www.nature.com/scientificreports

scientific reports

OPEN

Genome-wide variance quantitative trait locus analysis suggests small interaction effects in blood pressure traits

Gang Shi

Genome-wide variance quantitative trait loci (vQTL) analysis complements genome-wide association study (GWAS) and has the potential to identify novel variants associated with the trait, explain additional trait variance and lead to the identification of factors that modulate the genetic effects. I conducted genome-wide analysis of the UK Biobank data and identified 27 vQTLs associated with systolic blood pressure (SBP), diastolic blood pressure (DBP) and pulse pressure (PP). The top single-nucleotide polymorphisms (SNPs) are enriched for expression QTLs (eQTLs) or splicing QTLs (sQTLs) annotated by GTEx, suggesting their regulatory roles in mediating the associations with blood pressure (BP). Of the 27 vQTLs, 14 are known BP-associated QTLs discovered by GWASs. The heteroscedasticity effects of the 13 novel vQTLs are larger than their genetic main effects, which were not detected by existing GWASs. The total R-squared of the 27 top SNPs due to variance heteroscedasticity is 0.28%, compared with 0.50% owing to their main effects. The overall effect size of the variance heteroscedasticity is small in GWAS SNPs compared with their main effects. For the 411, 384 and 285 GWAS SNPs associated with SBP, DBP and PP, respectively, their heteroscedasticity effects were 0.52%, 0.43%, and 0.16%, and their main effects were 5.13%, 5.61%, and 3.75%, respectively. The number and effects of the vQTLs are small, which suggests that the effects of gene–environment and gene–gene interactions are small. The main effects of the SNPs remain the major source of genetic variance for BP, which would probably be true for other complex traits as well.

Variance quantitative trait locus (vQTL) refers to a locus that is associated with the difference in the variance in a quantitative trait[1–4]. Such variance heteroscedasticity may be induced by gene–environment interactions[2,5–8], gene–gene interactions[4], multiple linked functional variants at the locus[9,10] or scale effects[11,12]. Since standard genome-wide association studies (GWASs) focus on testing differences in the means across genotypes, genetic variance attributable to variance heteroscedasticity is missed. Genome-wide vQTL analysis complements GWASs and has the potential to identify novel variants associated with a trait. In addition, the variance heteroscedasticity of a vQTL is genotype-dependent, therefore, could explain additional trait variance that is genetically related. Without the need to measure environmental factors, genome-wide vQTL analysis has been used to screen potential gene–environment interactions and search for factors that modulate the genetic effects[2,5,6,8].

To date, the identified vQTLs are largely for obesity-related traits. Yang et al. found that the *FTO* gene locus was not only associated with the mean of body mass index (BMI), but also the variance of it[13]. In the analysis of BMI with UK biobank data, Young et al. identified 48 genome-wide significant loci that demonstrated smaller P values from the test including variance heteroscedasticity effect than from the test of additive effect in trait mean only[5]. In the analysis of 13 quantitative traits from the UK Biobank, Wang et al. discovered 75 significant vQTLs for 9 traits, 60 of which were for those related to obesity[6]. Their interaction analysis showed that the vQTLs were enriched with gene-environment interactions. In the analysis of the genetic risk score of BMI with 376 variants, Sulc et al. demonstrated that while the genetic risk score explained 5.2% of BMI variance, its interactions with environmental factors explained an additional 1.9%[7]. Marderstein et al. showed that the discovery and replication rates of gene-environment interactions for BMI were significantly higher when prioritizing variants in vQTLs compared to when accessing all genetic variants. They also demonstrated strong gene-environment interactions mediated the genetic contribution to body weight and diabetes risk.

School of Telecommunications Engineering, Xidian University, 2 South Taibai Road, Xi'an 710071, Shaanxi, China. email: gshi@xidian.edu.cn

A recent study of blood pressure (BP) shows that some portions of the BP variance could be attributed to gene–environment interactions[14]. In this work, I propose statistical methods for vQTL analysis at the biobank scale that are based on a linear mixed model and regressions. I conducted genome-wide vQTL analysis of BP data in the UK Biobank to search for novel single-nucleotide polymorphisms (SNPs) associated with BP and evaluated additional BP variance explained by their variance heteroscedasticity.

Material and methods

Mixed model analysis.

A variety of heteroscedasticity tests have been suggested for finding vQTLs, which have been reviewed previously[11,15]. In this work, I employ a linear mixed model and test the variance heteroscedasticity of a vQTL using the maximum likelihood approach, similar to that in (Sulc et al., 2020)[7]. Nevertheless, it focuses on testing the variance heteroscedasticity due to the interactions between a polygenic score and unknown environmental factors. The method is statistically powerful when the quantitative trait approximately follows a normal distribution.

Suppose that a quantitative trait y is associated with a genetic factor G and n covariates X_i, $i = 1,\ldots,n$, as follows:

$$y = \sum_{i=1}^{n} \beta_i^{\mathrm{C}} X_i + \beta^{\mathrm{G}} G + \sum_{i=1}^{m} \beta_i^{\mathrm{E}} E_i + \sum_{i=1}^{m} \beta_i^{\mathrm{I}} E_i G + \varepsilon,$$

where β^{G} and β_i^{C}, $i = 1,\ldots,n$, are the genetic and covariate effects, respectively, and $\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$ is the random error. Here, G is assumed to be additive, which could be the dosage or coded genotype of a SNP. E_i, $i = 1,\ldots,m$ are m environmental factors that modulate the genetic effect, and β_i^{E} and β_i^{I} are the environmental and interaction effects, respectively. Without loss of generality, E_i, $i = 1,\ldots,m$ may include other factors that modulate the genetic effect.

