Statistical Methods for Testing Genetic Pleiotropy

Daniel J. Schaid,*,1 Xingwei Tong,† Beth Larrabee,* Richard B. Kennedy,‡ Gregory A. Poland,‡
and Jason P. Sinnwell*

*Department of Health Sciences Research and ‡Mayo Clinic Vaccine Research Group, Mayo Clinic, Rochester, Minnesota 55905,
and †School of Statistics, Beijing Normal University, 100875 Beijing, China

ABSTRACT Genetic pleiotropy is when a single gene influences more than one trait. Detecting pleiotropy and understanding its causes can improve the biological understanding of a gene in multiple ways, yet current multivariate methods to evaluate pleiotropy test the null hypothesis that none of the traits are associated with a variant; departures from the null could be driven by just one associated trait. A formal test of pleiotropy should assume a null hypothesis that one or no traits are associated with a genetic variant. For the special case of two traits, one can construct this null hypothesis based on the intersection-union (IU) test, which rejects the null hypothesis only if the null hypotheses of no association for both traits are rejected. To allow for more than two traits, we developed a new likelihood-ratio test for pleiotropy. We then extended the testing framework to a sequential approach to test the null hypothesis that $k + 1$ traits are associated, given that the null of k traits are associated was rejected. This provides a formal testing framework to determine the number of traits associated with a genetic variant, while accounting for correlations among the traits. By simulations, we illustrate the type I error rate and power of our new methods; describe how they are influenced by sample size, the number of traits, and the trait correlations; and apply the new methods to multivariate immune phenotypes in response to smallpox vaccination. Our new approach provides a quantitative assessment of pleiotropy, enhancing current analytic practice.

KEYWORDS constrained model; likelihood-ratio test; multivariate analysis; seemingly unrelated regression; sequential testing

GENETIC pleiotropy is when a single gene influences more than one trait. Detecting pleiotropy, and understanding its causes, can improve the biological understanding of a gene in multiple ways: (1) There is potential to expand understanding of the medical impact of a gene, such as in phenome-wide association studies (Denny *et al.* 2013); (2) the pharmacologic genetic target could affect multiple traits or diseases, allowing a drug developed for a disease to be repurposed for other diseases or suggesting that a toxicity should be monitored for multiple traits; and (3) joint analysis of multiple traits can increase accuracy of phenotype prediction (Maier *et al.* 2015). Yet, understanding pleiotropy can be challenging. A gene can be associated with more than one trait for many reasons, such as when a single genetic variant directly influences multiple traits or when different variants within a gene influence different traits. Alternatively, the association of a gene with some of the traits can be indirect,

such as when a gene directly influences a trait, and that trait directly influences a second trait; the gene and the second trait are indirectly associated. The association of a gene with multiple traits can also result from spurious associations. One cause of spurious association is when subjects with more than one disease symptom are more likely ascertained for a study than if they had only one symptom—called Berkson's bias (Berkson 1946). A second cause is misclassification between two similar traits, a common problem for some psychiatric conditions. A third cause is when a genetic marker is in linkage disequilibrium with each of two causal loci (Gianola *et al.* 2015). These types of biases, and a thorough review of pleiotropy with numerous examples, are nicely summarized elsewhere (Solovieff *et al.* 2013). Despite the great deal of attention given to pleiotropy, most statistical tests do not formally test pleiotropy. Rather, they test the null hypothesis that no trait is associated with a variant; rejecting this null could be due to just one associated trait, not a situation of pleiotropy. The aim of this report is to provide a formal statistical method to assess pleiotropy to infer the number of traits associated with a variant.

Statistical methods to evaluate pleiotropy have been developed from different angles, ranging from comparison of univariate marginal associations of a genetic variant with

multiple traits, to multivariate analyses with simultaneous regression of all traits on a genetic variant, to reversed regression of a genetic variant on all traits. A brief survey of statistical methods for pleiotropy is provided here, with more details provided elsewhere (Schriner 2012; Yang and Wang 2012; Solovieff *et al.* 2013; Zhang *et al.* 2014). Univariate analyses are often based on comparison of variant-specific *P*-values across multiple traits. Although simple and feasible for meta-analyses, this approach ignores correlation among the traits and is based on *post hoc* analyses. More formal meta-analysis methods aggregate *P*-values to test whether any traits are associated with a variant, yet a significant association could be driven by just one trait. A slightly more sophisticated approach, also based on summary *P*-values, tests whether the distribution of *P*-values differs from the null distribution of no associations beyond those already detected (Cotsapas *et al.* 2011). Descriptions of additional univariate methods are given elsewhere (Solovieff *et al.* 2013).

Multivariate methods have been popular for quantitative traits. Although different statistical methods have been proposed, some of them result in the same statistical tests. The following three approaches to analyze quantitative traits result in the same *F*-statistic to test whether any of the traits are associated with a genetic variant: (1) simultaneous regression of all traits on a single variant [for example, using the statistical software R function $\text{lm}(\mathbf{Y} \sim \mathbf{g})$, where $\mathbf{Y}$ is a matrix of traits and $\mathbf{g}$ a vector for a single genetic variant coded as 0, 1, 2 for the dose of the minor allele], (2) regression of the minor allele dose on all traits ($\text{lm}(\mathbf{g} \sim \mathbf{Y})$), and (3) canonical correlation of $\mathbf{Y}$ with $\mathbf{g}$ [using either plink.multivariate (Ferreira and Purcell 2009) or R code given in *Appendix A*]. The regression of the dose of the minor allele on all traits is a convenient approach, particularly if some of the traits are binary. A slightly different approach is to account for the categorical nature of the dose of the minor allele: Instead of using linear regression, use ordinal logistic regression of the dose on the traits [R MultiPhen package (O'Reilly *et al.* 2012)]. An advantage of this approach is that it allows for binary traits, unlike most methods that assume traits are quantitative with a multivariate normal distribution. However, score tests for generalized linear models, based on estimating equations, have been developed as a way to simultaneously test multiple traits, some of which could be binary (Xu and Pan 2015). An approach somewhat between univariate and multivariate is based on reducing the dimension of the multiple traits by principal components (PC) and using a reduced set of PCs as either the dependent or the independent variables in regression. A comparison of univariate and multivariate approaches found that multivariate methods based on multivariate normality {*e.g.*, canonical correlation, linear regression of traits on minor allele dose, reverse regression, MultiPhen, and Bayes methods [BIMBAM (Stephens 2013) and SNPTEST (Marchini *et al.* 2007)]} all had similar power and were generally more powerful than univariate methods (Galesloot *et al.* 2014).
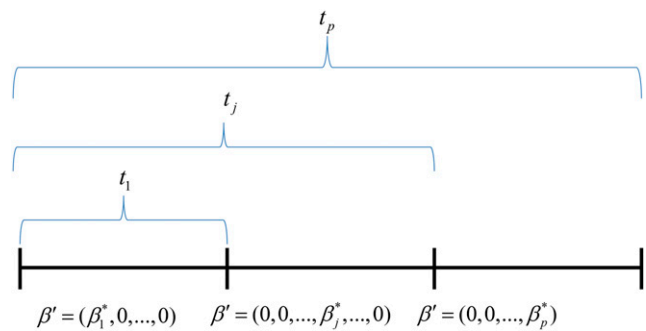
Figure 1 Example to illustrate why the pleiotropy LRT has an approximate χ^2_{p-1} distribution when only one β_j differs from zero. For this example, $\beta_1 \neq 0$ and $\beta_j = 0$ ($j \neq i$). Then, t_1 will be the minimum because it measures the sum of squared differences of the fitted values for the unconstrained ordinary least-squares model and the constrained model, which in this case is correctly specified. Because all other $t_j (j \neq 1)$ represent misspecified models, their values can become arbitrarily large as n increases. Hence, the correctly specified model will have the smallest values of t_j. And the distribution of t_j for a correctly specified model is χ^2_{p-1}.

The power advantage of multivariate over univariate methods occurs when the direction of the residual correlation is opposite from that of the genetic correlation induced by the causal variant (Liu *et al.* 2009; Galesloot *et al.* 2014). In addition to the methods discussed above, a few new approaches have been proposed, but have not yet been compared with others. An interesting approach is to scale the different traits by their standard deviation and then assume that the effect of a single-nucleotide polymorphism (SNP) is constant across all traits to construct a test of association with 1 d.f.—so-called "scaled marginal models" (Roy *et al.* 2003; Schifano *et al.* 2013). Finally, an approach based on kernel machine regression extended the sequential kernel association test (Wu *et al.* 2010) to multiple traits, providing a simultaneous test of multiple traits with multiple genetic variants in a genomic region (Maity *et al.* 2012).

