NeuroImage 115 (2015) 256–268

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Fast and powerful heritability inference for family-based neuroimaging studies

Habib Ganjgahi [a], Anderson M. Winkler [b,c], David C. Glahn [c,d], John Blangero [e], Peter Kochunov [f], Thomas E. Nichols [a,b,g,*]

[a] *Department of Statistics, The University of Warwick, Coventry, UK*
[b] *Centre for Functional MRI of the Brain, University of Oxford, Oxford, UK*
[c] *Department of Psychiatry, Yale University School of Medicine, New Haven, USA*
[d] *Olin Neuropsychiatry Research Center, Institute of Living, Hartford Hospital, Hartford, CT, USA*
[e] *Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA*
[f] *Maryland Psychiatric Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, MD, USA*
[g] *WMG, The University of Warwick, Coventry, UK*

ARTICLE INFO

Article history:
Received 19 December 2014
Accepted 3 March 2015
Available online 23 March 2015

Keywords:
Heritability
Permutation test
Multiple testing problem

ABSTRACT

Heritability estimation has become an important tool for imaging genetics studies. The large number of voxel- and vertex-wise measurements in imaging genetics studies presents a challenge both in terms of computational intensity and the need to account for elevated false positive risk because of the multiple testing problem. There is a gap in existing tools, as standard neuroimaging software cannot estimate heritability, and yet standard quantitative genetics tools cannot provide essential neuroimaging inferences, like family-wise error corrected voxel-wise or cluster-wiseP-values. Moreover, available heritability tools rely on P-values that can be inaccurate with usual parametric inference methods.

In this work we develop fast estimation and inference procedures for voxel-wise heritability, drawing on recent methodological results that simplify heritability likelihood computations (Blangero et al., 2013). We review the family of score and Wald tests and propose novel inference methods based on explained sum of squares of an auxiliary linear model. To address problems with inaccuracies with the standard results used to find P-values, we propose four different permutation schemes to allow semi-parametric inference (parametric likelihood-based estimation, non-parametric sampling distribution). In total, we evaluate 5 different significance tests for heritability, with either asymptotic parametric or permutation-basedP-value computations. We identify a number of tests that are both computationally efficient and powerful, making them ideal candidates for heritability studies in the massive data setting. We illustrate our method on fractional anisotropy measures in 859 subjects from the Genetics of Brain Structure study.

Introduction

Combining neuroimaging data with genetic analyses is an increasingly active area of research aimed at improving our understanding of the genetic and environmental control over brain structure and function in health and illness (see, e.g., Glahn et al., 2007). The foundation of any genetic analysis is establishing that a trait is heritable, that is, that a substantial fraction of its variability can be explained by genetic factors. Significant and reproducible heritability has been established for many neuroimaging traits assessing brain structure and function, including, for instance, location and strength of task-related brain activation (Blokland et al., 2008; Koten et al., 2009; Matthews et al., 2007; Polk et al., 2007), white matter integrity (Kochunov et al., 2014a, b; Jahanshad et al., 2013; Brouwer et al., 2010; Chiang et al., 2009, 2011; Kochunov et al., 2010), cortical and subcortical volumes, cortical thickness and density (Winkler et al., 2010; Rimol et al., 2010; Kochunov et al., 2011a, b; Kremen et al., 2010; den Braber et al., 2013).

Variance component models are the best-practice approach for deriving heritability estimates based on familial data (Almasy and Blangero, 1998; Blangero and Almasy, 1997; Amos, 1994; Hopper and Mathews, 1982), for allowing great flexibility in modeling of genetic additive and dominance effects, as well as common and unique environmental influences. The model also allows the inclusion of additional terms that allow linkage analysis, yet remaining relatively simple and requiring the estimation of only a few parameters. Estimation of parameters typically uses maximum likelihood under the assumption that the additive error follows a multivariate normal distribution. The iterative optimization of the likelihood function requires computationally

[*] Corresponding author at: Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK. Fax + 44 24765 24532.
E-mail address: t.e.nichols@warwick.ac.uk (T.E. Nichols).

http://dx.doi.org/10.1016/j.neuroimage.2015.03.005
1053-8119/© 2015 Published by Elsevier Inc.

intensive procedures, that are prone to convergence failures, something particularly problematic when fitting data at every voxel/element.

Typically a likelihood ratio test (LRT) is used for heritability hypothesis testing. As the null hypothesis value is on the boundary of the parameter space, the asymptotic distribution of LRT is not χ^2 with 1 degree of freedom (DF), but rather approximately as a 50:50 mixture of χ^2 distributions with 1 and 0 DF, where a 0 DFχ^2 is a point mass at 0 (Chernoff, 1954; Self and Liang, 1987; Stram and Lee, 1994; Dominicus et al., 2006; Verbeke and Molenberghs, 2003). However, this result depends on the assumption of independent and identically distributed (i.i.d.) data (Crainiceanu, 2008; Crainiceanu and Ruppert, 2004a, b, c), which is violated in the heritability problem. It has been shown that 0 values occur at a rate greater than 50%, producing conservative inferences (Blangero et al., 2013; Crainiceanu and Ruppert, 2004a; Shephard, 1993; Shephard and Harvey, 1990).

As with most statistical models, the quantitative genetic models used here are based on an assumption of multivariate Gaussianity, and this assumption is the basis of the estimation and test procedures described above. However, the heritability test statistic's null distribution may be inaccurate even when Gaussianity is perfectly satisfied, due to the limitations of the 50:50 χ^2 result just mentioned. Further, for neuroimaging spatial statistics, like family-wise error (FWE) corrected inference with either voxel- or cluster-wise inference, the relevant parametric null distributions are intractable. While random field theory (Worsley et al., 1992; Friston et al., 1994; Nichols and Hayasaka, 2003) results exist for χ^2 images (Cao, 1999), they are not directly applicable here as the test statistic image cannot be expressed as a linear combination of component error fields.

Hence, there is a compelling need for alternative inference procedures that make fewer assumptions. Permutation tests are a type of nonparametric test that can provide exact control — or approximately exact when there are nuisance variables — over false positive rates. These tests depend only on minimal assumptions, namely, that under the null hypothesis the data is exchangeable, that is, that the joint distribution of the data remains unaltered after permutation (Nichols and Holmes, 2002; Winkler et al., 2014).

There is relatively little work on permutation tests for variance component inference. The typical application of variance components models is not in quantitative genetics, but in hierarchical linear models where observational units are nested in clusters, such repeated measures designs. Of the few permutation methods proposed in this setting, they all permute the residuals (after removing the covariate effects) between and within clusters while fixing the model structure. While these procedures use different test statistics, e.g. Fitzmaurice and Lipsitz (2007) used the LRT as the statistic, while Lee and Braun (2012) used the sample variance of estimated random effect, they generally require iterative optimization of the likelihood function, and thus as permutation procedures they are yet more computationally demanding.

Samuh et al. (2012) presented a fast permutation test, though it is only applicable to the random intercept model. And recently Drikvandi et al. (2013) introduced a fast permutation test based on the variance least square estimator, which in essence fits a regression model to squared residuals. However, this approach is not based on maximum likelihood, and is only intended for a standard repeated measures model, where independent subjects are recorded multiple times, not multiple dependent subjects as in a pedigree study.

Our group presented a method to accelerate maximum likelihood estimation by applying an orthonormal data transformation that diagonalizes the phenotypic covariance, transforming a correlated heritability model into an independent but heterogeneous variance model (Blangero et al., 2013). However, this advance doesn't eliminate iterative optimization nor possible convergence problems.

In the present work, we expanded upon this work to derive approximate, non-iterative estimates and test statistics based on the first iteration of Newton's method. These procedures can be constructed with an auxiliary model based on regressing squared residuals on the kinship matrix eigenvalues. Then the Wald and score hypothesis tests can then be seen as generalized and ordinary explained sum of squares of the auxiliary model. In addition, as the null hypothesis of no heritability corresponds to homogeneous variance of the transformed phenotype, we draw from the statistical literature on tests of heteroscedasticity for a new and completely different test for heritability detection. We develop permutation test procedures for each of these methods, thus providing FWE-corrected voxel- and cluster-wise inferences.

The remainder of this paper is organized as follows. In the next section we detail the statistical model used and describe each of our proposed methods. The simulation framework used to evaluate the methods, and the real data analysis used for illustration are described in the Evaluation section. We then present and interpret results, and offer concluding remarks.

Theory

In this section we detail the statistical models used, introduce our fast heritability estimators and tests, and then propose several permutation strategies for these tests.

Original and eigensimplified polygenic models

At each voxel/element, a polygenic model for the phenotype Y measured on N individuals can be written as

$$Y = X\beta + g + \epsilon \tag{1}$$

where X is an $N \times p$ matrix consisting of an intercept and covariates, like ageand sex; β is the p-vector of regression coefficients; g is the N-vector of latent (unobserved) additive genetic effect; and ϵ is the N-vector of residual errors. In this study we consider the most common variance components model, with only additive and unique environmental components.

The trait covariance, $Var(Y) = Var(g + \epsilon) = \Sigma$ can be written as

$$\Sigma = 2\sigma_A^2\Phi + \sigma_E^2 I, \tag{2}$$

where Φ is the kinship matrix; σ_A^2 and σ_E^2 are the additive genetic and the environmental variance components, respectively; and I is the identity matrix. The kinship matrix is comprised of kinship coefficients, half the expected proportion of genetic material shared between each pair of individuals (Lange, 2003).

The narrow sense heritability is

$$h^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}. \tag{3}$$

Maximum likelihood is used for parameter estimation with the assumption that the data follows a multivariate normal distribution. The log likelihood for the untransformed model (Eqs. (1) & (2)) is

$$\ell(\beta, \Sigma; Y, X) = -\frac{1}{2}Nlog(2\pi) - \frac{1}{2}log(|\Sigma|) - \frac{1}{2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta). \tag{4}$$

For large datasets with arbitrary family structure, the computational burden of evaluating of the likelihood can be substantial. In particular, a quadratic form of the inverse covariance, Σ^{-1}, must be computed, along with the determinant of Σ. We take the approach of Blangero et al. (2013), who proposed an orthogonal transformation based on the eigenvectors of the kinship matrix, thus diagonalizing the covariance and simplifying the computation of the likelihood (Eq. (4)).

The eigensimplified polygenic model is obtained by transforming the data and model with a matrix S, the matrix of eigenvectors of Φ which are the same as the eigenvectors of Σ, Eq. (2). Applying this transformation to Eq. (1) gives the transformed model

$$S'Y = S'X\beta + S'g + S'\varepsilon$$

which we write as

$$Y^* = X^*\beta + \varepsilon^*, \tag{5}$$

where Y^* is the transformed data, X^* are the transformed covariates and ε^* is the transformed random component, where ε^* now encompasses both the genetic and non-genetic random variation. The diagonalizing property of the eigenvectors then gives a simplified form for the variance:

$$\mathrm{Var}(\varepsilon^*) = \Sigma^* = \sigma^2_{\mathrm{A}} D_g + \sigma^2_{\mathrm{E}} I, \tag{6}$$

where Σ^* is the variance of the transformed data and $D_g = \mathrm{diag}\{\lambda_{gi}\}$ is a diagonal matrix of the eigenvalues of 2Φ.

The log likelihood takes on the exact same form as Eq. (4) for Y^*, X^*, β and Σ^*, except is much easier to work with since Σ^* is diagonal:

$$\ell(\beta^*, \sigma^*_{\mathrm{A}}, \sigma^*_{\mathrm{E}}; Y^*, X^*) = -\frac{1}{2} N log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} log\left(\sigma^2_{\mathrm{A}}\lambda_{gi} + \sigma^2_{\mathrm{E}}\right) - \frac{1}{2}\sum_{i=1}^{N} \frac{\varepsilon_i^{*2}}{\sigma^2_{\mathrm{A}}\lambda_{gi} + \sigma^2_{\mathrm{E}}}.$$

Note that, while S' can be seen as a semi-whitening step, the transformed model can also be seen as a change of variables, where the variance is reparametrized as $\Sigma = S\Sigma^* S'$. As a reparametrization, the invariance property of maximum likelihood guarantees that the same values of β, σ^2_{A} and σ^2_{E} optimize both the original and transformed likelihoods.