Letting $\gamma_1 = \sum_{i=1}^{m} \beta_i^{\mathrm{E}} E_i$ and $\gamma_2 = \sum_{i=1}^{m} \beta_i^{\mathrm{I}} E_i$, I have

$$y = \sum_{i=1}^{n} \beta_i^{\mathrm{C}} X_i + \beta^{\mathrm{G}} G + \gamma_1 + \gamma_2 G + \varepsilon.$$

Assume that E_i, $i = 1,\ldots,m$ are centralized and uncorrelated, conditional on the covariates and genetic factor, trait variance depends on the genetic factor in a quadratic manner

$$\mathrm{var}\left(y\right) = \tau_1 + 2\tau_2 G + \tau_3 G^2 + \sigma^2,$$

where

$$\tau_1 = \mathrm{var}(\gamma_1) = \sum_{i=1}^{m} \left(\beta_i^{\mathrm{E}}\right)^2 \mathrm{var}(E_i)$$

$$\tau_2 = \mathrm{cov}(\gamma_1, \gamma_2) = \sum_{i=1}^{m} \beta_i^{\mathrm{E}} \beta_i^{\mathrm{I}} \mathrm{var}(E_i),$$

$$\tau_3 = \mathrm{var}(\gamma_2) = \sum_{i=1}^{m} \left(\beta_i^{\mathrm{I}}\right)^2 \mathrm{var}(E_i).$$

I further assume that γ_1 and γ_2 follow a bivariate normal distribution

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1 & \tau_2 \\ \tau_2 & \tau_3 \end{pmatrix}\right].$$

Note that σ^2 and τ_1 are not identifiable in the present model. I merge the error ε with the random effect γ_1 as the new γ_1 and have the linear mixed model

$$y = \sum_{i=1}^{n} \beta_i^{\mathrm{C}} X_i + \beta^{\mathrm{G}} G + \gamma_1 + \gamma_2 G, \tag{1}$$

whose mean and variance are

$$\mathrm{E}\left(y\right) = \sum_{i=1}^{n} \beta_i^{\mathrm{C}} X_i + \beta^{\mathrm{G}} \beta_{\mathrm{G}},$$

$$\mathrm{var}\left(y\right) = \tau_1 + 2\tau_2 G + \tau_3 G^2.$$

In this case, τ_1 is the variance of $\gamma_1 + \varepsilon$ in the original model, and τ_2 is the covariance between γ_2 and $\gamma_1 + \varepsilon$.

I solve the linear mixed model (1) numerically by the maximum likelihood method based on its profiled likelihood function and Newton–Raphson algorithm[16]. To alleviate computational burden, one can also first regress the trait on the covariates and genetic factor and then use the residual $\hat{e}$ to solve the variance component model

$$\hat{e} = \gamma_1 + \gamma_2 G. \tag{2}$$

To test the variance heteroscedasticity, the null hypothesis is H_0: $\tau_2 = \tau_3 = 0$, and the alternative hypothesis is that at least one of them is nonzero. The likelihood ratio statistics follows a 0.5:0.5 mixture distribution of a chi-square with 1 degree of freedom (df) and a chi-square with 2 df[17]. All tests proposed in this paper were implemented in the program "heter", which is available at https://github.com/eat1000/heter.

Regression analysis. To solve the linear mixed model (1) or the variance component model (2), it is usually desirable for the starting values of the iterating parameters to be close to their estimates. I extend the Breusch–Pagan test for heteroscedasticity[18] into a quadratic form, and the squared residual is regressed on the genetic factor as follows:

$$\hat{e}^2 = \tau_1 + 2\tau_2 G + \tau_3 G^2 + \varepsilon, \tag{3}$$

where ε is the random error. The estimated $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}_3$ by linear regression can be used as the starting values of the variance parameters for solving model (1) or (2). Alternatively, the heteroscedasticity test can also be performed by linear regression (3) by testing τ_2 and τ_3 jointly, and the likelihood ratio statistics follows a Chi-square distribution with 2 df.

Since the squared residual $\hat{e}^2$ is nonnegative, linear regression (3), which assumes normally distributed ε, could be suboptimal. An improved Chi-square regression analysis is formulated as follows:

$$\hat{e}^2 = \left(\tau_1 + 2\tau_2 G + \tau_3 G^2\right)\chi_1^2, \tag{4}$$

where χ_1^2 is a random variable following a Chi-square distribution with 1 df. I iteratively solve the Chi-square regression (4) by the maximum likelihood method with starting values estimated by linear regression (3). The test for variance heteroscedasticity is conducted based on testing τ_2 and τ_3 jointly using the likelihood ratio test, and the test statistics follows a Chi-square distribution with 2 df. It can be shown that the likelihood function of Chi-square regression (4) coincides with that of the variance component model (2) and that of the gamma regression with a shape parameter of 1/2 and a scale parameter modeled as $2\tau_1 + 4\tau_2 G + 2\tau_3 G^2$.

Notably, τ_1, τ_2 and τ_3 are unconstrained when solving regression models (3) and (4), and no distributions of γ_1 and γ_2 have to be assumed. Nevertheless, γ_1 and γ_2 are assumed to follow a bivariate normal distribution in the linear mixed model (1) and variance component model (2), and $\begin{pmatrix} \tau_1 & \tau_2 \\ \tau_2 & \tau_3 \end{pmatrix}$ are constrained to be positive semidefinite.

UK Biobank data and analyses. The UK Biobank is a large propective study in the United Kingdom with more than 500,000 participants aged between 40 and 69 years at the time of recruitment. The study design was described previously[19]. Deep phenotyping, genomic, and health-related data have been collected and are available for research investigating a wide range of diseases caused by a combination of genes, lifestyles, and environmental factors[20]. The UK Biobank has obtained informed consent from all participants and has obtained Research Tissue Bank approval from its ethics committee. This research has been approved by the UK Biobank under application number 44080 and I have complied with all relevant ethical regulations in this work.

Genome-wide genotype data of 487,422 individuals were imputed with the Haplotype Reference Consortium (HRC) reference panel by the UK Biobank, resulting in 93,095,623 autosomal SNPs, short indels and large structural variants[20]. I conducted genome-wide vQTL analysis of BP data in the UK Biobank using imputed genotype data. I restricted the analysis to autosomal SNPs that have minor allele frequencies higher than 0.01 and information scores larger than 0.9, obtaining 9,117,915 SNPs that passed the filters. I excluded quality control outliers for heterozygosity or missingness and samples with sex discordance between the self-reported and genetically inferred sex according to the sample quality control files provided by the UK Biobank. I also excluded non-European samples, samples from pregnant women, one sample in each related pair up to second-degree relatives, and samples from participants who had withdrawn consent.

Systolic blood pressure (SBP) and diastolic blood pressure (DBP) in the genome-wide vQTL analysis were averaged over multiple measurements assessed at baseline, which were further adjusted for antihypertensive medication use by adding 15 and 10 mm Hg[21] to SBP and DBP, respectively. Pulse pressure (PP) was computed as the difference between SBP and DBP and then logarithmically transformed. Covariates in the analysis included sex, age, age squared, BMI and the top 10 principal components (PCs). Samples with BP or BMI values that were 5 or more standard deviations outside the respective means and those with missing BP values or covariates were also excluded. The total sample sizes in the analyses were 396,077, 396,079 and 396,077 for SBP, DBP and PP, respectively.

