator that demeans the outcome, genotype, and other observed covariates by the mean within a family ([67]). Examining the trait simulated to have *neither mean nor variance effects*, Fig A in S1 File shows that in the presence of any unobserved confounding, other estimators (a random effects model that assumes zero correlation between unobserved confounders and observed covariates; a pooled regression model with one randomly sampled sibling) return biased estimates of the effect of an additional minor allele on the trait's mean when there is unobserved confounding. More specifically, it shows that across the 1000 replicates, we see upward bias in the random effects and pooled regression estimators when we move from the case of no confounding or a very low amount of confounding  $\rho = 0.01$  to the case of some confounding ( $\rho = 0.05$  and  $\rho = 0.1$ ). Table A in S1 File, which presents the results of a Hausman test comparing the null hypothesis that the estimated random effects coefficients are equal to the estimated fixed effects coefficients, rejects the null at levels of confounding greater than  $\rho = 0.01$ .

These results show how a fixed effects estimator that identifies the effect from sibling deviations from a family's mean count of alleles recovers an unbiased estimate. How can we translate these findings from the case of investigating the effect of an additional minor allele on the trait's mean (QTL) to investigating the effect of an additional minor allele on the trait's variance (vQTL)?

One approach is to use existing methods for variance detection on a transformed version of the data. In particular, the fixed effects estimator is generated by demeaning the outcome, genotype, and other covariates by the mean across the grouping unit (family in this case), as represented as follows, where  $i$  indexes an individual,  $j$  indexes a family,  $k$  indexes a SNP, and  $X$  represents genotype and other covariates.  $\alpha$  represents an intercept specific to each family,  $\beta$  represents coefficients on covariates that vary across individuals, families, and variants (e.g., minor allele count), and  $\epsilon$  represents an error term. The first equation shows the demeaning of each of the terms and the second equation shows that this demeaning removes the family-level



intercept from the regression.

$$y_{ijk} = (\alpha_j - \bar{\alpha}_j) + \beta_k(X_{ijk} - \bar{X}_{jk}) + (\epsilon_{ijk} - \bar{\epsilon}_{jk})$$

$$y_{ijk} = \beta_k(X_{ijk} - \bar{X}_{jk}) + (\epsilon_{ijk} - \bar{\epsilon}_{jk})$$

We can represent this demeaned version of the data more compactly as follows:

$$y_{ijk} = \beta_k \ddot{X}_{ijk} + \ddot{\epsilon}_{ijk} \tag{1}$$

One approach to translating the FE estimator to the vQTL context is to use the transformed version of the data represented in Eq 1 with two existing methods for vQTL detection: the squared Z-score method ([40]) and the DGLM ([59]). We explore the properties of each approach in the sections that follow on three versions of the dependent variable:

1. *Dependent variable simulated to have neither mean nor variance effects:* the dependent variable is simulated to have no relationship between the count of minor alleles and either the mean or variance of the outcome (see [Methods](#))
2. *Dependent variable simulated to have mean effects only:* the dependent variable is simulated to have a positive relationship between the count of minor alleles and the mean of the outcome (but not the outcome's variance) (see [Methods](#))
3. *Dependent variable simulated to have variance effects only:* the dependent variable is simulated to have a positive relationship between the count of minor alleles and the variance of the outcome (but not the outcome's mean) (see [Methods](#))
4. *Dependent variable simulated to have mean and variance effects:* the dependent variable is simulated to have a positive relationship between the count of minor alleles and both the mean and the variance of the outcome (see [Methods](#))

**Squared Z-score results.** Table B in [S1 File](#) presents the results of regressing the squared Z-score (estimated separately by sex) on the minor allele count with controls for sex, age, and ancestry across the 1000 replicates. The table summarizes the percentage of replicates where the coefficient on the minor allele count  $\neq 0$  at  $p < 0.05$ . We want this percentage to be low when the dependent variable is simulated to *have no* variance effects: so either null effects or mean effects only in order to control for type I error. We want this percentage to be high when the dependent variable is simulated to *have* variance effects: so either variance effects only or mean and variance effects. The results shows that in the *absence* of an unobserved confounder, we see moderately inflated type I error rates (estimating variance effects when only mean effects are present in 5.9% of simulations). In the *presence* of an unobserved confounder, we see more inflated type I error rates: 8% of simulations estimate variance effects when only mean effects are present. Table C in [S1 File](#) shows that demeaning the data does not address these inflated type I error rates.

**DGLM results.** The squared Z-score method generates one coefficient of interest that represents an allele's contribution to variability in the form of increasing an individual's squared Z-score. The DGLM method, which estimates a linear regression for the trait mean and a gamma regression from the squared residuals from the first model, generates two coefficients of interest: ( $\beta$  refers to coefficients on mean effects;  $\gamma$  refers to coefficients on variance effects). [Table 1](#) summarizes the true  $\beta$  and  $\gamma$  for different types of simulated traits:

**Table 1. Outcomes versus coefficients.**

Simulated dependent variable	Coefficient on minor allele count from DGLM
Neither mean nor variance effects	$\beta = 0; \gamma = 0$
Mean effects only	$\beta \neq 0; \gamma = 0$
Variance effects only	$\beta = 0; \gamma \neq 0$

<https://doi.org/10.1371/journal.pone.0194541.t001>

Table D and Table E in [S1 File](#) look at these three cases. For the trait with null effects, DGLM should estimate  $\beta = 0$  for mean effects and  $\gamma = 0$  for variance effects. For the trait with mean effects, DGLM should estimate  $\beta \neq 0$  and  $\gamma = 0$ . For the trait with variance effects, DGLM should estimate  $\beta = 0$  and  $\gamma \neq 0$ . Therefore, the table summarizes the percentage of  $\beta \neq 0$  and  $\gamma \neq 0$  at the  $p < 0.05$  level.

The results show that in the presence of between-family confounding between the genotype and outcome variable, the method has two types of type I error. First, when the outcome has *neither* mean nor variance effects and there is confounding, the method detects mean effects when none are present in over 77% of simulations. Second, when the outcome only has variance effects and there is confounding, the method detects mean effects when none are present in over 70% of simulations. Transforming the data via de-meaning reduces these type I error rates but fails to detect variance effects when these are present (type II error).

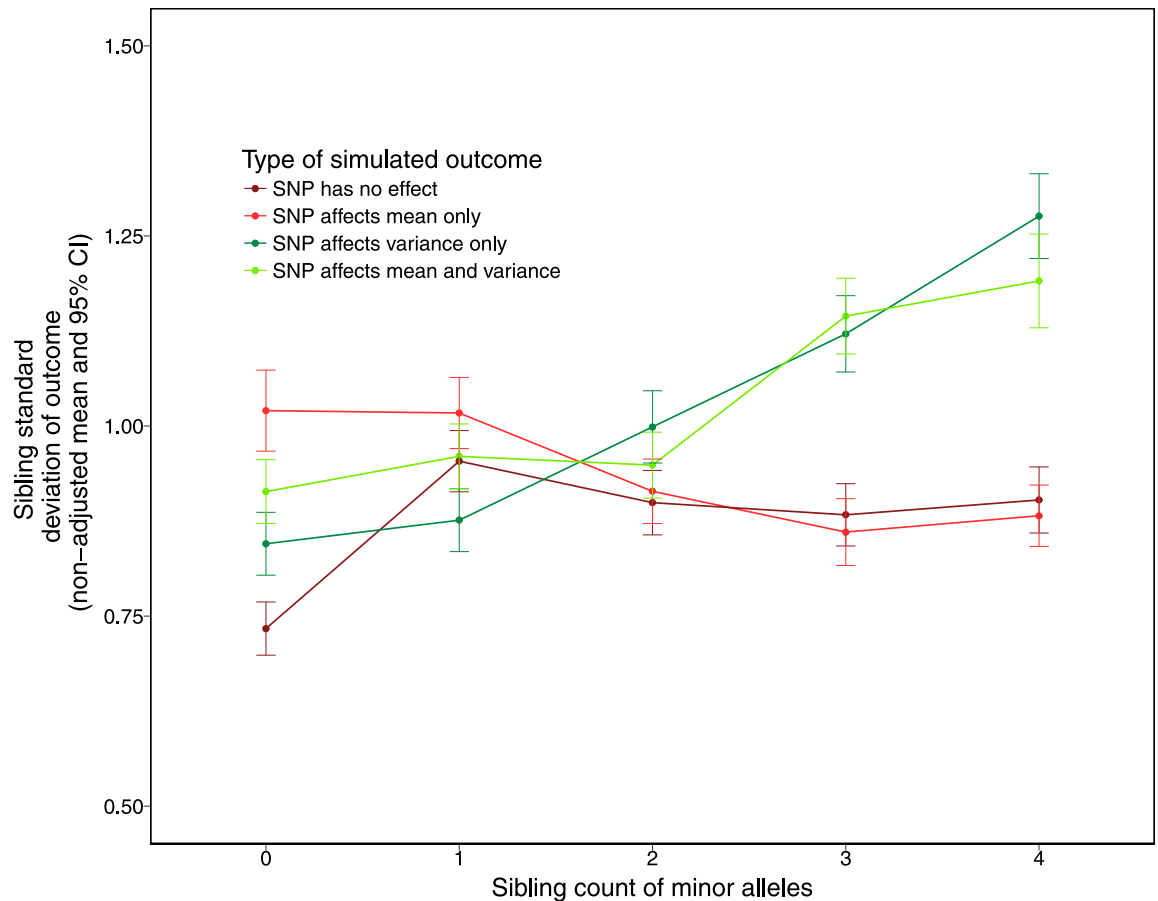
**Problems part one reveals.** The previous section reveals a problem when choosing a method for detecting the effects of an additional minor allele on the mean or variance of a trait in the presence of unobserved confounding between an individual's genotype and the trait. In particular, while the fixed effects estimator provides an unbiased estimate of the effect of minor alleles on the *mean* of a trait, de-meaning the data and then trying to estimate variance effects using the squared Z-score or DGLM approach leads to inflated rates of type I error when the loci has null or mean effects but no variance effects. In addition, the DGLM estimated on demeaned data fails to detect variance effects when these effects are present.

These problems point to the need for a method that corrects for bias caused by unobserved confounding but that is also able to detect effects of minor alleles on a trait's variance when these effects are present. The next section investigates properties of the sibling standard deviation method as a proposed solution.

**Part two: Properties of the sibling standard deviation method.** To examine variance effects, we estimate the sibling standard deviation method (see [Methods](#)). Before moving to results across the 1000 replicates, [Fig 1](#), which presents the sibling count of minor alleles and non-adjusted mean sibling standard deviation for pairs with that count for one randomly chosen replicate, highlights that the standard deviation of the trait increases ( $y$  axis) as we see an increase in the sibling count of minor alleles ( $x$  axis) for the 1. trait simulated to have variance effects only and 2. trait simulated to have both mean and variance effects. In contrast, the standard deviation appropriately does not increase for the 1. trait simulated to have mean effects only and 2. the trait simulated to have neither mean nor variance effects.

[Fig 2](#) shows that the results from the [Fig 1](#) generalize across the replicates. The figure shows the distribution of  $\gamma$  on sibling minor allele count from regressing the sibling standard deviation of a trait on this count for two traits: the trait simulated to have mean effects only and the trait simulated to have variance effects only. The figure shows the distribution is properly centered around zero for the trait with mean effects and properly centered around  $\gamma \neq 0$  for the trait with variance effects.