Use of this transformation has two major benefits. First, optimization time is substantially reduced, as the inverse and determinant of the transformed covariance are now trivial. Second, applying standard statistical inference procedures, including the score and the Wald test, to the eigensimplified polygenic model produces simple algebraic forms that can be harnessed for fast approximations. Both of these speed improvements facilitate the use of permutation tests that avoid asymptotic approximations.

Heritability estimation and test statistics

We segregate the transformed model parameters into fixed β and random $\theta = (\sigma^2_{\mathrm{A}}, \sigma^2_{\mathrm{E}})$ terms, and estimate them by maximizing the likelihood function via iterative numerical methods. Here, we consider Newton's method because it leads to computationally efficient heritability estimators and associated tests. Newton's method requires the score and expected information matrix of the transformed model, which are

$$S(\beta,\theta) = \begin{bmatrix} X^{*\prime}\Sigma^{*-1}\varepsilon^* \\ -\frac{1}{2}\left[U'\Sigma^{*-1}1 - U'\Sigma^{*-2}\varepsilon^{*2}\right] \end{bmatrix} \tag{7}$$

and

$$I(\beta,\theta) = \begin{bmatrix} X^{*\prime}\Sigma^{*-1}X^* & 0 \\ 0 & \frac{1}{2}U'\Sigma^{*-2}U \end{bmatrix}, \tag{8}$$

respectively, where $U = [\mathbf{1}, \lambda_g]$ is a $N \times 2$ matrix, $\mathbf{1}$ is a $N \times 1$ vector of ones and $\lambda_g = \{\lambda_{gi}\}$ is a $N \times 1$ vector of kinship matrix eigenvalues. It is useful to write f^* for the vector with elements $f_i^* = \hat{\varepsilon}_i^{*2}$, where $\hat{\varepsilon}^* = Y^* - X^*\hat{\beta}$ are the transformed model residuals. Newton's method gives update equations for $\hat{\beta}$ and $\hat{\theta}$ at iteration $j+1$ as:

$$\hat{\beta}_{j+1} = \left(X^{*\prime}\left(\hat{\Sigma}^*_j\right)^{-1}X^*\right)^{-1} X^{*\prime}\left(\hat{\Sigma}^*_j\right)^{-1}Y^* \tag{9}$$

$$\hat{\theta}_{j+1} = max\left\{0, \left(U'\left(\hat{\Sigma}^{*2}_j\right)^{-1}U\right)^{-1} U'\left(\hat{\Sigma}^{*2}_j\right)^{-1} f^*_j\right\}, \tag{10}$$

where j indexes iteration; the variance parameters θ must be positive, hence the maximum operator. When these updates are iterated until convergence as usual, we denote the estimates with a ML subscript, e.g. $\hat{\beta}_{\mathrm{ML}}, \hat{\theta}_{\mathrm{ML}}$ and $\hat{h}^2_{\mathrm{ML}} = \hat{\sigma}^2_{\mathrm{A,ML}} / \left(\hat{\sigma}^2_{\mathrm{A,ML}} + \hat{\sigma}^2_{\mathrm{E,ML}}\right)$.

To allow for potential improvements on speed, we also consider a one-step estimator. First, observe that since Σ^* is diagonal, Eq. (9) is the Weighted Least Squares (WLS) regression of Y^* on X^*, and Eq. (10) is based on the WLS regression of f^* on U. This immediately suggests initial values based on Ordinary Least Squares (OLS),

$$\hat{\beta}_{\mathrm{OLS}} = \left(X^{*\prime}X^*\right)^{-1}X^{*\prime}Y^*$$

$$\hat{\theta}_{\mathrm{OLS}} = max\left\{0, \left(U'U\right)^{-1}U' f^*_{\mathrm{OLS}}\right\}, \tag{11}$$

where f^*_{OLS} is the square of the OLS residuals

$$\hat{\varepsilon}_{\mathrm{OLS}} = Y^* - X^*\hat{\beta}_{\mathrm{OLS}}; \tag{12}$$

while not recommended as a final estimate, it also produces $\hat{h}^2_{\mathrm{OLS}} = \hat{\sigma}^2_{\mathrm{A,OLS}} / \left(\hat{\sigma}^2_{\mathrm{A,OLS}} + \hat{\sigma}^2_{\mathrm{E,OLS}}\right)$. Finally, our proposed one-step estimators are:

$$\hat{\beta}_{\mathrm{WLS}} = \left(X^{*\prime}\left(\hat{\Sigma}^*_{\mathrm{OLS}}\right)^{-1}X^*\right)^{-1} X^{*\prime}\left(\hat{\Sigma}^*_{\mathrm{OLS}}\right)^{-1}Y^*$$

$$\hat{\theta}_{\mathrm{WLS}} = max\left\{0, \left(U'\left(\hat{\Sigma}^{*2}_{\mathrm{OLS}}\right)^{-1}U\right)^{-1} U'\left(\hat{\Sigma}^{*2}_{\mathrm{OLS}}\right)^{-1} f^*_{\mathrm{OLS}}\right\}, \tag{13}$$

where $\hat{\Sigma}^*_{\mathrm{OLS}}$ is formed by $\hat{\theta}_{\mathrm{OLS}} = \left(\sigma^2_{\mathrm{A,OLS}}, \sigma^2_{\mathrm{E,OLS}}\right)$, also producing $\hat{h}^2_{\mathrm{WLS}} = \hat{\sigma}^2_{\mathrm{A,WLS}} / \left(\hat{\sigma}^2_{\mathrm{A,WLS}} + \hat{\sigma}^2_{\mathrm{E,WLS}}\right)$.

Amemiya (1977) showed that such one-step maximum likelihood estimators are asymptotically normal and consistent. Going forward, we will use "ML" to refer to the maximum-likelihood, iterated estimator and "WLS" to refer to this one-step estimator.

Test statistics

In this section we describe likelihood ratio tests (LRTs), Wald tests, and score test for hypothesis tests of nonzero heritability; we also add an additional test based on detecting heterogeneous variance structure to detect non-zero heritability. We only consider the transformed model, and tests on $H_0 : \sigma^2_{\mathrm{A}} = 0$ vs. $H_1 : \sigma^2_{\mathrm{A}} > 0$, equivalent to inference for heritability (Eq. (3)). Table 1 organizes the models and test statistics we consider.

Likelihood ratio test

The LRT (Neyman and Pearson, 1933) statistic is twice the difference of the log-likelihoods, unrestricted minus H_0-restricted. For ML this requires optimizing the likelihood function twice, once under the null $H_0 : \sigma^2_{\mathrm{A}} = 0$, and once under the alternative (though the null model is trivial, equivalent to OLS). We denote the test statistic for this test $T_{\mathrm{L,ML}}$. In addition, a LRT can be constructed for the transformed model in terms of the one-step WLS estimator; we denote this statistic as $T_{\mathrm{L,WLS}}$.

Wald test

The Wald test consists of a quadratic form of the parameter estimate (minus its null value) and its inverse asymptotic variance (i.e. expected Fisher's information matrix). Both the estimate and its variance are computed under the full, alternative model.

Table 1
Comparison of model and test statistic properties. Usual P-values and CI's (confidence Intervals) refer to the best practice inference tools used with maximum likelihood estimation.

Model Name	Model Expression	Estimation Method	Test statistics			
			LRT	WALD	Score	GQ
Original	$Y = X\beta + g + \epsilon$	ML	*(usual P-values)*	*(usual CI's)*		
Transformed	$Y^* = X^*\beta + \epsilon^*$	WLS, "1 Step"	$T_{\mathrm{L,WLS}}$	$T_{\mathrm{W,WLS}}$	T_{S}	
		ML, fully converged	$T_{\mathrm{L,ML}}$	$T_{\mathrm{W,ML}}$		
Transformed, Split	$Y_A^* = X_A^*\beta_A + \epsilon_A^*$ $Y_B^* = X_B^*\beta_B + \epsilon_B^*$	OLS, "0 Step"				T_{GQ}

The Wald test for the ML estimator (Rao, 2008) is

$$T_{\mathrm{W,ML}} = \frac{1}{2}\left(\hat{\sigma}^2_{\mathrm{A,ML}}\right)^2\left[C\left(U'\hat{\Sigma}^{*-2}_{\mathrm{ML}}U\right)^{-1}C'\right]^{-1}$$

$$= \frac{1}{2}\left(N - \left(\mathbf{1}'\hat{\Sigma}^{*-1}_{\mathrm{ML}}\mathbf{1}\right)^2\left(\mathbf{1}'\hat{\Sigma}^{*-2}_{\mathrm{ML}}\mathbf{1}\right)^{-1}\right)$$

where $C = [0 \;\; 1]$ is a contrast row vector, and the latter is a simpler form found in Buse (1984). Iterative optimization is required for $T_{\mathrm{W,ML}}$, though it can be more amenable to compute than LRT because the likelihood function is optimized only once.

The Wald test for our one-stepWLS estimator can be written as

$$\begin{aligned} T_{\mathrm{W,WLS}} &= \frac{1}{2}\left(\hat{\sigma}^2_{\mathrm{A,WLS}}\right)^2\left[C\left(U'\hat{\Sigma}^{*-2}_{\mathrm{WLS}}U\right)^{-1}C'\right]^{-1} \\ &= \frac{1}{2}\left(\hat{\sigma}^2_{\mathrm{A,WLS}}\right)^2 \\ &\quad \times\left(\hat{\Sigma}^{*-1}_{\mathrm{OLS}}\lambda_g\right)'\left(I - \hat{\Sigma}^{*-1}_{\mathrm{OLS}}\mathbf{1}\left(\left(\hat{\Sigma}^{*-1}_{\mathrm{OLS}}\mathbf{1}\right)'\left(\hat{\Sigma}^{*-1}_{\mathrm{OLS}}\mathbf{1}\right)\right)^{-1}\mathbf{1}'\hat{\Sigma}^{*-1}_{\mathrm{OLS}}\right)\hat{\Sigma}^{*-1}_{\mathrm{OLS}}\lambda_g. \end{aligned}$$

where the second line shows the computation to be half the generalized explained sum of squares (Buse, 1973, 1979) of an auxiliary model, the weighted least squares regression of f^*_{OLS} on λ_g, with weights determined by $\hat{\Sigma}^*_{\mathrm{OLS}}$.

Score test

The score test (Rao, 2008), also known as the Lagrange multiplier test, is a quadratic form of the score (the gradient of the log likelihood) and the expected Fisher's information, each evaluated under the null hypothesis. Among the tests that we consider, the score test is the least computationally demanding procedure, as it only requires estimation of the null model. For $H_0 : \sigma^2_{\mathrm{A}} = 0$, the score test with the transformed likelihood function is:

$$T_{\mathrm{S}} = \frac{\lambda_g'\hat{\Sigma}^{*-2}_{\mathrm{OLS}}f^*_{\mathrm{OLS}} - \lambda_g'\hat{\Sigma}^{*-1}_{\mathrm{OLS}}\mathbf{1}}{CU'\hat{\Sigma}^{*-2}_{\mathrm{OLS}}UC'}$$

$$= \frac{1}{2}\left(\frac{\hat{\sigma}^2_{\mathrm{A,OLS}}}{\hat{\sigma}^2_{\mathrm{OLS}}}\right)^2\lambda_g'\left(I - \frac{\mathbf{1}'\mathbf{1}}{N}\right)\lambda_g,$$

where $\hat{\sigma}^2_{\mathrm{OLS}} = (\hat{\varepsilon}_{\mathrm{OLS}})'\hat{\varepsilon}_{\mathrm{OLS}}/N$ is the OLS naive residual variance estimator. Similar to the Wald test, T_{S} can be obtained as half the regression sum of squares of an auxiliary model, the (unweighted) regression of $f^*/\hat{\sigma}^2_{\mathrm{A,OLS}}$ on λ_g. As it only involves the fitted null model, it isn't associated with a WLS or ML estimate.