Simulation studies. I conducted simulation studies to examine the distributions of the four test statistics under the null hypothesis and their statistical power to detect a vQTL with different heteroscedasticity effects. Covariates and genotype data in the UK Biobank were used for the simulation studies.

I simulated an SBP trait under the null hypothesis as follows:

$$\mathrm{SBP} = 86.5 + 5.65\mathrm{sex} + 0.00827\mathrm{age}^2 + 0.909\mathrm{BMI} + 0.0647\mathrm{PC}_4 + 0.0349\mathrm{PC}_9 + \sum_{i=1}^{54} \beta_i \mathrm{SNP}_i + \varepsilon$$

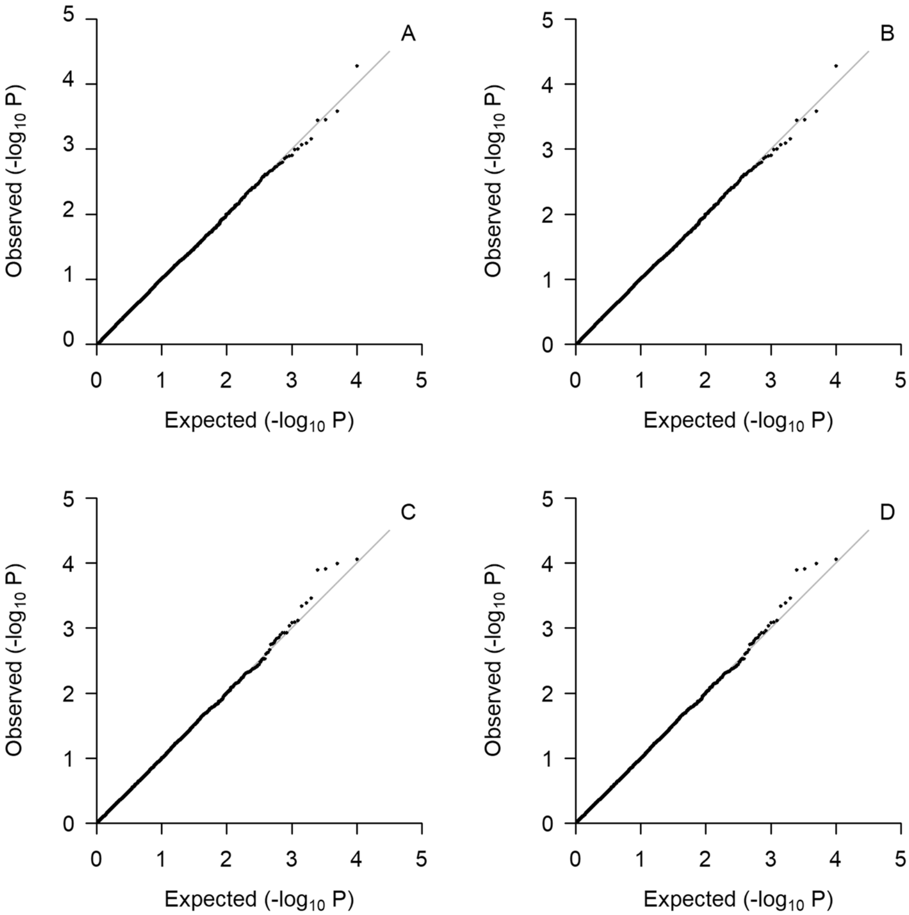

Figure 1. Quantile–quantile plots for the four test statistics under the null hypothesis. (**A**) Linear mixed model, (**B**) variance component model, (**C**) linear regression, (**D**) Chi-square regression.

in which the intercept, covariate effects and standard deviation of the random error ε were from the fitted model by using the SBP and covariates in the UK Biobank, the 54 SBP-associated SNPs and their effects were as reported in the literature[22], and ε was normally distributed with a zero mean and standard deviation of 18.44. Genotypes of the SNPs were called as those with the largest probabilities and then coded as being additive. For the 396,387 samples who passed quality controls and had non-missing covariates, their SBP values were simulated. Since the simulated SNP effects are constants across samples, there is no variance heteroscedasticity in this simulation. Then, I carried out four vQTL analyses on 10,000 SNPs randomly chosen from SNPs on 22 autosomes that passed the quality controls.

To evaluate the statistical power of the vQTL tests, I assumed that the effect of one SBP-associated SNP, rs880315, was random, following a normal distribution $\mathrm{N}\left(-0.475, \sigma^2\right)$ in the simulated samples. Here, the mean −0.475 is the average effect per reference allele as reported in the literature[22], and σ^2 is the variance of the effect. The effects of covariates and the other 53 SNPs remained the same as previously described. I considered 5 levels of heteroscedasticity with $\sigma^2 = 1$, 2, 3, 4, and 5. For each level, I simulated the SBP of 396,387 samples with 1000 replicates. I conducted the four vQTL analyses on SNP rs880315. The results of P values smaller than the genome-wide significance threshold of 5×10^{-8} [23] were considered to be significant, and a portion of the significant results among the 1000 replicates was the empirical power.

Results

Simulation results.

Under the null hypothesis, quantile–quantile (QQ) plots of the four vQTL tests are presented in Fig. 1. The observed P values from testing the 10,000 SNPs are shown as the vertical coordinates on a negative $\log_{10}$ scale, and the horizontal coordinates are their expected values from a uniform distribution between 0 and 1. As can be seen, the empirical distributions and their expected distributions align well for the four tests. The QQ plots for the linear mixed model (1) and variance component model (2) are almost identical. I compared the likelihood ratio statistics of the two tests, the test statistics based on the linear mixed model (1) and the variance component model (2) are almost identical as well. This is probably because the maximum likelihood estimates of fixed effects in linear mixed models are robust to the misspecification of their covariance structure[24]. Even though the estimates of the fixed effects are updated in each iteration when solving the linear mixed model (1), the changes are minuscule. Without the loss of much precision, the fixed effects can be estimated and kept

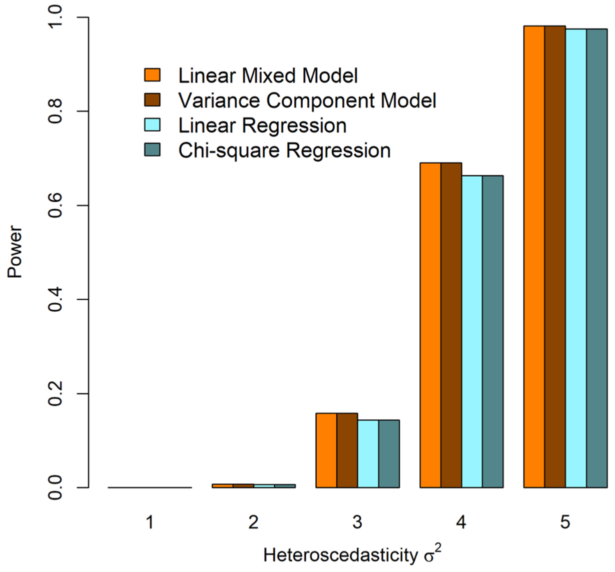

Figure 2. Statistical power of the four tests with different heteroscedasticity effects.