Table F in [S1 File](#) summarizes the percentage of simulations that reject the null of  $\gamma$  on minor allele count equaling zero. The null hypothesis in this case, rather than a specific value for the  $\gamma$  on minor allele count, follows other methods in posing the null of whether the

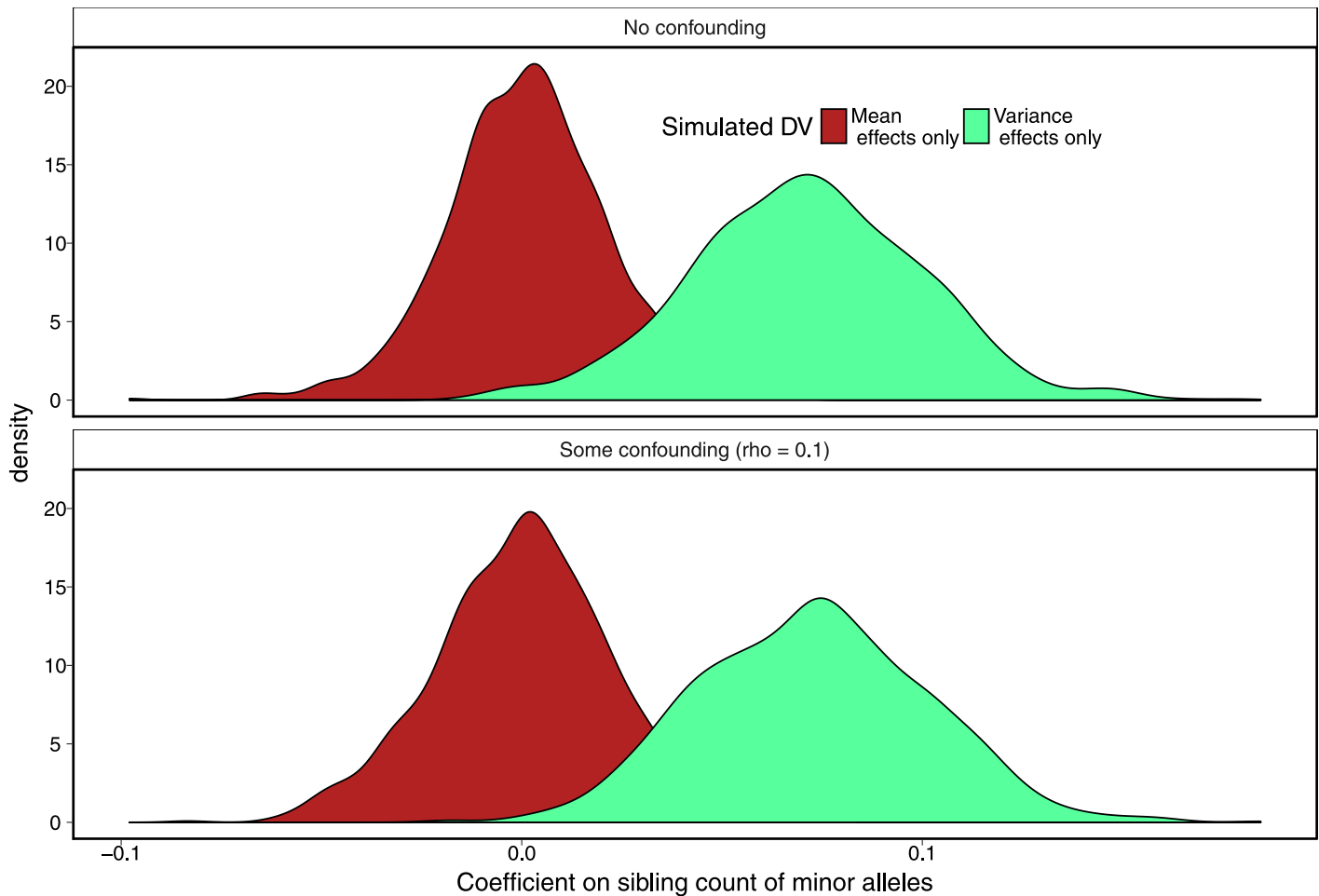


**Fig 1. Raw means of sibling standard deviations (before age, sex, mean of trait, and parental genotype controls) by count of minor alleles.** The graph shows that the sibling standard deviation increases with the count of minor allele snps that have effects on variance only, or effects on both the mean and variance of a trait, while stays flat for snp's that only effect the mean or that not associated with the trait.

<https://doi.org/10.1371/journal.pone.0194541.g001>

coefficient is equal to zero [51]. In the presence of *no unobserved confounding*, the sibling SD approach has the expected type I error rate of correctly failing to detect between 95.4 and 95.8% of variance effects when only mean effects are present depending on the specification, and correctly failing to detect between 94.6 and 95.7% of variance effects when neither mean nor variance effects are present. The squared Z-score method correctly fails to detect 94.1% of variance effects when only mean effects are present, indicating that the sibling SD method performs equally well as that existing method in the no confounding case.

The lower type I error rate of the sibling SD method comes with comparable type II error rates as other methods when we do not control for parental genotype: the sibling SD method detects variance effects when these are present in 81.6% of cases, while the squared Z-score method detects variance effects when these are present in roughly 87% of cases and the DGLM method detects variance effects in 75.3% of cases. When we control for parental genotype in the sibling SD method as the most stringent form of control on confounding, the method has a higher type II error rate than the squared Z-score and DGLM. At the sample size used in the simulation (1000 sibling pairs), the method is under-powered to detect variance effects when we control for parental genotype, detecting these effects in roughly 40% of cases. Later, we



**Fig 2. Results of sibling standard deviation method across 1000 replicates.** The figure shows that both in the presence and absence of family-level confounding between the genotype and outcome variable, the method, which examines the effect of an additional minor allele in the sibling pair on the trait’s standard deviation, correctly estimates no variance effects ( $\beta = 0$ ) when the outcome is simulated to have mean effects only, and correctly detects variance effects ( $\beta \neq 0$ ) when the outcome is simulated to have variance effects only.

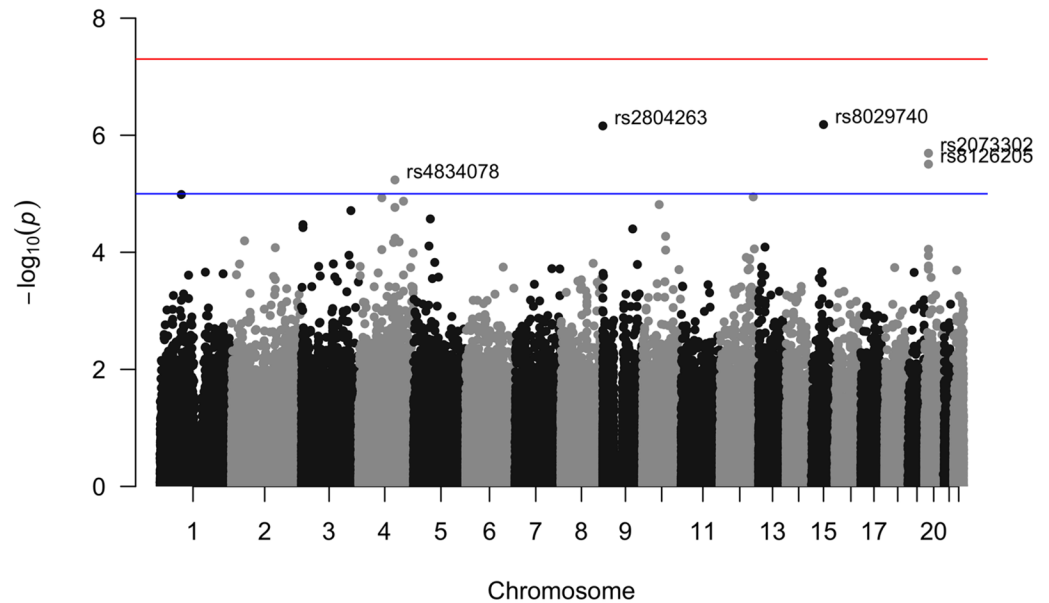
<https://doi.org/10.1371/journal.pone.0194541.g002>

discuss larger family-based study designs that are large enough to control for parent genotype and still have sufficient power to detect variance effects.

In the presence of *some unobserved confounding*, the sibling SD method greatly outperforms the other approaches in controlling type I error rates. While DGLM and the squared Z-score method detect mean effects when only variance effects are present in 8-17% of simulations depending on whether ancestry controls are included and whether the data are demeaned, the sibling SD method detects mean effects when only variance effects are present at rates of between 4.2% and 5.6% of cases depending on the specification. In addition, the method correctly detects variance effects when these effects are present in 80.8% (no parent genotype control), with the power dropping to 44.1% (parent genotype control) at the simulation sample size of 1000 sibling pairs.

### Empirical application to height and BMI

**Sibling SD results: Height and BMI.** The simulation results show that in the absence of an unobserved confounder, the sibling SD method performs equally well as existing



**Fig 3. Manhattan plot of sibling variation in height among FHS 3rd generation sibling pairs.** Results for the pairwise sibling standard deviation in height regressed against the sibling-pair minor count of alleles with controls for sex of sibship, mean age of siblings, age difference of siblings, sibling mean height, parental genotype.

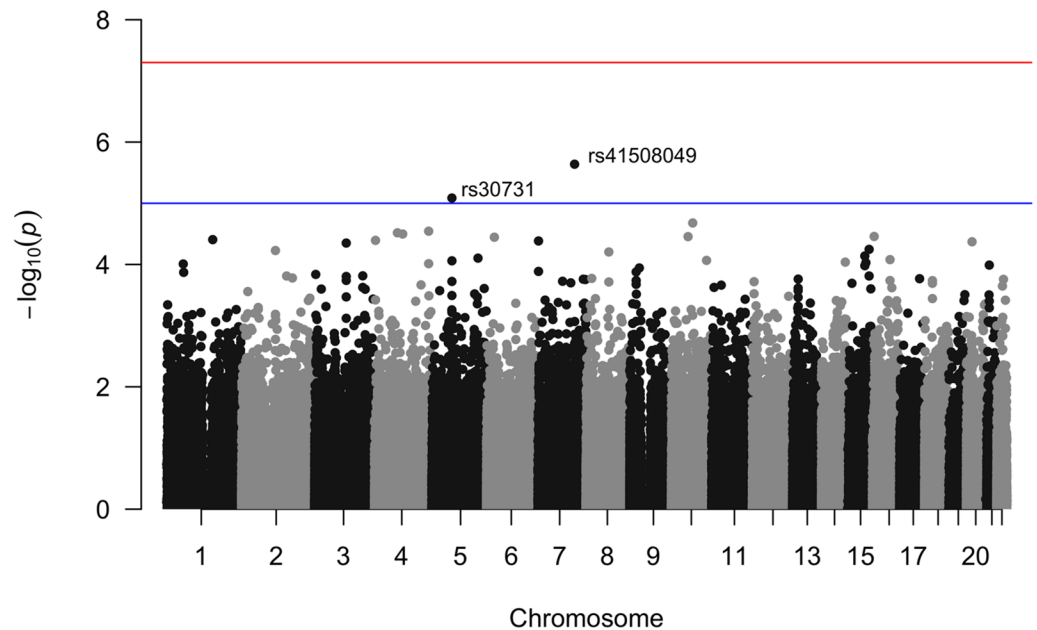
<https://doi.org/10.1371/journal.pone.0194541.g003>

approaches to variance detection (squared Z-score; DGLM); in the presence of a small degree of unobserved confounding ( $\rho = 0.1$  between a family indicator and offspring genotype), the sibling SD method performs significantly better than these two approaches in correctly failing to detect variance effects when only mean effects are present. We now apply this method to two phenotypes: height and BMI. When we use data from quartets in the FHS and control for mean sibling-pair height, sex, mean pair age, within-pair age difference, and parental genotype in genome-wide regressions on the standard deviation of the sibling pair height we find four SNPs that are genome-wide suggestively significant ( $p < 10^{-5}$ ) and meet other Hardy-Weinberg equilibrium (HWE) and minor-allele frequency (MAF) controls (see [Methods](#)):

rs2804263 (MAF 30.8%); rs2073302 (MAF 39.1%); rs8126205 (MAF 37.1%) and rs4834078 (MAF 24.0%) ([Fig 3](#)). For BMI, there were two SNPs that meet genome-wide suggestive significance: rs30731 (MAF 48.7%) and rs41508049 (MAF 10.3%) ([Fig 4](#), with a regional linkage map of the SNP that replicates in [Fig B](#) in [S1 File](#)). Our method, as expected, controls well for population structure: QQ plots for the height and BMI p-values do not show the telltale “early lift-off” typical of failure to control this confounder ([Fig C](#) in [S1 File](#)).