We note that Wald and score tests for a null hypothesis value lying on the boundary of parameter space can take a special form (Freedman, 2007; Molenberghs and Verbeke, 2007; Morgan et al., 2007; Verbeke and Molenberghs, 2007; Silvapulle, 1992; Silvapulle and Silvapulle, 1995; Verbeke and Molenberghs, 2003). However, for our model (Eq. (1)), the standard version is appropriate if the score function is positive at the boundary value and otherwise set to zero. As any negative score values are suppressed by our non-negative constrained estimates $\hat{\theta}_{\mathrm{OLS}}$ (Eq. (11)) and $\hat{\theta}_{\mathrm{WLS}}$ (Eq. (13)), our tests are implicitly zero when needed, and thus the appropriate Wald and score tests are as given above.

All three of the LRT, Wald, and score test procedures are asymptotically equivalent but have different small-sample performance, which we evaluate below. These tests are considered to follow asymptotically a 50 : 50 mixture of χ^2 distributions with 1 and 0 DF, where 0 a DF χ^2 is a point mass at 0 (Chernoff, 1954; Self and Liang, 1987; Stram and Lee, 1994; Dominicus et al., 2006; Verbeke and Molenberghs, 2003), although it has been shown that 0 values can occur with a higher frequency, and the standard 50:50 result will tend to produce conservative inferences (Blangero et al., 2013; Crainiceanu and Ruppert, 2004a; Shephard, 1993; Shephard and Harvey, 1990).

Goldfeld and Quandt (GQ) test

Instead of standard likelihood theory, an alternative approach to heritability hypothesis testing can be derived from tests of heteroscedasticity. This follows for the transformed model, since the null hypothesis of no heritability corresponds to homoscedasticity of the transformed phenotype variance ($\mathrm{Var}(\varepsilon^*) = \sigma^2 I$). Thus, rejection of the hypothesis of homoscedasticity implies a rejection of the hypothesis of zero heritability. One class of such tests requires an explicit, hypothesized form for the heterogeneous variance. Another type called "nonconstructive" does not require such explicit models; one example is the Goldfeld and Quandt (1965) (GQ) test, which we propose as a test for non-zero heritability.

The GQ test partitions observations into 2 groups, A & B, based on a variable that should explain any heterogeneous variance. The test statistic then compares the ratio of OLS residual mean squares:

$$T_{\mathrm{GQ}} = \frac{\hat{\varepsilon}^{*\prime}_A\hat{\varepsilon}^*_A/(n_A - 1)}{\hat{\varepsilon}^{*\prime}_B\hat{\varepsilon}^*_B/(n_B - 1)} \tag{14}$$

where subscript A refers to the high variance group, subscript B to low variance group, $\hat{\varepsilon}^*_A$refers to the residuals from regressing elements of Y^* in group A on corresponding rows of X^*, and likewise for $\hat{\varepsilon}^*_B$, finally, n_A and n_B are the number of observations in each respective group. With Gaussian errors and under a null hypothesis of homoscedasticity, T_{GQ} follows a F-distribution with degrees of freedom $\nu_1 = n_B - p$ and $\nu_2 = n_A - p$, where p is the number of columns in X^*.

For the transformed data Y^*, the kinship eigenvalues order the variance of the observations when $\sigma^2_{\mathrm{A}} > 0$. Thus we propose to define the two groups based on $\lambda_{gi} > 1$ and $\lambda_{gi} \leq 1$, where we make use of the fact $\sum_i \lambda_{gi}/N = trace(2\Phi)/N = 1$.

This test is only able to detect non-zero heritability and cannot produce estimates of h^2. On the other hand, the parametric null distribution of (Eq. (14)) does not depend on the mixture approximation and large sample properties, and its implementation is straightforward. To our knowledge, this is the first proposed use of a heteroscedasticity test to create an exact (non-asymptotic), non-iterative test of heritability.

Permutation test for heritability inference

Permutation methods can be used to construct the null sampling distribution which can be used to produce P-values and thresholds. For the model with only additive genetic and environmental variance components, the null hypothesis of no heritability implies fully independent data. Thus, if there were no nuisance variables (X), a permutation test could be conducted by freely permuting the data (Y). With covariates, we must permute suitable residuals, as detailed below.

To conduct inference on σ_A^2 in the presence of the nuisance parameters β and σ_E^2, we draw inspiration from various methods for permutation methods for the GLM (Winkler et al., 2014). For example, there are several different permutation schemes when testing a strict subset of all GLM regression parameters. One approach is to permute only the column of interest in the design matrix. This approach, due to Draper and Stoneman (1966) could be restated as isolating the portion of the model affected by the null hypothesis, and then only permuting that portion. This is the motivation for our first permutation strategy (P1), where we repeatedly fit the model, but randomly permute kinship each time.

Another approach is to use the reduced, null hypothesis model to generate residuals, permute these residuals, and use them as surrogate null data to be re-analyzed (Freedman and Lane, 1983). For the GLM, this is the recommended approach (Winkler et al., 2014), and corresponds to an ideal test where nuisance effects are removed from the data, leaving what should be only unstructured data (under the null) ready to be permuted. This is the motivation for permutation scheme (P2).

Finally, another approach to GLM permutation testing is to use the full, alternative hypothesis model to generate residuals, and then use these residuals as surrogate null data to be re-fit (ter Braak, 1992). This approach has the merit of removing all systematic variation from the data before permutation. This is the motivation for our third and fourth strategies (P3 & P4).

Partial model permutation (P1)

We implement approach P1 by permuting just the aspect of the model tested by the H_0. For the untransformed model this corresponds to permuting the model's covariance term to be

$$2\sigma_A^2 P\Phi P' + \sigma_E^2 I,$$

where P is one of N ! possible $N \times N$ permutation matrices. For the transformed model, the permutated covariance takes the form

$$\sigma_A^2 P D_g P' + \sigma_E^2 I.$$

Null model residual permutation (P2)

For P2 we generate residuals under $H_0 : \sigma_A^2 = 0$, i.e. OLS residuals $\hat{\varepsilon}_{\text{OLS}}$ (Eq. (12)). Then we permute these residuals, and add-back nuisance (fixed) effects to generate new H_0 realizations $\tilde{Y}^*$:

$$\tilde{Y}^* = X^* \hat{\beta}_{\text{OLS}} + P \hat{\varepsilon}^*_{\text{OLS}}, \tag{15}$$

where the tilde (˜) accent denotes one of many realizations, which in turn are fit with the model under consideration.

Full model residual permutation (P3)

For P3, we start with full model residuals, i.e. either $\hat{\varepsilon}_{\text{ML}}$ or $\hat{\varepsilon}_{\text{WLS}}$, depending on the estimator used. Then we permute these residuals, and add-back nuisance to generate new null hypothesis realizations; e.g., for WLS:

$$\tilde{Y}^* = X^* \hat{\beta}_{\text{WLS}} + P \hat{\varepsilon}^*_{\text{WLS}}. \tag{16}$$

and analogously for ML. Again, each realization $\tilde{Y}$ is fit to the current model.

Full model whitened residual permutation (P4)

P4 is like P3, but we go a step further and create residuals that are whitened before permutation. For example, for WLS:

$$\tilde{Y}^* = P\left(\hat{\Sigma}^{*-1/2} \hat{\varepsilon}^*_{\text{WLS}}\right), \tag{17}$$

and analogously for ML. Again, each realization is fit to the current model.

In total we have introduced five different test procedures and four permutation strategies, summarized in Table 2.

Multiple testing correction

Whether inference is conducted voxel-wise or cluster-wise, the use of use of an uncorrected $\alpha = 5\,\%$ level leads to an excess of false positives. False discovery rate correction, controlling the expected proportion of false positives among all detections, is easily applied based on uncorrected P-values alone (Genovese et al., 2002). As uncorrected permutation cluster-wiseP-values require an assumption of stationarity (though see Salimi-Khorshidi et al. (2010)), FDR is generally only applied with voxel-wiseP-values. Familywise error rate (FWE) correction, controlling the chance of one or more false positives across the whole set (family) of tests (Nichols and Hayasaka, 2003) requires the distribution of the maximum statistic, easily computed for either voxels or cluster size with permutation (Nichols and Holmes, 2002).

Evaluation

Simulation studies

We conduct various simulation studies to evaluate proposed methods for heritability inference on the transformed model. The first study considers estimator bias and variance for the different methods. The second study measures the accuracy of parametric and permutation inference methods. Finally, the third study evaluates FWE control in an image-wise setting for voxel and cluster-wise inferences.

In all simulations, the response variable is assumed to be $Y = X\beta + \varepsilon$ where ε follows $N(0, \Sigma)$, $\Sigma = h^2(2\Phi) + (1 - h^2)I$. The design matrix X consists of an intercept, a linear trend vector X_1 and a quadratic vector X_2 between 1 and -1, with $\beta = [0, 0, 10]$. Kinship structure Φ is based on real pedigrees (each described below), and the simulations considered a range of true heritabilities ($h^2 = 0, 0.2, 0.4, 0.6, 0.8$).

Simulation 1

This simulation evaluates the bias, standard deviation and mean squared error (MSE) of the heritability estimators (ML and WLS). The pedigrees and sample sizes used are shown in Table 3; we used pedigrees from the 10th Genetics Analysis Workshop (GAW10) (Mac-Cluer et al., 1997) and from the GOBS dataset (described below). Univariate data Y was simulated as per the Gaussian model described above, and 10,000 realizations were used.

Simulation 2

This simulation assesses the false positive rates for each method, on the basis of both parametric and permutation methods. For this analysis we used 2 pedigrees from the GAW10 dataset with 138 subjects; the small sample size was used to 'stress test' the methods. Univariate data Y was simulated as per the Gaussian model described above, 10,000 realizations were used, and 500 permutations for each

Table 2
Comparison of tests for heritability inference.

Tests	h^2 estimates	Distribution	Type	Optimization	Permutation
$T_{\text{L,ML}}$	✓	50:50 χ_1^2 and 0	Asymptotic	ML	P1, P2, P3, P4
$T_{\text{W,ML}}$	✓	50:50 χ_1^2 and 0	Asymptotic	ML	P1, P2, P3, P4
$T_{\text{W,WLS}}$	✓	50:50 χ_1^2 and 0	Asymptotic	WLS	P1, P2, P3, P4
T_{S}	✓	50:50 χ_1^2 and 0	Asymptotic	OLS	P1, P2, P3, P4
T_{GQ}	✗	F_{n_2-p,n_1-p}	Exact	OLS	P1, P2, P3, P4

Proposed test procedures: The score test (T_{S}), the Wald test and its variants in terms of WLS estimators ($T_{\text{W,WLS}}$) and ML estimators ($T_{\text{W,ML}}$), and the LRTs in terms of the transformed model ($T_{\text{L,ML}}$). ML optimization denotes iterative optimization until convergence; WLS a 1-step of Newton's method; and OLS an estimate based on (unweighted) least squares.

Table 3
Datasets used in simulation 1.

Datasets	Number of pedigrees	Sample size
GAW10	2	138
GAW10	9	626
GOBS	73	858
GAW10	23	1497

nonparametric procedure. On the basis of Simulations 1 and 2, 'winner' tests and a permutation strategy were chosen and fed into the 3rd simulation study.