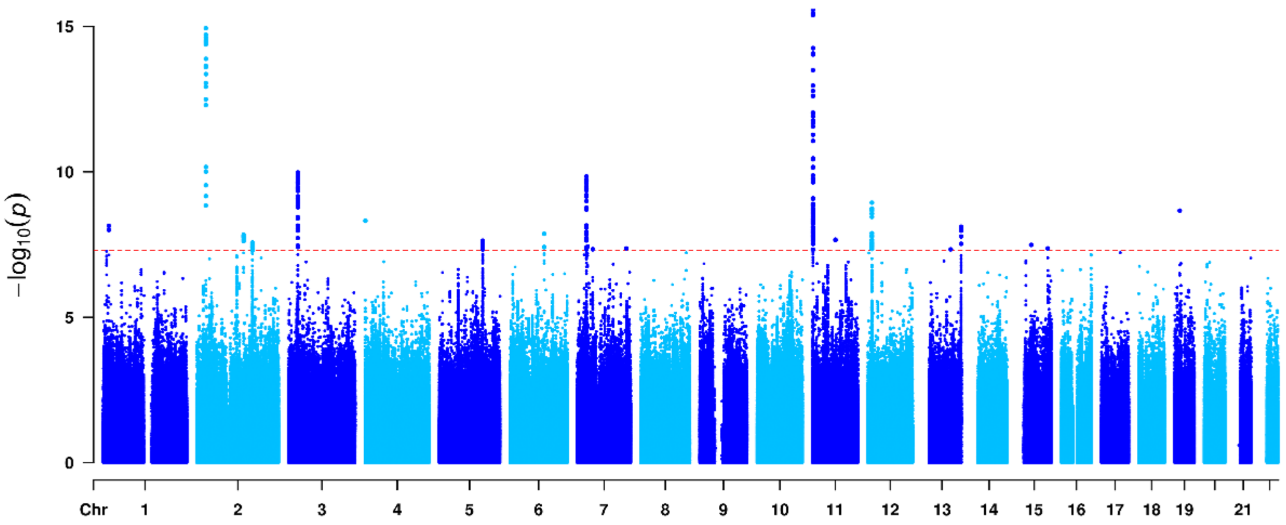

Figure 3. Manhattan plot visualizing genome-wide results from the vQTL analysis of SBP. Dots denote the SNPs in the genome-wide vQTL analysis, whose P values on a negative $\log_{10}$ scale are plotted against their physical positions. The dashed line represents the genome-wide significance level ($P = 5 \times 10^{-8}$).

unchanged, and the parameters of the random effects are estimated iteratively, which is equivalent to solving the variance component model (2). Similarly, the QQ plots for linear regression (3) and Chi-square regression (4) are very close, which is possibly due to the robustness of the linear regression.

The statistical power of the four tests is displayed in Fig. 2. Apparently, the power of the four tests becomes large with the increase in the heteroscedasticity effect. The power based on the linear mixed model (1) and the variance component model (2) are the same, and the test statistics based on the linear mixed model (1) and the variance component model (2) are almost identical. The powers of the tests based on linear regression and Chi-square regression are also the same. Nevertheless, test statistics based on Chi-square regression (4) are slightly larger than those based on linear regression (3) for some replicates when the heteroscedasticity effect is large. Compared with the linear mixed model (1) and variance component model (2), linear regression (3) and Chi-square regression (4) are statistically less powerful. This is because the test statistics based on regressions have larger dfs. Since the variance component method has the same power as the linear mixed model and is computationally much less expensive, I used it as the primary analysis in the analysis of BP data in the UK Biobank.

Genome-wide vQTL analysis of blood pressures. I conducted genome-wide vQTL analysis of BP data in the UK Biobank, and the results of the variance component model are displayed as Manhattan plots in Figs. 3, 4 and 5. I detected 20, 6 and 1 vQTLs that are associated with SBP, DBP and PP, respectively, at the genome-wide

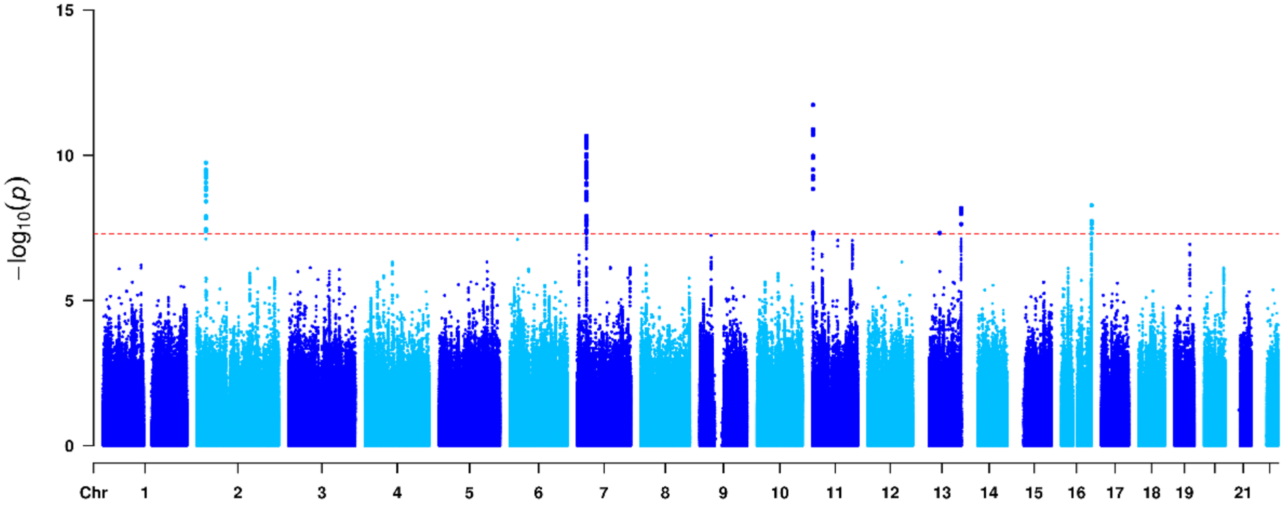

Figure 4. Manhattan plot visualizing genome-wide results from the vQTL analysis of DBP. Dots denote the SNPs in the genome-wide vQTL analysis, whose P values on a negative $\log_{10}$ scale are plotted against their physical positions. The dashed line represents the genome-wide significance level ($P = 5 \times 10^{-8}$).