The absence of genome-wide significant SNPs but presence of genome-wide suggestive SNPs may indicate that the method’s statistical power is low. A power analysis for the sample size utilized in the discovery dataset indeed supports this suggestion. For example, a single SNP would need to explain over 4.8% of the variation in the trait for the study to achieve 80% statistical power at the FHS sample size and a significance threshold of  $p < 10^{-5}$  ([Fig D](#) in [S1 File](#)). An effect size of this magnitude is not expected for human complex traits (e.g., the largest effect of a single SNP for mean height explains approximately 1% of the variation). Power is very low near  $R^2 = 0.01$  ([Fig D](#) in [S1 File](#)) in this particular sample. However, the figure shows how newly-released samples are adequately powered to detect the effects and highlight the method’s potential utility for better-powered studies.

Apparent effects on variance can be generated if effects of alleles at a locus are not additive (i.e., there is dominance) or if means and variances are correlated. Addressing the first issue,

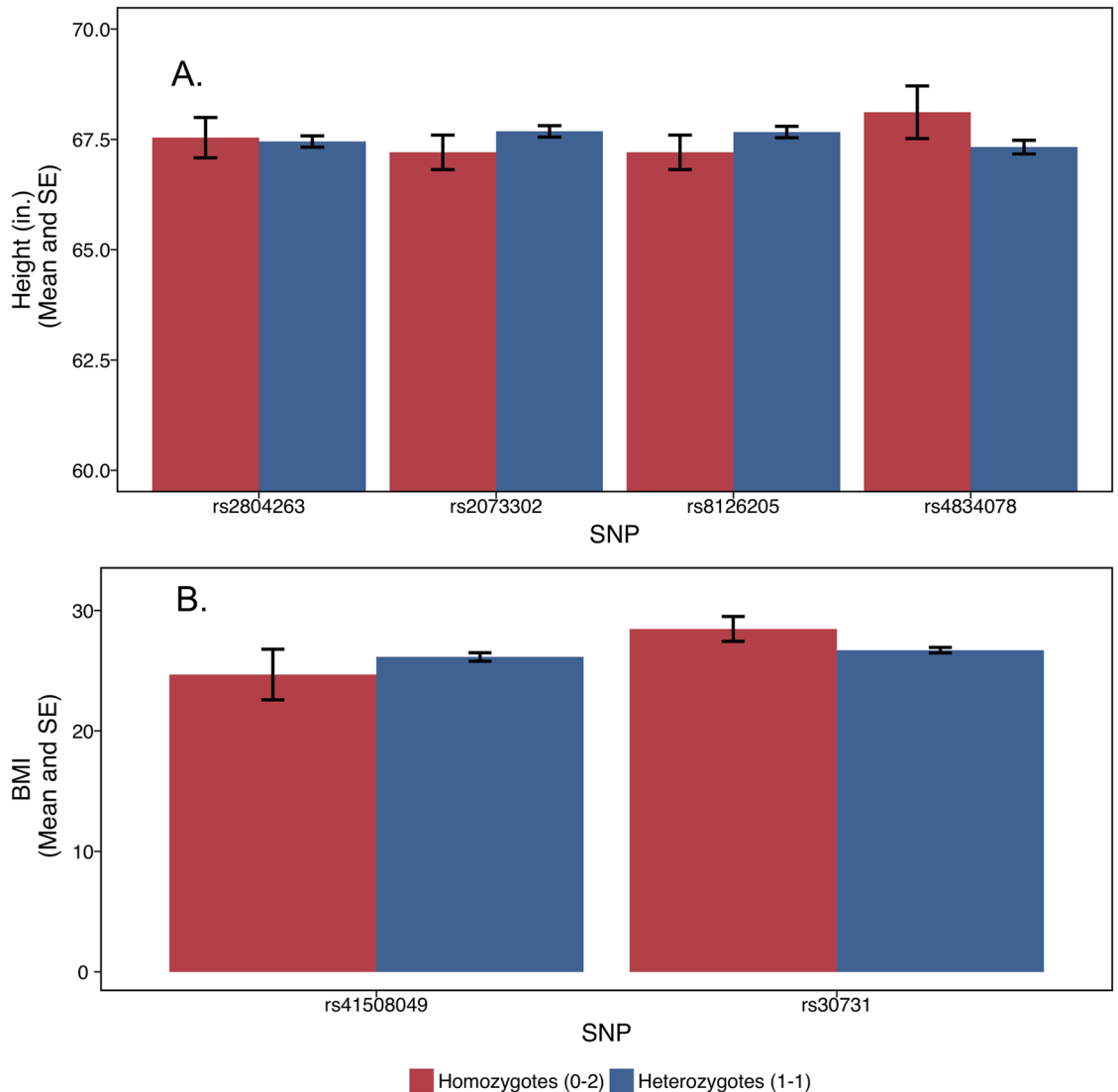


**Fig 4. Manhattan plot of sibling variation in BMI among FHS 3rd generation sibling pairs.** Results for the pairwise sibling standard deviation in BMI regressed against the sibling minor allele count with controls for sex of sibship, mean age of siblings, age difference of siblings, sibling mean BMI, and parental genotype.

<https://doi.org/10.1371/journal.pone.0194541.g004>

our approach does not control for non-additivity of allele effects at a locus, as it assumes a linear model. However, it does allow a test of whether an effect on variation net of mean was actually an artifact of non-linear effects on average rather than an actual variance effect. If the true relationship between phenotype and a sibling's minor allele dosage were non-linear (i.e. revealed dominance effects) our initial findings could be entirely driven by divergence among those sibling pairs with two minor alleles. For example, if an individual with two minor alleles were significantly taller than an individual with either one or zero minor alleles (recessive effect) then when we collapsed the sibling pairs with two minor alleles, we could generate artifactual variation effects because among those sibships with two minor alleles, some would be distributed 0-2 (and thus one sibling would be taller than the other) while other sibships would be 1-1 (and thus would be the same height). Put together, it would appear that two minor alleles increased the variation net of mean effects. And if strong enough, such a misspecified effect could exert enough leverage to make a linear effect on variation appear across all allele numbers (zero to four for the sibship). These concerns appear not to apply to our analysis. A two-sample t-test of equality of means that compares homozygotes and heterozygotes for each significant SNP on the respective trait finds no differences in levels for either trait at the  $p < 0.05$ . Fig 5, which presents the mean and standard errors, shows the lack of significant differences.

For the second issue, we can investigate correlated mean and variance effects in the present data. Others have shown this correlation depends on the minor allele frequency of the SNP and the effect size (the effect of the SNP on  $z^2$  is  $(1 - 2p)a^2$ , where  $p$  represents the minor allele frequency and  $a$  represents the additive effect [40]). To test these in an independent sample, the top SNP for variance in BMI found by Yang et al. ([40]) shows clear association with mean BMI in the same data set from the GIANT consortium (Fig E in S1 File), whereas the four suggestively significant SNPs in our analysis do not show an association with mean height in the GIANT consortium data (Fig F in S1 File). Further highlighting the importance of our



**Fig 5. Test for spurious association with variance due to non-linear effects on mean levels.** Mean and standard error for height (inches) and BMI among with two minor alleles is shown separately for homozygotes (one sibling with zero minor alleles and the other sibling with two) and heterozygotes (each sibling has one minor allele), for each genome-wide suggestively significant SNP for the respective trait (A. height; B. BMI). One significant SNP for height (rs8029740) is not depicted because there is only one sibling pair with the 1-1 allele combination and 0 sibling pairs with the 0-2 combination. A two-sample t-test for equality of means, estimated separately for each SNP, revealed no significant differences between the two groups for the top hits for each trait.

<https://doi.org/10.1371/journal.pone.0194541.g005>

regression-based control for sibling-pair mean are the observations that: 1) sibling-pair standard deviation shows a strong positive correlation with sibling-pair mean, and 2) this correlation is not eliminated by using the sibling-pair coefficient of variation rather than the standard deviation (Fig G in S1 File).

For replication analysis, we used respondents from the MTFs. (For list of proxy MTFs SNPs with information on MAF and linkage with FHS SNPs see Table G in S1 File) These families included phenotypic and genotypic information on pairs of twins as well as their parents, allowing us to replicate the sibling-based analysis with parental genetic controls so as to mimic random assignment of alleles. The MTFs has both dizygotic (DZ) and monozygotic (MZ) twins. Because MZ sibships do not vary in terms of cryptic genetic variation and may



experience much more similar environments to each other than do genetically distinct siblings, we also repeated our replication analysis only with DZ twin sets but found that exclusion of MZ twins did not affect results. Another concern is that twins (even DZ twins) may experience more similar environments than singleton siblings; thus, our replication analysis may suffer from attenuation bias to the extent that the cause of variation is environmental and not cryptic genetic differences (which should, by contrast, be equivalent for singleton full siblings and DZ twins). Among the SNPs that were genome-wide suggestively significant for the height analysis, only two of the four had viable proxy SNPs in the MTF5 dataset after quality control: rs2804263 and rs4834078 both had proxies whereas rs2073302 and rs8126205 did not (Table G in [S1 File](#)). When we ran the analysis for the proxy SNPs in the MTF5 dataset, none achieved statistical significance. Among the SNPs that were genome-wide suggestively significant for the BMI analysis, both SNPs had viable proxy SNPs in the MTF5 dataset after quality control (Table G in [S1 File](#)). When we ran the analysis for the proxy SNPs in the MTF5 dataset, one SNP achieved statistical significance: rs30731 (proxy in MTF5: rs28636).

In addition to checking whether the significant SNPs from the FHS analysis replicate in the MTF5 sample, we also 1) took the top hits from the FHS sample, 2) found proxy SNPs in Yang et al., and 3) investigated the p-values for these top hits with the squared Z-score in their analysis. None of the top hits from our study significantly predicted the squared Z-score in their study. There could be multiple reasons for this lack of overlap. First, the Yang et al. hits could be false positives due to failing to control for unobserved confounding. Alternately, the present analysis may have had false negatives that miss top hits from Yang et al. due to the lack of statistical power. We cannot distinguish between these two observations in the present analysis.

**Investigating rs30731/rs28636 for within-sibling variation in BMI.** rs30731/rs28636, a SNP that significantly affects within-sibling variation in BMI and that replicated in the MTF5 sample, is located on the MAST4 gene, which encodes a member of the microtubule-associated serine/threonine protein kinases. The proteins in this family contain a domain that gives the kinase the ability to determine its own scaffold to control the effects of their kinase activities.