A limitation of all current approaches is that they test whether any traits are associated with a genetic variant, and small *P*-values could be driven by the association of the genetic variant with a single trait. Hence, *post hoc* analyses are required to interpret the possibility of pleiotropy. This can be quite challenging when scaling up to a large number of genetic variants. Another significant challenge is to distinguish direct from indirect associations. When there is evidence that a secondary trait is associated with a genetic marker, and one wishes to distinguish whether the same genetic marker has a direct effect on a primary trait *vs.* an indirect effect, with the secondary trait acting as a mediator between the genetic marker and the primary trait, ideas from causal modeling have proved useful. For example, disentangling direct from indirect effects can be achieved by regressing the primary trait on the secondary trait, the genetic marker, and all other covariates shared between the primary and secondary traits. Results from this regression can be used to construct an adjusted primary trait that can then be used in

Table 1 Empirical type I error rate for common correlation structure when $\beta_1 = 1$ and all other $\beta_j = 0$ ($j \neq 1$), based on multivariate normal distribution

Sample size	No. traits	Trait correlation	Nominal type I error rate	
			0.05	0.01
100	4	0.2	0.072	0.020
		0.5	0.070	0.016
		0.8	0.066	0.014
	10	0.2	0.105	0.036
		0.5	0.092	0.032
		0.8	0.094	0.029
500	4	0.2	0.056	0.017
		0.5	0.058	0.011
		0.8	0.058	0.009
	10	0.2	0.061	0.010
		0.5	0.056	0.011
		0.8	0.068	0.019
1000	4	0.2	0.052	0.012
		0.5	0.058	0.008
		0.8	0.051	0.010
	10	0.2	0.057	0.012
		0.5	0.052	0.010
		0.8	0.046	0.007

Underlined values are for when empirical type I error exceeds the upper 95% C.I.

Table 2 Empirical type I error rate for random correlation structure when $\beta_1 = 1$ and all other $\beta_j = 0$ ($j \neq 1$), based on multivariate normal distribution

Sample size	No. traits	Trait correlation	Nominal type I error rate	
			0.05	0.01
100	4	0–0.2	0.066	0.011
		0.2–0.5	0.065	0.013
		0.5–0.8	0.055	0.011
	10	0–0.2	0.077	0.018
		0.2–0.5	0.106	0.029
		0.5–0.8	0.112	0.039
500	4	0–0.2	0.065	0.007
		0.2–0.5	0.049	0.015
		0.5–0.8	0.047	0.008
	10	0–0.2	0.049	0.011
		0.2–0.5	0.067	0.016
		0.5–0.8	0.053	0.012
1000	4	0–0.2	0.042	0.007
		0.2–0.5	0.055	0.009
		0.5–0.8	0.043	0.009
	10	0–0.2	0.072	0.012
		0.2–0.5	0.056	0.012
		0.5–0.8	0.056	0.011

Underlined values are for when empirical type I error exceeds the upper 95% C.I.

subsequent analyses (Vansteelandt *et al.* 2009). Another approach is based on Bayesian methods to partition associations into unassociated, indirect, and direct associations. However, it is difficult to accurately classify the type of causal association, particularly when residual correlations are large (*e.g.*, it is difficult to discriminate between direct and indirect effects) (Stephens 2013).

The above methods are used to test whether a single genetic variant is associated with multiple traits. When scaling up to genome-wide data, it has been useful to use all the genetic markers to estimate the marker-predicted heritability of a trait. This has recently been extended to multiple traits to estimate pleiotropy as the genetic correlation of multiple traits. Mixed models are used to partition the phenotype correlations into genetic correlation (*i.e.*, correlation of polygenic total genetic values) and environmental correlation (Korte *et al.* 2012; Lee *et al.* 2012; Zhou and Stephens 2014; Furlotte and Eskin 2015). Although this approach does not evaluate whether particular SNPs or particular genomic regions are the cause for phenotype correlations, it has the potential to guide design of studies that focus on pleiotropy. For example, the correlation of two phenotypes can be partitioned as $r_P = h_1 h_2 r_g + e_1 e_2 r_e$, where h_i^2 is the heritability of trait i, $e_i^2 = 1 - h_i^2$, r_g is the genetic correlation, and r_e is the environmental correlation (Falconer and Mackay 1996, p. 314). Heritability in the narrow sense is the percentage of the variance of the trait explained by additive genetic factors. This illustrates that if both traits have low heritability, the phenotype correlation is primarily due to environmental correlation (and nonadditive genetic effects that are missed by r_g), implying that large sample sizes would be needed to test pleiotropy when there are small genetic effects.

We have emphasized that current methods to evaluate pleiotropy do not perform a formal test of the null hypothesis of no pleiotropy. For the special case of two traits, one can construct a null hypothesis of no pleiotropy based on the intersection-union (IU) test (Silvapulle and Sen 2004). Consider the regression equation $y_j = \beta_{o,j} + \beta_{1,j} g + e_j$, where y_j is the vector of values for the jth trait, $\beta_{o,j}$ is the intercept, $\beta_{1,i}$ is the slope association parameter of interest, g is the vector of doses for the minor allele, and e_j is a vector of residuals. The union null hypothesis is $H_0 : \beta_{1,1} = 0$ or $\beta_{1,2} = 0$, and the intersection alternative hypothesis is $H_1 : \beta_{1,1} \neq 0$ and $\beta_{1,2} \neq 0$. Testing each $\beta_{1,i}$ at a desired type I error, say $\alpha = 0.05$, the null is rejected only if both tests reject. There is no need to correct for multiple testing, because the type I error rate is not inflated by this procedure. But this approach can be conservative, particularly if the two tests are uncorrelated. The IU test can be extended to $p > 2$ traits, but rejection of the null would occur only when all p tests are significant at the specified α. For our situation, we wish to reject the null if at least two of the p tests reject. One approach would be to apply the IU test to each pair of traits and reject the null if at least one of the IU tests rejects. But this would entail many pairs of tests, and for this situation one would need to correct for testing multiple pairs. Bonferroni correction would lead to an overly conservative test.

Because of current limitations, we developed a likelihood-ratio test for testing the null hypothesis of no pleiotropy—the null hypothesis that one or no traits are associated with a genetic variant *vs.* the alternative hypothesis that two or more traits are associated. We then extended the testing framework to test the null hypothesis that k or fewer traits are associated *vs.* the alternative hypothesis that more than k

Table 3 Empirical type I error rate for common correlation structure when all $\beta_j = 0$, based on multivariate normal distribution

Sample size	No. traits	Trait correlation	Nominal type I error rate	
			0.05	0.01
100	4	0.2	0.005	0.002
		0.5	0.006	0.001
		0.8	0.009	0.002
	10	0.2	0.015	0
		0.5	0.011	0
		0.8	0.014	0.003
500	4	0.2	0.003	0
		0.5	0.005	0.001
		0.8	0.011	0.001
	10	0.2	0.005	0.001
		0.5	0.003	0.001
		0.8	0.009	0.001
1000	4	0.2	0.005	0
		0.5	0.008	0
		0.8	0.004	0.001
	10	0.2	0.011	0.002
		0.5	0.012	0
		0.8	0.009	0.002

Table 4 Empirical type I error rate for random correlation structure when all $\beta_j = 0$, based on multivariate normal distribution

Sample size	No. traits	Trait correlation	Nominal type I error rate	
			0.05	0.01
100	4	0–0.2	0.005	0
		0.2–0.5	0.012	0.001
		0.5–0.8	0.014	0.001
	10	0–0.2	0.018	0.002
		0.2–0.5	0.01	0.002
		0.5–0.8	0.029	0.004
500	4	0–0.2	0.004	0
		0.2–0.5	0.009	0
		0.5–0.8	0.007	0.001
	10	0–0.2	0.009	0
		0.2–0.5	0.009	0.001
		0.5–0.8	0.019	0.005
1000	4	0–0.2	0.006	0.001
		0.2–0.5	0.006	0
		0.5–0.8	0.004	0
	10	0–0.2	0.01	0.002
		0.2–0.5	0.01	0
		0.5–0.8	0.01	0

traits are associated ($k = 0, 1, ...p - 1$). By this generalization, we propose sequential testing to test the null hypothesis that $k + 1$ traits are associated, given that the null hypothesis of k traits are associated was rejected. This sequential approach provides a refined approach to evaluate how many traits, and which traits, are associated with a genetic variant, accounting for correlation among the traits and possibly adjusting for covariates that could differ across the traits.

Methods

Likelihood-ratio test of pleiotropy: null of one or fewer traits

Suppose that p traits are measured on each of n subjects. Let $y_j' = (y_{j1}, ..., y_{jn})$ denote the vector of measures on the jth trait for n subjects. Assume that each trait is modeled by linear regression, denoted

$$y_j = x\beta_j + \varepsilon_j,$$

where x is the dose of the minor allele for n subjects. Also assume that all y_j and x are centered, so intercepts can be ignored. For simplicity of presentaiton, we ignore adjusting covariates, but our methods are general and allow for trait-specific covariates. By stacking vectors, we can express the model as $y = X\beta + \varepsilon$, where $y' = (y_1', ..., y_p')$, $X = \text{diag}(x)$, $\beta' = (\beta_1', ..., \beta_p')$, and $\varepsilon' = (\varepsilon_1', ..., \varepsilon_p')$. The error term $\varepsilon \sim N(0, \Omega)$, where $\Omega = \Sigma \otimes I$, I is an $n \times n$ identity matrix, $\otimes$ is the Kronecker product, and the $p \times p$ matrix Σ is the covariance matrix for the within-subject covariances of the errors. Under this model, the log-likelihood function of (β, Σ) is given by

$$l_n(\beta, \Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}(y - X\beta)'\left(\Sigma^{-1} \otimes I\right)(y - X\beta).$$

Suppose that the covariance Σ is known; otherwise, we can obtain a consistent estimate by maximum-likelihood estimation. For example, we can estimate β by using methods from seemingly unrelated regression, an approach called feasible generalized least squares. Separate ordinary linear regression for each trait can be used to obtain residuals to estimate $\widehat{\Sigma}$, and then this is used in the generalized least-squares (GLS) solution,

$$\widehat{\beta} = \left[X'\left(\widehat{\Sigma}^{-1} \otimes I\right)X\right]^{-1} X'\left(\widehat{\Sigma}^{-1} \otimes I\right)y.$$

Note that the feasible generalized least squares is asymptotically equivalent to maximum-likelihood estimation (MLE). There are two special cases when separate ordinary regressions and GLS result in the same solution: (1) when Σ is a diagonal matrix and (2) when the regressors in X_j are the same for all traits. Hence, for the case where each trait is regressed on the same x, without additional adjusting covariates, separate ordinary least-squares regression and GLS give the same results. The covariance matrix of the residuals then provides a consistent estimate of Σ. Then, the Cholesky decompositon of Ω is $\Omega = \Omega^{1/2}\Omega^{1/2}$, where $\Omega^{1/2} = \Sigma^{1/2} \otimes I$ and $\Omega^{-1/2} = \Sigma^{-1/2} \otimes I$. We then decorrelate the data by $\tilde{y} = \Omega^{-1/2}y$ and $\tilde{X} = \Omega^{-1/2}X$, to transform the model to $\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}$, where $\tilde{\varepsilon} = \Omega^{-1/2}\varepsilon \sim N(0, I_{np})$, which has log-likelihood $l_n(\beta) = -(1/2)(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta)$. Based on this log likelihood, we derived the likelihood-ratio test (LRT) to test the null hypothesis of no pleiotropy: One or no traits are associated with a genetic variant. Below we outline how to compute the LRT and provide details of the derivations in *Appendix B*.