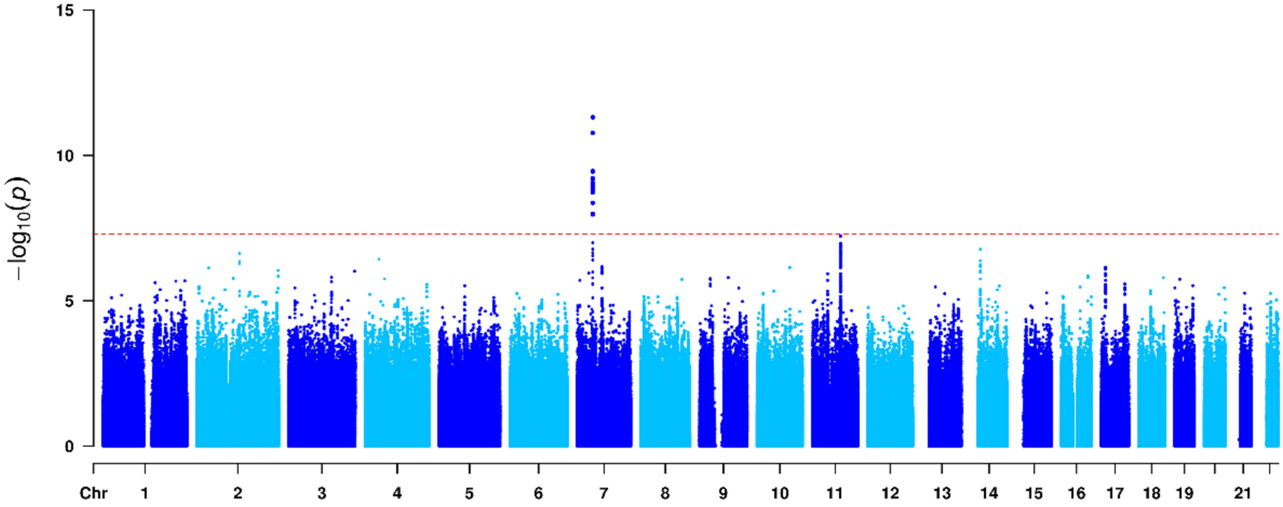

Figure 5. Manhattan plot visualizing genome-wide results from the vQTL analysis of PP. Dots denote the SNPs in the genome-wide vQTL analysis, whose P values on a negative $\log_{10}$ scale are plotted against their physical positions. The dashed line represents the genome-wide significance level ($P = 5 \times 10^{-8}$).

significance level ($P < 5 \times 10^{-8}$). The top SNPs that have the lowest P values at the loci are shown in Table 1. The results of the main effects of the SNPs were obtained from linear regression, as in standard GWAS. The R^2 values of the heteroscedasticity effects were computed as the likelihood-ratio based pseudo-R-squared[25].

The 27 significant vQTLs included the 23 top SNPs from 22 genomic loci, of which 10 SNPs from 9 loci showed significant SNP main effects ($P < 5 \times 10^{-8}$). rs1275984, rs11977526 and rs569550 are known BP-associated SNPs that are included in the GWAS catalog[26]. rs2029827 and rs12368847 are in linkage disequilibrium (LD $R^2 \geq 0.8$) with rs820430 and rs11609905, respectively, which are associated with BPs. rs7607074, rs115525024, rs10262140, rs376861852 and rs8032315 are in intermediate LD ($0.1 \leq$ LD $R^2 < 0.8$) with the BP-associated SNPs rs16849211, rs7812039, rs7812039, rs3934939, and rs17514846, respectively, and their LD R^2 ranges from 0.19 to 0.45.

The SNP main effects of the 27 vQTLs explain 0.50% of the BP variances collectively, and their heteroscedasticity accounts for an additional 0.28%, which is approximately half the variance attributed to the main effects. Because the vQTLs were discovered by the heteroscedasticity test, the relative contribution of the variance heteroscedasticity would be larger than that of all BP QTLs. I queried the GWAS catalog[26] for SNPs with genome-wide significance that were discovered among samples of European ancestry and were replicated. There were 440, 395 and 305 SNPs reported to be associated with SBP (EFO_0006335), DBP (EFO_0006336) and PP (EFO_0005763), respectively, of which 411, 384 and 285 SNPs were present in the analysis. The sums of the SNP main effects were 5.13%, 5.61%, and 3.75%, and the heteroscedasticity effects were 0.52%, 0.43%, and 0.16% for SBP, DBP and PP, respectively. The contributions of the heteroscedasticity relative to the genetic main effects are