GWAS studies have uncovered several significant associations between other SNPs on this gene and traits ranging from BMI to autism/PDD-NOS (Table H in [S1 File](#)). One exception is a GWAS of childhood obesity conducted by the Early Growth Genetics Consortium ([70]). The study found a genome-wide suggestively significant hit for rs28636 in the discovery sample that did not replicate. This non-replication could be due to the SNP being a variance-affecting locus that might show up in estimation of mean effects.

**Pathway and gene set analyses for all significant SNPs.** In addition to investigating the gene function for the replicated SNP, we also performed two analyses that pool SNPs: gene-based and pathway-based pooling (see [Methods](#)). For the gene analysis, we found using PASCAL that the gene on which the significant SNP for BMI variability was located (MAST4) was significantly enriched ( $p = 0.0015$  in the FHS data;  $p = 0.1$  in the replication data). Table I in [S1 File](#) shows other significant gene sets identified using PASCAL that replicated at various p-value thresholds in the MTF5 data. VEGAS1 yielded no further significant gene sets.

i-GSEA4GWAS and PASCAL for pathway analysis each yielded some pathways that appeared to be significantly enriched ([Table 2](#)). One of these pathways, associated with within-sibling variance in height in the FHS data, replicated in the MTF5 data: HSA04540 Gap Junction ( $p = 0.002$  for each data set)(Fig H in [S1 File](#)). HSA04540 includes members of several signaling pathways, including growth factors and their receptors, although any connection between these factors and organismal growth, as manifested in ultimate height, remains to be determined. Importantly, this pathway was not significant in the GWA for mean levels effects in either dataset. Table J in [S1 File](#) shows replicated pathways for BMI and height variability estimated using PASCAL. Important to note, however, is that while HSA-04540 was significant

**Table 2. Enriched canonical pathways for height and BMI sibling-pair standard deviations in FHS, estimated using i-GSEA4GWAS.**

Pathway/Gene set name	P-value	FDR	Sign. Genes/Slctd. Genes/All Genes
<b>HEIGHT</b>			
GLYCEROLIPID METABOLISM	< 0.001	6.0000000000000001E-3	20/35/45
HSA00561 GLYCEROLIPID METABOLISM	< 0.001	8.9999999999999993E-3	27/49/58
VEGFPATHWAY	< 0.001	1.5666665999999999E-2	47046
HSA03030 DNA POLYMERASE	1E-3	2.76E-2	45920
BILE ACID BIOSYNTHESIS	3.0000000000000001E-3	3.175E-2	46712
HSA04540 GAP JUNCTIONM*	2E-3	4.1666667999999997E-2	42/78/98
HSA00564 GLYCEROPHOSPHOLIPID METABOLISM	5.0000000000000001E-3	0.24111110999999999	20/49/68
<b>BMI</b>			
HSA00591 LINOLEIC ACID METABOLISM	< 0.001	6.0000000000000001E-3	11652

\*Replicates in MTFs data

<https://doi.org/10.1371/journal.pone.0194541.t002>

across datasets using the i-GSEA4GWAS tool, the PASCAL tool does not test for this pathway so we could not check its robustness across multiple pathway analysis softwares.

## Discussion

Our analysis extends earlier work that aimed to map variance-controlling loci in humans ([40]). Although the prior work enjoyed greater statistical power, it also had more potential for bias—due both to environmental confounding and to conflation of mean and variance effects. Indeed, Yang et al. identified a locus regulating BMI variability that is also strongly associated with mean levels and for which a gene-by-environment interaction effect on mean has been shown. In the present study, we first show via simulation that two widely-used methods for detecting variance-affecting loci—the squared Z-score method and DGLM—fail to adequately distinguish between a locus affecting the trait mean and a locus affecting the trait variance in the presence of unobserved confounding. The sibling SD method, by controlling for the sibling mean of a trait and identifying off of random, between-sibling variation in allele counts, does distinguish between these effects. Applying the method to data, we were able to perform within-family analysis on two samples of white Americans, completely free of population stratification, largely devoid of rGE confounding, and with controls for mean level effects as well as checks for non-linear (i.e. dominance) effects on mean levels. One SNP for BMI variability, located on the MAST4 gene, replicated. Notably, the SNP, unlike many on FTO, has not been found to affect mean levels of BMI. Like the latest methods to map variance-controlling loci in controlled crosses ([59]), our approach therefore avoids common confounds. At the same time it overcomes problems specifically associated with human traits, including the construction of variance-affecting environments, that existing regression-based methods for detecting vQTLs fail to address because they allow controls solely for observed confounders (e.g., [51, 58, 59, 71]).

Though underpowered in the FHS and MTFs sample sizes used in the present analysis (Fig D in S1 File), our results strongly support the benefits of approximating a randomized genetic experiment by analyzing within-family variation while controlling for parental genotype. Such an analysis addresses the possibility that it is merely cross-family environments interacting with a mean effect and/or population structure that produce apparent association with variability. Meanwhile, parameterizing the estimand as spread (SD) net of sibship mean levels provides a robust, flexible way to conceive of variation—that is, rather than parameterizing the relationship between mean and variance a priori by using the coefficient of variation or some

similar summary statistic. The trade-off inherent to our approach is that environmental and phenotypic variation within sibships may be attenuated, reducing statistical leverage; the extent of such a dynamic is wholly dependent on phenotype, of course. Although the within and between family components of the variation in the phenotype can be measured to determine whether or not the phenotype is suitable for such an approach, the extent of variation within and between sibships in the unmeasured environmental factors that matter is, of course, unknown.

Moving forward, we see three applications of the method: (1) combining the method with twin studies to better distinguish between  $G \times G$  versus  $G \times E$  effects that each contribute to trait variability; (2) examining the heritability of plasticity in a trait as a supplement to examining the heritability of levels of a trait; and (3) generating weights for polygenic scores to predict trait variance. We discuss each in turn.

### Variability across MZ versus DZ twins

First, we can combine the present within-family approach to measuring phenotypic variability with the classic twin comparison approach of behavior genetics, we obtain a method to distinguish between  $G \times E$  and  $G \times G$  interaction effects that may be revealed in a vGWAS for variation regulating loci. Namely, if a particular allele produces more variability among dizygotic twins than among monozygotic ones, we can infer that the difference between those allelic effects is attributable to two forces: 1. Putatively greater environmental differences within DZ twin pairs than within MZ twin-ships; and/or 2. The greater (cryptic) genetic variation within DZ pairs as compared to their MZ counterparts. Since prior work ([72]) shows that the equal environments assumption seems to hold for a wide range of outcomes, thus weakening support for 1 as the explanation, we can attribute the bulk of the difference in trait variation between DZ versus MZ twins to the theory that the allele in question is not only buffering the environment but also serving as a phenotypic capacitor (i.e. repressing cryptic genetic variation).

### Estimating the heritability of plasticity

Another way to combine the present approach with classical statistical genetic techniques is to supplement estimates of the heritability of *levels* of a trait with estimates of the heritability of *plasticity* in a trait. For instance, GREML involves partitioning the observed phenotypic distance in a trait between individuals into the sum of genetic and environmental contributors to this distance. If we switch from individuals as the unit of analysis when measuring this distance to sibling pairs as the unit of analysis, we can calculate the phenotypic distance between siblings in each pair and the average genotype at each locus across the two siblings. Then, we can place unrelated sibling pairs in a Genetic Relatedness Matrix and contrast the genetic distance between sibling pairs to the amount of variability for the phenotype the pairs display to recover estimates of the heritability of this variability.

### Constructing vPGS

Currently, researchers develop and use polygenic scores (PGS) that predict mean levels of a trait. We can extend the polygenic score approach to develop scores that predict variance in a trait (vPGS). Coefficients for a vPGS construction in a prediction sample would be obtained from a vGWAS done within families with sibling sets in the discovery sample to obtain estimates that better distinguish between mean and variance effects, but the application of coefficients could then be to the individual person. Fig D in [S1 File](#) shows that the method is adequately powered to detect effects of SNPs that explain fewer than 1% of the variation in traits in samples like the UK Biobank that could be used at this discovery stage.

Having a polygenic risk score that predicts particular forms of phenotypic variability may be helpful for researchers hoping for non-null results with respect to a given phenotypic measure who are therefore looking to recruit sensitive subjects for experimentation that involves specific environmental exposures. Per the earlier discussion, if calculated from pairs of MZ twins, such a polygenic score would capture only environmental sensitivity. But if a vGWAS of MZ twins and DZ twins were conducted, the results could be differentiated to provide a measure of phenotypic capacitance—i.e. regulation of internal, genetic variation.

Beyond telling experimenters which subjects may be more genetically sensitive, such a phenotypic capacitance score may have important predictive power in terms of disease. Namely, cancer, autoimmune diseases, metabolic syndrome and other irregularities of cell or system stability may themselves be predicted by a genetic architecture that is less robust. Thus, a genetic screening for a tendency toward developmental plasticity (i.e. if the plasticity score was calculated on developmental indicators such as height) may be diagnostic. If applied to behavioral phenotypes, such a score could be predictive of mental disorders that reflect a lack of canalization of mind, so to speak, such as schizophrenia.

In light of this discussion, we think that there is benefit to combining prior, pedigree-based approaches with newer GWAS methods to better estimate variance effects (as well as levels effects). Thus, we recommend that consortia of cohorts with genome-wide data on sibling pairs at the minimum, quartets ideally, be formed to advance GWA to a more solid foundation of inference that approximates the unbiased estimates of lab-based genetic manipulations by taking advantage of random differences in sibling genotypes.

## Materials and methods

### Simulation study

**Generating genotype and trait data.** The simulation proceeds in four steps. First, we generate genotypes for parents and offspring. Second, we generate an unobserved family-level confounder that is correlated to varying degrees with the observed sibling genotype. Third, we use the genotype to generate four traits:

1. Trait with neither mean nor variance effects
2. Trait with mean effects but not variance effects
3. Trait with variance effects but not mean effects
4. Trait with both mean and variance effects

In this third step, we generate four versions of each of the four traits: 1) a version that is not affected by between-family confounding; and three versions affected by varying degrees of between-family confounding. Fourth, we explore how the process in steps one through three generates between-family confounding that is not fully addressed by controls for the subpopulation that generates the genotype. Finally, we compare the performance of the sibling standard deviation approach to other methods. We describe the first four steps in the present section, and summarize the results of step five in the results.

All steps were repeated for 1000 replicates of size 2000 (1000 sibling pairs with 2 offspring each, generated from 4 ancestral sub-populations).