Table 5 Empirical type I error rate when $\beta_1 = 1$ and all other $\beta_j = 0$ $(j \neq 1)$, based on multivariate t distribution with 3 d.f., with common correlation structure

Sample size	No. traits	Trait correlation	Nominal type I error rate	
			0.05	0.01
100	4	0.2	0.042	0.011
		0.5	<u>0.067</u>	<u>0.019</u>
		0.8	0.057	0.015
	10	0.2	<u>0.088</u>	<u>0.018</u>
		0.5	<u>0.104</u>	<u>0.028</u>
		0.8	<u>0.094</u>	<u>0.030</u>
500	4	0.2	0.059	0.011
		0.5	0.038	0.007
		0.8	0.043	0.009
	10	0.2	0.041	0.016
		0.5	0.059	<u>0.021</u>
		0.8	0.050	0.010
1000	4	0.2	0.052	0.006
		0.5	0.047	0.006
		0.8	0.054	0.009
	10	0.2	0.054	0.010
		0.5	0.058	0.015
		0.8	0.055	0.016

Underlined values are for when empirical type I error exceeds the upper 95% C.I.

Table 6 Empirical type I error rate when all $\beta_j = 0$, based on multivariate t distribution with 3 d.f., with common correlation structure

Sample size	No. traits	Trait correlation	Nominal type I error rate	
			0.05	0.01
100	4	0.2	0.009	0
		0.5	0.007	0
		0.8	0.011	0.002
	10	0.2	0.015	0.001
		0.5	0.015	0.002
		0.8	0.013	0.003
500	4	0.2	0.003	0
		0.5	0.005	0
		0.8	0.005	0
	10	0.2	0.009	0.001
		0.5	0.008	0
		0.8	0.008	0.001
1000	4	0.2	0.004	0.001
		0.5	0.009	0.001
		0.8	0.007	0.001
	10	0.2	0.011	0.001
		0.5	0.002	0
		0.8	0.011	0.001

The null hypothesis of no pleiotropy can be expressed as

H_0 : Of the parameters $\beta_1, \dots, \beta_p$, there exists at most one that is nonzero $\leftrightarrow H_1$: otherwise.

The null hypothesis is equivalent to testing whether one of the following $p + 1$ tests holds,

$$H_{k0} : \beta_k \neq 0, \beta_j = 0 (j \neq k),$$

for $k = 0, \dots, p$. Note that H_{00} represents all $\beta_k = 0$ $(k = 1, \dots, p)$, while for $k > 0$, H_{k0} allows $\beta_k \neq 0$ while all other $\beta_j = 0$ $(j \neq k)$. To represent these $p + 1$ hypotheses, we use $H_{k0} : V_k\beta = 0$. Let V_0 be a matrix such that $H_{00} : V_0\beta = 0$ tests whether all $\beta_j = 0$. This is the usual multivariate test. In this case, V_0 is the identity matrix of dimension p. To construct V_k $(k > 0)$, create an identity matrix of dimension p and then remove the kth row. This results in $V_k\beta = (\beta_1, \dots, \beta_{k-1}, \beta_{k+1}\beta_p)'$. Then, the null hypothesis is equivalent to

H_0 : There exists one of $H_{k0} : V_k\beta = 0$, for $k = 0, \dots, p$.

To construct the LRT, center y and x about their means, use ordinary least squares to estimate β, use the residuals to estimate Σ, and then use Σ to decorrelate y and X according to $\tilde{y} = \Omega^{-1/2}y$, $\tilde{X} = \Omega^{-1/2}X$, where $\Omega^{-1/2} = \Sigma^{-1/2} \otimes I$. Then, for each $k = 0, \dots, p$, compute

$$t_k = \tilde{y}'\tilde{X}\left(\tilde{X}'\tilde{X}\right)^{-1}V_k\left[V_k\left(\tilde{X}'\tilde{X}\right)^{-1}V_k'\right]^{-1}V_k\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{y}.$$

An alternative way to express t_k is $t_k = \left\|\tilde{X}\beta_n - \tilde{X}\beta_{V_k}\right\|^2$, the squared l_2 norm between the fitted values based on the ordinary least-squares estimates, β_n, and the fitted values based on the constrained estimates, β_{V_k} (see *Appendix B*).

As shown in *Appendix B*, the LRT is

$$T = \min_{k = 0, \dots, p} t_k.$$

Because t_j is based on the sum of squared differences of the fitted values between the unconstrained and constrained models, for a correctly specified constrained model, t_j has a χ^2 distribution. But the distribution of T is more complicated. The statistic T has two different asymptotic distributions depending on when $\beta = 0$ or not. When $\beta = 0$, the asymptotic distribution of each t_j is a χ^2 distribution, yet the distribution of the minimum of them, T, is unknown. Alternatively, when $\beta = 0$, we can use the commonly used χ^2 test for the null hypothesis that all $\beta_j = 0$. This motivates us to do the test by two stages. The first stage is to just test $H_{00} : \beta = 0$, using the statistic $t_0 \sim \chi_p^2$ as the test statistic, so we reject H_{00} if $t_0 > \chi_p^2(\alpha)$, where $\chi_p^2(\alpha)$ is the $1 - \alpha$ quantile of the χ^2 distribution with p d.f. If H_{00} cannot be rejected, then the H_0 of no pleiotropy cannot be rejected. If H_{00} is rejected, we turn to the second stage to test the null hypothesis that one H_{k0} holds for $k = 1, \dots, p$. For this we ignore t_0 and use the test statistic

$$T_1 = \min_{k = 1, \dots, p} t_k.$$

Since $T_1 \sim \chi_{p-1}^2$, we reject the null hypothesis that only one H_{k0} holds for $k = 1, \dots, p$ if $T_1 > \chi_{p-1}^2(\alpha)$. Then, the null hypothesis H_0 of no pleiotropy is rejected only if both H_{00} is rejected and the null hypothesis that only one H_{k0} holds is rejected $(k = 1, \dots, p)$.

To provide intuition why T_1 has a large sample χ^2 distribution with $(p - 1)$ d.f. when only one β_j differs from zero, while all others equal zero, we present an example in
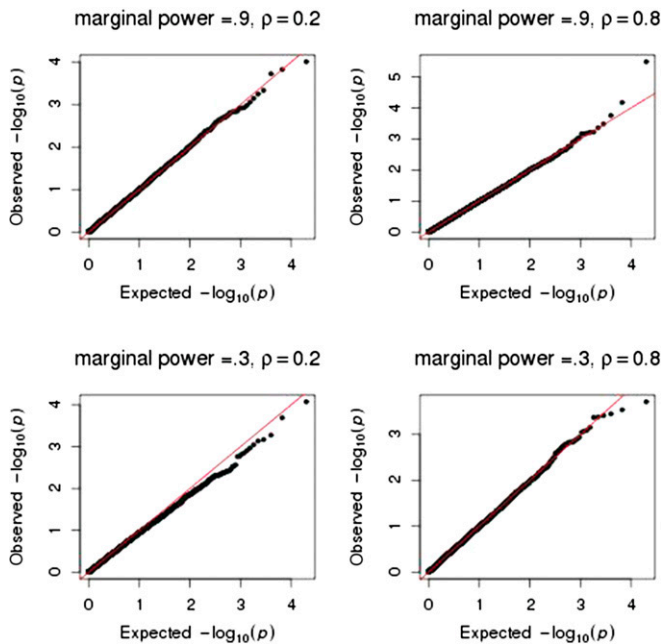
Figure 2 Quantile–quantile plots of *P*-values to test the null hypothesis of no pleiotropy. Sample size was 1000 with four equally correlated traits (ρ = 0.2 or 0.8). All but one β was zero; the nonzero β was chosen such that there was either 30% or 90% power to detect its marginal effect using α = 0.001. A total of 10,000 simulations were performed.

Figure 1. In this example $\beta_1 \neq 0$ and $\beta_j = 0$ $(j \neq 1)$. As shown in *Appendix B* (*Corollary 1*), the distribution of t_j for a correctly specified model is χ^2_{p-1}. In contrast, the incorrect models result in arbitrarily large values of t_j (see *Corollary 2* in *Appendix B*). This means that t_1 will be minimum and $T = t_1 \sim \chi^2_{p-1}$.