Simulation 3

Image simulations were conducted under the null hypothesis ($h^2 = 0$) on a 96 × 96 × 20 image that the response variable for each voxel are simulated as described above, smoothed with a Gaussian filter with a Full Width at Half Maximum of 4 mm. To avoid edge effects, larger images were simulated, smoothed and then truncated. For each realization we collected empirical null distributions of maximum statistic and maximum cluster size to compute FWEP-values; we considered different cluster forming thresholds (parametric uncorrected P-value = 0.05, 0.01, 0.005, 0.001). We generated 5000 realizations and used 500 permutations with each synthetic dataset.

Application in diffusion tensor imaging data

We used data from the Genetics of Brain Structure and Function Study (GOBS) (Olvera et al., 2011; McKay et al., 2014) to perform voxel and cluster-wiseFA heritability inference in healthy subjects. The sample comprised 859 Mexican–American individuals from 73 extended pedigrees (average size 17.2 people, range = 1–247). The sample was 59 % female (351 men/508 women) and had a mean age of 43.2 (SD = 15.0; range = 19–85). All participants provided written informed consent on forms approved by the Institutional Review Boards at the University of Texas Health Science Center San Antonio (UTHSCSA) and Yale University.

Table 4
Simulation 2 result, comparing parametric rejection rates (percent), 5% nominal. For GAW10 data with 2 families, 138 subjects, 10,000 realizations. GQ test has the most accurate false positive rate, LRT with ML ($T_{L,ML}$) is the most powerful; both GQ (T_{GQ}) and score (T_S) test have good power (95% MC CI for 0.05, i.e. for the null case is (4.57%, 5.42%)).

Test	True effect (h^2)				
	0	0.2	0.4	0.6	0.8
T_S	3.76	40.66	76.76	94.32	98.94
$T_{W,WLS}$	1.56	26.94	73.46	95.62	99.64
$T_{W,ML}$	2.50	33.00	77.74	94.84	97.54
$T_{L,ML}$	3.16	42.28	81.80	96.40	98.90
T_{GQ}	4.36	35.60	78.22	96.50	99.70

Diffusion imaging was performed at the Research Imaging Center, UTHSCSA, on a Siemens 3 T Trio scanner using a multi-channel phased array head coil. Asingle-shot single refocusing spin-echo, echo-planar imaging sequence was used to acquire diffusion-weighted data with a spatial resolution of 1.7 × 1.7 × 3.0 mm. The sequence parameters were: TE/TR = 87/8000 ms, FOV = 200 mm, 55 isotropically distributed diffusion weighted directions, two diffusion weighting values, b = 0 and 700 s/mm^2 and three b = 0 (non-diffusion-weighted) images.

ENIGMA–DTI protocols for extraction of tract-wise average FA values were used. These protocols are detailed elsewhere (Jahanshad et al., 2013) and are available online http://enigma.ini.usc.edu/protocols/dti-protocols/. Briefly, FA images from HCP subjects were

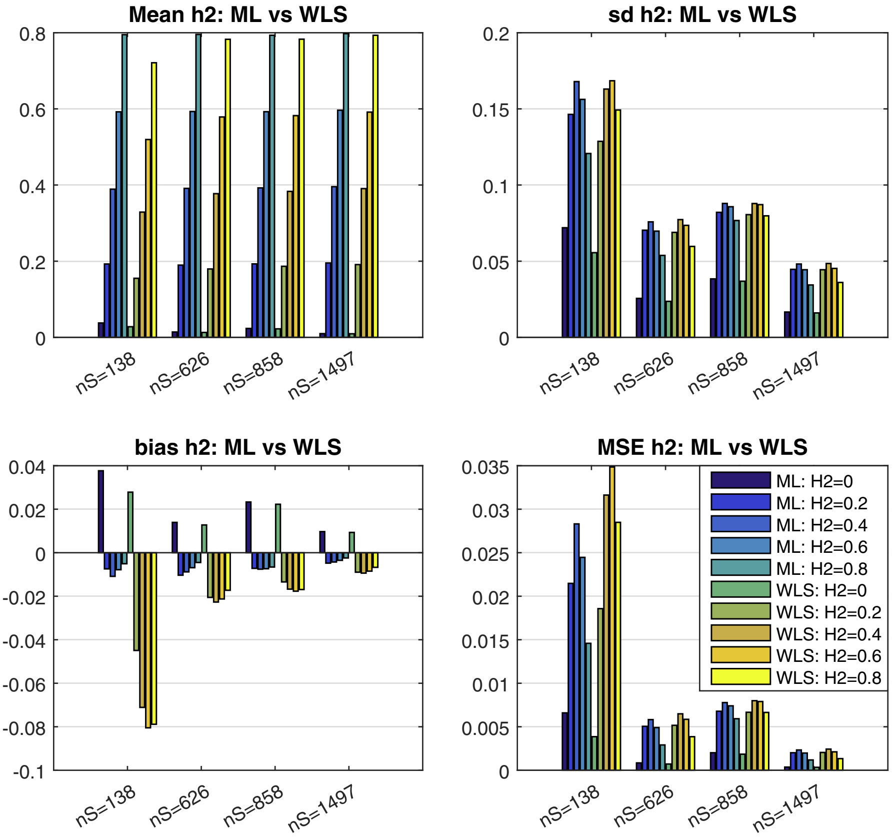


Fig. 1. Simulation 1 results, comparing ML and WLS behavior in terms of mean estimate (top left; true h^2 varies on abscissa within clusters), standard deviation (SD; top right), bias (lower left), and mean squared error (MSE; bottom right). See Table 3 for details of each pedigree; nS denotes number of subjects. WLS has worse bias than ML, but small in absolute magnitude, leading to quite similar MSE for large samples.

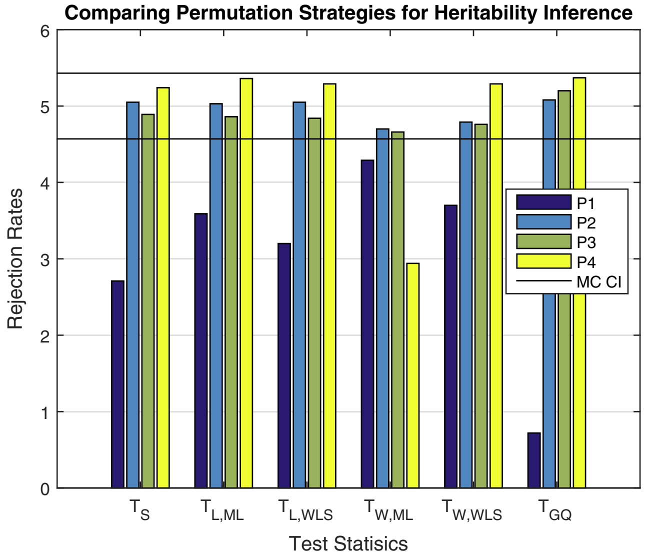


Fig. 2. Simulation 2 results, false positive rates for heritability permutation inference, 5% nominal. Based on GAW10 data with 2 families, 138 subjects, 10,000 realizations, and 500 permutations each realization. Monte Carlo confidence interval (MC CI) is (4.57%,5.43%). Permutation schemes P2–P4 generally seem to work well, while $T_{W,ML}$ tends to be conservative.

non-linearly registered to the ENIGMA–DTI target brain using FSL's FNIRT(Jahanshad et al., 2013). This target was created as a "minimal deformation target" based on images from the participating studies as previously described (Jahanshad et al., 2013b). The data were then processed using FSL's tract-based spatial statistics (TBSS; http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/TBSS) analytic method (Smith et al., 2006) modified to project individual FA values on the hand-segmented ENIGMA–DTI skeleton mask. The protocol, target brain, ENIGMA–DTI skeleton mask, source code and executables, are all publicly available (http://enigma.ini.usc.edu/ongoing/dti-working-group/). The FA values are normalized across individuals by inverse Gaussian transform (Servin & Stephens, 2007; Allison et al., 1999) to ensure normality assumption. Finally, we analyzed the voxel and cluster-wiseFA values with applying along the ENIGMA skeleton mask. To validate our proposed methods for heritability estimation and inference for imaging data, we applied them on GOBS dataset.

Results

Univariate heritability simulation results

Simulation 1

Fig. 1 compares WLS and ML heritability estimators for various designs and effect sizes, in terms of mean, standard deviation (SD) and mean squared error (MSE), for 10,000 Monte Carlo realizations. Large sample theory dictates that ML should provide best performance, and indeed it has least bias and smallest standard deviation, but the (non-iterative) WLS has MSEs that are only slightly larger. As expected, when the sample size is increased WLS and ML heritability estimators reach almost the same performance. While the WLS estimator bias is worse (more negative) than that of ML, the absolute magnitude of bias is small in large samples.

Simulation 2

This simulation assesses the accuracy of parametric null distributions, either a 50:50 χ^2 mixture or F distribution, and power. Under H_0, all false positive rates (Table 4) are conservative except T_{GQ}. The LRT and score tests have Type I error rates that are closer to the nominal level than the Wald tests for the simulated null data ($h^2 = 0$) but none

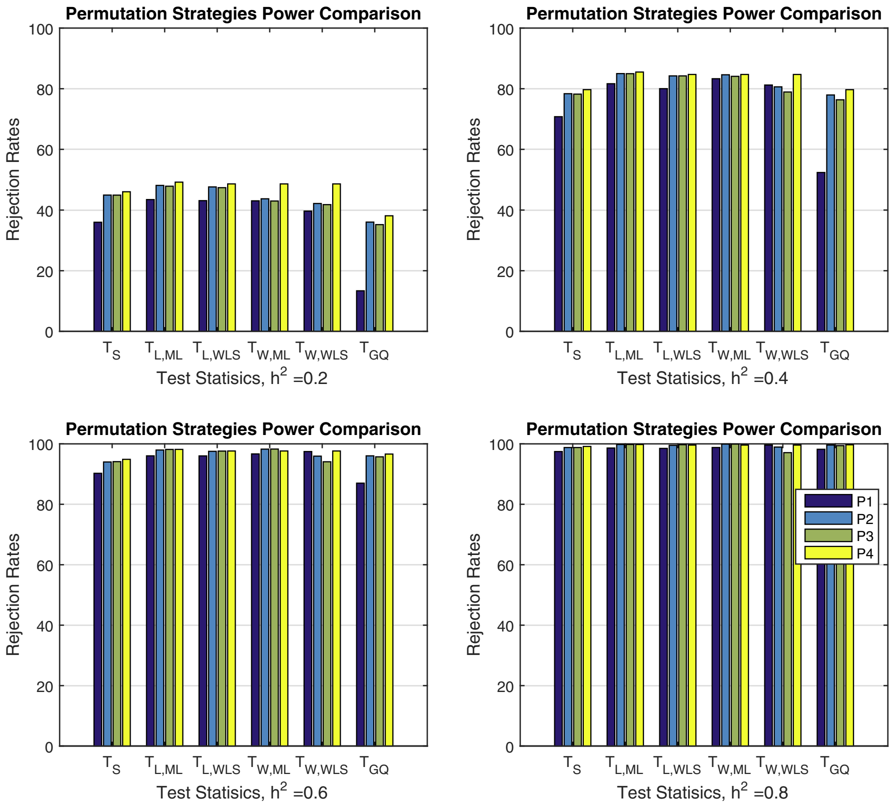


Fig. 3. Simulation 2 results, power for heritability permutation inference. For GAW10 data with 2 families, 138 subjects, 10,000 realizations, and 500 permutations each realization. Monte Carlo confidence interval varies with true rejection rate; for 40% it is (39.0%,41.0%), for 80% it is (79.2%,80.8%).

of them in the MC confidence interval (4.57%–5.42%). Also, the WLS Wald tests had lower error rates than ML Wald tests. In terms of power, the same pattern exists between tests and the LRT and T_{GQ} are the most powerful ones.