**Step one: Generating genotypes.** We use the following process to generate genotypes for parents and offspring, repeated separately for each of the 1000 replicates:

1. *Generate parent genotypes*: we use the function `simMD` taken from the source code for R's `popgen` package to generate parent genotypes, and use the following parameters in the present simulation:
  - $N = 2000$  parents,  $N = 1000$  families with 2 offspring per family
  - 4 subpopulations, with  $c = 0.01$  representing the extent to which each subpopulation differs in allele frequencies of SNPs from typical values
  - 1 causal snp. Traits with no effects were thus generated by setting the coefficient on the parameter that governs the relationship between 1) the allele on trait mean, and 2) the allele on trait variance, to zero.
  - Allele frequency ( $p$ ) of each SNP randomly drawn from a uniform distribution with bounds at  $[0.1, 0.9]$
2. *Generate offspring genotypes*: we use the parent genotypes generated in step 1 to generate offspring genotypes assuming random mating and segregation
3. Step one and step two result in parent and offspring genotypes we use in steps two and three of the simulation

**Step two: Generating a family-level confounder correlated with genotype.** Consider the following model for the relationship between a SNP and phenotype for an individual  $i$ , nested in family  $j$ , for snp  $k$ . For now, we just consider identifying the causal effect of an allele on the mean of  $Y$ :

$$y_{ijk} = \alpha + \beta_k X_{ijk} + \epsilon_{ijk}$$

Linear regressions such as those run in GWAS, which often restrict the sample to unrelated individuals, ignore the grouping structure of the family and estimate the following model:

$$y_{ik} = \alpha + \beta_k X_{ik} + \epsilon_{ik}$$

Random effects models acknowledge this grouping structure by positing that each family has its own intercept that shifts the outcome up or down. The random effects models then estimate these  $\alpha_j$  using a distribution that pulls some of the family-specific intercepts ( $\hat{\alpha}_j$ ) towards the mean intercept across families ( $\mu_\alpha$ ) depending on  $\sigma_\alpha$  ([73]):

$$y_{ijk} = \alpha_j + \beta_k X_{ijk} + \epsilon_{ijk} \quad \epsilon \sim N(\mu_\alpha, \sigma_\alpha^2)$$

Random effects methods generate unbiased estimates for  $\beta_k$  (the causal effect of an allele for snp  $k$  on a trait) when there is no unobserved confounding at the family level that is correlated with genotype of other observed covariates ( $cor(X_{ijk}, \alpha_j) = 0$ ) ([67]). However, these methods generate biased estimates when  $cor(X_{ijk}, \alpha_j) \neq 0$ . To generate simulations to test this bias in the genetics context, we use the following process to generate a family intercept that is correlated with offspring genotype, and through the process described in step three, is also correlated with the trait:

1. Operationalize genotype as the sibling pairs' summed minor allele count at that locus
2. Choose a  $\rho$  parameter for the degree of correlation, and construct a variance-covariance matrix  $Q$  representing the correlation between the genotype and family-level intercept. E.g.,

if  $\rho = 0.1$ :

$$\begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$$

3. Take the cholesky decomposition of Q
4. Generate a starting value for the family-level intercept that is not correlated with genotype ( $\alpha_j \sim N(0, 1)$ ); this starting value provides a baseline set of intercepts uncorrelated with genotype that we will then rescale to be correlated with genotype to varying degrees.
5. Multiply cholesky decomposition by matrix containing the genotype and the non-correlated intercept to generate  $\alpha_j$  correlated with the sibling pair's genotype
6. Repeat the previous steps for four values of  $\rho$  representing different degrees of confounding:
  - a.  $\rho = 0$  (no confounding)
  - b.  $\rho = 0.01$
  - c.  $\rho = 0.05$
  - d.  $\rho = 0.1$

**Step three: Generating traits.** To generate traits, we used a similar process for simulating traits as used in [58]. We generated four general types of traits (traits with neither mean nor variance effects; traits with mean effects only; traits with variance effects only; traits with both mean and variance effects) using the following general setup, and varying the  $\gamma$  and  $\epsilon$  parameters, where  $i$  represents an individual and  $k$  indexes a SNP. Sex was simulated from a binomial distribution with  $p = 0.5$ . Age was simulated from a normal distribution  $\sim N(\mu = 50, sd = 10)$ .  $G_1$  indicates heterozygotes, while  $G_2$  indicates minor allele homozygotes. Across all simulations with mean or variance effects, an additional minor allele results in *increases* in the mean or in the variance:

$$y_{ik} = G_{1ik}\gamma_1 + G_{2ik}\gamma_2 + 0.5 \times sex_i + 0.05 \times age_i + \epsilon_{ik} \quad (2)$$

1. Neither mean nor variance effects:
  - $\gamma_1 = \gamma_2 = 0$
  - $\epsilon \sim N(0, 1)$
2. Mean effects only
  - $\gamma_1 = 0.15; \gamma_2 = 0.35$
  - $\epsilon \sim N(0, 1)$
3. Variance effects only
  - $\gamma_1 = \gamma_2 = 0$

- $\epsilon \sim N(0, 1)$  for major allele homozygotes;  $\epsilon \sim N(0, 1.15^2)$  for heterozygotes;  $\epsilon \sim N(0, 1.4^2)$  for minor allele homozygotes

4. Mean and variance effects

- $\gamma_1 = 0.15; \gamma_2 = 0.35$
- $\epsilon \sim N(0, 1)$  for major allele homozygotes;  $\epsilon \sim N(0, 1.15^2)$  for heterozygotes;  $\epsilon \sim N(0, 1.4^2)$  for minor allele homozygotes

For the simulation to address the possibility of confounding by unobserved, between-family factors, we also modify Eq 2 to include the family-level intercept that is correlated with observed genotype at varying levels (no correlation, medium correlation, high correlation), and refer to the latter two outcomes as “confounded outcomes”, where  $i$  refers to an individual and  $j$  indexes the family that the sibling pair was generated from:

$$y_{ijk} = G_{1ijk}\gamma_1 + G_{2ijk}\gamma_2 + 0.5 \times sex_{ij} + 0.05 \times age_{ij} + \alpha_j + \epsilon_{ijk} \tag{3}$$

**Step four: Controlling for ancestry.** To control for ancestral background in the methods that follow, we control for an indicator for which of the four subpopulations generated the individual/pair’s genotype. We used an indicator for ancestry rather than ancestry derived from genotype because, and following the general simulation method used in [58] where traits were generated using one SNP because each SNP-trait association is estimated separately, the genotype has only SNP.

As the strength of between-family confounding increases, the correlation between these indicators for population stratification and the family-level intercept increases. To illustrate this increase, we run the following regression for each of the four degrees of correlation between family-level intercepts and observed genotypes:

$$\text{family intercept}_i = \alpha + \beta_1 \times \text{subpop}_i + \epsilon_i \tag{4}$$

We then calculated the percent of  $\beta_1 \neq 0$  with  $p < 0.05$  across the 1000 replicates with that simulated level of correlation. Table K in S1 File summarizes the results. It shows that as we move from zero confounding to some confounding, the relationship between the ancestral indicator and the family-level intercept becomes more significant. However, the results also show, and as we show in the results that follow, while this correlation between the population stratification indicators and family-level confounding *reduces* bias in estimates that this confounding causes, the control does not fully eliminate bias.

Eq 4 looks at the strength of relationship between ancestral background and the family intercept at various degrees of confounding, and shows how this strength increases as the degree of confounding increases. The reason this confounding biases estimates of the effect of the minor allele count on a trait is because this family intercept is also correlated with genotype. To show this, we run the following regressions at each level of confounding:

$$\text{sibling minor allele count}_i = \alpha + \beta_1 \times \text{family intercept}_i + \epsilon_i \tag{5}$$

Table L in S1 File presents the mean  $\beta_1$  across each level of confounding, and shows that as confounding increases, the relationship between family intercept and the minor allele count increases. In the presence of zero between-family confounding, the slightly positive  $\beta_1$  is caused by the correlation between parent and offspring genotypes that generates correlated genotypes between siblings of the same parents.



### Power

For the trait simulated to have variance effects, the average effect size was  $R^2 = 0.01$ . With the  $N = 1000$  sibling pairs sample sized used in the analysis, the power to detect these effects is 81%, results we see confirmed in Table F in [S1 File](#), which shows roughly 80% power to detect variance effects when these are present.

### Analysis

All analysis was performed using R. The following packages were used to estimate the models (see [Results](#))M with code that names all the arguments/parameters from the functions used in each package posted at the following link: <https://scholar.princeton.edu/dconley/princeton-biosociology-lab>

- Power analysis: `pwr` packages with `pwr.f2.test` command
- Pooled regressions: `lm` with default settings
- Random effects model: `plm` with `model = "random"`
- Fixed effects model: `plm` with `effects = "within"` and index of the family identifier
- DGLM: `dglm` with a gaussian-family link function and REML as estimation method
- Squared Z score method: after estimating the Z-score separately by sex, `lm` with default settings
- Sibling SD method: after estimating the sibling standard deviation, `lm` with default settings

### Empirical application

**Data.** Data for discovery analysis come from the Framingham Heart Study (FHS), second (parental) and third (sibling) generation respondents. (This dataset is publicly available through dbGaP <http://www.ncbi.nlm.nih.gov/gap>. QC code can be obtained from the FHS investigators ([74]) The FHS is, in fact, one of the cohorts included in the GIANT meta-analysis performed by Yang et al. ([40]). Height and weight were taken from clinical measurements and then BMI was calculated as (weight in kilograms)<sup>2</sup> / height in meters. Genotypes were assayed using the Affymetrix GeneChip Human Mapping 500K Array and the 50K Human Gene Focused Panel. Genotypes were determined using the BRLMM algorithm. Our analysis began with the original 500,568 SNPs, and resulted in 260,469 SNPs available for analysis after cleaning (e.g., HWE screens and a MAF cut-off of 0.05). The screens were conducted using all available individuals with genetic data, not only those that were included in this analysis. Genome-region association plots were produced using SNAP ([75]), except for those of published GIANT consortium data, which were produced using LocusZoom ([76]). Regional linkage maps were produced using SNAP ([75]) and data from the 1,000 Genomes CEU Panel ([77]).

Among third-generation respondents, the numbers in our sample by sibship size are presented in [Table 3](#). The 200 families with only one sibling in the data drop from the sibling

**Table 3. Distribution of third generation siblings included in data by sibship size.**

# 3rd-G Sibs in Family:	1	2	3	4	5	6	7	8	9
N (pairs):	–	292	483	504	250	150	189	28	36
Families:	200	292	161	84	25	10	9	1	1

Total actual N of sibling pairs after random selection of one per family: 583

<https://doi.org/10.1371/journal.pone.0194541.t003>

analysis. Those with more than two contribute multiple pairs to the data; however, our final analysis selects only one pair per second-generation family as the more complicated error structure with multiple pairs leads to early takeoff on QQ plots. The siblings are genetically related but are not DZ or MZ twins.

The Minnesota Twin Family Study (MTFS) replication data were genotyped on the Illumina 660W Quad array ([78]) and phenotypes can be found elsewhere ([79]) Quality control procedures were applied separately to each individual cohort. Individuals with a call rate  $< 0.95$  ( $N = 22$ ), estimated inbreeding coefficient  $> 0.15$  ( $N = 2$ ), or showing evidence of non-European descent from multidimensional scaling ( $N = 298$ , mainly individuals with Mexican ancestry) were removed. Individuals were considered outlying from European descent if one or more of the first four eigenvectors were more than three standard deviations removed from the mean. SNPs with MAF  $< 0.01$ , call rate  $< 0.95$  or HWE-test  $p$ -value  $< 0.001$  were removed. The reason for the lower MAF threshold in the MTFS sample rather than FHS sample is that for the latter data, internal QC procedures remove SNPs at this threshold prior to releasing the data to researchers. For our analysis, we included both the MZ and DZ twin pairs because restricting to DZ twin pairs that more closely approximate the sibships in the FHS discovery sample does not change the substantive findings. Because the discovery sample and the replication sample were genotyped on different arrays, we deployed SNAP to find corresponding SNPs ([75]). The resulting sample size for our analysis was 1,048 pairs.

The Boston University Medical Center IRB approved the use of the discovery analysis data (Framingham Heart Study). The University of Minnesota IRB approved the use of the replication analysis data (Minnesota Twin Family Study). The Princeton University IRB approved the secondary analysis of each data (Protocol 8322).