General likelihood-ratio sequential testing: null of K associated traits

The above sequential approach is based on testing the null hypothesis $H_{00} : \boldsymbol{\beta} = 0$, and then if this rejects, to turn to the second stage to test the null hypothesis that only one $H_{k0} : \beta_k \neq 0, \beta_j = 0$ $(j \neq k)$ holds for $k = 1, \ldots, p$. The advantage of this approach is that if H_{00} is rejected and the null hypothesis that only one H_{k0} holds is accepted, we can conclude that there is only one nonzero β. But if the null hypothesis that only one H_{k0} holds is rejected, we cannot make a firm conclusion about the number of traits associated with a genetic variant. To provide a more rigorous testing framework, we extended our approach to sequentially test the null hypothesis that a specified number of β's are nonzero. So, if the null hypothesis that k β's are nonzero is rejected, but the null hypothesis that $k + 1$ β's are nonzero is accepted, we can conclude there are $k + 1$ traits associated with a genetic variant. Furthermore, because the sequential testing is based on a likelihood-ratio framework, evaluating all possible combinations of nonzero β's, the combination that fails to reject the null hypothesis provides evidence of which traits are associated with the genetic variant. The de-

tails of the statistical procedures of this general sequential testing method are provided in *Appendix B*, as well as a proof that the type I error is controlled. In summary, this general sequential procedure provides a formal way to determine not only they number of traits associated with a genetic variant, but also which traits are associated.

Simulations

To evaluate the adequacy of the χ^2 distribution for the LRT, we performed simulations. For the pleiotropy null, we performed two sets of simulations. The first one assumed that all $\beta_j = 0$, the usual null for multivariate data. The second one fixed $\beta_1 = 1$ and all other $\beta_j = 0$ $(j = 2, \ldots, p)$. The value of $\beta_1 = 1$ was chosen because the power for detecting this marginal effect size was very large for our setup. We assumed three different sample sizes, $n = 100, 500, 1000$, and two different values of $p = 4, 10$. The small sample size of $n = 100$ was used to evaluate the adequacy of our asymptotic derivations for small samples. The variance of the errors was assumed to be 1, and the covariance was assumed to be either a constant ρ for all pairs of traits (*i.e.*, exchangeable correlation structure) or a range of covariances. For the range of covariances, we randomly chose covariances from a specified range, assuming a uniform distribution of the covariances. With a specified covariance structure, we simulated the random errors from either a multivariate normal distribution or a multivariate *t* distribution with 3 d.f., to evaluate the impact of heavy-tailed distributions. For all simulations, a single SNP was simulated, assuming a minor allele frequency of 0.2.

To evaluate the power of our proposed LRT for pleiotropy, we simulated 10 traits from a multivariate normal distribution with variances of 1 and equal covariances among the traits, set at $\rho = 0.2, 0.5, or 0.8$, for a total of $n = 500$ subjects. The number of traits associated with the SNP ranged over two, three, or five. The marginal effect of a trait was set at $\beta = 0.25$. This effect size explains 2% of the variation of a trait, and there is 90% power to detect a marginal effect of this size, using nominal $\alpha = 0.05$. We also set the marginal effect to $\beta = 0.2$, which corresponds to an explained 1.2% of the variation of a trait, and there is 70% power to detect a marginal effect of this size. All simulations were repeated 1000 times.

Data application

Our newly developed LRT for pleiotropy was applied to a data set that has 10 immunologic phenotypes measured in response to primary smallpox vaccination. These phenotypes included measures of humoral immunity (neutralizing antibody titer) and cellular immunity [two separate IFNγ ELISPOT assays and cytokine secretion upon viral stimulation as measured by ELISA (IL-1β, IL-2, IL-6, IL-12p40, IFNα, IFNγ, TNFα)]. All 645 subjects included in the presented analyses were of Caucasian ancestry. All subjects provided informed consent for use of their samples and this study was approved by the Mayo Clinic Institutional Review Board. A genome-wide association of the 10 phenotypes was performed, with each phenotype adjusted for relevant covariates

Table 7   Power to detect pleiotropy when associated traits have β = 0.25 (explain 2% trait variation; power = 90% for marginal effect)

No. associated traits (β = 0.25)	Trait correlation	Nominal type I error rate	
		0.05	0.01
2	0.2	0.503	0.237
	0.5	0.801	0.581
	0.8	0.971	0.900
3	0.2	0.859	0.677
	0.5	0.956	0.858
	0.8	0.999	0.998
5	0.2	0.980	0.928
	0.5	0.999	0.985
	0.8	1.000	1.000

Shown is a multivariate normal distribution with equal correlation structure, for a sample size of 500 subjects and minor allele frequency of genetic variant set to 0.20.

Table 8   Power to detect pleiotropy when associated traits have β = 0.2 (explain 1.2% trait variation; power = 70% for marginal effect)

No. associated traits (β = 0.2)	Trait correlation	Nominal type I error rate	
		0.05	0.01
2	0.2	0.220	0.069
	0.5	0.393	0.165
	0.8	0.964	0.850
3	0.2	0.494	0.255
	0.5	0.715	0.478
	0.8	1.000	0.991
5	0.2	0.842	0.636
	0.5	0.907	0.750
	0.8	1.000	1.000

Shown is a multivariate normal distribution with equal correlation structure, for a sample size of 500 subjects and minor allele frequency of genetic variant set to 0.20.

(*i.e.*, *P*-value <0.10 for association of a covariate with the phenotype, including eigenvectors to adjust for potential population stratification). Details of the study can be found in prior published reports (Kennedy *et al.* 2012a,b; Ovsyannikova *et al.* 2012a,b, 2013, 2014).

Data availability

Software implementing the proposed tests for pleiotropy for quantitative traits is available as an R package called "pleio" in the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/pleio/index.html).

Results

Simulation results

The type I error rates based on simulations are presented in Table 1, Table 2, Table 3, Table 4, Table 5, and Table 6. For all simulations, we show results from the two-stage test (using t_0 for stage 1 and $T_1 = \min\{t_k; k = 1, ..., p\}$ for stage 2), but in all cases, the results from the two-stage test were identical to those from the compound pleiotropy test $T = \min\{t_k; k = 0, ..., p\}$. The results for when only one β_j differs from zero (Table 1, Table 2, and Table 5) illustrate that the LRT can have inflated type I error rates for small sample sizes ($n = 100$), with more extreme inflation as p increased from 4 to 10. In contrast, for moderate to large sample sizes ($n = 500, 1000$), the type I error rates were close to the nominal level, with only an occasional slight inflation. The inflated type I error rate for small sample sizes seems to be caused by the need to estimate the covariance matrix of the residuals. When we simulated errors that were independent and used the identity matrix for the residual correlations, the simulated type I error rates were very close to the nominal rates for all sample sizes. In contrast, when all β_j were zero (Table 3, Table 4, and Table 6), the LRT has conservative type I error rates. This, however, is not of concern, because controlling the type I error rate when only one β_j differs from zero is the major error that should be controlled when testing pleiotropy. These results were consistent

for different amounts and patterns of residual correlations and for multivariate normal and multivariate *t* distributions.

To further evaluate the adequacy of our asymptotic approximations for large samples, we performed 10,000 simulations for 1000 subjects and four traits that had a common correlation structure. All but one β was zero; the nonzero β was chosen such that there was either 90% or 30% power to detect its marginal effect using $\alpha = 0.001$. This scenario reflects modern large-scale genomic studies that use more stringent significance thresholds. The quantile–quantile plots in Figure 2 show that the asymptotic χ^2 distribution to test pleiotropy provides adequate *P*-values over the entire range of *P*-values for when the marginal effect of one β is small (power of 30%) or large (power of 90%) and for when the correlation of the traits is small ($\rho = 0.2$) or large ($\rho = 0.8$).

The simulation-based power is illustrated in Table 7 and Table 8. The general patterns show that the power to detect two or more associated traits increases with the number of truly associated traits, the effect size of each trait, and larger residual correlations among the traits.

To provide insights into the properties of our proposed sequential testing of multiple traits, we simulated six traits with a common correlation structure such that three of the traits were associated with a genetic variant (*i.e.*, three true nonzero β's). The effect sizes of the associated traits were chosen to have marginal power of 0.3, 0.7, or 0.9 for a sample size of 1000 subjects. These marginal effect sizes correspond to 0.2%, 0.6%, and 1.0% explained variation of the marginal trait. A total of 1000 simulations were performed. The results are presented in Table 9. The frequency of accepting the null hypothesis that all β's = 0 (*e.g.*, no β's selected to be associated with the genetic variant) ranged from 0.646 for when power was 0.3 to 0.015 when power was 0.9—not surprising that greater power resulted in greater frequency of selecting at least one β to be nonzero. Table 9 also presents the frequency for which the three true nonzero β's were selected, conditional on at least one of the six β's was selected. For weak marginal power (*e.g.*, power of 0.3), the frequency of selecting all three nonzero β's was small (0.034–0.213,

Table 9 Properties of sequential testing for six traits: proportions out of 1000 simulations with no β's selected, proportions according to the number of true nonzero β's selected, and proportions according to the number of zero β's falsely selected