The conservative false positive rates are attributable to asymptotic null distributions. In particular, the 50:50 mixture approximation has recently been shown to be conservative, which we confirm here. On the other hand, parametric null distribution of T_{GQ} does not depend on a mixture approximation and, under a normality assumption, it follows F-distribution exactly; this is likely why GQ had the most accurate false positive rate (4.36%).

Figs. 2 and 3 show the performance of permutation inference, with rejection rates and power for different effect sizes under the various permutation strategies. Fig. 2 shows that, generally permutation strategy P1 is more conservative than P2, P3 and P4. Moreover the error rates in terms of P2 are close to the nominal level. Although the permutation strategy P4 has higher rejection rates, they still fall within the Monte Carlo confidence interval (4.57%–5.43%) except for $T_{W,ML}$.

With respect to power, Fig. 3 shows that again P2, P3 and P4 are generally superior to P1 for various effect sizes. In addition P2, P3 and P4 have almost same performance, all within the Monte Carlo confidence bounds.

Based on all of these results, we selected T_S, $T_{W,WLS}$ and T_{GQ} and P2 as the computationally most efficient tests to be considered in the image-wise simulations.

Image-wise simulation results

Simulation 3

This simulation evaluates false positive rate control in the more challenging image-wise setting, for both voxel and cluster-wise heritability inference. Fig. 4 shows the P–P plot of uncorrected P-values, plotted as $-\ log_{10}P$-values. Except for modest conservativeness ($P \approx 10^{-2.5}$), and of course the truncation due to limited permutations (500 permutations, minimal P-value of 0.002, maximum $-\ log_{10}P$-value of 2.69), the accuracy is quite good over-all. Fig. 5 show that FWE-correctedP-values are also accurate, with slight conservativeness with the GQ test. For the 5% level specifically, voxel-wiseFWE for the score, the Wald and the GQ tests were 5.08 %, 5.44 % and 5.4 % respectively, well within the Monte Carlo 95% CI, (4.40%–5.60%).

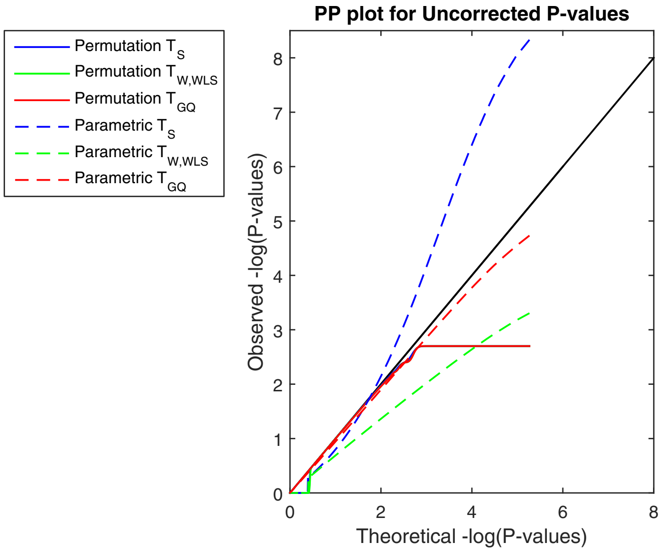


Fig. 4. Simulation 3 results, $-\ log_{10}$ PP Plot for uncorrected parametric and permutation P-values for our proposed test statistics. Permutation P-values are valid (solid lines), though are bounded below by 1/500 (above by 2.70 in $-\ log_{10}P$), the smallest possible permutation P-value for the 500 permutations used. The permutation P-values are overplotted here, and only the permutation T_{GQ} is visible. Parametric P-values for the non-asymptotic GQ test (dashed red line) perform well, while the parametric score test's P-values (dashed blue line) are severely anticonservative (invalid) and Wald test P-values (dashed green line) are severely conservative. Different behavior is seen for P-values larger than 0.5 (smaller than 0.70 in $-\ log_{10}P$) as tests giving ≈ 50 % zero values produce ≈ 50 % P-values of 1 (0 in $-\ log_{10}P$). Results based on GAW10 data with 2 families, 138 subjects, 5000 realizations, 500 permutations each realization, and 96 × 96 × 20 images with 4 mm FWHM smoothing.

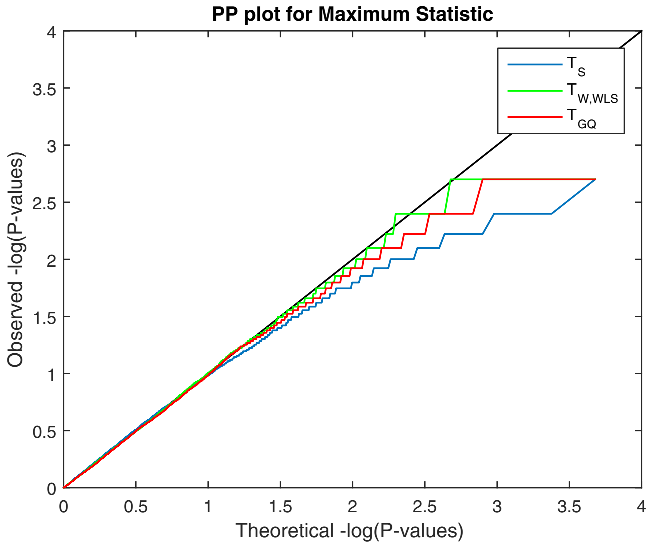


Fig. 5. Simulation 3 results, $-\ log_{10}$ PP plot for voxel-wise FWE permutation P-values under the null hypothesis, for three of our proposed test statistics. Each FWE P-value is for the maximum voxel-wise test statistic in each realized dataset. All three test statistics produce valid P-values, though are bounded below by 1/500 (above by 2.70 in $-\ log_{10}P$). The Wald test's FWE it slightly conservative, and score a bit more so. Results based on GAW10 data with 2 families, 138 subjects, 5000 realizations, 500 permutations each realization, and 96 × 96 × 20 images with 4 mm FWHM smoothing.

Fig. 6 shows cluster-wiseFWE rates for different cluster forming thresholds. All rates are nominal except for the higher cluster forming thresholds of $T_{W,WLS}$($P = 0.005$ & $P = 0.001$). The cluster-forming thresholds come from the parametric null distribution, and Fig. 4

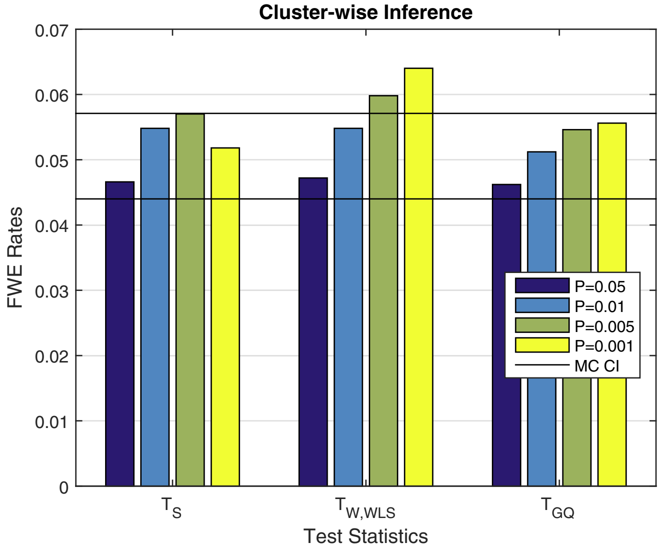


Fig. 6. Simulation 3 results, FWE error rates for cluster-wise permutation heritability inference under the null hypothesis, for three of our proposed test statistics. Score and GC test have nominal false positive rates, while the Wald test is anticonservative for high (uncorrected P of 0.005 & 0.001) clustering forming thresholds. This is likely due to use of parametric cluster-forming threshold; see text for more discussion. Results based on GAW10 data with 2 families, 138 subjects, 5000 realizations, 500 permutations each realization. Monte Carlo 95% confidence interval (4.40%,5.60%).

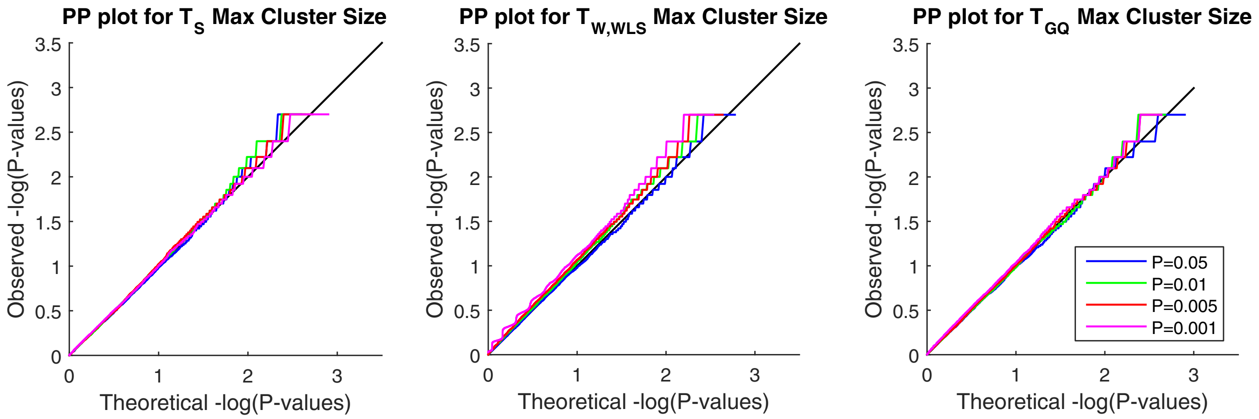


Fig. 7. Simulation 3 results, $-log_{10}$ PP plots for cluster-wise FWE permutation P-values under the null hypothesis, for three of our proposed test statistics. Each FWE P-value is for the maximum cluster size in each realized dataset. GQ has most accurate FWE P-values, followed by the score test; Wald is slightly anticonservative for high cluster forming thresholds; see text for discussion. For GAW10 data with 2 families, 138 subjects, 5000 realizations, 500 permutations each realization (MC CI = (4.40,5.60)).

shows severe conservativeness for $T_{W,WLS}$'s parametric P-values. For example, that figure shows that when a P = 0.001 uncorrected threshold is used for $T_{W,WLS}$, the actual false positive rate is less than 0.0001. This effect, combined with variation of effective false positive rate of the cluster-forming threshold over permutations, could explain this slight anticonservativeness.

Fig. 7 compares the selected test maximum cluster size P-values based on different cluster forming thresholds with their theoretical values; again $T_{W,WLS}$behavior for large cluster forming thresholds shows slightly inflated rejection rates.

Real data analysis

Voxel-wiseFA heritability estimation and inference for the GOBS study are shown with ML and WLS estimators, creating four test statistic images: $T_{L,ML}$, T_S, $T_{W,WLS}$, and T_{GQ}; permutation scheme P2 was used to compute uncorrected and FWE-correctedP-values. Fig. 8 shows histograms of h^2_{ML} (top) and h^2_{ML} (bottom), showing generally the same distribution of heritability over the white matter skeleton. Fig. 10 shows h^2 estimates on the TBSS skeleton. Fig. 9 directly compares WLS and ML heritability estimates with a scatter plot, showing a slight but consistent trend towards underestimation of h^2_{ML} relative to h^2_{ML}, consistent with simulation (Fig. 1).