In addition to the GWA analysis in the discovery and replication sample, we performed two sets of analyses that pooled SNPs across multiple software implementations. First, we investigated gene-based pooling using VEGAS1 ([80]) with the following parameters: CEU subpopulation specified, no assumed allele frequency difference by sex, and using all SNPs (not just best hits) within the gene region itself ( $\pm 0$  KB). We estimated pathway-based pooling using i-GSEA4GWAS ([81]). Because estimation of significant pathways and gene sets can be sensitive to different algorithms' methods of computing  $p$ -values and controlling type I error rates, we test the robustness of these findings with PASCAL ([82]). These analyses were performed for both BMI and height in both the discovery and replication sample. For the discovery sample, gene-based and pathway-based analyses were performed using 260,434 variants input; 239,526 variants used; 14,783 genes mapped; 221 gene sets selected. For the replication sample, gene-based and pathway-based analyses were performed using 522,726 variants input; 487,692 variants used; 16,840 genes mapped; 259 gene sets selected.

**Statistical analysis.** All analysis was performed using R. The power analysis depicted in Fig D in S1 File to estimate the power of the approach at varying putative effect sizes was performed using the *pwr* package in R [83], varying  $n$  to reflect the present sample sizes and recently-released samples with larger sibling populations and showing power separately for two  $p$  value thresholds ( $p < 10^{-5}$  for discovery analyses;  $p < 0.05$  for replication). More precisely, we used the following procedure:

1. Vary the putative  $R^2$  explained by a SNP from 0.0 to 0.01, incrementing by 0.0001
2. Translate  $R^2$  into effect size ( $\frac{R^2}{1-R^2}$ )
3. Since we are analyzing power to estimate the  $\beta$  in a linear regression, using the `pwr.f2.test` command to estimate power at varying effect sizes

Heteroscedasticity-robust standard errors should not substantially affect power under the present sample size ([84]).

For the main analysis, sibling-pair standard deviations (SD) were fit by linear regression using the `lm` command with default options to the following model, where the key regressor is the number of minor alleles for the pair of siblings at a given locus. Because this number is for two individuals, the range is 0 to 4:

$$SD_j = \alpha + \gamma_k \times \text{sibling minor allele dosage}_{jk} + \delta \times \text{parent minor allele dosage}_{jk} + \gamma Z_j + \epsilon_{jk}$$

$j$  indexes a sibling pair,  $k$  indexes a SNP, minor allele dosage is the total number of minor alleles in a sibling pair, parent minor allele dosage is the total number of minor alleles across the parents,  $Z_j$  is a vector of sibling pair-level controls that includes controls for the mean level of the trait in the sibling pair, pair sex (MM or FM or FF), mean pair age, and the within-pair age difference, and  $\epsilon_{jk}$  is the residual for sibling pair  $j$  at snp  $k$ . As we show in the simulation study, the method returns similar results if parental genotype is or is not included. And if we do not include parental genotype, we can include an additional term—the difference in minor allele counts between siblings—that helps increase the precision of the estimate of  $\gamma_k$ . Qualitative results do not change if we instead specify the mother’s and father’s genotypes separately. In the model,  $SD_j$  is the standard deviation of a trait within a sibling pair, calculated as follows, where  $i$  indexes an individual sibling in the pair,  $j$  indexes the pair,  $x$  refers to the trait, and  $\bar{x}$  refers to the mean of the trait across the two siblings:

$$SD_j = \sqrt{\frac{\sum_{i=1}^2 (x_{ij} - \bar{x}_j)^2}{2 - 1}} = \sqrt{\sum_{i=1}^2 (x_{ij} - \bar{x}_j)^2}$$

Huber-White standard errors robust to clustering on pedigree ID (to account for correlated errors among cousins: sibling pairs that share the same grandparents but not the same parents) were calculated for the FHS analysis in the following way, where  $i =$  sibling pair 1,  $j =$  sibling pair 2,  $n =$  total sibling pairs,  $g =$  pedigree grouping, and  $k =$  SNP.  $\sigma$  refers to the residual variance across the sample, and  $\sigma_i$  refers to the residual variance for an individual.

For the simple case where  $\sigma^2 = \sigma_i^2 \forall i$  and  $SST_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ :

$$var(\hat{\beta}_k) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{SST_x^2}$$

To account for errors that may be heteroskedastic and correlated within a shared pedigree, we adjust the variance to be robust to cases where  $\sigma^2 \neq \sigma_i^2$  and where for  $g = g'$  (two sibling pairs share same grandparent/pedigree ID).  $\hat{u}_{ig}$  represents the observed residual for participant  $i$  belonging to pedigree  $g$ . For the case of pedigree ID’s with two or more sibling pairs, this becomes the following variance robust to heteroskedastic and correlated errors, with an indicator function for when the sibling pairs belong to the same pedigree:

$$var(\hat{\beta}_k) = \frac{\sum_{j=1}^n \sum_{i=1}^n (x_i - \bar{x})(x_j - \bar{x}) \hat{u}_{ig} \hat{u}_{jg} \mathbf{1}[g = g']}{SST_x^2}$$

For the case of pedigree ID's with one sibling pair only, the above equation reduces to the following variance robust to heteroskedastic errors ([85]):

$$\text{var}(\hat{\beta}_k) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_{ig}^2}{SST_x^2}$$

## Supporting information

**S1 File. Fig A, Estimated coefficients on SNPs for simulated dependent variable with no effects and confounding between a family-level indicator, genotype, and outcome.** The red dashed line represents the true snp level effect ( $\beta = 0$ ), while the density curves show the range of estimated  $\hat{\beta}$  for each of the models. We see the fixed effects model correctly centers the  $\hat{\beta}$  near the  $\beta = 0$ , while the other family-level random effects (random intercept) and pooled regression show estimates with significant upward bias in the presence of confounding above  $\rho = 0.1$ . However, the random effects has the advantage of smaller sampling variance (more efficient estimator) across all levels of confounding because it pools estimates across families. **Fig B, Regional linkage map for FHS genome-wide suggestive SNPs for sibling-pair standard deviation in BMI from 1,000 Genomes, CEU Panel.** Maps produced by SNAP ([75]). **Fig C, QQ plots associated with Manhattan plots in Fig 1.** A) Observed versus expected p-value distributions for analysis of sibling-pair standard deviation in height for FHS generation-three respondents with controls for parental genotype, mean height of sibling pair, sex, and sex difference. B) Same as in (A) except for BMI instead of height. Shaded gray regions depict 95% confidence intervals. **Fig D, Power to detect an effect size of  $R^2$ .** The figure contrasts power at three potential sample sizes (defined as the number of sibling pairs in the data)(see [Methods](#)): 1) the Framingham Heart Study (FHS) sample used in the present analysis; 2) the Adolescent and Longitudinal Study of Health (AddHealth) sample; and 3) the UK Biobank sample. Likewise, the figure contrasts two potential p-value thresholds:  $p < 10^{-5}$  for the discovery analysis;  $p < 0.05$  for the confirmation analysis. The figure shows that although the sample used in the present analysis (FHS) is not adequately powered to detect realistic effect sizes of  $R^2 < 0.01$ , newly-released datasets with larger sibling subsamples are adequately powered to detect effects using the method. **Fig E, Regional Association Plot of rs7202116, top hit for variance in BMI found by Yang et al. (2012), on mean level of BMI from GIANT consortium data.** Figure produced using LocusZoom ([76]). **Fig F, Regional Association Plot of genome-wide suggestively significant ( $p < 10^{-5}$ ) hits from Fig 1 on mean height from GIANT consortium data.** (A—D) Plots for the SNPs rs2804263, rs2073302, rs8126205, and rs4834078, respectively, show no markers in the respective regions that approach even genome-wide suggestive significance ( $p < 10^{-5}$ ). Figures produced using LocusZoom ([76]). **Fig G, Relationship between sibling-pair mean BMI and sibling-pair standard deviation (SD) or coefficient of variation (CV).** A) Sibling-pair SD versus mean ( $\rho = 0.43$ ). B) Sibling-pair CV versus mean ( $\rho = 0.25$ ). **Fig H, Manhattan plots for enriched pathway HSA04540 Gap Junction for height variability.** A) FHS discovery sample; B) MTPS replication sample. **Table A, Results of Hausman test comparing  $\hat{\beta}_{FE}$  with  $\hat{\beta}_{RE}$  across the simulations.** The results show that at zero and very small amounts of confounding, the test fails to reject the null hypothesis that  $\hat{\beta}_{RE}$  is a consistent estimator, but at higher levels of confounding, the test rejects this null hypothesis. **Table B, Results of regressing squared Z-score of trait on minor allele count across 1000 replicates with non-demeaned data.** The results show an inflated type I error rate for the trait simulated to have either null effects or mean effects but no variance effects in the presence of an unobserved confounder between genotype and outcome (underlined rows). **Table C, Results of regressing squared Z-score of trait on minor allele count across 1000 replicates. Regressions are**

**estimated using the demeaned data.** The results show an inflated type I error rate for the trait simulated to have either null effects or mean effects but no variance effects in the presence of an unobserved confounder between genotype and outcome (underlined rows) even after transforming the data. **Table D, Results of DGLM on non-demeaned (non-transformed) for simulated DV with different types of effects across 1000 replicates.** The results show two types of inflated type I error rates. First, when a variant has null effects (neither effects on the mean nor effects on the variance) and there is confounding, the DGLM has an inflated type I error rate, detecting  $\beta \neq 0$  in 77.8% of simulations. Second, when a variant has variance effects but no mean effects, the method also has an inflated type I error rate, detecting  $\beta \neq 0$  in 11% of cases in the absence of confounding and 70% of cases in the presence of confounding. **Table E, Results of DGLM on demeaned (non-transformed) for simulated DV with different types of effects across 1000 replicates.** The results show an inflated type I error rate (estimate  $\beta \neq 0$  despite the presence of allele effects on the variance and not the mean) that is smaller but still present in the demeaned data. The results also show that while demeaning reduces the type I error rate (false detection of mean effects), the transformation leads to type II errors (fails to detect variance effects when these are present). **Table F Results of regressing sibling SD of trait on minor allele count across 1000 replicates.** The results show the percentage of simulations for which the coefficient on the minor allele count is significant at the  $p < 0.05$  level when we regress the sibling standard deviation of the trait on this count and controls. In order for the method to adequately control for Type I error, we want this percentage to be low for the traits simulated to have 1. neither mean nor variance effects or 2. mean effects only. In order for the method to be adequately powered, we want this percentage to be high for the traits simulated to have: 1. variance effects only and 2. both mean and variance effects. For the first goal, the results show that in contrast to the squared Z-score and DGLM, which each, in the presence of an unobserved confounder, display type I error rates of around 20% in detecting variance effects in traits simulated to have mean effects only, the sibling SD method avoids this type of error (underlined rows) both with and without controls for parental genotype. The results also illustrate that the method detects variance effects when the trait either has variance effects only or when the trait exhibits both mean and variance effects. The first half of the table also shows a lower type I error rate than squared Z-score when there is no unobserved confounder. **Table G, Proxy SNPs and results for replication analysis using Minnesota Twin Family Study data.** **Table H, Replicated GWAS hits for other SNPs on MAST4 Results are from the NCBI Phenotype-Genotype Integrator.** **Table I, Gene set analysis results using PASCAL. The table shows significant gene sets in FHS that replicated in MTF5 at different p-value thresholds (MAST4, the location of the replicated SNP, does not appear because although it was  $p < 0.01$  in the FHS dataset, it was  $p = 0.1$  in the MTF5 dataset).** **Table J, Pathway analysis results using PASCAL The table shows significant pathways in FHS that replicated in MTF5 at different p-value thresholds. The pathway that replicated using the i-GSEA4G-WAS tool is not among those tested by PASCAL.** **Table K, Illustrating correlation between population indicators and family-level intercept across 1000 replicates at four degrees of family-level confounding.** The results show that at low levels of confounding ( $\rho \leq 0.1$ ), the broad ancestry indicators are correlated with the indicator for family genotype but are too broad to fully capture the confounding. As the confounding increases beyond low levels, the ancestry indicators better capture the confounding. **Table L, Relationship between family intercept and observed genotype.** The table shows that as the degree of between-family confounding increases, there is a stronger relationship between the intercept that shifts levels of a trait up or down between families and the genotype. (PDF)

## Acknowledgments

For helpful feedback on this work, we would like to thank members of the Siegal Lab at the NYU Center for Genomics and Systems Biology as well as Justin Blau, Richard Bonneau, and Michael Purugganan of the NYU Department of Biology, Tomas Kirchoff of NYU Medical School, and Boriana Pratt of Princeton's Office of Population Research. The authors are also grateful to participants in the University of Colorado's 2012 workshop on Integrating Genetics and the Social Sciences for helpful comments.