Marginal power	β	Explained variation (%)	Trait correlation	Probability no β's selected	No. true β's selected				No. β's falsely selected		
					0	1	2	3	1	2	3
0.3	0.081	0.2	0.2	0.646	0.138	0.644	0.184	0.034	0.209	0.008	0.000
			0.5	0.830	0.235	0.471	0.229	0.065	0.247	0.053	0.006
			0.8	0.850	0.253	0.267	0.267	0.213	0.127	0.140	0.040
0.7	0.139	0.6	0.2	0.128	0.048	0.361	0.412	0.179	0.119	0.011	0.002
			0.5	0.169	0.113	0.178	0.337	0.372	0.140	0.057	0.012
			0.8	0.169	0.105	0.011	0.034	0.851	0.024	0.019	0.093
0.9	0.180	1.0	0.2	0.014	0.007	0.109	0.413	0.472	0.071	0.005	0.001
			0.5	0.015	0.030	0.028	0.172	0.770	0.048	0.017	0.015
			0.8	0.015	0.043	0.001	0.000	0.956	0.034	0.005	0.044

Marginal power is for detecting the effect of a single trait in a sample size of 1000 subjects ($\alpha = 0.05$), and the corresponding β and trait explained variation are provided.

depending on the trait correlation). Yet the frequency of selecting at least one of the three nonzero β's was reasonable (0.747–0.862). The frequency of correctly selecting all three nonzero β's increased as either marginal power increased or trait correlation increased. For example, for marginal power of 0.7, the frequency of selecting all three nonzero β's was 0.179 for weak correlation ($\rho = 0.2$), and was 0.851 for strong correlation ($\rho = 0.8$). For marginal power of 0.9, the frequency of selecting all three nonzero β's was 0.472 for weak correlation ($\rho = 0.2$), and was 0.956 for strong correlation ($\rho = 0.8$). In contrast to selecting true nonzero β's, we also present in Table 9 the frequency of wrongly selecting β's that are truly zero. Not surprisingly, when marginal power is weak (power of 0.3), if at least one β is selected, there is a significant chance of wrongly selecting a true-zero β (*e.g.*, frequency of 0.209 when $\rho = 0.2$). This type of error decreased as the marginal power for traits increased and the trait correlation increased. For example, when power was 0.9 and trait correlation was $\rho = 0.8$, the frequency of selecting one true-zero β was 0.034, approaching the nominal type I error rate of 0.05. Table 9 illustrates that although there is a chance of wrongly selecting one true-zero β, the frequency of selecting more than one true-zero β was small.

Data application results

Based on the traditional multivariate regression of all 10 traits on each SNP, we found a strong association of at least one of the traits with SNPs in a region on chromosome 5 (see Figure 3). Figure 4 illustrates the traditional multivariate test (t_0) in a small region of chromosome 5 (left panel) and the LRT of pleiotropy in the same region (right panel). The test for pleiotropy provides strong evidence that the signal of association was driven by a single phenotype. This is confirmed qualitatively in Figure 5, which shows the individual marginal trait associations for the chromosome 5 region. Although the individual marginal associations in Figure 5 give the visual impression that only one trait is strongly associated with the chromosome 5 SNPs, the LRT of pleiotropy provides a formal statistical test that accounts for the correlations among the traits.

Discussion

Genetic pleiotropy has been of scientific interest since the time of Gregor Mendel, as he described different traits in peas controlled by genes, such as pea coat color and texture, color of flowers, and whether there were axial spots. In current research, understanding pleiotropy can aid the understanding of complex biological mechanisms of genes (as shown in our vaccine response data), as well as aid the development of pharmacologic and vaccine targets. Yet the statistical methods to assess pleiotropy have resorted to *ad hoc* comparison of univariate statistical tests or multivariate methods that test the null hypothesis of no trait associations. Because a formal statistical test of pleiotropy was lacking, we developed a novel LRT statistic. The statistic is easy to compute, based
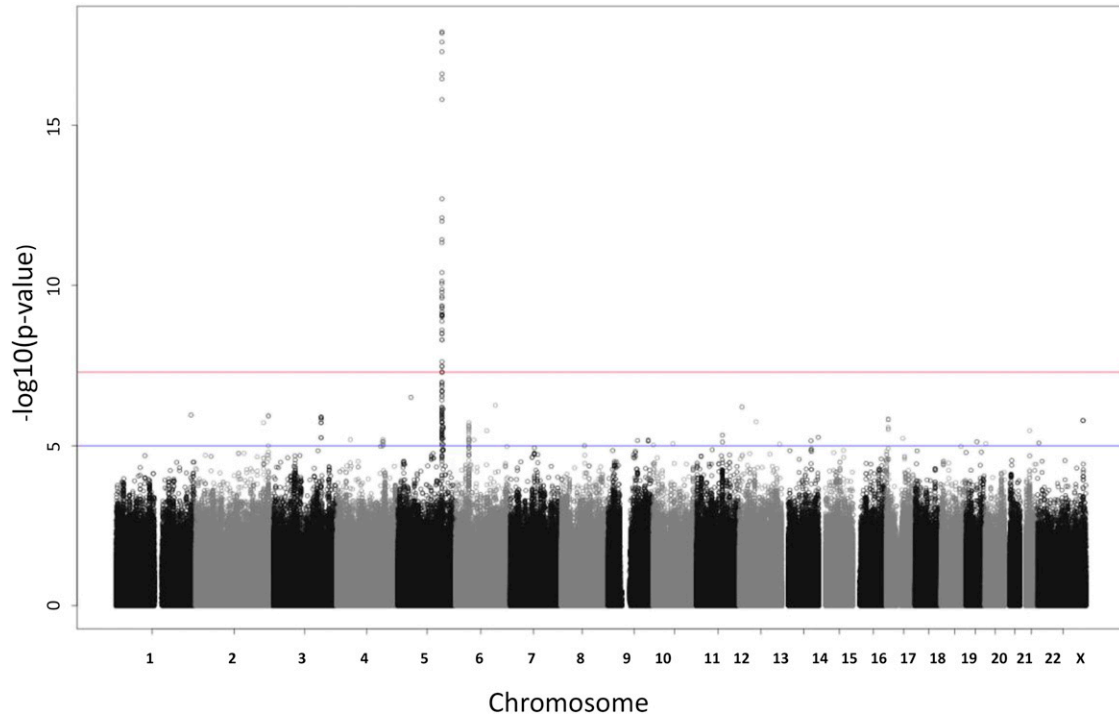
Figure 3 Manhattan plot of the multivariate regression of 10 traits on each SNP, using statistic t_0, to test whether any of the traits are associated with an SNP. The upper red horizontal line corresponds to a *P*-value of 5×10^{-8} and the lower blue line to a *P*-value of 10^{-5}.

on well-known linear regression methods for quantitative traits. Our simulations show that the LRT closely follows a χ^2 distribution when only one trait is associated with a genetic variant and that the LRT tends to be conservative when no traits are associated. We proposed a sequential testing procedure, where the null hypothesis of no associated traits could be tested first (using standard multivariate regression methods) and, if significant, be followed by a test of whether only one trait is associated. If the test of only one associated trait rejects, we proposed sequential testing the null of j associated traits ($j = 2, \ldots, p - 1$), until the sequential test fails to reject the null hypothesis. This approach provides a way to assess the number of traits associated with a genetic variant, accounting for the correlations among the traits. A limitation of our approach, and most other methods for associations of genetic variants with multiple traits, is that it has limited power when an allele is rare. An alternative approach is to compare the similarity of multiple traits with the similarity of rare-variant genotypes across a genetic region, for pairs of subjects (Broadaway *et al.* 2016). The benefit of this approach is balanced with the limitation of not knowing which genotypes are associated with which traits. Our proposed sequential testing might provide a worthy follow-up procedure if some variants are not too rare.

Although our proposed methods assumed the subjects are independent, it is straightforward to extend our approach to pedigree data. To do so, the variance matrix of residuals for independent subjects, $V(\varepsilon) = (\Sigma \otimes I)$, would be replaced with $V(\varepsilon) = (\Sigma \otimes K)$. The matrix K contains diagonal elements

$K_{ii} = 1 + h_i$, where h_i is the inbreeding coefficient for subject i, and off-diagonal elements $K_{ij} = 2\varphi_{ij}$. The parameter φ_{ij} is the kinship coefficient between individuals i and j, the probability that a randomly chosen allele at a given locus from individual i is identical by descent to a randomly chosen allele from individual j, conditional on their ancestral relationship. For subjects from different pedigrees, $\varphi_{ij} = 0$, so K can be structured as a block-diagonal matrix, with diagonal block K_i for the ith pedigree. With this adjustment, our methods can be used for pedigree data or for data with population structure where matrix K is an estimate of genetic relationships (Schaid *et al.* 2013).