							Main effect			Heteroscedasticity		
Locus	Top SNP	Chr	Position	NCA/CA	CAF	Trait	Effect	SE	P Value	R^2	P Value	R^2
PADI2	rs116515879	1	17,422,797	G/T	0.02	SBP	−0.2868	0.1483	5.01×10^{-2}	9.45×10^{-6}	7.32×10^{-9}	9.17×10^{-5}
CIB4-KCNK3	rs1275984*†	2	26,911,509	A/C	0.62	SBP	−0.6121	0.0427	1.33×10^{-46}	5.19×10^{-4}	1.15×10^{-15}	1.71×10^{-4}
						DBP	−0.3205	0.024	1.03×10^{-40}	4.51×10^{-4}	1.80×10^{-10}	1.10×10^{-4}
HNMT	rs3100721†‡	2	138,755,054	T/C	0.35	SBP	0.0566	0.0433	1.91×10^{-1}	4.31×10^{-6}	1.47×10^{-8}	8.82×10^{-5}
FIGN-GRB14	rs7607074§†	2	164,898,256	T/A	0.51	SBP	−0.4084	0.0417	1.10×10^{-22}	2.43×10^{-4}	2.75×10^{-8}	8.50×10^{-5}
SLC4A7-EOMES	rs2029827*†‡	3	27,551,275	A/G	0.4	SBP	−0.4596	0.0424	2.40×10^{-27}	2.96×10^{-4}	1.06×10^{-10}	1.13×10^{-4}
SPON2	rs111822223†‡	4	1,179,966	A/G	0.07	SBP	0.217	0.0808	7.12×10^{-3}	1.82×10^{-5}	4.90×10^{-9}	9.37×10^{-5}
MINAR2-CHSY3	rs11955088†	5	129,206,827	A/G	0.57	SBP	0.0476	0.0419	2.56×10^{-1}	3.26×10^{-6}	2.36×10^{-8}	8.58×10^{-5}
GRIK2	rs144326314	6	101,863,870	G/GTCAA	0.11	SBP	−0.1517	0.0701	2.92×10^{-2}	1.18×10^{-5}	1.36×10^{-8}	8.86×10^{-5}
HOXA13-EVX1	rs115525024§	7	27,236,559	T/G	0.93	DBP	0.5121	0.0447	2.46×10^{-30}	3.31×10^{-4}	2.14×10^{-11}	1.21×10^{-4}
	rs10262140§†‡	7	27,256,464	T/C	0.94	SBP	0.977	0.0862	9.34×10^{-30}	3.24×10^{-4}	1.45×10^{-10}	1.11×10^{-4}
CRHR2	rs41413147†	7	30,723,336	C/T	0.01	SBP	−0.6249	0.19	9.93×10^{-4}	2.73×10^{-5}	3.75×10^{-8}	8.35×10^{-5}
IGFBP3-TNS3	rs11977526*†‡	7	46,008,110	G/A	0.4	SBP	−0.3802	0.0423	2.54×10^{-19}	2.04×10^{-4}	4.64×10^{-8}	8.24×10^{-5}
						PP	−0.0085	0.0005	6.33×10^{-66}	7.42×10^{-4}	4.86×10^{-12}	1.29×10^{-4}
TPK1-CNTNAP2	rs574733666	7	145,677,500	AT/A	0.09	SBP	0.1299	0.0742	7.42×10^{-2}	7.74×10^{-6}	4.37×10^{-8}	8.27×10^{-5}
LSP1	rs569550*†‡	11	1,887,068	T/G	0.39	SBP	0.6291	0.0427	3.49×10^{-49}	5.48×10^{-4}	4.29×10^{-17}	1.87×10^{-4}
						DBP	0.276	0.024	1.16×10^{-30}	3.34×10^{-4}	1.83×10^{-12}	1.33×10^{-4}
PPP6R3	rs547525853	11	68,243,714	C/T	0.01	SBP	0.3915	0.2036	5.14×10^{-2}	9.33×10^{-6}	2.17×10^{-8}	8.62×10^{-5}
DUSP16	rs12368847*†‡	12	12,682,123	G/A	0.27	SBP	−0.2912	0.0469	5.25×10^{-10}	9.74×10^{-5}	1.16×10^{-9}	1.01×10^{-4}
DLEU1-DLEU7	rs77827164	13	51,174,271	G/A	0.02	DBP	−0.0409	0.0869	6.38×10^{-1}	5.60×10^{-7}	4.73×10^{-8}	8.23×10^{-5}
LINC00564-SLITRK1	rs141943794	13	83,713,849	A/T	0.02	SBP	0.0189	0.1596	9.06×10^{-1}	3.56×10^{-8}	4.69×10^{-8}	8.24×10^{-5}
LINC00552-TMEM255B	rs376861852§	13	114,454,081	G/T	0.36	SBP	0.0799	0.0432	6.03×10^{-2}	8.64×10^{-6}	7.79×10^{-9}	9.14×10^{-5}
						DBP	0.1385	0.0243	1.13×10^{-8}	8.23×10^{-5}	6.63×10^{-9}	9.22×10^{-5}
TYRO3-MGA	rs182202119	15	41,918,628	A/T	0.01	SBP	0.3606	0.1964	6.22×10^{-2}	8.51×10^{-6}	3.29×10^{-8}	8.41×10^{-5}
FURIN	rs8032315§†‡	15	91,418,297	T/A	0.32	SBP	0.7136	0.0444	4.58×10^{-58}	6.51×10^{-4}	4.35×10^{-8}	8.27×10^{-5}
SPIRE2	rs34169212	16	89,912,736	C/CT	0.55	DBP	0.1212	0.0237	3.00×10^{-7}	6.63×10^{-5}	5.31×10^{-9}	9.33×10^{-5}
WIZ	rs113267381	19	15,550,532	G/A	0.02	SBP	−0.1549	0.175	3.76×10^{-1}	1.98×10^{-6}	2.18×10^{-9}	9.78×10^{-5}

Table 1. Genome-wide significant vQTLs associated with SBP, DBP, or PP. Novel loci are highlighted in bold. Positions are given in GRCh37 coordinates. SNP effects are in mm Hg per copy of the coded allele. *The SNP is, or in linkage disequilibrium (LD $R^2 \geq 0.8$) with, a BP-associated SNP in the GWAS catalog. †The SNP is an eQTL SNP in GTEx. ‡The SNP is an sQTL SNP in GTEx. §The SNP is in intermediate LD ($0.1 \leq$ LD $R^2 < 0.8$) with a BP-associated SNP in the GWAS catalog. Chr, chromosome; NCA, noncoded allele; CA, coded allele; CAF, coded allele frequency, SE, standard error.

approximately 0.1 for SBP and DBP and 0.04 for PP. Similarly, the relative contributions of the heteroscedasticity would be underestimated compared with that of all BP SNPs since the GWAS SNPs were largely detected by their genetic main effects. PP has the smallest relative contribution from the heteroscedasticity effects, possibly because PP is a derived trait and its random variation is larger than that in SBP and DBP.

Out of the 27 significant vQTLs, 13 are novel and are highlighted in Table 1. The P values of their genetic main effects ranged from 3.0×10^{-7} to 0.91 in this study and were not previously reported to be associated with BPs. The sizes of their SNP main effects, in terms of explained BP variance, are smaller than those of their heteroscedasticity effects. The total R^2 values of the SNP main effects and heteroscedasticity effects are 0.017% and 0.11%, respectively. Clearly, it would be difficult to identify this type of locus in standard GWASs that focus on testing SNP main effects only.

The 27 vQTLs are enriched for genetic loci that affect the expression (eQTL) or splicing (sQTL) of protein-coding genes. Querying the GTEx data version 8[27], 12 of the 23 top SNPs are eQTLs, and 8 top SNPs are both eQTLs and sQTLs. All of the eQTLs and sQTLs had target genes in cis. As many GWAS loci are associated with complex traits, the vQTLs likely play regulatory roles that mediate the BP associations as well. Interestingly, 8 of the 12 eQTLs were annotated as being both eQTLs and sQTLs, while cis-eQTLs have only a 12% overlap with cis-sQTLs in GTEx[27]. This agreed with the finding that sQTLs in GWAS results display stronger associations with complex traits than variants that exclusively affect gene expression[28,29].