## Author Contributions

**Conceptualization:** Dalton Conley, Ben Domingue, Christopher Dawes, Jason Boardman, Mark Siegal.

**Formal analysis:** Dalton Conley, Ben Domingue, Christopher Dawes, Jason Boardman, Mark Siegal.

**Funding acquisition:** Dalton Conley.

**Investigation:** Dalton Conley, Ben Domingue, Christopher Dawes, Jason Boardman, Mark Siegal.

**Methodology:** Rebecca Johnson, Ben Domingue, Mark Siegal.

**Writing – original draft:** Dalton Conley, Ben Domingue, Christopher Dawes, Jason Boardman, Mark Siegal.

**Writing – review & editing:** Dalton Conley, Rebecca Johnson, Mark Siegal.

## References

1. Dongen SV. Fluctuating asymmetry and developmental instability in evolutionary biology: past, present and future. *Journal of Evolutionary Biology*. 2006; 19(6):1727–1743. <https://doi.org/10.1111/j.1420-9101.2006.01175.x> PMID: 17040371
2. Levy SF, Siegal ML. The robustness continuum. *Advances in Experimental Medicine and Biology*. 2012; 751:431–452. [https://doi.org/10.1007/978-1-4614-3567-9\\_20](https://doi.org/10.1007/978-1-4614-3567-9_20) PMID: 22821470
3. Geiler-Samerotte KA, Bauer CR, Li S, Ziv N, Gresham D, Siegal ML. The details in the distributions: why and how to study phenotypic variability. *Current Opinion in Biotechnology*. 2013; 24(4):752–759. <https://doi.org/10.1016/j.copbio.2013.03.010> PMID: 23566377
4. Gibson G, Dworkin I. Uncovering cryptic genetic variation. *Nature Reviews Genetics*. 2004; 5(9): nrg1426. <https://doi.org/10.1038/nrg1426>
5. Masel J, Siegal ML. Robustness: mechanisms and consequences. *Trends in Genetics*. 2009; 25(9):395–403. <https://doi.org/10.1016/j.tig.2009.07.005> PMID: 19717203
6. Berger D, Bauerfeind SS, Blanckenhorn WU, Schäfer MA. High Temperatures Reveal Cryptic Genetic Variation in a Polymorphic Female Sperm Storage Organ. *Evolution*. 2011; 65(10):2830–2842. <https://doi.org/10.1111/j.1558-5646.2011.01392.x> PMID: 21967425
7. Gibson G. Decanalization and the origin of complex disease. *Nature Reviews Genetics*. 2009; 10(2): nrg2502. <https://doi.org/10.1038/nrg2502>
8. Gibson G, Reed LK. Cryptic genetic variation. *Current Biology*. 2008; 18(21):R989–R990. <https://doi.org/10.1016/j.cub.2008.08.011> PMID: 19000804
9. Wagner A. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society of London B: Biological Sciences*. 2008; 275(1630):91–100. <https://doi.org/10.1098/rspb.2007.1137>
10. Siegal ML, Leu JY. On the Nature and Evolutionary Impact of Phenotypic Robustness Mechanisms. *Annual Review of Ecology, Evolution, and Systematics*. 2014; 45:496–517. <https://doi.org/10.1146/annurev-ecolsys-120213-091705> PMID: 26034410
11. Jarosz DF, Lindquist S. Hsp90 and Environmental Stress Transform the Adaptive Value of Natural Genetic Variation. *Science*. 2010; 330(6012):1820–1824. <https://doi.org/10.1126/science.1195487> PMID: 21205668



12. Queitsch C, Sangster TA, Lindquist S. Hsp90 as a capacitor of phenotypic variation. *Nature*. 2002; 417(6889):618–624. <https://doi.org/10.1038/nature749> PMID: 12050657
13. Rutherford SL, Lindquist S. Hsp90 as a capacitor for morphological evolution. *Nature*. 1998; 396(6709):24550. <https://doi.org/10.1038/24550>
14. Yeyati PL, Bancewicz RM, Maule J, Heyningen Vv. Hsp90 Selectively Modulates Phenotype in Vertebrate Development. *PLOS Genetics*. 2007; 3(3):e43. <https://doi.org/10.1371/journal.pgen.0030043> PMID: 17397257
15. Hsieh YY, Hung PH, Leu JY. Hsp90 regulates nongenetic variation in response to environmental stress. *Molecular Cell*. 2013; 50(1):82–92. <https://doi.org/10.1016/j.molcel.2013.01.026> PMID: 23434373
16. Sangster TA, Salathia N, Undurraga S, Milo R, Schellenberg K, Lindquist S, et al. HSP90 affects the expression of genetic variation and developmental stability in quantitative traits. *Proceedings of the National Academy of Sciences*. 2008; 105(8):2963–2968. <https://doi.org/10.1073/pnas.0712200105>
17. Siegal ML. Crouching variation revealed. *Molecular Ecology*. 2013; 22(5):1187–1189. <https://doi.org/10.1111/mec.12195> PMID: 23437837
18. Bauer CR, Li S, Siegal ML. Essential gene disruptions reveal complex relationships between phenotypic robustness, pleiotropy, and fitness. *Molecular Systems Biology*. 2015; 11(1):773. <https://doi.org/10.15252/msb.20145264> PMID: 25609648
19. Levy SF, Siegal ML. Network Hubs Buffer Environmental Variation in *Saccharomyces cerevisiae*. *PLOS Biology*. 2008; 6(11):e264. <https://doi.org/10.1371/journal.pbio.0060264> PMID: 18986213
20. Richardson JB, Uppendahl LD, Traficante MK, Levy SF, Siegal ML. Histone Variant HTZ1 Shows Extensive Epistasis with, but Does Not Increase Robustness to, New Mutations. *PLOS Genetics*. 2013; 9(8):e1003733. <https://doi.org/10.1371/journal.pgen.1003733> PMID: 23990806
21. Geiler-Samerotte KA, Zhu YO, Goulet BE, Hall DW, Siegal ML. Selection Transforms the Landscape of Genetic Variation Interacting with Hsp90. *PLoS biology*. 2016; 14(10):e2000465. <https://doi.org/10.1371/journal.pbio.2000465> PMID: 27768682
22. Ansel J, Bottin H, Rodriguez-Beltran C, Damon C, Nagarajan M, Fehrmann S, et al. Cell-to-Cell Stochastic Variation in Gene Expression Is a Complex Genetic Trait. *PLOS Genetics*. 2008; 4(4):e1000049. <https://doi.org/10.1371/journal.pgen.1000049> PMID: 18404214
23. Ayroles JF, Buchanan SM, O'Leary C, Skutt-Kakaria K, Grenier JK, Clark AG, et al. Behavioral idiosyncrasy reveals genetic control of phenotypic variability. *Proceedings of the National Academy of Sciences*. 2015; 112(21):6706–6711. <https://doi.org/10.1073/pnas.1503830112>
24. Fehrmann S, Bottin-Duplus H, Leonidou A, Mollereau E, Barthelaix A, Wei W, et al. Natural sequence variants of yeast environmental sensors confer cell-to-cell expression variability. *Molecular Systems Biology*. 2013; 9(1):695. <https://doi.org/10.1038/msb.2013.53> PMID: 24104478
25. Fraser HB, Schadt EE. The Quantitative Genetics of Phenotypic Robustness. *PLOS ONE*. 2010; 5(1):e8635. <https://doi.org/10.1371/journal.pone.0008635> PMID: 20072615
26. Hall MC, Dworkin I, Ungerer MC, Purugganan M. Genetics of microenvironmental canalization in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*. 2007; 104(34):13717–13722. <https://doi.org/10.1073/pnas.0701936104>
27. Hulse AM, Cai JJ. Genetic Variants Contribute to Gene Expression Variability in Humans. *Genetics*. 2013; 193(1):95–108. <https://doi.org/10.1534/genetics.112.146779> PMID: 23150607
28. Janhunen M, Kause A, Vehviläinen H, Järvisalo O. Genetics of Microenvironmental Sensitivity of Body Weight in Rainbow Trout (*Oncorhynchus mykiss*) Selected for Improved Growth. *PLOS ONE*. 2012; 7(6):e38766. <https://doi.org/10.1371/journal.pone.0038766> PMID: 22701708
29. Jimenez-Gomez JM, Corwin JA, Joseph B, Maloof JN, Kliebenstein DJ. Genomic Analysis of QTLs and Genes Altering Natural Variation in Stochastic Noise. *PLOS Genetics*. 2011; 7(9):e1002295. <https://doi.org/10.1371/journal.pgen.1002295> PMID: 21980300
30. Perry GML, Nehrke KW, Bushinsky DA, Reid R, Lewandowski KL, Hueber P, et al. Sex Modifies Genetic Effects on Residual Variance in Urinary Calcium Excretion in Rat (*Rattus norvegicus*). *Genetics*. 2012; 191(3):1003–1013. <https://doi.org/10.1534/genetics.112.138909> PMID: 22554889
31. Rönnegård L, Valdar W. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genetics*. 2012; 13:63. <https://doi.org/10.1186/1471-2156-13-63> PMID: 22827487
32. Shen X, Pettersson M, Ronnegard L, Carlborg O. Inheritance Beyond Plain Heritability: Variance-Controlling Genes in *Arabidopsis thaliana*. *PLOS Genetics*. 2012; 8(8):e1002839. <https://doi.org/10.1371/journal.pgen.1002839> PMID: 22876191
33. Stinchcombe JR, Caicedo AL, Hopkins R, Mays C, Boyd EW, Purugganan MD, et al. Vernalization sensitivity in *Arabidopsis thaliana* (Brassicaceae): the effects of latitude and FLC variation. *American Journal of Botany*. 2005; 92(10):1701–1707. <https://doi.org/10.3732/ajb.92.10.1701> PMID: 21646087