Application of our new approach to a study of immune phenotypes in response to smallpox vaccination strongly
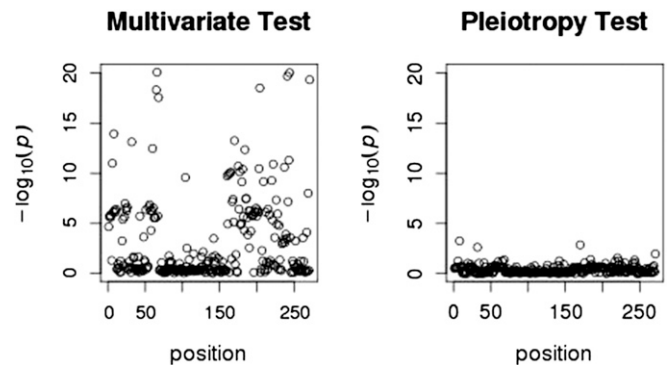
Figure 4 Zoomed-in region of chromosome 5, comparing the traditional multivariate regression of 10 traits on each SNP (left) with the pleiotropy LRT for the same traits and region (right).
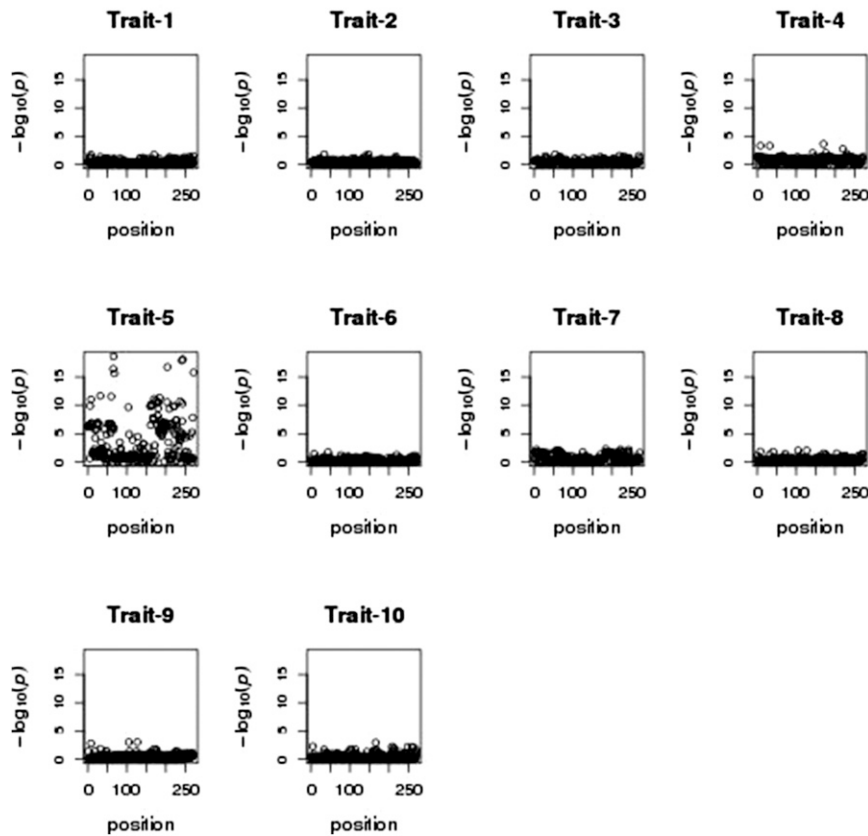
Figure 5 Zoomed-in region of chromosome 5 for each of the univariate regression results for the 10 traits. Trait 5 is the only trait showing strong associations with SNPs in this region of chromosome 5.

suggests that only 1 of 10 correlated traits is statistically associated with SNPs in a region on chromosome 5. The benefit of this type of analysis is that it provides strong guidance on follow-up functional studies for genome-wide association studies with multiple traits. In our case, it allowed investigators to focus on the single immunologic trait truly associated with the chromosome 5 SNPs, rather than conducting labor-intensive, expensive, and time-consuming experiments on unrelated immune response traits.

We recognize that our proposed LRT depends on the assumption that residuals have a multivariate normal distribution. Our simulations with a multivariate t distribution (3 d.f.) suggest that the LRT is robust to heavy-tailed distributions. To ensure robustness with the traditional multivariate regression, it is common practice to transform the data to have at least normally distributed marginal distributions, such as use of normal quantile transformation. This is a reasonable approach for our proposed LRT.

A limitation of our method is that each of the traits is assumed to be quantitative. If all traits are binary, or if there is a mixture of quantitative and binary traits, then the dependence of the LRT on an assumed likelihood would need to be reconsidered. One approach is to consider a general multivariate exponential family of models (Prentice and Zhao 1991; Zhao *et al.* 1992; Sammel *et al.* 1997). Another approach would be to consider the reverse regression of an SNP dose on all traits, like the ordinal logistic MultiPhen approach of O'Reilly *et al.* (2012), yet develop an LRT for

pleiotropy whereby one of the β's is allowed to be unconstrained under the null. An alternative approach that we are developing is based on generalized linear models and generalized estimating equations. The theoretical underpinnings of these alternate approaches, and their computational challenges, are topics of future research.

Acknowledgments

Merck Research Laboratories to study immune responses to mumps vaccine. These activities have been reviewed by the Mayo Clinic Conflict of Interest Review Board and are conducted in compliance with Mayo Clinic Conflict of Interest policies. This research has been reviewed by the Mayo Clinic Conflict of Interest Review Board and was conducted in compliance with Mayo Clinic Conflict of Interest policies.

Literature Cited

Berkson, J., 1946   Limitations of the application of fourfold table analysis to hospital data. Biom. Bull. 2(3): 47–53.

Broadaway, K. A., D. J. Cutler, R. Duncan, J. L. Moore, E. B. Ware et al., 2016   A statistical approach for testing cross-phenotype effects of rare variants. Am. J. Hum. Genet. 98(3): 525–540.

Cotsapas, C., B. F. Voight, E. Rossin, K. Lage, B. M. Neale et al., 2011   Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet. 7(8): e1002254.

Denny, J. C., L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink et al., 2013   Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. 31(12): 1102–1110.

Falconer, D., and T. Mackay, 1996   Introduction to Quantitative Genetics. Pearson Prentice Hall, New York.

Ferreira, M. A., and S. M. Purcell, 2009   A multivariate test of association. Bioinformatics 25(1): 132–133.

Furlotte, N., and E. Eskin, 2015   Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. Genetics 200: 59–68.

Galesloot, T. E., K. van Steen, L. A. Kiemeney, L. A. Janss, and S. H. Vermeulen, 2014   A comparison of multivariate genome-wide association methods. PLoS One 9(4): e95923.

Gianola, D., G. de los Campos, M. H. N. Toro, C. Schon, and D. Sorensen, 2015   Do molecular markers inform about pleiotropy? Genetics 201: 23–29.

Kennedy, R. B., I. G. Ovsyannikova, V. S. Pankratz, I. H. Haralambieva, R. A. Vierkant et al., 2012a   Genome-wide genetic associations with IFNgamma response to smallpox vaccine. Hum. Genet. 131 (9): 1433–1451.

Kennedy, R. B., I. G. Ovsyannikova, V. S. Pankratz, I. H. Haralambieva, R. A. Vierkant et al., 2012b   Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. Hum. Genet. 131(9): 1403–1421.

Korte, A., B. J. Vilhjalmsson, V. Segura, A. Platt, Q. Long et al., 2012   A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat. Genet. 44(9): 1066–1071.

Lee, S., J. Yang, M. Goddard, P. Visscher, and N. Wray, 2012   Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics 28: 2540–2542.

Liu, J., Y. Pei, C. J. Papasian, and H. W. Deng, 2009   Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genet. Epidemiol. 33(3): 217–227.

Maier, R., G. Moser, G. B. Chen, S. Ripke, W. Coryell et al., 2015   Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am. J. Hum. Genet. 96(2): 283–294.

Maity, A., P. F. Sullivan, and J. Y. Tzeng, 2012   Multivariate phenotype association analysis by marker-set kernel machine regression. Genet. Epidemiol. 36(7): 686–695.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly, 2007   A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39(7): 906–913.

O'Reilly, P. F., C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott et al., 2012   MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS One 7(5): e34861.

Ovsyannikova, I. G., I. H. Haralambieva, R. B. Kennedy, V. S. Pankratz, R. A. Vierkant et al., 2012a   Impact of cytokine and cytokine receptor gene polymorphisms on cellular immunity after smallpox vaccination. Gene 510(1): 59–65.

Ovsyannikova, I. G., R. B. Kennedy, M. O'Byrne, R. M. Jacobson, V. S. Pankratz et al., 2012b   Genome-wide association study of antibody response to smallpox vaccine. Vaccine 30(28): 4182–4189.

Ovsyannikova, I. G., I. H. Haralambieva, R. B. Kennedy, M. M. O'Byrne, V. S. Pankratz et al., 2013   Genetic variation in IL18R1 and IL18 genes and Interferon gamma ELISPOT response to smallpox vaccination: an unexpected relationship. J. Infect. Dis. 208(9): 1422–1430.

Ovsyannikova, I. G., V. S. Pankratz, H. M. Salk, R. B. Kennedy, and G. A. Poland, 2014   HLA alleles associated with the adaptive immune response to smallpox vaccine: a replication study. Hum. Genet. 133(9): 1083–1092.

Prentice, R. L., and L. P. Zhao, 1991   Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 47: 825–839.

Roy, J., X. Lin, and L. M. Ryan, 2003   Scaled marginal models for multiple continuous outcomes. Biostatistics 4(3): 371–383.

Sammel, M., L. Ryan, and J. Legler, 1997   Latent variable models for mixed discrete and continuous outcomes. J. R. Stat. Soc. B 59 (3): 667–678.

Schaid, D. J., S. K. McDonnell, J. P. Sinnwell, and S. N. Thibodeau, 2013   Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet. Epidemiol. 37(5): 409–418.

Schifano, E. D., L. Li, D. C. Christiani, and X. Lin, 2013   Genome-wide association analysis for multiple continuous secondary phenotypes. Am. J. Hum. Genet. 92(5): 744–759.

Schriner, D., 2012   Moving toward system genetics through multiple trait analysis in genome-wide association studies. Front. Genet. 16(7): 1–7.

Silvapulle, M. J., and P. K. Sen, 2004   Constrained Statistical Inference: Order, Inequality, and Shape Constraints. John Wiley & Sons, New York.

Solovieff, N., C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller, 2013   Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. 14(7): 483–495.

Stephens, M., 2013   A unified framework for association analysis with multiple related phenotypes. PLoS One 8(7): e65245.