Voxel-wise uncorrected $-$ log_{10}P-values from T_S, $T_{W,WLS}$, T_{GQ} and $T_{L,ML}$ based on P2 are compared in Fig. 11. Considering $T_{L,ML}$ as a reference (on the abscissa), $T_{W,WLS}$ and T_{GQ} are generally less sensitive than $T_{L,ML}$(Fig. 11 middle and right panels), consistent with the simulations above. However, T_S was more comparable with $T_{L,ML}$(Fig. 11 left panel). Level 5% FWE-corrected statistic thresholds for T_S, $T_{W,WLS}$, $T_{L,ML}$ and T_{GQ} are 39.92, 18.31, 24.27 and 1.72, respectively, producing significant voxel counts of 8521, 1043, 7418 and 2446, respectively, out of 117,139 voxels.

Cluster-wise inference results for cluster forming thresholds corresponded to uncorrected P-value = 0.01 % are shown in Table 5 the tests that we consider. Level 5% FWE-corrected cluster size thresholds for T_S, $T_{W,WLS}$, $T_{L,ML}$ and T_{GQ} are 265, 98, 142 and 135 voxels, respectively. For voxel-wise inference, Fig. 12, the score test was most similar to ML's LRT, and likewise for cluster-wise inference, Fig. 13.

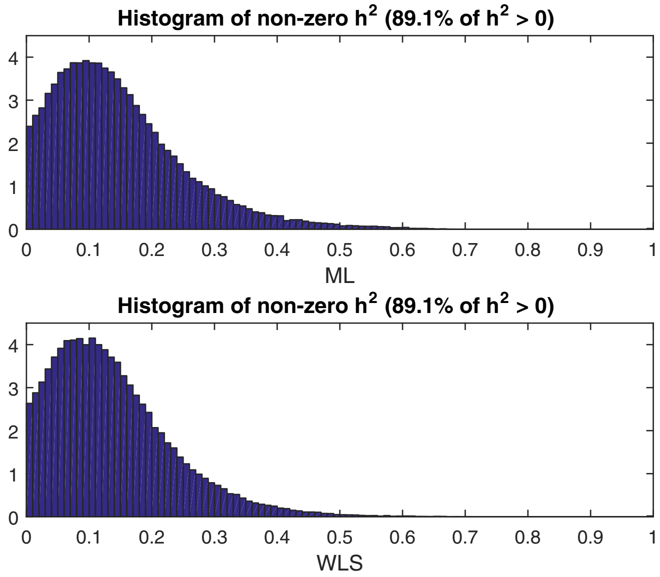


Fig. 8. Real data results, comparison of voxel-wise heritability estimates from ML and WLS estimates. The histograms show that the estimates from the two methods are largely similar.

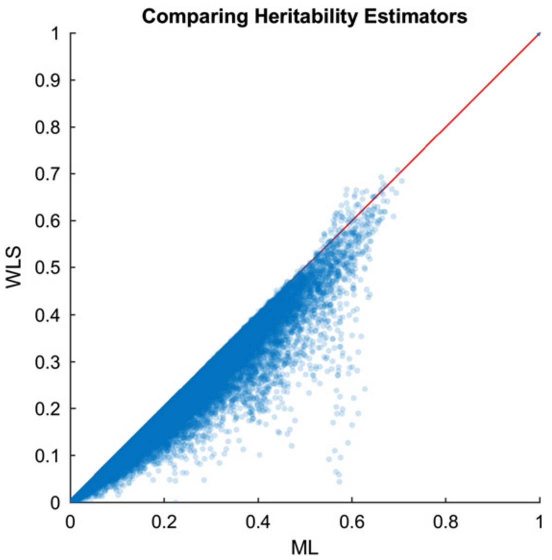
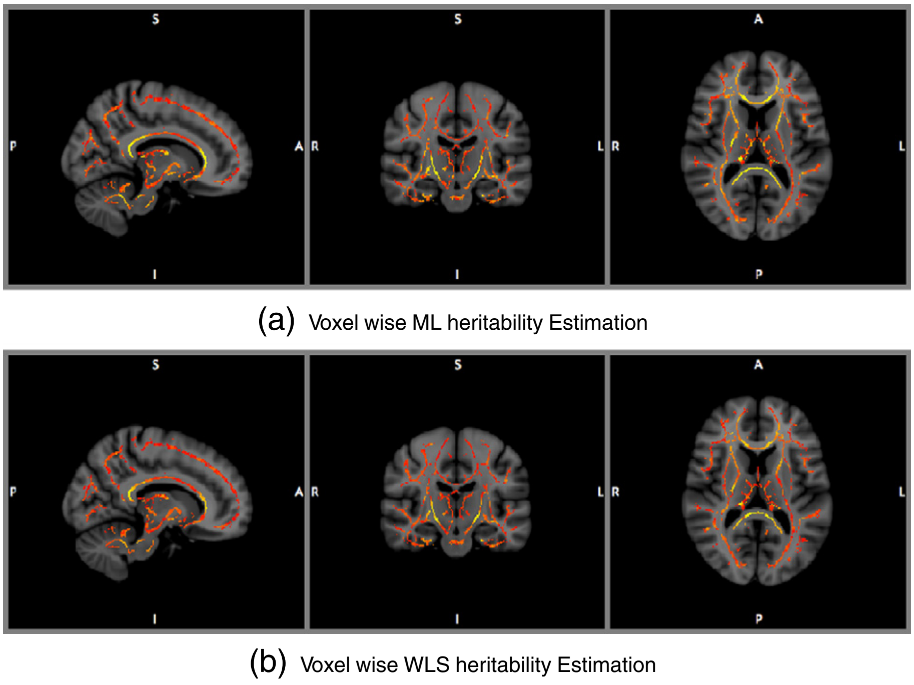


Fig. 9. Real data results, scatterplot of voxel-wise heritability estimates from ML and WLS estimates. The two methods are largely similar, though ML is almost always larger than WLS estimates.



(a) Voxel wise ML heritability Estimation

(b) Voxel wise WLS heritability Estimation

Fig. 10. Real data results, voxel-wise heritability estimates for ML (top) and WLS (bottom). Heritability shown in hot-metal color scale, intensity range [0,0.5] for both, overlaid on MNI reference brain. Differences only apparent in highest FA areas.

Discussion & conclusions

We have proposed a number of computationally efficient tests for heritability with family data. To our knowledge this is the first work that enables practitioners to study brain phenotype heritability in each voxel without confronting an intense computational burden. Our methods are based on the eigensimplified model of Blangero et al. (2013), most of which can be implemented with auxiliary models, corresponding to regressing squared OLS residuals on the kinship matrix eigenvalues.

For heritability estimation our WLS method, based on one step of Newton's method, was a fast and reasonable approximation to fully iterated ML, ideal for application to brain image data.

For heritability inference, we found that parametric P-values for LRT, Wald and score methods were all conservative, likely due to the untenable i.i.d. assumption underlying the 50:50 χ^2 mixture approximation. As an alternative, permutation test error rates were much closer than parametric one to the nominal level. Notably, all of our simulations included fixed effects covariates (X).

The GQ heteroscedasticity test, adapted here for heritability detection, had good performance in simulation, with the best false positive control and respectable power, but on the real data was dramatically different (see Fig. 12) and apparently less powerful.

Image wise simulation results showed that FWE-correctedvoxel- and cluster-wise inference was valid at the 5% level for T_S and T_{GQ}, permutation scheme P2. In real data, the P-values for T_{GQ} were less similar to the LRT results than the score or Wald test, and was less sensitive over all. The GQ test's power depends on the cut point used to define the two groups, though we did not investigate further. On balance we suggest the use of T_S for standard neuroimaging inference tool including voxel and cluster-wise inference.

Running time for different test statistics that were presented in Table 6 based on a benchmark with Intel(R) core(TM) i7-2600CPU @ 3.4 GH and 16 GB RAM feature confirms that the empirical null

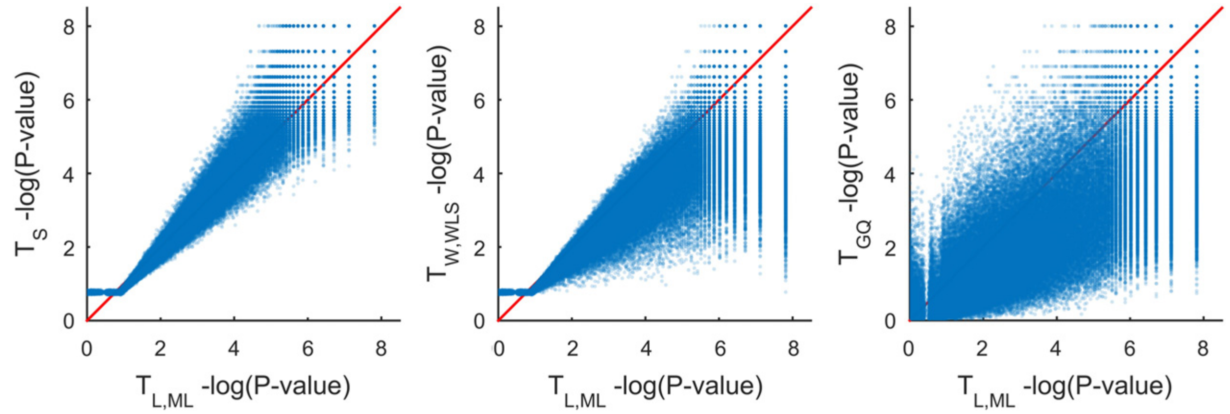


Fig. 11. Real data results, scatter plots of voxel-wise uncorrected — $log_{10}P$-values for score, WLS Wald and GQ tests vs. the ML LRT test. Score P-values are most faithful representation of the ML LRT P-values, while WLS Wald P-values tend to be more conservative; GQ P-values are much more different and generally more conservative.

Table 5
Real data results, cluster-wise inferences with different methods.

Method	Total # of clusters	# of significant clusters	Largest cluster size	Smallest corrected P-value
$T_{L,ML}$	1770	22	24,246	0.0005
$T_{W,WLS}$	1725	19	3643	0.0003
T_S	1689	11	31,250	0.0003
T_{GQ}	1751	20	4383	0.0003

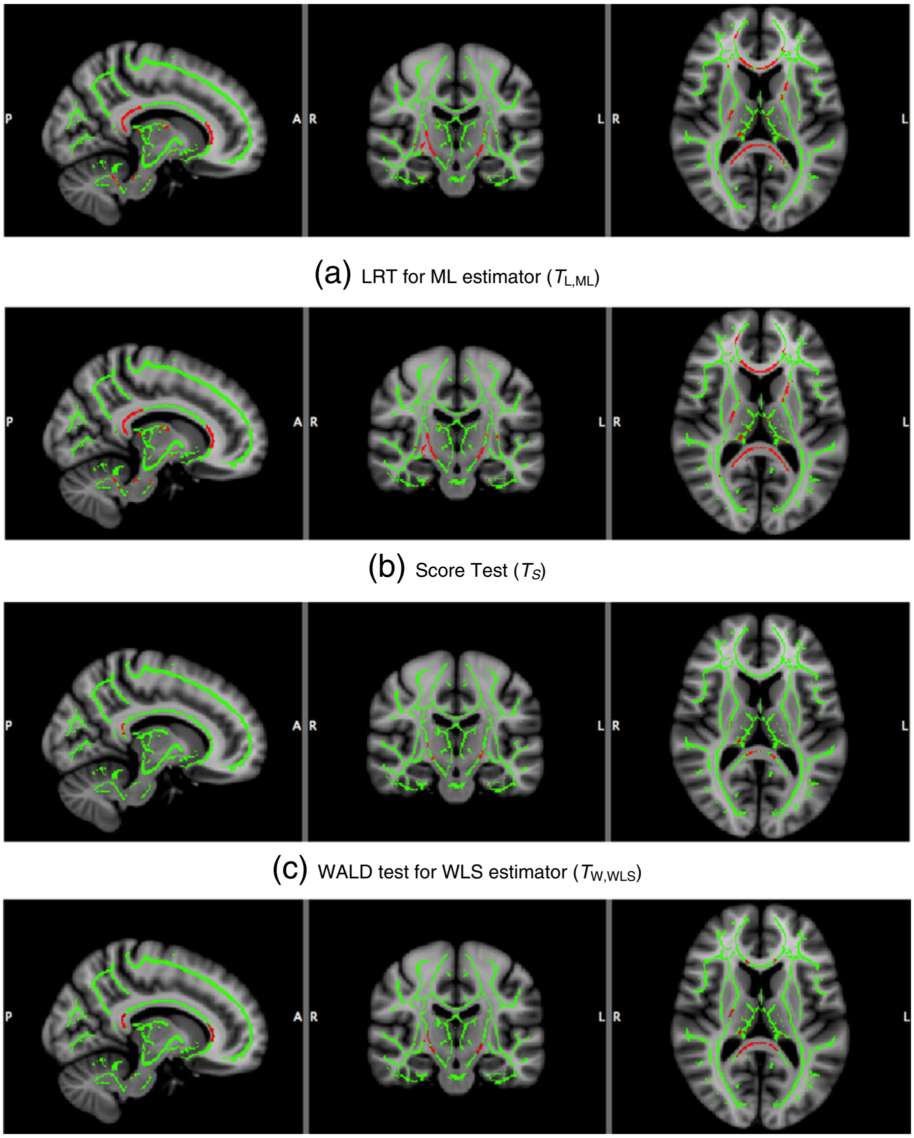
Cluster-wise inference for $T_{L,ML}$, $T_{W,WLS}$, T_S and T_{GQ}. Based on 858 subjects from GOBS and 3000 permutations.