34. Takahashi KH, Okada Y, Teramura K. Genome-Wide Deficiency Mapping of the Regions Responsible for Temporal Canalization of the Developmental Processes of *Drosophila melanogaster*. *Journal of Heredity*. 2011; 102(4):448–457. <https://doi.org/10.1093/jhered/esr026> PMID: 21525178
35. Takahashi KH, Okada Y, Teramura K. Deficiency Screening for Genomic Regions with Effects on Environmental Sensitivity of the Sensory Bristles of *Drosophila Melanogaster*. *Evolution*. 2012; 66(9):2878–2890. <https://doi.org/10.1111/j.1558-5646.2012.01636.x> PMID: 22946809
36. Tonsor SJ, Elnaccash TW, Scheiner SM. Developmental Instability Is Genetically Correlated with Phenotypic Plasticity, Constraining Heritability, and Fitness. *Evolution*. 2013; 67(10):2923–2935. PMID: 24094343
37. Chen B, Wagner A. Hsp90 is important for fecundity, longevity, and buffering of cryptic deleterious variation in wild fly populations. *BMC Evolutionary Biology*. 2012; 12:25. <https://doi.org/10.1186/1471-2148-12-25> PMID: 22369091
38. Sgrò CM, Wegener B, Hoffmann AA. A naturally occurring variant of Hsp90 that is associated with decanalization. *Proceedings of the Royal Society of London B: Biological Sciences*. 2010; 277(1690):2049–2057. <https://doi.org/10.1098/rspb.2010.0008>
39. Siegal ML, Masel J. Hsp90 depletion goes wild. *BMC Biology*. 2012; 10:14. <https://doi.org/10.1186/1741-7007-10-14> PMID: 22369621
40. et al Y. *FTO* genotype is associated with phenotypic variability of body mass index. *Nature*. 2012; 490(7419):nature11401.
41. Paré G, Cook NR, Ridker PM, Chasman DI. 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 Genetics*. 2010; 6(6). <https://doi.org/10.1371/journal.pgen.1000981> PMID: 20585554
42. Andreasen CH, Stender-Petersen KL, Mogensen MS, Torekov SS, Wegner L, Andersen G, et al. Low Physical Activity Accentuates the Effect of the *FTO* rs9939609 Polymorphism on Body Fat Accumulation. *Diabetes*. 2008; 57(1):95–101. <https://doi.org/10.2337/db07-0910> PMID: 17942823
43. Cha SW, Choi SM, Kim KS, Park BL, Kim JR, Kim JY, et al. Replication of Genetic Effects of *FTO* Polymorphisms on BMI in a Korean Population. *Obesity*. 2008; 16(9):2187–2189. <https://doi.org/10.1038/oby.2008.314> PMID: 18551112
44. Chang YC, Liu PH, Lee WJ, Chang TJ, Jiang YD, Li HY, et al. Common Variation in the Fat Mass and Obesity-Associated (*FTO*) Gene Confers Risk of Obesity and Modulates BMI in the Chinese Population. *Diabetes*. 2008; 57(8):2245–2252. <https://doi.org/10.2337/db08-0377> PMID: 18487448
45. Dina C, Meyre D, Gallina S, Durand E, Körner A, Jacobson P, et al. Variation in *FTO* contributes to childhood obesity and severe adult obesity. *Nature Genetics*. 2007; 39(6):ng2048. <https://doi.org/10.1038/ng2048>
46. Hunt SC, Stone S, Xin Y, Scherer CA, Magness CL, Iadonato SP, et al. Association of the *FTO* Gene With BMI. *Obesity*. 2008; 16(4):902–904. <https://doi.org/10.1038/oby.2007.126> PMID: 18239580
47. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A Common Variant in the *FTO* Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science*. 2007; 316(5826):889–894. <https://doi.org/10.1126/science.1141634> PMID: 17434869
48. Haworth CMA, Carnell S, Meaburn EL, Davis OSP, Plomin R, Wardle J. Increasing Heritability of BMI and Stronger Associations With the *FTO* Gene Over Childhood. *Obesity*. 2008; 16(12):2663–2668. <https://doi.org/10.1038/oby.2008.434> PMID: 18846049
49. Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J, et al. Genome-Wide Association Scan Shows Genetic Variants in the *FTO* Gene Are Associated with Obesity-Related Traits. *PLoS Genetics*. 2007; 3(7):e115. <https://doi.org/10.1371/journal.pgen.0030115> PMID: 17658951
50. Tönjes A, Zeggini E, Kovacs P, Böttcher Y, Schleinitz D, Dietrich K, et al. Association of *FTO* variants with BMI and fat mass in the self-contained population of Sorbs in Germany. *European Journal of Human Genetics*. 2009; 18(1):ejhg2009107.
51. Cao Y, Wei P, Bailey M, Kauwe JSK, Maxwell TJ, for the Alzheimer’s Disease Neuroimaging Initiative. A Versatile Omnibus Test for Detecting Mean and Variance Heterogeneity. *Genetic Epidemiology*. 2014; 38(1):51–59. <https://doi.org/10.1002/gepi.21778> PMID: 24482837
52. Struchalin MV, Dehghan A, Witteman JC, van Duijn C, Aulchenko YS. Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC Genetics*. 2010; 11:92. <https://doi.org/10.1186/1471-2156-11-92> PMID: 20942902
53. Soave D, Corvol H, Panjwani N, Gong J, Li W, Boëlle PY, et al. A Joint Location-Scale Test Improves Power to Detect Associated SNPs, Gene Sets, and Pathways. *The American Journal of Human Genetics*. 2015; 97(1):125–138. <https://doi.org/10.1016/j.ajhg.2015.05.015> PMID: 26140448

54. Aschard H, Zaitlen N, Tamimi RM, Lindström S, Kraft P. A non-parametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes. *Genetic epidemiology*. 2013; 37(4):323–333. <https://doi.org/10.1002/gepi.21716> PMID: 23512279
55. Hong C, Ning Y, Wei P, Cao Y, Chen Y. A semiparametric model for vQTL mapping. *Biometrics*. 2017; 73(2):571–581. <https://doi.org/10.1111/biom.12612> PMID: 27861717
56. Wei WH, Bowes J, Plant D, Viatte S, Yarwood A, Massey J, et al. Major histocompatibility complex harbors widespread genotypic variability of non-additive risk of rheumatoid arthritis including epistasis. *Scientific Reports*. 2016; 6.
57. Caplin A. Genes, Addiction, and Economics;. Available from: <https://bfi.uchicago.edu/events/biological-basis-preferences-and-behavior>.
58. Cao Y, Maxwell TJ, Wei P. A Family-Based Joint Test for Mean and Variance Heterogeneity for Quantitative Traits. *Annals of Human Genetics*. 2015; 79(1):46–56. <https://doi.org/10.1111/ahg.12089> PMID: 25393880
59. Rönnegård L, Valdar W. Detecting Major Genetic Loci Controlling Phenotypic Variability in Experimental Crosses. *Genetics*. 2011; 188(2):435–447. <https://doi.org/10.1534/genetics.111.127068> PMID: 21467569
60. Hamer DH. Beware the chopsticks gene. *Molecular psychiatry*. 2000; 5(1):11–13. <https://doi.org/10.1038/sj.mp.4000662> PMID: 10673763
61. Day RL, Laland KN, Odling-Smee FJ. Rethinking Adaptation: The Niche-Construction Perspective. *Perspectives in Biology and Medicine*. 2003; 46(1):80–95. <https://doi.org/10.1353/pbm.2003.0003> PMID: 12582272
62. Laland KN, Brown GR. Niche construction, human behavior, and the adaptive-lag hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*. 2006; 15(3):95–104. <https://doi.org/10.1002/evan.20093>
63. Laland KN, Odling-Smee FJ, Feldman MW. Evolutionary consequences of niche construction and their implications for ecology. *Proceedings of the National Academy of Sciences*. 1999; 96(18):10242–10247. <https://doi.org/10.1073/pnas.96.18.10242>
64. Odling-Smee FJ, Laland KN, Feldman MW. Niche Construction. *The American Naturalist*. 1996; 147(4):641–648. <https://doi.org/10.1086/285870>
65. Odling-Smee FJ, Laland KN, Feldman MW. *Niche Construction: The Neglected Process in Evolution*. Princeton University Press; 2003.
66. Wang G, Yang E, Brinkmeyer-Langford CL, Cai JJ. Additive, Epistatic, and Environmental Effects Through the Lens of Expression Variability QTL in a Twin Cohort. *Genetics*. 2014; 196(2):413–425. <https://doi.org/10.1534/genetics.113.157503> PMID: 24298061
67. Dieleman JL, Templin T. Random-Effects, Fixed-Effects and the within-between Specification for Clustered Data in Observational Health Studies: A Simulation Study. *PLOS ONE*. 2014; 9(10):e110257. <https://doi.org/10.1371/journal.pone.0110257> PMID: 25343620
68. Ahmad S, Rukh G, Varga TV, Ali A, Kurbasic A, Shungin D, et al. Gene× physical activity interactions in obesity: combined analysis of 111,421 individuals of European ancestry. *PLoS genetics*. 2013; 9(7): e1003607. <https://doi.org/10.1371/journal.pgen.1003607> PMID: 23935507
69. Sun X, Elston R, Morris N, Zhu X. What Is the Significance of Difference in Phenotypic Variability across SNP Genotypes? *The American Journal of Human Genetics*. 2013; 93(2):390–397. <https://doi.org/10.1016/j.ajhg.2013.06.017> PMID: 23910463
70. Consortium EGGE, et al. A genome-wide association meta-analysis identifies new childhood obesity loci. *Nature genetics*. 2012; 44(5):526–531. <https://doi.org/10.1038/ng.2247>
71. Dumitrascu B, Darnell G, Ayroles J, Engelhardt BE. A Bayesian test to identify variance effects. *arXiv:151201616 [q-bio, stat]*. 2015;.
72. Conley D, Rauscher E, Dawes C, Magnusson PK, Siegal ML. Heritability and the equal environments assumption: Evidence from multiple samples of misclassified twins. *Behavior Genetics*. 2013; 43(5):415–426. <https://doi.org/10.1007/s10519-013-9602-1> PMID: 23903437
73. Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models*. vol. 1. Cambridge University Press New York, NY, USA; 2007.
74. dbGaP | phs000007.v29.p10 | Framingham Cohort;. Available from: [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000007.v29.p10](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v29.p10).
75. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, Bakker D, et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. 2008; 24(24):2938–2939. <https://doi.org/10.1093/bioinformatics/btn564> PMID: 18974171

76. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010; 26(18):2336–2337. <https://doi.org/10.1093/bioinformatics/btq419> PMID: 20634204
77. A map of human genome variation from population scale sequencing. *Nature*. 2010; 467(7319):1061–1073. <https://doi.org/10.1038/nature09534> PMID: 20981092
78. Miller MB, Basu S, Cunningham J, Eskin E, Malone SM, Oetting WS, et al. The Minnesota Center for Twin and Family Research Genome-Wide Association Study. *Twin research and human genetics: the official journal of the International Society for Twin Studies*. 2012; 15(6):767–774. <https://doi.org/10.1017/thg.2012.62>
79. Iacono WG, McGue M. Minnesota Twin Family Study. *Twin Research and Human Genetics*. 2002; 5(5):482–487. <https://doi.org/10.1375/136905202320906327>
80. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A Versatile Gene-Based Test for Genome-wide Association Studies. *The American Journal of Human Genetics*. 2010; 87(1):139–145. <https://doi.org/10.1016/j.ajhg.2010.06.009> PMID: 20598278
81. Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Research*. 2010; 38(suppl\_2):W90–W95. <https://doi.org/10.1093/nar/gkq324> PMID: 20435672
82. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS computational biology*. 2016; 12(1): e1004714. <https://doi.org/10.1371/journal.pcbi.1004714> PMID: 26808494
83. Champely S. pwr: Basic functions for power analysis. R package version 1.1. 1. The R Foundation. 2009;.
84. Imbens GW, Kolesár M. Robust Standard Errors in Small Samples: Some Practical Advice. *The Review of Economics and Statistics*. 2015; 98(4):701–712. [https://doi.org/10.1162/REST\\_a\\_00552](https://doi.org/10.1162/REST_a_00552)
85. Cameron AC, Miller DL. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*. 2015; 50(2):317–372. <https://doi.org/10.3368/jhr.50.2.317>