Vansteelandt, S., S. Goetgeluk, S. Lutz, I. Waldman, H. Lyon et al., 2009   On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects. Genet. Epidemiol. 33(5): 394–405.

Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock et al., 2010   Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. 86(6): 929–942.

Xu, Z., and W. Pan, 2015   Approximate score-based testing with application to multivariate trait association analysis. Genet. Epidemiol. 39(6): 469–479.

Yang, Q., and Y. Wang, 2012   Methods for analyzing multivariate phenotypes in genetic association studies. J. Probab. Stat. 2012: 652569.

Zhang, Y., Z. Xu, X. Shen, and W. Pan, 2014   Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. Neuroimage 96: 309–325.

Zhao, L. P., R. L. Prentice, and S. G. Self, 1992   Multivariate mean parameter estimation by using a partly exponential model. J. R. Stat. Soc. B 54(3): 805–811.

Zhou, X., and M. Stephens, 2014   Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods 11(4): 407–409.

Communicating editor: G. A. Churchill

Appendices

Appendix A

R code to compute F-statistic for canonical correlation of matrix **Y** with vector **x**.

```
library(CCA)
cc.fit <- cc(Y,x)
cc.fstat <- function(cc.fit){
rho <- cc.fit$cor
lambda <- 1 - rho^2
dimx <- max(dim(cc.fit$xcoef))
dimy <- max(dim(cc.fit$ycoef))
k <- max(c(dimx, dimy))
n <- nrow(Y)
fstat <- ((1-lambda)/lambda) * ((n-k-1)/k)
pval <- 1-pf(fstat, k, n-k-1, ncp=0)
return(list(fstat=fstat, pval=pval))
}
```

Appendix B: Hypothesis Tests for Linear Model

Notation and model

Based on the regression model described in the main text, suppose that p traits are measured on each of n subjects, with $y'_j = (y_{j1}, ..., y_{jn})$ the vector of measures on the jth trait for n subjects, and stack the vectors as $y' = (y'_1, ..., y'_p)$. Let $X = \text{diag}(x)$, where x is a vector of length n. We assume that y and x are centered on their means. We can express the model as $y = X\beta + \varepsilon$, where $\varepsilon' = (\varepsilon'_1, ..., \varepsilon'_p)$. The error term $\varepsilon \sim N(0, \Omega)$, where $\Omega = \Sigma \otimes I$ and the $p \times p$ matrix Σ is the covariance matrix for the within-subject covariances of the errors. Then, the Cholesky decompositon of Ω is $\Omega = \Omega^{1/2}\Omega^{1/2}$, where $\Omega^{1/2} = \Sigma^{1/2} \otimes I$ and $\Omega^{-1/2} = \Sigma^{-1/2} \otimes I$. Using $\tilde{y} = \Omega^{-1/2}y$, $\tilde{X} = \Omega^{-1/2}X$, the model can transform to independent standard normal random variables, $\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}$, where $\tilde{\varepsilon} = \Sigma^{-1/2}\varepsilon \sim N(0, I_{np})$, and with log likelihood $l_n(\beta) = -(1/2)(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta)$.

Theorem 1. *Let V be a $k \times p$ matrix of rank k ($k \leq p$). Then the minimizer of $\left\| \tilde{y} - \tilde{X}\beta \right\|^2$ under the constraint $V\beta = 0$ is*

$$\beta_V = \beta_n - \beta_V^*,$$

where $\beta_n = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}$ is the ordinary least-squares (OLS) estimate and

$$\beta_V^* = \left(\tilde{X}'\tilde{X}\right)^{-1}V'\left[V\left(\tilde{X}'\tilde{X}\right)^{-1}V'\right]^{-1}V\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{y}.$$

Furthermore,

$$\left\|\tilde{y} - \tilde{X}\beta_V\right\|^2 = \left\|\tilde{y} - \tilde{X}\beta_n\right\|^2 + \left\|\tilde{X}\beta_V^*\right\|^2. \tag{B1}$$

Proof. Denote $\beta^* = \beta_n - \beta$. Note that

$$\left\|\tilde{y} - \tilde{X}\beta\right\|^2 = \left\|\tilde{y} - \tilde{X}\beta_n + \tilde{X}(\beta_n - \beta)\right\|^2 = \left\|\tilde{y} - X\beta_n\right\|^2 + \beta'^*\tilde{X}'\tilde{X}\beta^* + 2(\tilde{y} - \tilde{X}\beta_n)'\tilde{X}\beta^* = \left\|\tilde{y} - \tilde{X}\beta_n\right\|^2 + \beta'^*\tilde{X}'\tilde{X}\beta^*.$$

The last above step results from $2(\tilde{y} - \tilde{X}\beta_n)'\tilde{X}\beta^* = 0$, because $\tilde{y}\tilde{X} = \beta'_n\tilde{X}'\tilde{X}$.

Under the constraint $V\beta = 0$, $V\beta^* = V\beta_n$. Applying the Lagrange multiplier method, we minimize

$$Q\left(\beta^*, \lambda\right) = \left\|\tilde{y} - \tilde{X}\beta_n\right\|^2 + \beta^*\tilde{X}'\tilde{X}\beta^* + 2\lambda\left(V\beta^* - V\beta_n\right).$$

By taking the derivative of Q with respect to β and λ, we obtain the solution $\beta_V^* = (\tilde{X}'\tilde{X})^{-1}V'[V(\tilde{X}'\tilde{X})^{-1}V']^{-1}V(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}$ and the estimate of β is $\beta_V = \beta_n - \beta_V^*$.

Therefore, we have

$$\left\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_V\right\|^2 = \left\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_n\right\|^2 + \boldsymbol{\beta}_V^{*\prime}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\boldsymbol{\beta}_V^*.$$

This completes the *Proof*.

Remark. Equation B1 illustrates that the residual sums of squares (ssq) for the constrained model $\left(\left\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_V\right\|^2\right)$ are partitioned into two parts: (1) the ssq for the OLS fit $\left(\left\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_n\right\|^2\right)$ and (2) the sum of squared differences of the fitted values for the OLS model and the constrained model $\left(\left\|\tilde{\boldsymbol{X}}\boldsymbol{\beta}_V^*\right\|^2 = \left\|\tilde{\boldsymbol{X}}\boldsymbol{\beta}_n - \tilde{\boldsymbol{X}}\boldsymbol{\beta}_V\right\|^2\right)$.

Corollary 1. *Under the null hypothesis,* $V\beta = 0$,
$$\left\|\tilde{\boldsymbol{X}}\boldsymbol{\beta}_V^*\right\|^2 = \left[\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}'\right]^{-1}\left\|\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\varepsilon}}\right\|^2 \sim \chi_k^2.$$

Proof.

$$\left\|\tilde{\boldsymbol{X}}\boldsymbol{\beta}_V^*\right\|^2 = \left\|\tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}'\left[\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}'\right]^{-1}\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}\right\|^2$$

$$= \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}'[\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}']^{-1}\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}$$

$$\times \left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}'[\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}']^{-1}\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}$$

$$= \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}'\left[\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}'\right]^{-1}\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}}$$

$$= \tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}'[\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}']^{-1}\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\varepsilon}}.$$

The substitution of $\tilde{\boldsymbol{y}}$ with $\tilde{\boldsymbol{\varepsilon}}$ in the last step of the above *Proof* can be made because by the assumed linear model, $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\beta + \tilde{\boldsymbol{\varepsilon}}$, we find that

$$\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} = \boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\left(\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}\right) = \boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\varepsilon}} = \boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{\varepsilon}}.$$

The last step above results because $\boldsymbol{V}\boldsymbol{\beta} = 0$ under the null hypothesis.

It is easy to verify that the matrix $\boldsymbol{P}_V = \tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}'[\boldsymbol{V}\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\boldsymbol{V}']^{-1}\boldsymbol{V}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}'$ is idempotent and is of rank k if the rank of $\boldsymbol{V}$ is of rank k. Because $\tilde{\boldsymbol{\varepsilon}} = \sim N(0, \boldsymbol{I}_{np})$ and $\boldsymbol{P}_V$ is idempotent, $\boldsymbol{\varepsilon}'\boldsymbol{P}_V\tilde{\boldsymbol{\varepsilon}} \sim \chi_k^2$, completing the *Proof*.

Corollary 2. *If* $\boldsymbol{V}\boldsymbol{\beta}_0 \neq 0$, *then*
$$\left\|\tilde{\boldsymbol{X}}\boldsymbol{\beta}_V^*\right\|^2 = O(n) \to \infty.$$

Proof. It follows from the *Proof* of *Corollary 1* that $\left\|\tilde{\boldsymbol{X}}\boldsymbol{\beta}_V^*\right\|^2 = \tilde{\boldsymbol{y}}'\boldsymbol{P}_V\tilde{\boldsymbol{y}}$. By the linear model, we have

$$\left\|\tilde{\boldsymbol{X}}\boldsymbol{\beta}_V^*\right\| = \tilde{\boldsymbol{y}}'\boldsymbol{P}_V\tilde{\boldsymbol{y}} = \boldsymbol{\beta}_0'\tilde{\boldsymbol{X}}'\boldsymbol{P}_V\tilde{\boldsymbol{X}}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}'\boldsymbol{P}_V\boldsymbol{\varepsilon} + 2\boldsymbol{\beta}_0'\tilde{\boldsymbol{X}}'\boldsymbol{P}_V\boldsymbol{\varepsilon}.$$

It is clear that $\boldsymbol{\varepsilon}'\boldsymbol{P}_V\boldsymbol{\varepsilon} \sim \chi_k^2 = O_p(1)$ and $\boldsymbol{\beta}_0'\tilde{\boldsymbol{X}}'\boldsymbol{P}_V\boldsymbol{\varepsilon} = O_p(n^{1/2})$. In addition, $\boldsymbol{\beta}_0'\tilde{\boldsymbol{X}}'\boldsymbol{P}_V\tilde{\boldsymbol{X}}\boldsymbol{\beta}_0 = (\boldsymbol{V}\boldsymbol{\beta}_0)'[\boldsymbol{V}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\boldsymbol{V}']^{-1}\boldsymbol{V}\boldsymbol{\beta} = n(\boldsymbol{V}\boldsymbol{\beta}_0)'[\boldsymbol{V}(n^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\boldsymbol{V}']^{-1}\boldsymbol{V}\boldsymbol{\beta} = O(n)$ since $\boldsymbol{V}(n^{-1}\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\boldsymbol{V}' \to \boldsymbol{V}(E\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}\boldsymbol{V}' = O(1)$ in probability. Combining all three facts yields *Corollary 2*.