distribution of explained sum of squares of auxiliary model (T_S) under the permutation scheme P2 can be derived substantially faster than $T_{L,ML}$, the classic test statistic for heritability inference. Although the sample size plays an important role in running time, we believe that T_S can be derived significantly faster than the other tests, since it does not depend on numerical optimization. Hence, the whole permutation distribution can be derived easily, either for a univariate trait or a multivariate spatially dependent neuroimaging data accounting explicitly for family wise error.

Finally, we note that yet-more computationally efficient estimates can be obtained by conditioning on the over-all variance estimate, $\hat{\sigma}^2$, which leads to a 1-parameter variance model. However, in initial simulations we found that this lead to greater bias in h^2 and specifically h^2 estimates in excess of 1.0. Thus we retained the 2-parameter variance model.

In conclusion, our results present a novel inference technique to be implemented in the genetic imaging analysis software like SOLAR-Eclipse(http://www.nitrc.org/projects/se_linux). These methods provide fast inference procedure on millions of phenotypes, filtering a small number of elements for further investigation with more computational intense tools. In future work we will extend these tools for inference on covariates, in particular permutation-based tests for voxel-wiseGWAS analysis for family based data.

The methods in this work will soon be found in the SOLAR and SOLAR-Eclipse packages, and a Matlab implementation is available at http://warwick.ac.uk/tenichols/FPHI.



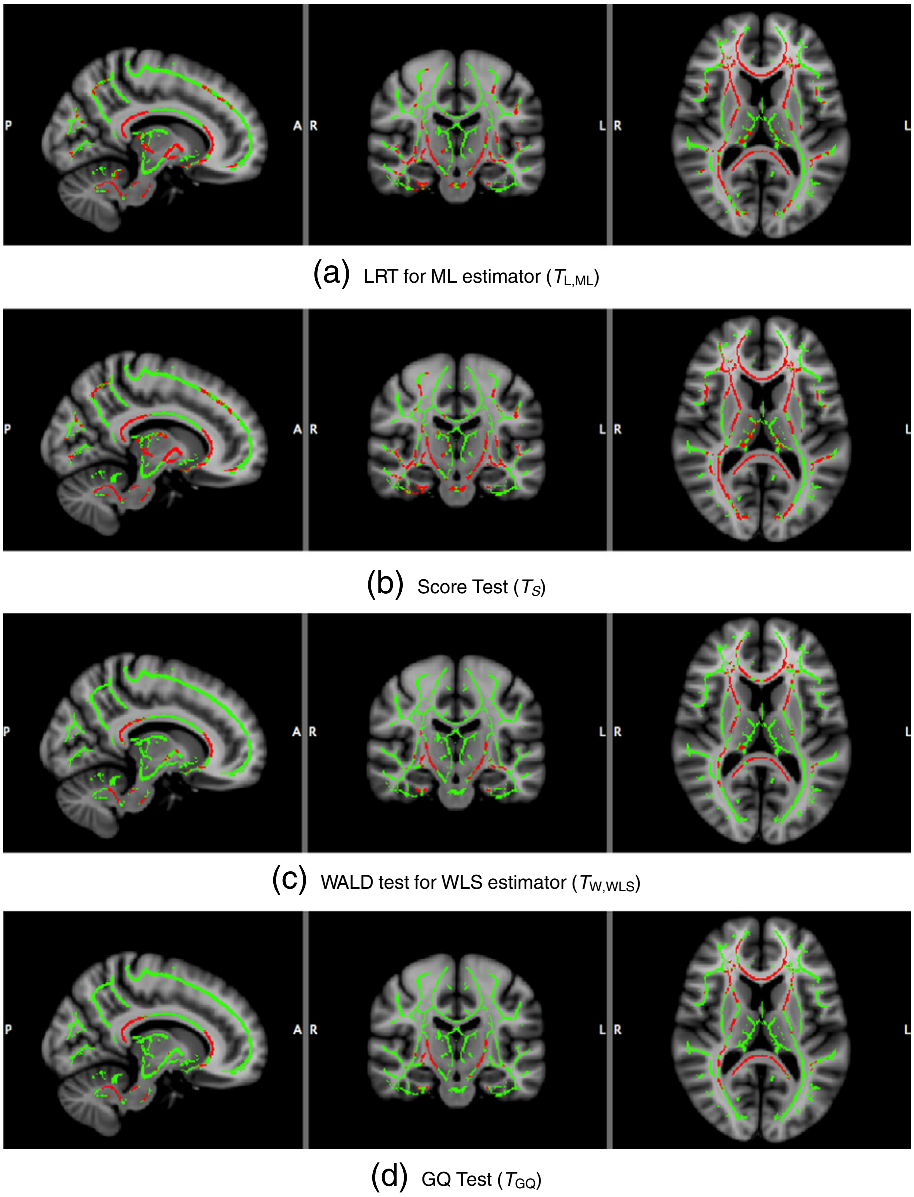
(a) LRT for ML estimator ($T_{L,ML}$)

(b) Score Test (T_S)

(c) WALD test for WLS estimator ($T_{W,WLS}$)

(d) GQ Test (T_{GQ})

Fig. 12. Real data results, voxel-wise 5% FWE significant heritability, for 4 different methods. Full skeleton and significant voxels are in green and red, respectively. The non-iterative score test gives very similar results to the ML (fully iterated) LRT, with the other 2 methods being less sensitive.



(a) LRT for ML estimator ($T_{L,ML}$)

(b) Score Test (T_S)

(c) WALD test for WLS estimator ($T_{W,WLS}$)

(d) GQ Test (T_{GQ})

Fig. 13. Real data results, cluster-wise 5% FWE significant heritability, for 4 different methods, cluster-forming threshold parametric uncorrected P = 0.01. Full skeleton and significant voxels are in green and red, respectively. Methods appear more similar, but again the non-iterative score test is most similar to the ML LRT result.

Acknowledgments

This study was supported by R01 EB015611 (PK,TN), MH0708143 and MH083824 grants to DCG and by MH078111 and MH59490 to JB. This work was also supported in part by a Consortium grant (U54 EB020403) from the NIH Institutes contributing to the Big Data to Knowledge (BD2K) Initiative, including the NIBIB and NCI. TN is supported by the Wellcome Trust.

Table 6
Computation times. Comparison of running times for a dataset with 138 subjects, 2 families, (GAW10 kinship) and 184,320 voxels. Run on Intel(R) core(TM) i7-2600 CPU @ 3.4 GH and 16 GB RAM.

Statistics	Univariate trait	Image-wise trait
$T_{L,ML}$	1 s	8 h
$T_{W,WLS}$	0.005 s	2 s
T_S	0.005 s	2 s
T_{GQ}	0.004 s	1.5 s

References

Allison, D.B., Neale, M.C., Zannolli, R., Schork, N.J., Amos, C.I., Blangero, J., 1999. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am. J. Hum. Genet. 65 (2), 531–544.

Almasy, L., Blangero, J., 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. 62 (5), 1198–1211.

Amemiya, T., 1977. A note on a heteroscedastic model. J. Econ. 6 (3), 365–370.

Amos, C.I., 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. Am. J. Hum. Genet. 54 (3), 535–543.

Blangero, J., Almasy, L., 1997. Multipoint oligogenic linkage analysis of quantitative traits. Genet. Epidemiol. 14 (6), 959–964.

Blangero, J., Diego, V.P., Dyer, T.D., Almeida, M., Peralta, J., Kent, J.W., Williams, J.T., Almasy, L., Göring, H.H.H., 2013. A Kernel of Truth: Statistical Advances in Polygenic Variance Component Models for Complex Human Pedigrees. vol. 81. Academic Press.

Blokland, G.A., McMahon, K.L., Hoffman, J., Zhu, G., Meredith, M., Martin, N.G., Thompson, P.M., de Zubicaray, G.I., Wright, M.J., 2008. Quantifying the heritability of task-related brain activation and performance during the n-back working memory task: a twin fMRI study. Biol. Psychol. 79 (1), 70–79.

Brouwer, R.M., Mandl, R.C., Peper, J.S., van Baal, G.C.M., Kahn, R.S., Boomsma, D.I., Pol, H.E.H., 2010. Heritability of {DTI} and {MTR} in nine-year-old children. NeuroImage 53 (3), 1085–1092.

Buse, A., 1973. Goodness of fit in generalized least squares estimation. Am. Stat. 27 (3), 106–108.

Buse, A., 1979. Goodness-of-fit in the seemingly unrelated regressions model: a generalization. J. Econ. 10.

Buse, A., 1984. Tests for additive heteroskedasticity: Goldfeld and Quandt revisited. Empir. Econ. 9 (4), 199–216.

Cao, J., 1999. The size of the connected components of excursion sets of X^2, t and F fields. Adv. Appl. Probab. 31 (3), 579–595.

Chernoff, H., 1954. On the distribution of the likelihood ratio. Ann. Math. Stat. 25 (3), 573–578.

Chiang, M.-C., Barysheva, M., Shattuck, D.W., Lee, A.D., Madsen, S.K., Avedissian, C., Klunder, A.D., Toga, A.W., McMahon, K.L., de Zubicaray, G.I., Wright, M.J., Srivastava, A., Balov, N., Thompson, P.M., 2009. Genetics of brain fiber architecture and intellectual performance. J. Neurosci. 29 (7), 2212–2224.

Chiang, M.-C., McMahon, K.L., de Zubicaray, G.I., Martin, N.G., Hickie, I., Toga, A.W., Wright, M.J., Thompson, P.M., 2011. Genetics of white matter development: a {DTI} study of 705 twins and their siblings aged 12 to 29. NeuroImage (ISSN: 1053-8119) 54 (3), 2308–2317.

Crainiceanu, C., 2008. Likelihood ratio testing for zero variance components in linear mixed models. In: Dunson, D. (Ed.), Random Effect and Latent Variable Model Selection. Lecture Notes in Statistics volume 192. Springer, New York, pp. 3–17.

Crainiceanu, C.M., Ruppert, D., 2004a. Restricted likelihood ratio tests in nonparametric longitudinal models. Stat. Sin. 14 (3), 713–730.