Nonparametric methods are widely available for heteroscedasticity testing and have been used in vQTL analysis. They are not based on particular assumptions about trait distributions and are applicable for analyzing a wide range of quantitative traits. Of the BPs in this work, SBP and DBP approximately follow normal distributions. PP is highly skewed with a long and fat right tail; hence, I used logarithmic transformation. After adjusting for the covariate effects, residuals of SBP, DBP and the transformed PP have skewness values of 0.55, 0.36, 0.03 and kurtosis values of 0.61, 0.25, 0.23, respectively, although their normality tests were rejected given the large sample size of this study. I further conducted Levene's (Brown-Forsythe) test[30,31] on the 27 vQTLs identified by the variance component test. SNP genotypes were called those with the largest probabilities and residuals

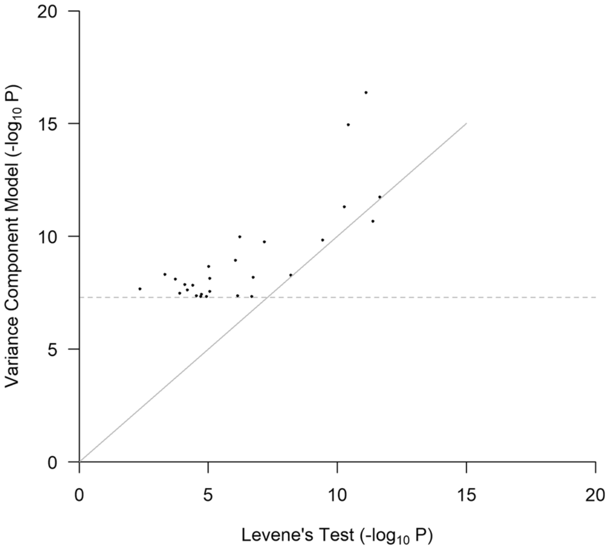

Figure 6. P values of the 27 vQTLs by Levene's test. The dashed line represents the genome-wide significance level ($P=5\times10^{-8}$).

after adjusting for covariate effects, and the SNP main effect was used to test the equality of variance in the three genotype groups. The P values of Levene's tests and the variance component test are compared in Fig. 6. It can be seen that Levene's tests provide varying levels of support for the 27 vQTLs. Unsurprisingly, the results of the variance component model were more significant than those of Levene's test, except for one vQTL. This agrees with the general case that parametric methods are more powerful when their underlying assumptions are approximately satisfied.

Discussion

I carried out genome-wide vQTL analysis by testing the allelic substitution effects on trait variance and identified 27 vQTLs associated with BPs. Such heteroscedasticity effects explained additional BP variance that was missed by GWASs. The heteroscedasticity effects of the 13 novel vQTLs were larger than their genetic main effects and were not detected by existing GWASs. In addition, 9 novel vQTLs demonstrated almost no genetic main effects, and their P values for testing SNP main effects were larger than 0.05 in this biobank-scale analysis. Complementary to GWAS, vQTL analysis has the potential to discover more variants associated with other complex traits.

On the other hand, if the heteroscedasticity effects harbor some gene–environment or gene–gene interactions, the overall interaction effects appear to be small compared with the additive main effects of GWAS SNPs. The heteroscedasticity effects of the 411 SBP-associated SNPs that were discovered by GWASs added up to 0.52%, compared with the 5.13% attributable to the SNP main effects. For the 20 vQTLs that were associated with SBP in the analysis, the total heteroscedasticity effects and genetic main effects were 0.20% and 0.30%, respectively. This is consistent with that additive main effects are the major sources of genetic variance[32,33], and detecting gene–environment interactions usually requires much larger sample sizes[34]. Considering the large sample size used in this study, the number and effect sizes of the vQTLs are small. Hence, the gene–environment and gene–gene interactions are unlikely to explain a major part of the "missing heritability"[35] of BPs.

While I provide statistical evidence supporting the vQTLs of BP traits, the results should be considered preliminary. In particular, I lack independent replication. Because of the so-called "winner's curse"[36], the reported effect sizes of the vQTLs in the discovery samples tend to be overestimated. Replication in external datasets would require much larger sample sizes, which implies that meta-analysis is necessary. To the best of my knowledge, methods and software that facilitate meta-analysis of results from variance component analysis are presently lacking. Alternatively, regression analysis of the vQTLs can be performed in the replication samples. Estimated τ_2 and τ_3, together with their variance and covariance, can be synthesized by the generalized least squares method[37]. In this way, the meta-analysis of vQTL is methodologically equivalent to the meta-analysis that jointly tests SNP main effects and interaction effects[38,39].

My discovery of the vQTLs is limited by the diversity of population. Samples in the vQTL analysis were restricted to individuals of European ancestry. Since the causal alleles that by chance have reached higher frequencies are more likely to be identified[40], analyzing samples of a single ancestry not only limits the transferability of results across populations, but results in ascertainment bias and missing the vQTLs that differ among diverse populations. My study is also limited by the geographical and environmental diversity. Complex traits are known to have a strong geographical component involving genetic predisposition and environmental exposure[41]. Effect sizes of the gene-environment interactions may be smaller in the study samples than in geographically and environmentally more diversified samples. Hence, many potential vQTLs could be missed in this study.

In this work, I focused on genome-wide vQTL analysis of BP data in the UK Biobank. The reported vQTLs may include some interaction effects that were not previously identified. As suggested in the literature[2,5,6,8], vQTL analysis can be used as a screening tool for prioritizing variants that may harbor interaction effects. Factors modulating the genetic effects can be hypothesized and tested thereafter, which is beyond the scope of this paper. Any interactions that can be detected and confirmed will positively contribute to the understanding of complex traits or diseases.

Data availability

This research has been conducted using the UK Biobank resource under application number 44080. The genetic and phenotype datasets are not publicly available but can be accessed via the UK Biobank data access process. More details are available at http://www.ukbiobank.ac.uk/register-apply/.