Hypothesis tests

Now we consider the null hypothesis of no pleiotropy:

H_0 : *Of the parameters* $\beta_1, ..., \beta_p$, *there exists at most one that is nonzero* $\leftrightarrow H_1$: otherwise.

The null hypothesis is equivalent to testing whether one of the following $p + 1$ tests holds:

$$H_{k0} : \beta_k \neq 0, \beta_j = 0 \ (j \neq k),$$

for $k = 0, \ldots, p$. Note that H_{00} represents all $\beta_k = 0$ $(k = 1, \ldots, p)$, while for $k > 0$, H_{k0} allows $\beta_k \neq 0$ while all other $\beta_j = 0$ $(j \neq k)$.

To represent these $p + 1$ hypotheses, we use $H_{k0} : V_k \beta = 0$. Let V_0 be a matrix such that $H_{00} : V_0 \beta = 0$ tests whether all $\beta_j = 0$. In this case, V_0 is the identity matrix of dimension p. To construct V_k $(k > 0)$, create an identity matrix of dimension p and then remove the kth row. This results in $V_k \beta = (\beta_1, \ldots, \beta_{k-1}, \beta_{k+1} \beta_p)'$. Then, the null hypothesis is equivalent to

$$H_0 : \text{there exists one of } H_{k0} : V_k \beta = 0, \text{ for } k = 0, \ldots, p .$$

For $k = 0, 1, \ldots, p$, set $t_k = \tilde{y}' P_{V_k} \tilde{y}$. Then it follows from *Theorem 1* that

$$t_k = \left\| \tilde{X} \beta_{V_k} \right\|^2 - \left\| \tilde{X} \beta_n \right\|^2,$$

where β_{V_k} is the least-squares estimate under the constraint $V_k \beta = 0$ and

$$P_{V_k} = \tilde{X} \left(\tilde{X}' \tilde{X} \right)^{-1} V_k' \left[V_k \left(\tilde{X}' \tilde{X} \right)^{-1} V_k' \right]^{-1} V_k \left(\tilde{X}' \tilde{X} \right)^{-1} \tilde{X}'.$$

Then we have the following corollary.

Corollary 3. *The LRT, $-2 \times$ log of ratio of likelihoods, is given by*

$$T = \min_{k = 0, \ldots, p} t_k.$$

If H_{00} holds, then

$$t_k = \varepsilon' P_{V_k} \varepsilon,$$

and $t_0 \sim \chi_p^2$, $t_k \sim \chi_{p-1}^2$ for $k = 1, \ldots, p$.

If only one H_{k0} $(k > 0)$ holds, then

$$T \sim \chi_{p-1}^2.$$

From *Corollary 3*, we can see that the test statistic T has two different asymptotic distributions when $\beta = 0$ or not. When $\beta = 0$, the asymptotic distribution of T is unknown. Alternatively, when $\beta = 0$, we can use the commonly used χ_p^2 test for the null hypothesis that all $\beta_j = 0$. This motivates us to do the test by two stages. The first stage is just test $H_{00} : \beta = 0$, using the statistic $t_0 \sim \chi_p^2$ as the test statistic, so we reject H_{00} if $t_0 > \chi_p^2(\alpha)$, where $\chi_p^2(\alpha)$ is the $1 - \alpha$ quantile of a χ^2 distribution with p d.f. If H_{00} cannot be rejected, then H_0 cannot be rejected. If H_{00} is rejected, we turn to the second stage to test the null hypothesis that one H_{k0} holds for $k = 1, \ldots, p$. Then we can use the test statistic $T_1 = \min_{k=1, \ldots, p} t_k$. Since $T_1 \sim \chi_{p-1}^2$, we reject the null hypothesis that one H_{k0} holds for $k = 1, \ldots, p$ if $T_1 > \chi_{p-1}^2(\alpha)$. Then, the null hypothesis H_0 is rejected only if both H_{00} is rejected and the null hypothesis that one H_{k0} holds is rejected $(k = 1, \ldots, p)$. Since both tests are conducted at type I error rate of α, and this is based on the principal of the IU test (Silvapulle and Sen 2004), the type I error rate for rejecting H_0 is no more than α.

Remark. If p is too large, it might be beneficial to ignore the t_0 and directly use $T_1 \sim \chi_{p-1}^2$ to construct the rejection region.

Sequential test of nonzero betas

The above solutions can be easily extended to test the following null hypothesis:

$$H_0 : \text{There exist at most } K \text{ nonzero components of } \beta \leftrightarrow H_1 : \textit{otherwise} .$$

With appropriately defined V matrices (there are C_K^p matrices), the LRT reduces to

$$T = \min_{k=1,\dots,C_K^p} \|\boldsymbol{P}_{V_k}\tilde{\boldsymbol{y}}\|^2,$$

where $\boldsymbol{P}_{V_k} = \tilde{X}(\tilde{X}'\tilde{X})^{-1}\boldsymbol{V}_k'[\boldsymbol{V}_k(\tilde{X}'\tilde{X})^{-1}\boldsymbol{V}']_k^{-1}\boldsymbol{V}_k(\tilde{X}'\tilde{X})^{-1}\tilde{X}'$.

In the above, $\boldsymbol{V}_k$ is a $(p-K) \times p$ matrix. For example, if for indexes $1 \le i_1 < \cdots < i_K \le p$, we test

$$\beta_{i_1} \neq 0, \dots, \beta_{i_K} \neq 0 \quad \text{and} \quad \beta_j = 0, j \neq i_1, \dots, i_K,$$

then we can constitute the corresponding matrix $\boldsymbol{V}_{i_1,\dots,i_K}$ as follows: (i) Constitute a $p \times p$ identity matrix and (ii) delete the rows for indexes $i_1, \dots, i_K$.

Then we can use the following multistage test:

i. First test $\mathrm{H}_{00}: \boldsymbol{\beta} = 0$. Reject if $t_0 > \chi_p^2(\alpha)$. If reject, go to the next stage; otherwise stop and conclude H_{00} is true.
ii. For $s = 1, \dots, K-1$, test H_{s0} : there are only s components of $\boldsymbol{\beta} \neq 0$. Reject H_{s0} if $T_s > \chi_{p-s}^2(\alpha)$, where $T_s = \min_{1 \le i_1 < \cdots < i_s \le p}\tilde{\boldsymbol{y}}'\boldsymbol{P}_{V_{i_1,\dots,i_s}}\tilde{\boldsymbol{y}}$. The indexes range over the C_s^p choices. If reject, continue testing by incrementing s by 1. If fail to reject H_{s0}, stop testing and conclude there are s traits associated with x.

The type I error rate of this sequential testing is no greater than the nominal α level. To understand this, suppose there are K nonzero β's, and define the type I error as concluding there are $> K$ nonzero β's. Note that the test statistic T_s at each stage is based on the minimum of statistics, where each statistic is based on $\|\tilde{X}\boldsymbol{\beta}_V^*\|^2 = \|\tilde{X}\boldsymbol{\beta}_n - \tilde{X}\boldsymbol{\beta}_V\|^2$, a measure of distance between fitted values based on the unconstrained OLS model and the constrained model determined by $\boldsymbol{V}$. If one of the constrained models is correct, then by *Corollaries 1* and *2*, $T_s \sim \chi^2$. If, however, none of the constrained models are correct, $T_s = O(n) \to \infty$. This means that at testing stage $j < K$, the probability of rejecting the null hypothesis at stage j depends on the power to detect the misspecified models, which approach 1 as n increases. With this background, we can formally evaluate the type I error rate. Define r_j as an indicator of whether the null hypothesis is rejected at stage j. The probability of a type I error is the probability of rejecting the sequential stage testing up to and including stage K. This joint probability can be expressed as

$$P(r_0, r_1, r_2, \dots, r_K) = P(r_0)P(r_1 \mid r_0)P(r_2 \mid r_0, r_1)..P(r_K \mid r_0, r_1, \dots, r_{K-1}). \tag{B2}$$

The last term in expression (B2) represents the probability of rejecting the null hypothesis when the null hypothesis is true, so one of the constrained models is correct. The test statistic at stage K follows a χ_{p-K}^2 distribution, so $P(r_K \mid r_0, r_1, \dots, r_{K-1}) = \alpha$. All other terms in (B2) approach 1 as n increases, because all stages $j < K$ represent misspecified models. This proves that the type I error rate is no greater than the specified α level.