Crainiceanu, C.M., Ruppert, D., 2004b. Likelihood ratio tests in linear mixed models with one variance component. J. R. Stat. Soc. Ser. B (Stat Methodol.) 66 (1), 165–185.

Crainiceanu, C.M., Ruppert, D., 2004c. Likelihood ratio tests for goodness-of-fit of a nonlinear regression model. J. Multivar. Anal. 91 (1), 35–52.

den Braber, A., Bohlken, M.M., Brouwer, R.M., Ent, D. van 't, Kanai, R., Kahn, R.S., de Geus, E.J.C., Hulshoff Pol, H.E., Boomsma, D.I., 2013. Heritability of subcortical brain measures: a perspective for future genome-wide association studies. NeuroImage 83C, 98–102.

Dominicus, A., Skrondal, A., Gjessing, H., Pedersen, N., Palmgren, J., 2006. Likelihood ratio tests in behavioral genetics: problems and solutions. Behav. Genet. (ISSN: 0001-8244) 36 (2), 331–340.

Draper, N., Stoneman, D., 1966. Testing for the inclusion of variables in linear regression by a randomisation technique. Technometrics 8 (4), 695–699.

Drikvandi, R., Verbeke, G., Khodadadi, A., Partovi Nia, V., 2013. Testing multiple variance components in linear mixed-effects models. Biostatistics 14 (1), 144–159 (Oxford, England).

Fitzmaurice, G.M., Lipsitz, S.R., Ibrahim, J.G., Sept. 2007. A note on permutation tests for variance components in multilevel generalized linear mixed models. Biometrics 63 (3), 942–946.

Freedman, D., Lane, D., 1983. A nonstochastic interpretation of reported significance levels. J. Bus. Econ. Stat. 1 (4), 292–298.

Freedman., D.A., Nov. 2007. How can the score test be inconsistent? Am. Stat. 61 (4), 291–295.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1994. Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1 (3), 210–220.

Genovese, C.R., Lazar, N.A., Nichols, T.E., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 870–878.

Glahn, D.C., Thompson, P.M., Blangero, J., 2007. Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function. Hum. Brain Mapp. 28 (6), 488–501.

Goldfeld, S., Quandt, R., 1965. Some tests for homoscedasticity. J. Am. Stat. 60 (310), 539–547.

Hopper, J.L., Mathews, J.D., 1982. Extensions to multi-variate normal models for pedigree analysis. Ann. Hum. Genet. 46, 373–383.

Jahanshad, N., Kochunov, P.V., Sprooten, E., Mandl, R.C., Nichols, T.E., Almasy, L., Blangero, J., Brouwer, R.M., Curran, J.E., de Zubicaray, G.I., Duggirala, R., Fox, P.T., Hong, L.E., Landman, B.A., Martin, N.G., McMahon, K.L., Medland, S.E., Mitchell, B.D., Olvera, R.L., Peterson, C.P., Starr, J.M., Sussmann, J.E., Toga, A.W., Wardlaw, J.M., Wright, M.J., Pol, H.E.H., Bastin, M.E., McIntosh, A.M., Deary, I.J., Thompson, P.M., Glahn, D.C., 2013. Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: a pilot project of the ENIGMA–DTI working group. NeuroImage 81, 455–469.

Kochunov, P., Glahn, D., Lancaster, J., Winkler, A., Smith, S., Thompson, P., Almasy, L., Duggirala, R., Fox, P., Blangero, J., 2010. Genetics of microstructure of cerebral white matter using diffusion tensor imaging. NeuroImage 53 (3), 1109–1116.

Kochunov, P., Glahn, D., Lancaster, J., Thompson, P., Kochunov, V., Rogers, B., Fox, P., Blangero, J., Williamson, D., 2011a. Fractional anisotropy of cerebral white matter and thickness of cortical gray matter across the lifespan. NeuroImage 58 (1), 41–49.

Kochunov, P., Glahn, D., Nichols, T., Winkler, A., Hong, E., Holcomb, H., Stein, J., Thompson, P., Curran, J., Carless, M., Olvera, R., Johnson, M., Cole, S., Kochunov, V., Kent, J., Blangero, J., 2011b. Genetic analysis of cortical thickness and fractional anisotropy of water diffusion in the brain. Front. Neurosci. 5 (120).

Kochunov, P., Jahanshad, N., Marcus, D., Winkler, A., Sproote, E., Nichols, T., Hong, L., Behrens, T., Andersson, E., Yacoub, J., Ugurbil, K., Brouwer, C., Landman, B., Braber, A., Almassy, L., Fox, P., Olvera, R., Blangero, J., DC, G., Van Essen, D., 2014a. Heritability of fractional anisotropy in human white matter: a comparison of Human Connectome Project and ENIGMA–DTI data. NeuroImage 111, 300–311.

Kochunov, P., Jahanshad, N., Sprooten, E., Nichols, T.E., Mandl, R.C., Almasy, L., Booth, T., Brouwer, R.M., Curran, J.E., de Zubicaray, G.I., Dimitrova, R., Duggirala, R., Fox, P.T., Hong, L.E., Landman, B.A., Lemaitre, H., Lopez, L.M., Martin, N.G., McMahon, K.L., Mitchell, B.D., Olvera, R.L., Peterson, C.P., Starr, J.M., Sussmann, J.E., Toga, A.W., Wardlaw, J.M., Wright, M.J., Wright, S.N., Bastin, M.E., McIntosh, A.M., Boomsma, D.I., Kahn, R.S., den Braber, A., de Geus, E.J., Deary, I.J., Pol, H.E.H., Williamson, D.E., Blangero, J., Ent, D. van 't, Thompson, P.M., Glahn, D.C., 2014b. Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: comparing meta and mega analytical approaches for data pooling. NeuroImage 95, 136–150.

Koten, J.W., Wood, G., Hagoort, P., Goebel, R., Propping, P., Willmes, K., Boomsma, D.I., 2009. Genetic contribution to variation in cognitive function: an fMRI study in twins. Science 323 (5922), 1737–1740.

Kremen, W.S., Prom-Wormley, E., Panizzon, M.S., Eyler, L.T., Fischl, B., Neale, M.C., Franz, C.E., Lyons, M.J., Pacheco, J., Perry, M.E., Stevens, A., Schmitt, J.E., Grant, M.D., Seidman, L.J., Thermenos, H.W., Tsuang, M.T., Eisen, S.A., Dale, A.M., Fennema-Notestine, C., 2010. Genetic and environmental influences on the size of specific brain regions in midlife: the VETSA MRI study. NeuroImage 49 (2), 1213–1223.

Lange, K., 2003. Mathematical and Statistical Methods for Genetic Analysis. 2nd ed. Springer.

Lee, O.E., Braun, T.M., 2012. Permutation tests for random effects in linear mixed models. Biometrics 68 (2), 486–493.

MacCluer, J.W., Blangero, J., Dyer, T.D., Speer, M.C., Jan. 1997. GAW10: simulated family data for a common oligogenic disease with quantitative risk factors. Genet. Epidemiol. 14 (6), 737–742.

Matthews, S.C., Simmons, A.N., Strigo, I., Jang, K., Stein, M.B., Paulus, M.P., 2007. Heritability of anterior cingulate response to conflict: an fMRI study in female twins. NeuroImage 38 (1), 223–227.

McKay, D., Knowles, E., Winkler, A., Sprooten, E., Kochunov, P., Olvera, R., Curran, J., Kent, J., Jack, W., Carless, M., Göring, H., Dyer, T., Duggirala, R., Almasy, L., Fox, P., Blangero, J., Glahn, D., 2014. Influence of age, sex and genetic factors on the human brain. Brain Imaging Behav. 8 (2), 143–152.

Molenberghs, G., Verbeke, G., Feb. 2007. Likelihood ratio, score, and Wald tests in a constrained parameter space. Am. Stat. 61 (1), 22–27.

Morgan, B.J.T., Palmer, K.J., Ridout, M.S., Nov. 2007. Negative score test statistic. Am. Stat. 61 (4), 285–288.

Neyman, J., Pearson, E.S., 1933. On the problem of the most efficient tests of statistical hypotheses, Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character 231, 289–337.

Nichols, T.E., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat. Methods Med. Res. 12 (5), 419–446.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15 (1), 1–25.

Olvera, R., Bearden, C., Velligan, D., Almasy, L., Carless, M., Curran, J., Williamson, D., Duggirala, R., Blangero, J., Glahn, D., 2011. Common genetic influences on depression, alcohol, and substance use disorders in Mexican–American families. Am. J. Med. Genet. B Neuropsychiatr. Genet. 156 (5), 561–568.

Polk, T.A., Park, J., Smith, M.R., Park, D.C., 2007. Nature versus nurture in ventral visual cortex: a functional magnetic resonance imaging study of twins. J. Neurosci. 27 (51), 13921–13925.

Rao, C.R., 2008. Linear Statistical Inference and its Applications. John Wiley & Sons, Inc.

Rimol, L.M., Panizzon, M.S., Fennema-Notestine, C., Eyler, L.T., Fischl, B., Franz, C.E., Hagler, D.J., Lyons, M.J., Neale, M.C., Pacheco, J., Perry, M.E., Schmitt, J.E., Grant, M.D., Seidman, L.J., Thermenos, H.W., Tsuang, M.T., Eisen, S.A., Kremen, W.S., Dale, A.M., 2010. Cortical thickness is influenced by regionally specific genetic factors. Biol. Psychiatry 67 (5), 493–499.

Salimi-Khorshidi, G., Smith, S.M., Nichols, T.E., 2010. Adjusting the effect of nonstationarity in cluster-based and TFCE inference. NeuroImage 54 (3), 2006–2019.

Samuh, M.H., Grilli, L., Rampichini, C., Salmaso, L., Lunardon, N., 2012. The use of permutation tests for variance components in linear mixed models. Commun. Stat. - Theory and Methods 41 (16-17), 3020–3029.

Self, S.G., Liang, K.-Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. 82 (398), 605–610.

Servin, B., Stephens, M., 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. Plos Genet. 3 (7).

Shephard, N., 1993. Maximum likelihood estimation of regression models with stochastic trend components. J. Am. Stat. Assoc. 88 (422), 590–595.

Shephard, N.G., Harvey, A.C., 1990. On the probability of estimating a deterministic component in the local level model. J. Time Ser. Anal. 11 (4), 339–347.

Silvapulle, M.J., 1992. Robust Wald-type tests of one-sided hypotheses in the linear model. J. Am. Stat. Assoc. 87 (417), 156–161.

Silvapulle, M.J., Silvapulle, P., 1995. A score test against one-sided alternatives. J. Am. Stat. Assoc. 90 (429), 342–349.

Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., Ciccarelli, O., Cader, M.Z., Matthews, P.M., Behrens, T.E., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. NeuroImage 31 (4), 1487–1505.

Stram, D.O., Lee, J.W., 1994. Variance components testing in the longitudinal mixed effects model. Biometrics 50 (4), 1171–1177.

ter Braak, C.J., 1992. Permutation Versus Bootstrap Significance Tests in Multiple Regression and ANOVA, volume 376 of Lecture Notes in Economics and Mathematical Systems. Springer, Berlin Heidelberg.

Verbeke, G., Molenberghs, G., 2003. The use of score tests for inference on variance components. Biometrics 59 (2), 254–262.

Verbeke, G., Molenberghs, G., Nov. 2007. What can go wrong with the score test? Am. Stat. 61 (4), 289–290.

Winkler, A.M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P.T., Duggirala, R., Glahn, D.C., 2010. Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. NeuroImage 53 (3), 1135–1146.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., Feb. 2014. Permutation inference for the general linear model. NeuroImage (ISSN: 1095-9572) 92C, 381–397.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., Nov. 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12 (6), 900–918.