Received: 5 May 2022; Accepted: 18 July 2022
Published online: 25 July 2022

References

1. Rönnegård, L., Felleki, M., Fikse, F., Mulder, H. A. & Strandberg, E. Genetic heterogeneity of residual variance - estimation of variance components using double hierarchical generalized linear models. *Genet. Sel. Evol.* **42**(1), 8. https://doi.org/10.1186/1297-9686-42-8 (2010).
2. Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the Women's Genome Health Study. *PLoS Genet.* **6**(6), e1000981. https://doi.org/10.1371/journal.pgen.1000981 (2010).
3. Struchalin, M. V., Dehghan, A., Witteman, J. C., van Duijn, C. & Aulchenko, Y. S. Variance heterogeneity analysis for detection of potentially interacting genetic loci: Method and its limitations. *BMC Genet.* **11**, 92. https://doi.org/10.1186/1471-2156-11-92 (2010).
4. Rönnegård, L. & Valdar, W. Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics* **188**(2), 435–447. https://doi.org/10.1534/genetics.111.127068 (2011).
5. Young, A. I., Wauthier, F. L. & Donnelly, P. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nat. Genet.* **50**(11), 1608–1614. https://doi.org/10.1038/s41588-018-0225-6 (2018).
6. Wang, H. *et al.* Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci. Adv.* **5**(8), eaaw3538. https://doi.org/10.1126/sciadv.aaw3538 (2019).
7. Sulc, J. *et al.* Quantification of the overall contribution of gene-environment interaction for obesity-related traits. *Nat Commun.* **11**(1), 1385. https://doi.org/10.1038/s41467-020-15107-0 (2020).
8. Marderstein, A. R. *et al.* Leveraging phenotypic variability to identify genetic interactions in human phenotypes. *Am. J. Hum. Genet.* **108**(1), 49–67. https://doi.org/10.1016/j.ajhg.2020.11.016 (2021).
9. Forsberg, S. K. *et al.* The multi-allelic genetic architecture of a variance-heterogeneity locus for molybdenum concentration in leaves acts as a source of unexplained additive genetic variance. *PLoS Genet.* **11**(11), e1005648. https://doi.org/10.1371/journal.pgen.1005648 (2015).
10. Ek, W. E. *et al.* Genetic variants influencing phenotypic variance heterogeneity. *Hum Mol Genet.* **27**(5), 799–810. https://doi.org/10.1093/hmg/ddx441 (2018).
11. Rönnegård, L. & Valdar, W. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genet.* **13**, 63. https://doi.org/10.1186/1471-2156-13-63 (2012).
12. Sun, X., Elston, R., Morris, N. & Zhu, X. What is the significance of difference in phenotypic variability across SNP genotypes?. *Am. J. Hum. Genet.* **93**(2), 390–397. https://doi.org/10.1016/j.ajhg.2013.06.017 (2013).
13. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490**(7419), 267–272. https://doi.org/10.1038/nature11401 (2012).
14. Kerin, M. & Marchini, J. Inferring Gene-by-environment interactions with a Bayesian whole-genome regression model. *Am. J. Hum. Genet.* **107**(4), 698–713. https://doi.org/10.1016/j.ajhg.2020.08.009 (2020).
15. Dumitrascu, B., Darnell, G., Ayroles, J. & Engelhardt, B. E. Statistical tests for detecting variance effects in quantitative trait studies. *Bioinformatics* **35**(2), 200–210. https://doi.org/10.1093/bioinformatics/bty565 (2019).
16. Wolfinger, R., Tobias, R. & Sall, J. Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM J. Sci. Comput.* **15**(6), 1294–1310 (1994).
17. Self, S. G. & Liang, K. Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**(398), 605–610 (1987).
18. Breusch, T. S. & Pagan, A. R. Simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47**(5), 1287–1294 (1979).
19. Sudlow, C. *et al.* UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779 (2015).
20. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z (2018).
21. Tobin, M. D., Sheehan, N. A., Scurrah, K. J. & Burton, P. R. Adjusting for treatment effects in studies of quantitative traits: Antihypertensive therapy and systolic blood pressure. *Stat. Med.* **24**(19), 2911–2935. https://doi.org/10.1002/sim.2165 (2005).
22. Ehret, G. B. *et al.* The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nat. Genet.* **48**(10), 1171–1184. https://doi.org/10.1038/ng.3667 (2016).
23. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**(4), 381–385 (2008).
24. Liang, K. & Zeger, S. Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1), 13–22 (1986).
25. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78**(3), 691–692 (1991).
26. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120 (2019).
27. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776 (2020).
28. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**(6285), 600–604. https://doi.org/10.1126/science.aad9417 (2016).
29. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**(1), 727. https://doi.org/10.1038/s41467-020-20578-2 (2021).
30. Levene, H. *Robust tests for equality of variances* 278–292 (Stanford University Press, Stanford, 1960).
31. Brown, M. B. & Forsythe, A. B. Robust tests for the equality of variances. *J. Am. Stat. Assoc.* **69**(346), 364–367 (1974).
32. Hivert, V. *et al.* Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *Am. J. Hum. Genet.* **108**(5), 786–798. https://doi.org/10.1016/j.ajhg.2021.02.014 (2021).

33. Pazokitoroudi, A., Chiu, A. M., Burch, K. S., Pasaniuc, B. & Sankararaman, S. Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. *Am. J. Hum. Genet.* **108**(5), 799–808. https://doi.org/10.1016/j.ajhg.2021.03.018 (2021).
34. Thomas, D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu. Rev. Public Health.* **31**, 21–36. https://doi.org/10.1146/annurev.publhealth.012809.103619 (2010).
35. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**(7265), 747–753. https://doi.org/10.1038/nature08494 (2009).
36. Yu, K. *et al.* Flexible design for following up positive findings. *Am. J. Hum. Genet.* **81**(3), 540–551. https://doi.org/10.1086/520678 (2007).
37. Becker, B. J. & Wu, M. J. The synthesis of regression slopes in meta-analysis. *Stat. Sci.* **22**(3), 414–429 (2007).
38. Manning, A. K. *et al.* Meta-analysis of gene-environment interaction: Joint estimation of SNP and SNP × environment regression coefficients. *Genet. Epidemiol.* **35**(1), 11–18. https://doi.org/10.1002/gepi.20546 (2011).
39. Jin, Q. & Shi, G. Meta-analysis of joint test of SNP and SNP-environment interaction with heterogeneity. *Hum Hered.* **86**(1–4), 1–9. https://doi.org/10.1159/000519098 (2021).
40. Amorim, C. E. G. *et al.* The population genetics of human disease: The case of recessive, lethal mutations. *PLoS Genet.* **13**(9), e1006915 (2017).
41. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2**, 9 (2019).

Acknowledgements

This work was supported by the national Thousand Youth Talents Plan. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Author contributions

G.S. developed the methods, conducted the analysis and wrote the manuscript.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to G.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.